МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)

А.Г. ХРАМОВ

МЕТОДЫ И АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Рекомендовано редакционно-издательским советом федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева» в качестве учебного пособия для обучающихся по основной образовательной программе высшего образования по направлению подготовки 01.04.02 Прикладная математика и информатика

САМАРА
Издательство Самарского университета
2019

УДК 004.8(075)
ББК 32.973.2я7
Х895

Рецензенты: зам. директора ИПУСС РАН по научной работе С.В. Смирнов;
д-р техн. наук, проф. В.А. Фурсов

Храмов, Александр Григорьевич
Х895 Методы и алгоритмы интеллектуального анализа данных: учеб. пособие / *А.Г. Храмов.* – Самара: Изд-во Самарского университета, 2019. – 176 с.: ил.

ISBN 978-5-7883-1414-3

В данном учебном пособии содержится описание основных методов и алгоритмов интеллектуального анализа данных (Data Mining, Machine Learning). Разбираются примеры учебных практических задач, в том числе с использованием системы R статистического анализа данных и графики.

Пособие может быть использовано для подготовки магистров по направлению 01.04.02 Прикладная математика и информатика в рамках соответствующих магистерских специальностей.

Подготовлено на кафедре технической кибернетики факультета информатики института информатики, математики и электроники Самарского университета.

УДК 004.8(075)
ББК 32.973.2я7

ISBN 978-5-7883-1414-3

Оглавление

Введение 6

1 Корреляционный анализ зависимости 9
1.1 Коэффициент корреляции (Пирсона) 9
1.2 Ранговые коэффициенты корреляции 13
1.2.1 Коэффициент ранговой корреляции Спирмена 14
1.2.2 Коэффициент ранговой корреляции Кендалла 16
1.3 Примеры расчёта коэффициентов ранговой корреляции и проверки гипотезы о статистической независимости факторов .. 18

2 Линейный регрессионный анализ 23
2.1 Схема регрессионного эксперимента 24
2.2 Классическая модель линейной регрессии 27
2.3 Метод наименьших квадратов 28
2.4 Примеры использования метода наименьших квадратов 29
2.5 Свойства МНК-оценки 35
2.6 Теорема Гаусса–Маркова 36
2.7 Оценивание дисперсии шума наблюдения 38
2.8 Интервальные оценки параметров линейной регрессионной модели и проверка гипотез о коэффициентах регрессии 39

3 Дисперсионный анализ 43
3.1 Вероятностные распределения хи-квадрат и Фишера 43
3.2 Статистика Фишера 44
3.3 МНК-оценка линейной векторной функции коэффициентов регрессии 46
3.4 Однофакторный двухуровневый дисперсионный анализ 48
3.5 Однофакторный многоуровневый дисперсионный анализ 53
3.6 Двухфакторный дисперсионный анализ при произвольном числе наблюдений 60
3.7 Общий подход к дисперсионному анализу 63
3.8 Пример двухфакторного дисперсионного анализа 66

4 Методы и алгоритмы кластеризации данных 70
4.1 Задача кластеризации 70
4.2 Метод k-средних (k-means) 76

4.3 Алгоритм кластеризации Isodata (ИСОМАД) 82
4.4 Иерархическая кластеризация 83

5 Распознавание образов 90
5.1 Статистическая проверка гипотез 90
5.2 Минимизация функции риска 101
5.3 Критерий Неймана–Пирсона 102
5.4 ROC-кривая 108
5.5 Наивный байесовский классификатор 109

6 Линейные классификаторы 113
6.1 Байесовский линейный классификатор 113
6.2 Линейная разделяющая функция, минимизирующая вероятность ошибки 116

7 Дискриминантный анализ 127
7.1 Матрицы рассеяния и критерий разделимости 127
7.2 Разделимость выборки 129
7.3 Выбор признаков, максимизирующих критерий $\mathbf{J}_1$ 130
7.3.1 Линейное преобразование случайных векторов 130
7.3.2 Диагонализирующее и декоррелирующее преобразования 131
7.3.3 Одновременная диагонализация двух ковариационных матриц 132
7.3.4 Оптимизация сокращения размерности пространства признаков 135

8 Метод опорных векторов (SVM) 137
8.1 Линейно разделимая выборка 137
8.2 Линейно неразделимая выборка 142
8.3 Ядра и спрямляющие пространства 146

9 Задачи для самостоятельного решения 149

Список литература и интернет-ресурсы 157

Приложение А Статистические функции 158
А.1 Квантиль распределения 158
А.2 Нормальное распределение 158
А.3 Распределение хи-квадрат 160

А.4 Распределение Стьюдента 163
А.5 Распределение Фишера 166

Приложение Б Сведения из теории матриц 170
Б.1 Спектральное разложение симметричной матрицы 170
Б.2 Свойства операции *trace* (след) 170
Б.3 Линейное преобразование случайного вектора 171

Приложение В Десять основных алгоритмов анализа данных 172

ВВЕДЕНИЕ

Data Mining (дословно *добыча данных*) – на русский язык чаще переводится как «анализ данных». Для перевода «*Data Mining*» на русский язык используют словосочетания: *просев информации*, *добыча данных*, *извлечение данных*, *интеллектуальный анализ данных*, *обнаружение знаний в базах данных*. Не существует «неинтеллектуального» анализа данных [А.Г. Дьяконов; 2012].

Data Mining – собирательное название для обозначения совокупности методов обнаружения в данных, доступных для интерпретации, ранее неизвестных, нетривиальных и практически полезных знаний, необходимых для принятия решений. Методы Data Mining: *классификация, моделирование, прогнозирование, деревья решений, искусственные нейронные сети, генетические алгоритмы, эволюционное программирование, ассоциативная память, нечёткая логика, статистические методы* (**дескриптивный анализ**, **корреляционный** и **регрессионный анализ**, **факторный анализ**, **дисперсионный анализ**, компонентный анализ, **дискриминантный анализ**, **анализ временных рядов**, анализ выживаемости, анализ связей). Выделенные разделы будут изучаться в рамках настоящего курса «*Интеллектуальный анализ данных*» или изучались ранее в курсе «*Тория случайных процессов*».

Интеллектуальный анализ данных – это раздел информатики, изучающий процессы обработки данных с целью получения полезной информации и принятия решений. Основная польза разрабатываемых методов анализа данных заключается в некоторой предсказательной способности: проанализировав неко-

торый набор данных, информационная система анализа данных должна обучиться для дальнейшего распознавания или прогнозирования некоторых участков данных в ситуациях, когда часть данных утеряна или неизвестна. Кроме этого, системы анализа данных могут решать задачи редуцирования объёма данных с целью устранения избыточности, визуализации данных для их удобного восприятия человеком, моделирования новых данных по имеющимся данным и др.

Впервые понятие Data Mining появилось в 1989 году. Изначально оно было связано с автоматизацией и оптимизацией запросов к крупным базам данных. Между тем понятие анализ данных (*англ.* Data Analysis) существовало намного раньше и означало обработку и интерпретацию данных, полученных в ходе экспериментов, в основном научных. С развитием науки и техники эти понятия расширялись и обобщались, стали очень близки друг к другу и в настоящий момент тесно связаны как с анализом больших объёмов данных (*англ.* Big Data), так и с понятием машинного обучения (*англ.* Machine Learning).

Интеллектуальный анализ данных – это во многом прикладная теория, число приложений которой к реальным промышленным задачам растёт с каждым годом. В настоящее время методы и средства интеллектуального анализа данных используются при веб-разработке, в биоинформатике, в системах компьютерного зрения, в разработке компьютерных игр, в маркетинге, в медицинской диагностике, в методах оптимизации, при разработке поисковых систем, при распознавании образов, изображений, речи и сигналов и т.д. Востребованность специалистов по интеллектуальному анализу данных постоянно возрастает, как и доля финансирования разработок в этой области. Появляется всё больше программных решений для анализа данных, в том числе с открытым исходным кодом. Всё это свидетельствует о необходимости включения курса интеллектуального анализа данных в учебные программы по боль-

шинству технических специальностей, связанных с информатикой.

Смежная область – **Машинное обучение** (*Machine Learning*) – раздел искусственного интеллекта, математическая дисциплина, использующая математическую статистику, численные методы оптимизации, теорию вероятностей, выделяющая знания из данных. Различают два типа обучения: 1. *Обучение по прецедентам*, или *индуктивное обучение,* – основано на выявлении закономерностей в эмпирических данных. 2. *Дедуктивное обучение* – предполагает формализацию знаний экспертов и их перенос в компьютер в виде базы знаний, относится к области экспертных систем, поэтому термины *машинное обучение* и *обучение по прецедентам* можно считать синонимами.

В табл. В.1 (приложение В) приведены десять основных алгоритмов анализа данных. Мы их изучим в течение настоящего курса.

1 КОРРЕЛЯЦИОННЫЙ АНАЛИЗ ЗАВИСИМОСТИ

1.1 Коэффициент корреляции (Пирсона)

Простейшим способом проверки наличия статистической зависимости между двумя факторами является анализ их коэффициента корреляции (**Пирсона**). Допустим, что известна выборка $\left\{x^i, y^i\right\}_{i=1}^{N}$ из генеральной совокупности $\left(X, Y\right)$, имеющей двумерное нормальное распределение с плотностью вероятности:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-r^2}} \exp\left[-\frac{\left(x-m_X\right)^2}{2\sigma_x^2} + \frac{r\left(x-m_X\right)\left(y-m_Y\right)}{\sigma_X\sigma_Y} - \frac{\left(y-m_Y\right)^2}{2\sigma_Y^2}\right],$$

где $m_X = \mathbf{M}X$, $m_Y = \mathbf{M}Y$, $\sigma_x^2 = \mathbf{D}X$, $\sigma_Y^2 = \mathbf{D}Y$,

$$r = \frac{\mathbf{M}XY - \mathbf{M}X \cdot \mathbf{M}Y}{\sqrt{\mathbf{D}X \cdot \mathbf{D}Y}}.$$

Наличие линейной статистической зависимости между факторами X и Y характеризуется отличием от нуля коэффициента корреляции r. Точечной оценкой коэффициента корреляции r является выборочный коэффициент корреляции

$$\hat{r} = \frac{\sum_{i=1}^{N}\left(x^i - \overline{x}\right)\left(y^i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x^i - \overline{x}\right)^2 \sum_{i=1}^{N}\left(y^i - \overline{y}\right)^2}} = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{s_x \cdot s_y}, \qquad (1.1)$$

где $\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x^i$, $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y^i$, $\overline{xy} = \frac{1}{N}\sum_{i=1}^{N} x^i y^i$, $\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N}\left(x^i\right)^2$,

$\overline{y^2} = \frac{1}{N}\sum_{i=1}^{N}\left(y^i\right)^2$, $s_x^2 = \overline{x^2} - \overline{x}^2$, $s_y^2 = \overline{y^2} - \overline{y}^2$.

Точное вероятностное распределение выборочного коэффициента корреляции $\hat{r}$ получено Р. Фишером в 1915 году [1]. Плотность вероятности имеет вид:

$$f_{\hat{r}}(\rho) = \frac{\left(1-r^2\right)^{(N-1)/2}}{\pi\,\Gamma\left(N-2\right)}\left(1-\rho^2\right)^{(N-4)/2}\frac{d^{N-2}}{d\left(\rho r\right)^{N-2}}\left\{\frac{\arccos\left(-\rho r\right)}{\sqrt{1-r^2\rho^2}}\right\}, |\rho| < 1.$$

В частности, при $r = 0$ (некоррелированность факторов X и Y) плотность распределения $\hat{r}$ имеет вид:

$$f_{\hat{r}}(\rho) = \frac{1}{B\left(N/2-1,1/2\right)}\left(1-\rho^2\right)^{(N-4)/2}, \quad |\rho| < 1.$$

Здесь $\Gamma(s) = \int_0^\infty t^{s-1}e^{-t}dt$ – гамма-функция, $\mathrm{B}(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$ – бета-функция.

Сложный вид вероятностного распределения выборочного коэффициента корреляции не позволяет использовать его в практических целях для получения интервальных оценок коэффициента корреляции и для проверки гипотезы о некоррелированности. Поэтому для этих целей используются различные асимптотические приближения [1]. В частности, для проверки гипотезы о независимости факторов X и Y ($H_0 : r = 0$) используется статистика:

$$t = \frac{\hat{r}\sqrt{N-2}}{\sqrt{1-\hat{r}^2}} \sim t\left(N-2\right), \qquad (1.2)$$

которая имеет t-распределение Стьюдента с числом степеней свободы $(N-2)$. Факторы X и Y считаются линейно зависимыми, когда величина выборочного коэффициента корреляции $\hat{r}$ отлична от нуля на уровне значимости α, то есть если выполняется неравенство

$$\hat{r}^2 > \left[1+\frac{N-2}{\left(t_{N-2}^{1-\alpha/2}\right)^2}\right]^{-1}, \qquad (1.3)$$

где $t_{N-2}^{1-\alpha/2}$ – квантиль на уровне $(1-\alpha/2)$ t-распределения Стьюдента с (N – 2) степенями свободы. Чем с меньшим уровнем значимости отвергается гипотеза о некоррелированности, тем меньше вероятность того, что это сделано ошибочно. Таким образом, *уровень значимости* статистической гипотезы – это вероятность отвергнуть эту гипотезу при условии, что на самом деле она верна.

На рис. 1.1 изображены критические области для проверки гипотез (см. курс математической статистики).

Если решить неравенство (1.3) относительно α, то можно найти граничный уровень значимости α_0, для которого гипотеза о некоррелированности всё ещё отвергается при том, что при чуть меньшем уровне значимости она уже была бы принята при заданных выборочном коэффициенте корреляции $\hat{r}$ и объёме выборки N:

$$\alpha < \alpha_0 = 2-2F_{t(N-2)}\left(\frac{|\hat{r}|\sqrt{N-2}}{\sqrt{1-\hat{r}^2}}\right),$$

где $F_{t(N-2)}(\cdot)$ – интегральная функция t-распределения Стьюдента с $(N-2)$ степенями свободы. Значение граничного уровня значимости α_0 в западной литературе носит название **p-*value***.

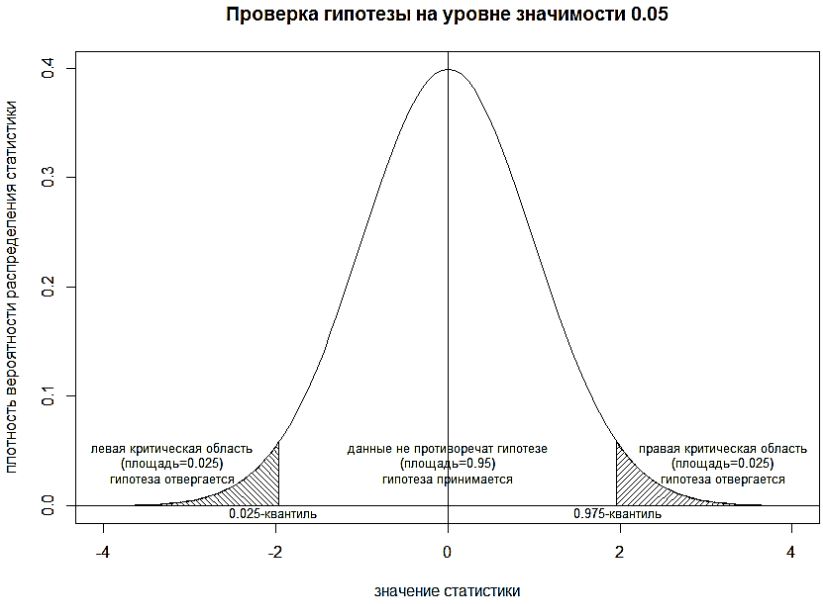


Рис. 1.1. Проверка гипотезы при двусторонней критической области

Эту же статистику (1.2) можно использовать для построения доверительных интервалов коэффициента корреляции, однако при значениях r, существенно отличных от нуля, более точные значения границ интервалов получаются при использовании *z-преобразования Фишера* коэффициента корреляции:

$$z = \operatorname{arcth} r = \frac{1}{2}\ln\frac{1+r}{1-r}; \quad r = \operatorname{th} z = \frac{e^{2z}-1}{e^{2z}+1},$$

где th и arcth – функции гиперболического тангенса и гиперболического арктангенса.

Статистика Фишера

$$z = \left(N-3\right)^{1/2}\left(\operatorname{arcth}\hat{r} - \operatorname{arcth} r\right) \sim \mathrm{N}\left(0;1\right) \qquad (1.4)$$

распределена по стандартному нормальному закону $\mathrm{N}(0;1)$ асимптотически при $N \to \infty$.

Из (1.4) получаем доверительный интервал для коэффициента корреляции:

$$\mathbf{Pr}\left\{\left|\operatorname{arcth}\hat{r} - \operatorname{arcth} r\right| < (N-3)^{-1/2}\, u^{(1+p)/2}\right\} = p\,,$$

$$\mathbf{Pr}\left\{\operatorname{th}\left(\operatorname{arcth}\hat{r} - (N-3)^{-1/2}\, u^{(1+p)/2}\right) < r < \operatorname{th}\left(\operatorname{arcth}\hat{r} + (N-3)^{-1/2}\, u^{(1+p)/2}\right)\right\} = p\,, \quad (1.5)$$

где p – доверительная вероятность; $u^{(1+p)/2}$ – квантиль на уровне $(1+p)/2$ стандартного нормального распределения.

1.2 Ранговые коэффициенты корреляции

Иногда возникает потребность в статистическом анализе нечисловых факторов. Например, требуется проверить наличие статистической зависимости между результатами двух кругов чемпионата по футболу. Так может быть и в случае, когда признаками являются некоторые натуральные порядковые номера или другие целые числа, абсолютные значения которых несут только информацию о порядке объектов. В этом случае целесообразно вычислять корреляцию не между самими признаками, а между порядковыми номерами объектов, упорядоченных по этим признакам.

В этом случае вместо коэффициента корреляции Пирсона используются коэффициенты ранговой корреляции **Кендалла** и **Спирмена.**

1.2.1 Коэффициент ранговой корреляции Спирмена

Коэффициент ранговой корреляции Спирмена ρ (*англ.* Spearman's rank correlation coefficient) является частным случаем коэффициента корреляции Пирсона, если в качестве значений факторов *X, Y* использовать значения их **рангов** x^i, y^i , то есть индексов в упорядоченных последовательностях их значений:

$$X'_1 \le X'_2 \le X'_3 \le \ldots \le X'_{N-1} \le X'_N,\ x^i \in [1;N],$$

$$Y'_1 \le Y'_2 \le Y'_3 \le \ldots \le Y'_{N-1} \le Y'_N,\ y^i \in [1;N].$$

Докажем, что

$$\rho = 1 - \frac{6}{N\left(N^2-1\right)} \sum_{i=1}^{N} \left(x^i - y^i\right)^2 ,$$

где x^i – ранг фактора X_i, y^i – ранг фактора Y_i, i – номер объекта (наблюдения).

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x^i = \frac{1}{N}\left(1+2+3+\ldots+N\right) = \frac{N+1}{2},$$

$$\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y^i = \frac{1}{N}\left(1+2+3+\ldots+N\right) = \frac{N+1}{2},$$

$$\overline{x^2} = \frac{1}{N}\sum_{i=1}^{N} \left(x^i\right)^2 = \frac{1}{N}\left(1^2+2^2+3^2+\ldots+N^2\right) =$$

$$= \frac{1}{N}\cdot\frac{N\left(N+1\right)\left(2N+1\right)}{6},\ \overline{x^2} = \overline{y^2} = \frac{\left(N+1\right)\left(2N+1\right)}{6},$$

$$s_x^2 = s_y^2 = \frac{\left(N+1\right)\left(2N+1\right)}{6} - \left(\frac{N+1}{2}\right)^2 = \frac{N^2-1}{12},$$

$$\sum_{i=1}^{N}\left(x^i-y^i\right)^2=\sum_{i=1}^{N}\left(x^i\right)^2+\sum_{i=1}^{N}\left(y^i\right)^2-2\sum_{i=1}^{N}x^iy^i=$$

$$=\frac{N\left(N+1\right)\left(2N+1\right)}{6}+\frac{N\left(N+1\right)\left(2N+1\right)}{6}-2\sum_{i=1}^{N}x^iy^i.$$

Из последнего уравнения следует, что

$$\sum_{i=1}^{N}x^iy^i=\frac{N\left(N+1\right)\left(2N+1\right)}{6}-\frac{1}{2}\sum_{i=1}^{N}\left(x^i-y^i\right)^2,$$

$$\overline{xy}=\frac{1}{N}\sum_{i=1}^{N}x^iy^i=\frac{\left(N+1\right)\left(2N+1\right)}{6}-\frac{1}{2N}\sum_{i=1}^{N}\left(x^i-y^i\right)^2.$$

Тогда по определению выборочного коэффициента корреляции Пирсона из (1.1) получаем

$$\rho=\frac{\overline{xy}-\overline{x}\cdot\overline{y}}{s_x\cdot s_y}=\frac{\left[\frac{\left(N+1\right)\left(2N+1\right)}{6}-\frac{1}{2N}\sum_{i=1}^{N}\left(x^i-y^i\right)^2\right]-\left(\frac{N+1}{2}\right)^2}{\frac{N^2-1}{12}}=$$

$$=1-\frac{6}{N\left(N^2-1\right)}\sum_{i=1}^{N}\left(x^i-y^i\right)^2.$$

Для проверки гипотезы о независимости факторов X и Y, как и для коэффициента корреляции Пирсона, используется статистика (1.2) $t=\rho\sqrt{N-2}\Big/\sqrt{1-\rho^2}$, которая имеет t-распределение Стьюдента с числом степеней свободы $\left(N-2\right)$.

Граничный (предельный) уровень значимости:

$$\alpha_0=2-2F_{t(N-2)}\left(\frac{\left|\rho\right|\sqrt{N-2}}{\sqrt{1-\rho^2}}\right).$$

1.2.2 Коэффициент ранговой корреляции Кендалла

Коэффициент ранговой корреляции Кендалла (*англ.* Kendall's rank correlation coefficient) вычисляют на основе ранжирования признаков.

Если рассмотреть все возможные пары объектов (i, j), $i, j = 1, 2, 3, \ldots, N$, то про каждую пару можно сказать, одинаково ли или нет упорядочены признаки в этой паре. Обозначим через P – количество пар объектов, у которых признаки ***x*** и ***y*** упорядочены одинаково, а через Q – количество пар объектов, у которых признаки ***x*** и ***y*** упорядочены по-разному. Тогда

$$P - Q = \sum_{i=1}^{N} \sum_{j=1}^{N} \operatorname{sgn}\left(x_k - x_i\right) \operatorname{sgn}\left(y_k - y_i\right),$$

где x_i – ранг признака ***x***, y_i – ранг признака ***y***, i – номер объекта,

$$\operatorname{sgn}(u) = \begin{cases} 1, & u > 0; \\ 0, & u = 0; \\ -1, & u < 0. \end{cases}$$

Коэффициент ранговой корреляции вычисляют с использованием показателя $(P - Q)$:

$$\tau = \frac{2(P - Q)}{N(N - 1)}.$$

При «ручном» расчёте параметров P и Q можно использовать следующий алгоритм:

1. Значения фактора X выставляют в порядке возрастания и присваивают ранги.

2. Ранжируют значения показателя Y.
3. Рассчитывают: P – суммарное число наблюдений, следующих за текущими наблюдениями, с большим значением рангов Y; Q – суммарное число наблюдений, следующих за текущими наблюдениями с меньшим значением рангов Y (равные ранги не учитываются).
4. Рассчитывают коэффициент корреляции:

$$\tau = \frac{2\left(P-Q\right)}{N\left(N-1\right)} .$$

$$\operatorname{sgn}\left(u\right) = \begin{cases} 1, & u > 0; \\ 0, & u = 0; \\ -1, & u < 0. \end{cases}$$

При отсутствии зависимости факторов X и Y коэффициент ранговой корреляции τ Кендалла имеет при $N \geq 10$ имеет распределение, близкое к нормальному с параметрами $\mathbf{M}\tau = 0$, $\mathbf{D}\tau = \frac{2\left(2N+5\right)}{9N\left(N-1\right)}$, то есть статистика $u = \tau\sqrt{\frac{9N\left(N-1\right)}{2\left(2N+5\right)}}$ имеет приблизительно стандартное нормальное распределение.

Коэффициенты корреляции τ и ρ изменяются от –1 до +1. Если коэффициент корреляции равен +1, то это означает, что ранжировки одинаковы; если он равен –1, то противоположны (ранжировки обратны друг другу). Равенство коэффициента корреляции нулю означает, что ранжировки линейно независимы (некоррелированы).

1.3 Примеры расчёта коэффициентов ранговой корреляции и проверки гипотезы о статистической независимости факторов

Пример 1. Требуется найти коэффициенты ранговой корреляции Спирмена и Кендалла между местом команды после первого круга и местом команды по результатам сезона. Проверить гипотезу о независимости рассматриваемых факторов на уровне значимости **0.1**.

Таблица 1.1 – **Итоги чемпионата России по футболу 2002 года**

	Локомотив	ЦСКА	Спартак	Торпедо	Крылья Советов	Сатурн
Итог сезона	1	2	3	4	5	6
Итог I круга	1	2	3	10	6	4
	Шинник	Динамо	Ротор	Зенит		
	7	8	9	10		
	8	7	5	9		

Количество объектов (команд): $N = 10.$

Коэффициент ранговой корреляции Спирмена

$$\sum_{i=1}^{N}\left(x_i - y_i\right)^2 = 0^2 + 0^2 + 0^2 + 6^2 + 1^2 + 2^2 + 1^2 + 1^2 + 4^2 + 1^2 = 60,$$

$$\rho = 1 - \frac{6}{N\left(N^2 - 1\right)}\sum_{i=1}^{N}\left(x_i - y_i\right)^2 = 1 - \frac{6}{10 \cdot 99} \cdot 60 = \frac{\mathbf{7}}{\mathbf{11}} \approx 0.636.$$

Значение *t*-статистики Стьюдента:

$$t = \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho^2}} = \frac{(7/11)\sqrt{10-2}}{\sqrt{1-(7/11)^2}} = \frac{7\sqrt{8}}{\sqrt{72}} = \frac{7}{3} \approx 2.333333.$$

Критическое значение – квантиль на уровне 0.95 t-распределения Стьюдента с 8 степенями свободы $t_8^{0.95} \approx 1.859548$. Так как $2.333333 > 1.859548$, то гипотезу о независимости рассматриваемых факторов можно отвергнуть на уровне значимости **0.1**, то есть факторы считаем статистически зависимыми на этом уровне значимости.

Граничный уровень значимости (*p-value*):

$$\alpha_0 = 2 - 2F_{N-2}^t\left(\frac{|\rho|\sqrt{N-2}}{\sqrt{1-\rho^2}}\right) = 2 - 2F_8^t\left(\frac{7}{3}\right) \approx 0.0479,$$

где $F_8^t(\cdot)$ – интегральная функция t-распределения Стьюдента с 8 степенями свободы.

Коэффициент ранговой корреляции Кендалла

$$S = P - Q = (9+8+7+0+3+4+1+1+1+0) -$$
$$-(0+0+0+6+2+0+2+1+0+0) = 34 - 11 = 23,$$

$$\tau = \frac{2S}{N(N-1)} = \frac{2 \cdot 23}{10 \cdot 9} = \frac{\mathbf{23}}{\mathbf{45}} \approx \mathbf{0.511}.$$

Находим значение статистики

$$u = \tau\sqrt{\frac{9N(N-1)}{2(2N+5)}} = \frac{23}{45}\sqrt{\frac{90 \cdot 9}{2 \cdot 25}} = \frac{23}{5\sqrt{5}} \approx 2.057183.$$

Критическое значение – квантиль на уровне 0.95 стандартного нормального распределения $u^{0.95} \approx 1.644854$. Так как $2.057183 > 1.644854$, то гипотезу о независимости рассматриваемых факторов можно отвергнуть на уровне значимости 0.1.

Граничный уровень значимости (*p-value*):

$$\alpha_0 = 2 - 2F_u\left(\tau\sqrt{\frac{9N(N-1)}{2(2N+5)}}\right) = 2 - 2F_u\left(\frac{23}{5\sqrt{5}}\right) \approx 0.0397,$$

где $F_u(\cdot)$ – интегральная функция стандартного нормального распределения.

Видим, что p-value при проверке с помощью ранговых коэффициентов корреляции гипотезы о некоррелированности между факторами (местом команды после первого круга и местом команды по результатам сезона) лежит в пределах 4–5%.

Пример 2. Вычислить выборочные коэффициенты корреляции Пирсона r, Спирмена ρ и Кендалла τ двух ***бинарных*** факторов: X – пациенту проведена или не проведена вакцинация от гриппа, $\boldsymbol{Y}$ **–** клиент заболел или не заболел гриппом. Протокол прививок и наблюдений задан следующей таблицей (число объектов $N = 9$):

№ пациента		1	2	3	4	5	6	7	8	9
X	Прививка	1	1	1	0	1	0	0	1	0
Y	Заболевание	0	1	1	1	1	0	0	1	0

***Коэффициент корреляции Пирсона* r**

$$r = \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{s_x \cdot s_y} = \frac{\frac{4}{9} - \frac{5}{9}\cdot\frac{5}{9}}{\frac{5}{9} - \left(\frac{5}{9}\right)^2} = \frac{11}{20} = 0.55.$$

***Коэффициент корреляции Спирмена* ρ**

Для вычисления ранговой корреляции Спирмена проведём ранжирование факторов *X* и *Y*:

№ пациента		1	2	3	4	5	6	7	8	9
x	Прививка (ранг)	5	6	7	1	8	2	3	9	4
y	Заболевание (ранг)	1	5	6	7	8	2	3	9	4

$$\rho = 1 - \frac{6}{N\left(N^2-1\right)}\sum_{i=1}^{N}\left(x_i - y_i\right)^2 =$$

$$= 1 - \frac{6}{9\cdot 80}\left(4^2+1^2+1^2+6^2+0^2+0^2+0^2+0^2+0^2\right) = 1 - \frac{6\cdot 54}{9\cdot 80} = \frac{11}{20} = 0.55.$$

Коэффициент корреляции Кендалла τ

Для вычисления ранговой корреляции Кендалла расположим один их факторов (например, фактор *X*) в порядке возрастания рангов и запишем соответствующие ранги фактора *Y*:

№ пациента		4	6	7	9	1	2	3	5	8
x	Прививка (ранг)	1	2	3	4	5	6	7	8	9
y	Заболевание (ранг)	7	2	3	4	1	5	6	8	9

$$\tau = \frac{2\left(P-Q\right)}{N\left(N-1\right)} = \frac{2\cdot 18}{9\cdot 8} = 0.5;$$

$$P = 2+6+5+4+4+3+2+1+0 = 27,$$
$$Q = 6+1+1+1+0+0+0+0+0 = 9.$$

Пример 3. На уровне значимости $\alpha = 0.1$ проверить гипотезу о статистической независимости рабочего стажа и показателя травматизма с использованием рангового коэффициента корреляции Спирмена.

X – Рабочий стаж в годах	До 1 года	1–2 года	3–4 года	5–6 лет	7 лет и более
Y – Травматизм на 100 работающих	24	16	12	12	6

(Данные взяты на сайте http://forex365.ru/indicators/korrelyacionnyj-analiz-spirmena.html – *Корреляционный анализ Спирмена*).

Запишем ранги факторов (по возрастанию значений факторов, с корректировкой одинаковых рангов):

x – Ранги рабочего стажа	1	2	3	4	5
y – Рани травматизма	5	4	2,5	2,5	1

$$\rho = 1 - \frac{6}{N\left(N^2-1\right)}\sum_{i=1}^{N}\left(x_i - y_i\right)^2 = 1 - \frac{6}{5\cdot 24}\left(4^2 + 2^2 + 0.5^2 + 1.5^2 + 4^2\right) =$$

$$= -\frac{37}{40} = -0.925\,.$$

Значение t-статистики Стьюдента:

$$t = \frac{|\rho|\sqrt{N-2}}{\sqrt{1-\rho^2}} = \frac{(37/40)\sqrt{5-2}}{\sqrt{1-(37/40)^2}} = \frac{37}{\sqrt{77}} \approx 4.217.$$

Критическое значение – квантиль на уровне 0.95 t-распределения Стьюдента с 3 степенями свободы $t_3^{0.95} \approx 2.353363$. Так как $4.217 > 2.353363$, то гипотезу о независимости рассматриваемых факторов можно отвергнуть на уровне значимости $\mathbf{0.1}$, то есть факторы считаем статистически зависимыми на этом уровне значимости, то есть травматизм зависит от величины рабочего стажа.

Граничный уровень значимости (*p-value*):

$$\alpha_0 = 2 - 2F_{N-2}^t\left(\frac{|\rho|\sqrt{N-2}}{\sqrt{1-\rho^2}}\right) = 2 - 2F_3^t\left(\frac{37}{\sqrt{77}}\right) \approx 0.0244,$$

где $F_3^t(\cdot)$ – интегральная функция t-распределения Стьюдента с 3 степенями свободы.

2 ЛИНЕЙНЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Регрессионный анализ (*англ.* regression analysis) – это статистический метод исследования влияния одной или нескольких независимых переменных $\mathbf{x} = \left(x_1 \; x_2 \; x_3 \; \ldots \; x_n \right)^T$ на зависимую переменную y. Независимые переменные иначе называют **регрессорами** (*англ.* regressors) или **предикторами** (*англ.* predictors), а зависимую переменную – **критериальной** (*англ.* criterion variable). В частности, регрессионный анализ помогает нам понять, как значение критериальной переменной изменяется, если меняется какая-либо из независимых переменных, а другие независимые переменные фиксированы. В регрессионном анализе оценивается условное математическое ожидание (среднее значение) зависимой переменной при заданных значениях предикторов:

$$f\left(x_1, x_2, x_3, \ldots, x_n\right) = \mathbf{M}\left(y \mid x_1, x_2, x_3, \ldots, x_n\right).$$

Функция $y = f\left(x_1, x_2, x_3, \ldots, x_n\right)$ называется фунцией регрессии y на $x_1, x_2, x_3, \ldots, x_n$.

Для построения функции регресссии f необходимо задать её общий вид и параметры. В зависимости от вида функции и её неизвестных параметров различают **линейную** регрессию и **нелинейную** регрессию.

2.1 Схема регрессионного эксперимента

Для построения функции регрессии f статистическим методом зададим *схему регрессионного эксперимента*, которую в общем случае можно описать уравнением наблюдения:

$$y^i = f\left(\mathbf{x}^i, \boldsymbol{\theta}, \xi^i\right),\ i = \overline{1, N}, \qquad (2.1)$$

где i – номер опыта; N – число опытов; $\mathbf{x} = \left(x_1\ x_2\ x_3\ \ldots\ x_n\right)^T$ – вектор независимых переменных, $\mathbf{x}^i \subset \mathbf{R}^n$; n – размерность пространства независимых переменных; $\mathbf{x}^1, \mathbf{x}^2, \mathbf{x}^3, \ldots, \mathbf{x}^N$ – точки наблюдения: $\mathbf{x}^i = \left(x_1^i\ x_2^i\ x_3^i\ \ldots\ x_n^i\right)^T$, $i = \overline{1, N}$; y^i – значение выходного (наблюдаемого, измеряемого) фактора в i-м опыте; ξ^i – случайное значение шумового фактора (помехи) в i-м опыте; $\boldsymbol{\theta} = \left(\theta_1\ \theta_2\ \theta_3\ \ldots\ \theta_m\right)^T$ – вектор параметров объекта исследования, подлежащий оцениванию; m – число параметров. Параметры $\theta_1, \theta_2, \theta_3, \ldots, \theta_m$ считаются детерминированными (постоянными, неслучайными) величинами.

На рис. 2.1 приведена иллюстрация нелинейного регрессионного эксперимента. Здесь $\mathbf{x}^i = x_1^i = x^i\ (n = 1)$, $\boldsymbol{\theta} = \left(\omega\ a\ b\right)^T$, $y^i = a\sin\left(\omega x^i\right) + b + \xi^i$. Заметим, что если $\omega = \text{const}$ – известная детерминированная величина, то рассматриваемая регрессия является линейной по параметрам $\boldsymbol{\theta} = \left(a\ b\right)^T$.

На рис. 2.2 приведена иллюстрация простейшего линейного регрессионного эксперимента. Здесь $\mathbf{x}^i = x_1^i = x^i\ (n = 1)$, $\boldsymbol{\theta} = \left(a\ b\right)^T$, $y^i = ax^i + b + \xi^i$.

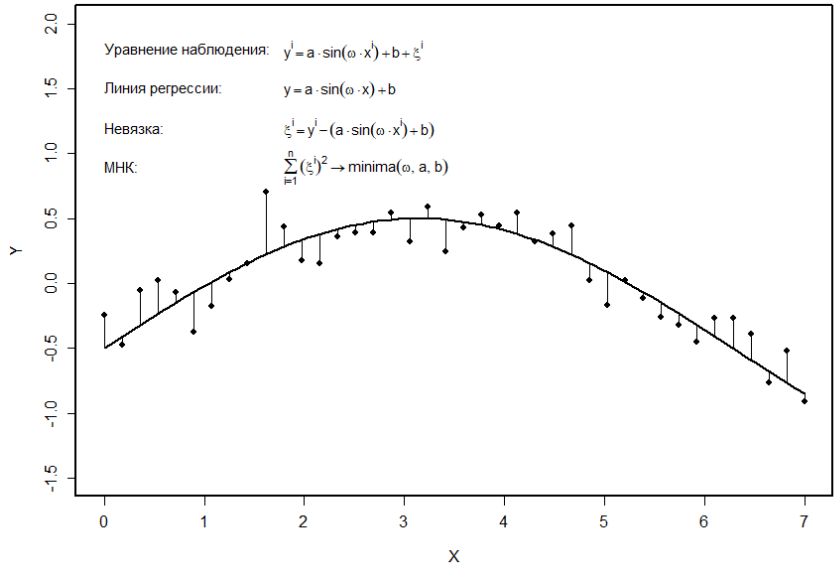


Рис. 2.1. Нелинейная регрессия

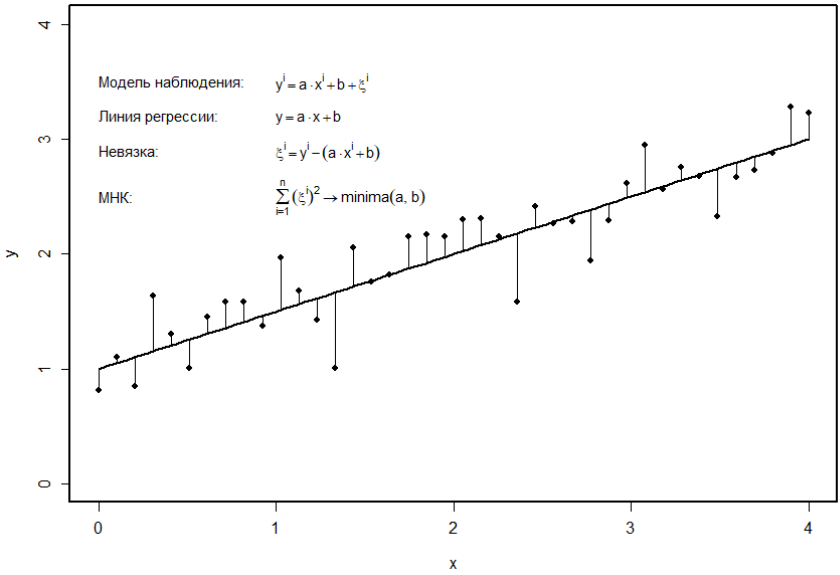


Рис. 2.2. Линейная регрессия

В общем случае модель наблюдения (2.1) *с линейной относительно параметров* функцией регрессии f и *с аддитивным шумом наблюдения* ξ можно единственным образом представить в виде:

$$y = f(\mathbf{x}, \boldsymbol{\theta}, \xi) = f_1(\mathbf{x})\theta_1 + f_2(\mathbf{x})\theta_2 + \ldots + f_m(\mathbf{x})\theta_m + \xi, \qquad (2.2)$$

где $f_1(\mathbf{x})$, $f_2(\mathbf{x})$, …, $f_m(\mathbf{x})$ – некоторые фунции, не зависящие от неизвестных параметров θ_1, θ_2, θ_3, …, θ_m и являющиеся коэффициентами перед этими неизвестными параметрами. Такие m функций будем называть *базовыми*.

Заметим, что в таком представлении базовые функции $f_k(\mathbf{x})$ и сама функция регрессии $f(\mathbf{x}, \boldsymbol{\theta}, \xi)$ могут быть нелинейными от независимых переменных $\mathbf{x}$. Например, $f(\mathbf{x}, \boldsymbol{\theta}, \xi) = a\sin(\omega x) + b + \xi$ – нелинейная функция x, когда параметры $\boldsymbol{\theta} = (a\ b)^T$, а $\omega = \text{const}$ (см. рис. 2.1).

Объединяя базовые функции в вектор-столбец $\mathbf{f}(\mathbf{x}) = \left(f(\mathbf{x})\ \ f(\mathbf{x})\ \ \ldots\ \ f(\mathbf{x})\right)^T$, приходим к матричной форме записи уравнения наблюдения:

$\mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi}$, то есть

$$\begin{pmatrix} y^1 \\ y^2 \\ y^3 \\ \ldots \\ y^N \end{pmatrix} = \begin{bmatrix} f_1^1 & f_2^1 & \ldots & f_m^1 \\ f_1^2 & f_2^2 & \ldots & f_m^2 \\ f_1^3 & f_2^3 & \ldots & f_m^3 \\ \ldots & \ldots & \ldots & \ldots \\ f_1^N & f_2^N & \ldots & f_m^N \end{bmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \ldots \\ \theta_m \end{pmatrix} + \begin{pmatrix} \xi^1 \\ \xi^2 \\ \xi^3 \\ \ldots \\ \xi^N \end{pmatrix}. \qquad (2.3)$$

Матрица

$$\mathbf{F} = \begin{bmatrix} f_1^1 & f_2^1 & \dots & f_m^1 \\ f_1^2 & f_2^2 & \dots & f_m^2 \\ f_1^3 & f_2^3 & \dots & f_m^3 \\ \dots & \dots & \dots & \dots \\ f_1^N & f_2^N & \dots & f_m^N \end{bmatrix} \tag{2.4}$$

размера $N \times m$ называется матрицей эксперимента.

Здесь $f_k^i = f_k\left(\mathbf{x}^i\right) = f_k\left(x_1^i \;\; x_2^i \;\; x_3^i \;\; \dots \;\; x_n^i\right)$; $i = \overline{1, N}$ – значение k-й базовой функции при i-м наблюдении; $\mathbf{Y} = \left(y^1 \;\; y^2 \;\; y^3 \dots y^N\right)^T$ – вектор наблюдения; $\boldsymbol{\xi} = \left(\xi^1 \;\; \xi^2 \;\; \xi^3 \dots \xi^N\right)^T$ – случайный вектор шума наблюдения.

2.2 Классическая модель линейной регрессии

Будем называть *классической моделью линейной регрессии* схему эксперимента (2.3), когда шумы наблюдения ξ^i являются некоррелированными (независимыми) для различных наблюдений случайными величинами, имеющими нулевое математическое ожидание и одинаковую дисперсию σ^2:

$\mathbf{M}\xi^i = 0$, $\mathbf{M}\xi^i\xi^j = \sigma^2\delta_{ij}$ (δ_{ij} – символ Кронекера).

Нулевое математическое ожидание шума наблюдения означает одинаковые в среднем отклонения наблюдаемых значений от линии регрессии, одинаковость дисперсии шума наблюдения в различных опытах, одинаковый в среднем разброс наблюдаемых значений от линии регрессии. Некоррелированность (независимость) шума наблюдения в различных опытах говорит о невлиянии опытов друг на друга. Таким образом, классическую модель наблюдения можно записать в матричном виде следующим образом:

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi}\,,\ \mathbf{M}\boldsymbol{\xi} = \mathbf{0}\,,\ \mathbf{D}\boldsymbol{\xi} = \sigma^2\mathbf{I}_N\,, \tag{2.5}$$

где $\mathbf{D}\boldsymbol{\xi} = \mathbf{M}\boldsymbol{\xi}\boldsymbol{\xi}^T$ – дисперсионная (корреляционная) матрица шума наблюдения, $\mathbf{I}_N$ – единичная матрица размера $N \times N$.

Классическая модель линейной регрессии *при гауссовом шуме наблюдения* имеет вид:

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi}\,,\ \boldsymbol{\xi} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_N\right), \tag{2.6}$$

где запись $\sim \mathrm{N}(\mathbf{a}, \mathbf{R})$ обозначает нормальный случайный вектор с математическим ожиданием $\mathbf{a}$ и корреляционной матрицей $\mathbf{R}$.

2.3 Метод наименьших квадратов

МНК-оценкой (*англ.* Least squares estimation) *параметров* линейной регрессии $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1\ \hat{\theta}_2\ \ldots\ \hat{\theta}_m\right)^T$ называется оценка параметров $\boldsymbol{\theta} = \left(\theta_1\ \theta_2\ \ldots\ \theta_m\right)^T$, минимизирующая сумму квадратов шумовых невязок:

$$R_0^2 = \sum_{i=1}^{N}\left[\, y^i - \left(f_1^i\theta_1 + f_2^i\theta_2 + f_3^i\theta_3 + \ldots + f_m^i\theta_m \right)\right]^2 \to \min_{\theta_1,\theta_2,\ldots,\theta_m}. \tag{2.7}$$

Величина R_0^2 называется остаточной суммой квадратов невязок.

В матричном виде: $R_0^2 = \left(\mathbf{Y} - \mathbf{F}\boldsymbol{\theta}\right)^T\left(\mathbf{Y} - \mathbf{F}\boldsymbol{\theta}\right) \to \min_{\boldsymbol{\theta}}$, то есть

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}}\left(\mathbf{Y} - \mathbf{F}\boldsymbol{\theta}\right)^T\left(\mathbf{Y} - \mathbf{F}\boldsymbol{\theta}\right).$$

Из условия минимума суммы квадратов невязок (2.7)

$$\frac{\partial R_0^2}{\partial \boldsymbol{\theta}} = \frac{\partial}{\partial \boldsymbol{\theta}}\left[\left(\mathbf{Y}-\mathbf{F}\boldsymbol{\theta}\right)^T\left(\mathbf{Y}-\mathbf{F}\boldsymbol{\theta}\right)\right] = \mathbf{F}^T\left(\mathbf{Y}-\mathbf{F}\boldsymbol{\theta}\right) = \mathbf{0}$$

получаем *систему нормальных уравнений* относительно МНК-оценки $\hat{\boldsymbol{\theta}}$:

$$\mathbf{F}^T\mathbf{F}\hat{\boldsymbol{\theta}} = \mathbf{F}^T\mathbf{Y}. \qquad (2.8)$$

Подробный вывод системы нормальных уравнений приведён в [2].

Здесь и в дальнейшем считаем, что система нормальных уравнений (2.8) является невырожденной, то есть $\operatorname{rank}\mathbf{F} = m$ и обратная матрица $\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$ существует. Тогда МНК-оценка параметров линейной регрессии $\hat{\boldsymbol{\theta}}$ может быть определена из следующего соотношения:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}. \qquad (2.9)$$

Если система нормальных уравнений (2.8) является вырожденной, то получить МНК-оценку можно, воспользовавшись аппаратом псевдообратных матриц и наложением дополнительных условий [2].

2.4 Примеры использования метода наименьших квадратов

Пример 1. Линейная регрессия на прямую.

Рассмотрим простейшую линейную регрессию на прямую, показанную на рис. 2.2. Наблюдению доступно N точек на плоскости $\left(x_i, y_i\right)$. Уравнение наблюдения:

$$y_i = ax_i + b + \xi_i.$$

Вектор неизвестных параметров: $\boldsymbol{\theta} = \begin{pmatrix} a \\ b \end{pmatrix}$.

$$\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \cdots & \\ x_N & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \\ \cdots \\ \xi_N \end{pmatrix} \quad \Rightarrow \quad \mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi},$$ матрица

эксперимента: $\mathbf{F} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \cdots & \\ x_N & 1 \end{pmatrix}$,

$$\mathbf{F}^{\mathrm{T}}\mathbf{F} = \begin{pmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & N \end{pmatrix} = N \begin{pmatrix} \overline{x^2} & \overline{x} \\ \overline{x} & 1 \end{pmatrix},$$

$$\left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\right)^{-1} = \frac{1}{N\left(\overline{x^2} - \overline{x}^2\right)} \begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \overline{x^2} \end{pmatrix},$$

$$\mathbf{F}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} y_i \end{pmatrix} = N \begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix},$$

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{Y} = \frac{1}{\overline{x^2} - \overline{x}^2} \begin{pmatrix} 1 & -\overline{x} \\ -\overline{x} & \overline{x^2} \end{pmatrix} \begin{pmatrix} \overline{xy} \\ \overline{y} \end{pmatrix}.$$

С использованием обозначений п.1.1 получаем:

$$\hat{a} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} = r\frac{s_y}{s_x}, \; \hat{b} = \frac{\overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy}}{\overline{x^2} - \overline{x}^2} = \frac{1}{\overline{x^2} - \overline{x}^2} \begin{pmatrix} \overline{xy} - \overline{x} \cdot \overline{y} \\ \overline{x^2} \cdot \overline{y} - \overline{x} \cdot \overline{xy} \end{pmatrix}.$$

Пример 2. Квадратичная аппроксимация.

Построить МНК-оценки параметров a, b, c квадратичной функции $y = ax^2 + bx + c$ по наблюдениям в пяти точках $x \in \{-2, -1, 0, 1, 2\}$.

Решение.

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 4 & -2 & 1 \\ 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \\ \xi_5 \end{pmatrix} \quad \Rightarrow \quad \mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi},$$

$$\mathbf{F} = \begin{pmatrix} 4 & -2 & 1 \\ 1 & -1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \end{pmatrix},$$

$$\mathbf{F}^{\mathrm{T}}\mathbf{F} = \begin{pmatrix} 34 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 5 \end{pmatrix},$$

$$\left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\right)^{-1} = \frac{1}{700}\begin{pmatrix} 50 & 0 & -100 \\ 0 & 70 & 0 \\ -100 & 0 & 340 \end{pmatrix} = \frac{1}{70}\begin{pmatrix} 5 & 0 & -10 \\ 0 & 7 & 0 \\ -10 & 0 & 34 \end{pmatrix},$$

$$\mathbf{F}^{\mathrm{T}}\mathbf{Y} = \begin{pmatrix} 4 & 1 & 0 & 1 & 4 \\ -2 & -1 & 0 & 1 & 2 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 4y_1 + y_2 + y_4 + 4y_5 \\ -2y_1 - y_2 + y_4 + 2y_5 \\ y_1 + y_2 + y_3 + y_4 + y_5 \end{pmatrix},$$

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{Y} = \frac{1}{70}\begin{pmatrix} 5 & 0 & -10 \\ 0 & 7 & 0 \\ -10 & 0 & 34 \end{pmatrix}\begin{pmatrix} 4y_1+y_2+y_4+4y_5 \\ -2y_1-y_2+y_4+2y_5 \\ y_1+y_2+y_3+y_4+y_5 \end{pmatrix},$$

$$\hat{\boldsymbol{\theta}} = \begin{pmatrix} \frac{1}{70}\left[5\left(4y_1+y_2+y_4+4y_5\right)-10\left(y_1+y_2+y_3+y_4+y_5\right)\right] \\ \frac{1}{10}\left(-2y_1-y_2+y_4+2y_5\right) \\ \frac{1}{70}\left[-10\left(4y_1+y_2+y_4+4y_5\right)+34\left(y_1+y_2+y_3+y_4+y_5\right)\right] \end{pmatrix},$$

$$\hat{a} = \frac{4y_1-y_2-2y_3-y_4+4y_5}{14}, \quad \hat{b} = \frac{-2y_1-y_2+y_4+2y_5}{10},$$

$$\hat{c} = \frac{-3y_1+12y_2+17y_3+12y_4-3y_5}{35}.$$

Пример 3. Билинейная аппроксимация по четырём точкам.

Построить МНК-оценки параметров a, b, c симметричной билинейной функции двух переменных $z(x,y) = axy + b(x+y) + c$ по наблюдениям в четырёх точках:

$z_1 = z(0;0)$, $z_2 = z(0;1)$, $z_3 = z(1;0)$, $z_4 = z(1;1)$.

Решение.

$$\boldsymbol{\theta} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}, \quad \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}\begin{pmatrix} a \\ b \\ c \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \\ \xi_3 \\ \xi_4 \end{pmatrix} \quad \Rightarrow \quad \mathbf{Z} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi},$$

$$\mathbf{F} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix}, \quad \mathbf{F}^{\mathrm{T}} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 2 \\ 1 & 1 & 1 & 1 \end{pmatrix}, \quad \mathbf{F}^{\mathrm{T}}\mathbf{F} = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 6 & 4 \\ 1 & 4 & 4 \end{pmatrix},$$

$$\left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\right)^{-1}=\frac{1}{2}\begin{pmatrix}8&-4&2\\-4&3&-2\\2&-2&2\end{pmatrix},\ \mathbf{F}^{\mathrm{T}}\mathbf{Y}=\begin{pmatrix}z_4\\z_2+z_3+2z_4\\z_1+z_2+z_3+z_4\end{pmatrix},$$

$$\hat{\boldsymbol{\theta}}=\left(\mathbf{F}^{\mathrm{T}}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{Y}=\frac{1}{2}\begin{pmatrix}8&-4&2\\-4&3&-2\\2&-2&2\end{pmatrix}\cdot\begin{pmatrix}z_4\\z_2+z_3+2z_4\\z_1+z_2+z_3+z_4\end{pmatrix}=$$

$$=\begin{pmatrix}z_1-z_2-z_3+z_4\\\frac{1}{2}\left(z_2+z_3-2z_1\right)\\z_1\end{pmatrix}.$$

Ответ: $\hat{a}=\dfrac{z_1-z_2-z_3+z_4}{2},\quad \hat{b}=\dfrac{z_2+z_3}{2}-z_1,\quad \hat{c}=z_1.$

Пример 4. Билинейная аппроксимация по девяти точкам.

Построить МНК-оценки параметров a, b, c, d билинейной функции двух переменных $z(x,y)=axy+bx+cy+d$ по наблюдениям в девяти точках на сетке 3×3:

$z_1 = z(\ 1;1),\ z_2 = z(\ 1;0),\ z_3 = z(\ 1;-1),$
$z_4 = z(\ 0;1),\ z_5 = z(\ 0;0),\ z_6 = z(\ 0;-1),$
$z_7 = z(-1;1),\ z_8 = z(-1;0),\ z_9 = z(-1;-1).$

Решение.

$$\begin{pmatrix}z_1\\z_2\\z_3\\z_4\\z_5\\z_6\\z_7\\z_8\\z_9\end{pmatrix}=\begin{pmatrix}1&1&1&1\\0&1&0&1\\-1&1&-1&1\\0&0&1&1\\0&0&0&1\\0&0&-1&1\\-1&-1&1&1\\0&-1&0&1\\1&-1&-1&1\end{pmatrix}\cdot\begin{pmatrix}a\\b\\c\\d\end{pmatrix}+\begin{pmatrix}\xi_1\\\xi_2\\\xi_3\\\xi_4\\\xi_5\\\xi_6\\\xi_7\\\xi_8\\\xi_9\end{pmatrix}\ \Rightarrow\ \mathbf{Y}=\mathbf{F}\boldsymbol{\theta}+\boldsymbol{\xi},$$

$$\mathbf{F} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ -1 & 1 & -1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 0 & -1 & 0 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix},$$

$$\mathbf{F}^{\mathbf{T}}\mathbf{F} = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 8 \end{pmatrix}, \left(\mathbf{F}^{\mathbf{T}}\mathbf{F}\right)^{-1} = \begin{pmatrix} 1/4 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 \\ 0 & 0 & 1/6 & 0 \\ 0 & 0 & 0 & 1/8 \end{pmatrix},$$

$$\mathbf{F}^{\mathbf{T}}\mathbf{Y} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \\ z_7 \\ z_8 \\ z_9 \end{pmatrix} =$$

$$= \begin{pmatrix} z_1 - z_3 - z_7 + z_9 \\ z_1 + z_2 + z_3 - z_7 - z_8 - z_9 \\ z_1 - z_3 + z_4 - z_6 + z_7 - z_9 \\ z_1 + z_2 + z_3 + z_4 + z_5 + z_6 + z_7 + z_8 + z_9 \end{pmatrix},$$

$$\hat{a} = \frac{1}{4}\left(z_1 - z_3 - z_7 + z_9\right), \qquad \hat{c} = \frac{1}{6}\left(z_1 - z_3 + z_4 - z_6 + z_7 - z_9\right),$$

$$\hat{b} = \frac{1}{6}\left(z_1 + z_2 + z_3 - z_7 - z_8 - z_9\right),$$

$$\hat{d} = \frac{1}{8}\left(z_1 + z_2 + z_3 + z_4 + z_5 + z_6 + z_7 + z_8 + z_9\right).$$

2.5 Свойства МНК-оценки

Здесь и далее будем рассматривать классическую модель наблюдения (2.5). В соответствии с этой моделью вектор параметров $\boldsymbol{\theta}$ считается неизвестной неслучайной (детерминированной) величиной.

Оценка некоторого вектора параметров называется несмещённой, если её математическое ожидание равно истинному значению вектора параметров. Докажем, что МНК-оценка $\hat{\boldsymbol{\theta}}$ является *несмещённой* оценкой вектора параметров, то есть:

$$\mathbf{M}\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}. \tag{2.10}$$

$$\mathbf{M}\hat{\boldsymbol{\theta}} = \mathbf{M}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y} = \mathbf{M}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\left(\mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi}\right) =$$

$$= \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\left(\mathbf{F}^T\mathbf{F}\right)\mathbf{M}\boldsymbol{\theta} + \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{M}\boldsymbol{\xi} = \boldsymbol{\theta} + \mathbf{0} = \boldsymbol{\theta}.$$

Докажем, что дисперсионная матрица МНК-оценки параметров (то есть ковариационная матрица ошибки оценивания $\boldsymbol{\varepsilon} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}$) равна $\sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$, где σ^2 – дисперсия шума наблюдения, $\mathbf{F}$ – матрица эксперимента.

$$\mathbf{D}\boldsymbol{\varepsilon} = \mathbf{D}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \mathbf{D}\hat{\boldsymbol{\theta}} = \mathbf{D}\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}\right] =$$

$$= \mathbf{D}\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\left(\mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi}\right)\right] = \mathbf{D}\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\boldsymbol{\xi}\right] =$$

$$= \left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\mathbf{D}\boldsymbol{\xi}\left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)^T =$$

$$= \left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\sigma^2\mathbf{I}_N\left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)^T = \sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}.$$

Индивидуальные дисперсии оценок параметров (дисперсии ошибок оценивания) определяются диагональными элементами матрицы:

$$\mathbf{D}\hat{\theta}_i = \mathbf{D}\left(\hat{\theta}_i - \theta_i\right) = \sigma^2\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii},\ i = 1, 2, 3, \ldots, m.$$

МНК-оценка параметров является *эффективной* оценкой (то есть обладающей минимальной дисперсией ошибки) в классе линейных несмещённых оценок, о чём утверждается в следующей теореме Гаусса–Маркова.

2.6 Теорема Гаусса–Маркова

Линейной оценкой вектора параметров $\boldsymbol{\theta}$ называется оценка $\hat{\boldsymbol{\theta}} = \mathbf{C}\mathbf{Y}$, где $\mathbf{C}$ – произвольная матрица размера $M \times N$. *Линейной несмещённой оценкой* вектора параметров $\boldsymbol{\theta}$ называется такая линейная оценка $\hat{\boldsymbol{\theta}}$, для которой $\mathbf{M}\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ (несмещённость МНК-оценки доказана выше в п.2.5). *Наилучшей линейной несмещённой оценкой (НЛН-оценкой)* вектора параметров $\boldsymbol{\theta}$ называется такая оценка $\hat{\hat{\boldsymbol{\theta}}}$, для кото-

рой $\mathbf{D}\hat{\hat{\boldsymbol{\theta}}} \le \mathbf{D}\hat{\boldsymbol{\theta}}$, где $\hat{\boldsymbol{\theta}}$ – произвольная линейная несмещённая оценка вектора параметров $\boldsymbol{\theta}$.

Далее будем использовать обозначения $\mathbf{A} < \mathbf{B}$, $\mathbf{B} > \mathbf{A}$, $\mathbf{B}-\mathbf{A} > \mathbf{0}$ ($\mathbf{A} \le \mathbf{B}$, $\mathbf{B} \ge \mathbf{A}$, $\mathbf{B}-\mathbf{A} \ge \mathbf{0}$) для указания факта положительной (неотрицательной) определённости матрицы $\mathbf{B}-\mathbf{A}$.

Теорема Гаусса–Маркова. Для классической модели наблюдения (2.5) МНК-оценка $\hat{\boldsymbol{\theta}} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}$ является НЛН-оценкой $\hat{\hat{\boldsymbol{\theta}}}$ вектора параметров $\boldsymbol{\theta}$.

Доказательство: Пусть $\hat{\boldsymbol{\theta}} = \mathbf{C}\mathbf{Y}$ – произвольная линейная оценка вектора параметров $\boldsymbol{\theta}$. Тогда $\mathbf{M}\hat{\boldsymbol{\theta}} = \mathbf{M}\mathbf{C}\mathbf{Y} = \mathbf{M}\mathbf{C}\left(\mathbf{F}\boldsymbol{\theta}+\boldsymbol{\xi}\right) = \mathbf{C}\mathbf{F}\boldsymbol{\theta}$. Условие несмещённости оценки $\hat{\boldsymbol{\theta}}$: $\mathbf{C}\mathbf{F}\boldsymbol{\theta} = \boldsymbol{\theta}$. Отсюда следует, что для несмещённости любой линейной оценки необходимо и достаточно, чтобы выполнялось условие: $\mathbf{C}\mathbf{F} = \mathbf{I}_m$.

Если $\mathbf{C} = \mathbf{C}_0 = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T$, то получаем МНК-оценку $\hat{\boldsymbol{\theta}} = \mathbf{C}_0\mathbf{Y} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}$ с дисперсией ошибки $\sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$ (см. п.2.5). Обозначив $\mathbf{A} = \mathbf{C}-\mathbf{C}_0$, заметим, что $\mathbf{A}\mathbf{C}_0^T = \mathbf{0}$:

$$\mathbf{A}\mathbf{C}_0^T = \mathbf{A}\left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)^T = \left(\mathbf{C}-\mathbf{C}_0\right)\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \left(\mathbf{C}\mathbf{F}-\mathbf{C}_0\mathbf{F}\right)\left(\mathbf{F}^T\mathbf{F}\right)^{-1} =$$

$$= \left(\mathbf{I}_m-\mathbf{I}_m\right)\left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \mathbf{0}.$$

Аналогично $\mathbf{C}_0\mathbf{A}^T = \mathbf{0}$.

$$\mathbf{D}\hat{\boldsymbol{\theta}} = \mathbf{D}\left(\mathbf{CY}\right) = \mathbf{C}\cdot\mathbf{DY}\cdot\mathbf{C}^T = \mathbf{C}\cdot\mathbf{D}\left(\mathbf{F}\boldsymbol{\theta}+\boldsymbol{\xi}\right)\cdot\mathbf{C}^T = \mathbf{C}\cdot\mathbf{D}\boldsymbol{\xi}\cdot\mathbf{C}^T =$$
$$= \mathbf{C}\cdot\sigma^2\mathbf{I}_N\cdot\mathbf{C}^T = \sigma^2\mathbf{CC}^T = \sigma^2\left(\mathbf{C}_0+\mathbf{A}\right)\left(\mathbf{C}_0+\mathbf{A}\right)^T =$$
$$= \sigma^2\left(\mathbf{C}_0\mathbf{C}_0^T + \mathbf{C}_0\mathbf{A}^T + \mathbf{A}\mathbf{C}_0^T + \mathbf{A}\mathbf{A}^T\right) =$$
$$= \sigma^2\left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)^T + \mathbf{A}\mathbf{A}^T\right) =$$
$$= \sigma^2\left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1} + \mathbf{A}\mathbf{A}^T\right) =$$
$$= \sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1} + \sigma^2\mathbf{A}\mathbf{A}^T = \mathbf{D}\hat{\boldsymbol{\theta}}^{МНК} + \sigma^2\mathbf{A}\mathbf{A}^T \geq \mathbf{D}\hat{\boldsymbol{\theta}}^{МНК},$$

так как $\mathbf{D}\hat{\boldsymbol{\theta}}^{МНК} = \sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$, $\mathbf{A}\mathbf{C}_0^T = \mathbf{C}_0\mathbf{A}^T = \mathbf{0}$, а матрица $\mathbf{A}\mathbf{A}^T$ неотрицательно определена.

Мы доказали, что произвольная линейная несмещённая оценка имеет не меньшую дисперсию, чем оценка МНК, что означает эффективность оценки МНК.

2.7 Оценивание дисперсии шума наблюдения

Здесь будем рассматривать классическую модель линейной регрессии *при гауссовом шуме наблюдения* (2.6). Докажем, что статистика

$$s^2 = \frac{R_0^2}{N-m}$$

является несмещённой оценкой дисперсии шума наблюдения σ^2, то есть $\mathbf{M}s^2 = \sigma^2$. Здесь обозначено, как и ранее: N – число опытов (наблюдений), m – число оцениваемых параметров, R_0^2 – остаточная сумма квадратов ошибок:

$$R_0^2 = \left(\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\theta}}\right)^T \left(\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\theta}}\right) = \sum_{i=1}^{N} \left[y^i - \left(f_1^i \hat{\theta}_1 + f_2^i \hat{\theta}_2 + f_3^i \hat{\theta}_3 + \ldots + f_m^i \hat{\theta}_m \right) \right]^2.$$

Доказательство основано на том факте, что статистика R_0^2/σ^2 имеет распределение хи-квадрат с $(N-m)$ степенями свободы и её математическое ожидание равно $(N-m)$. Доказательство распределения статистики R_0^2/σ^2 по закону хи-квадрат приведено в [2].

Находим

$$\mathbf{M}s^2 = \frac{1}{N-m}\mathbf{M}R_0^2 = \frac{\sigma^2}{N-m}\mathbf{M}\left(\frac{R_0^2}{\sigma^2}\right) = \frac{\sigma^2}{N-m}\mathbf{M}\chi_{N-m}^2 = \sigma^2$$

(см. приложение А4, формула (А7)).

2.8 Интервальные оценки параметров линейной регрессионной модели и проверка гипотез о коэффициентах регрессии

Для классической модели с гауссовским шумом наблюдения (2.6) вектор оценки $\hat{\boldsymbol{\theta}}$ имеет многомерное нормальное распределение с математическим ожиданием $\mathbf{M}\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$ и дисперсионной матрицей $\mathbf{D}\hat{\boldsymbol{\theta}} = \sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$(доказано выше в п.2.5): $\hat{\boldsymbol{\theta}} \sim \mathrm{N}\left(\boldsymbol{\theta}, \sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right)$. Вектор $\hat{\boldsymbol{\theta}}$ имеет многомерное нормальное распределение, так как $\hat{\boldsymbol{\theta}} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\left(\mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi}\right)$ – линейное преобразование нормального случайного вектора $\boldsymbol{\xi}$. Поэтому индивидуальные оценки параметров регрессии имеют одномерные нормальные распределения:

$$\hat{\theta}_i \sim \mathrm{N}\left(\theta_i, \sigma^2\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}\right), \quad i=1,2,\ldots,m,$$

то есть статистика $\dfrac{\hat{\theta}_i-\theta_i}{\sigma\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}}$ имеет стандартное нормальное распределение $\mathrm{N}(0,1)$.

Докажем, что статистика $t=\dfrac{\hat{\theta}_i-\theta_i}{s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}}$, $i=1,2,\ldots,m$, где $s^2=\dfrac{R_0^2}{N-m}$, имеет распределение Стьюдента с $(N-m)$ степенями свободы: $t\sim t(N-m)$.

Доказательство. Так как статистика R_0^2/σ^2 имеет хи-квадрат распределение с $(N-m)$ степенями свободы, то в соответствии с определением (А.9) получаем:

$$\frac{\hat{\theta}_i-\theta_i}{s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}}=\frac{\dfrac{\hat{\theta}_i-\theta_i}{\sigma}\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}^{-1/2}}{\sqrt{\dfrac{1}{N-m}\dfrac{R_0^2}{\sigma^2}}}\sim\frac{\mathrm{N}(0;1)}{\sqrt{\dfrac{\chi^2_{N-m}}{N-m}}}=t(N-m), \quad (2.11)$$

то есть статистика $t=\dfrac{\hat{\theta}_i-\theta_i}{s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}}$ имеет распределение Стьюдента с $(N-m)$ степенями свободы, что и требовалось доказать. При доказательстве учитывалось, что случайные величины $\dfrac{\hat{\theta}_i-\theta_i}{\sigma\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}}$ и R_0^2 являются независимыми.

Доказательство: Сначала докажем, что случайные векторы невязки $\left(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}}\right)$ и оценки $\hat{\boldsymbol{\theta}}$ не коррелированы.

$$\begin{aligned}
\mathbf{M}\left[\left(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}}\right)\hat{\boldsymbol{\theta}}^T\right] &= \mathbf{M}\left[\left(\mathbf{Y}-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}\right)\left(\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}\right)^T\right] = \\
&= \mathbf{M}\left[\left(\mathbf{I}_N-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\mathbf{Y}\mathbf{Y}^T\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right] = \\
&= \mathbf{M}\left[\left(\mathbf{I}_N-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\left(\mathbf{F}\boldsymbol{\theta}+\boldsymbol{\xi}\right)\left(\mathbf{F}\boldsymbol{\theta}+\boldsymbol{\xi}\right)^T\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right] = \\
&= \left(\mathbf{I}_N-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\mathbf{M}\left(\boldsymbol{\xi}\boldsymbol{\xi}^T\right)\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \\
&= \left(\mathbf{I}_N-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\mathbf{I}_N\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \left(\mathbf{I}_N-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \\
&= \mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \mathbf{0},
\end{aligned}$$

$$\begin{aligned}
\mathbf{M}\left(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}}\right) &= \mathbf{M}\left(\mathbf{Y}-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}\right) = \left(\mathbf{I}_N-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\mathbf{M}\mathbf{Y} = \\
&= \left(\mathbf{I}_N-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\mathbf{M}\left(\mathbf{F}\boldsymbol{\theta}+\boldsymbol{\xi}\right) = \left(\mathbf{I}_N-\mathbf{F}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\right)\mathbf{F}\boldsymbol{\theta} = \mathbf{F}\boldsymbol{\theta}-\mathbf{F}\boldsymbol{\theta} = \mathbf{0}, \\
\mathbf{R}\left(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}\right) &= \mathbf{M}\left[\left(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}}\right)\hat{\boldsymbol{\theta}}^T\right]-\mathbf{M}\left(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}}\right)\mathbf{M}\hat{\boldsymbol{\theta}}^T = \mathbf{0}-\mathbf{0}\cdot\mathbf{M}\hat{\boldsymbol{\theta}}^T = \mathbf{0}.
\end{aligned}$$

Так как случайные векторы $\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}}$ и $\hat{\boldsymbol{\theta}}$ являются нормально распределёнными, то из их некоррелированности следует их независимость. Тогда также независимыми будут их функции: $R_0^2=\left(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{Y}-\mathbf{F}\hat{\boldsymbol{\theta}}\right)$ и $\dfrac{\hat{\theta}_i-\theta_i}{\sigma\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}}$.

Из (2.11) следует, что соотношение для определения индивидуальных интервальных оценок параметров $\theta_1,\theta_2,\ldots,\theta_m$ имеет вид:

$$\Pr\left(\left|\hat{\theta}_i-\theta_i\right|<t_{N-m}^{(1+p)/2}s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}\right)=p,$$

$$\Pr\left(\hat{\theta}_i - t_{N-m}^{(1+p)/2} s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}} < \theta_i < \hat{\theta}_i + t_{N-m}^{(1+p)/2} s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}\right) = p,$$

$\left(\hat{\theta}_i - t_{N-m}^{(1+p)/2} s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}},\ \hat{\theta}_i + t_{N-m}^{(1+p)/2} s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}\right)$ – доверительный интервал для параметра θ_i; p – доверительная вероятность; $s^2 = \dfrac{R_0^2}{N-m}$, $t_{N-m}^{(1+p)/2}$ – квантиль на уровне $(1+p)/2$ *t*-распределения Стьюдента с $(N-m)$ степенями свободы.

Для проверки статистической значимости параметра регрессии θ_i (то есть его отличия от нуля) также воспользуемся *t*-статистикой Стьюдента. При выполнении нулевой гипотезы $(H_0 : \theta_i = 0)$ статистика $t = \dfrac{\hat{\theta}_i}{s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}}$ имеет распределение Стьюдента с $(N-m)$ степенями свободы.

Задавшись уровнем значимости α, определим границы критической области из соотношения

$$\Pr\left(\left|\hat{\theta}_{ii}\right| < t_{N-m}^{1-\alpha/2} s\sqrt{\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]_{ii}}\right) = 1-\alpha,$$

где $s^2 = \dfrac{R_0^2}{N-m}$, $t_{N-m}^{1-\alpha/2}$ – квантиль на уровне $(1-\alpha/2)$ *t*-распределения Стьюдента с $(N-m)$ степенями свободы.

3 ДИСПЕРСИОННЫЙ АНАЛИЗ

Дисперсионным анализом называется статистический метод выявления влияния отдельных факторов на результат эксперимента путём исследования *значимости* различий в средних значениях. В литературе, в частности, при описании математических пакетов часто встречается обозначение ***ANOVA*** (от английского ***AN****alysis* ***O****f* ***VA****riance* – дисперсионный анализ). Фундаментальная концепция дисперсионного анализа предложена английским статистиком Рональдом Фишером в 1920 году. Изобретённая им *SS*-технология статистиченского анализа (от английского ***S****um of* ***S****quares* – сумма квадратов) является достаточно удобной и наглядной. Эта технология основана на математической статистике с использованием вероятностных распределений *хи-квадрат* и *Фишера–Снедекора* (приложение А). Мы будем использовать полученные ранее результаты для линейного регрессионного анализа при гауссовском шуме наблюдения (см. главу 2).

3.1 Вероятностные распределения хи-квадрат и Фишера

Если независимые случайные величины $\xi_1, \xi_2, \ldots, \xi_k$ имеют нормальное распределение с нулевым математическим ожиданием и единичной дисперсией, то случайная величина $\chi_k^2 = \sum_{i=1}^{k} \xi_i^2$ имеет *хи-квадрат*-распределение с k степенями свободы:

$$\sum_{i=1}^{k} \xi_i^2 \sim \chi^2(k). \tag{3.1}$$

Если нормальный случайный вектор $\boldsymbol{\eta} = (\eta_1\ \eta_2\ \eta_3\ \ldots\ \eta_m)^T$ имеет нулевое математическое ожидание и дисперсионную матрицу $\mathbf{D}_{\boldsymbol{\eta}}$, то случайная величина $\chi_m^2 = \boldsymbol{\eta}^T \mathbf{D}_{\boldsymbol{\eta}}^{-1} \boldsymbol{\eta}$ имеет *хи-квадрат*-распределение с m степенями свободы:

$$\chi_m^2 = \boldsymbol{\eta}^T \mathbf{D}_{\boldsymbol{\eta}}^{-1} \boldsymbol{\eta} \sim \chi^2(m). \tag{3.2}$$

Если независимые случайные величины χ_k^2 и χ_l^2 имеют χ^2 -распределение соответственно с k и l степенями свободы, то случайная величина $F_{k,l} = \dfrac{\chi_k^2 / k}{\chi_l^2 / l}$ имеет распределение Фишера (*F*-распределение, распределение Фишера–Снедекора) с k и l степенями свободы:

$$\frac{\chi_k^2 / k}{\chi_l^2 / l} \sim F(k,l). \tag{3.3}$$

3.2 Статистика Фишера

Статистика

$$F_{m,N-m} = \frac{1}{ms^2} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \mathbf{F}^T \mathbf{F} \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \sim F(m, N-m) \tag{3.4}$$

имеет распределение Фишера с m и $(N-m)$ степенями свободы.

Доказательство. Из соотношения $\frac{R_0^2}{\sigma^2} \sim \chi^2\left(N-m\right)$ (см. п.2.7) следует, что величина $\frac{\left(N-m\right)s^2}{\sigma^2}$ имеет хи-квадрат распределение с $\left(N-m\right)$ степенями свободы. Из гауссовского распределения вектора оценки $\hat{\boldsymbol{\theta}} \sim \mathrm{N}\left(\boldsymbol{\theta}, \sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right)$ следует, что случайный вектор $\boldsymbol{\eta} = \frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\sigma}$ является гауссовским с нулевым математическим ожиданием и дисперсионной матрицей $\mathbf{D}_{\boldsymbol{\eta}} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}$. Из (3.2) следует, что $\boldsymbol{\eta}^T\left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]^{-1}\boldsymbol{\eta} \sim \chi^2\left(m\right)$, то есть величина $\left(\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\sigma}\right)^T \mathbf{F}^T\mathbf{F}\left(\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\sigma}\right)$ имеет распределение хи-квадрат с m степенями свободы. Так как остаточная сумма квадратов R_0^2 и вектор оценки $\hat{\boldsymbol{\theta}}$ вероятностно независимы (доказывается в [2]), то также будут независимыми и функции от них: $\frac{\left(N-m\right)s^2}{\sigma^2}$ и $\left(\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\sigma}\right)^T \mathbf{F}^T\mathbf{F}\left(\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\sigma}\right)$. В соответствии с (3.3) отношение двух независимых случайных величин, имеющих распределения хи-квадрат и нормированных на их число степеней свободы, подчиняется закону распределения Фишера:

$$\frac{\left(\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\sigma}\right)^T \mathbf{F}^T\mathbf{F}\left(\frac{\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}}{\sigma}\right) \Big/ m}{\frac{\left(N-m\right)s^2}{\sigma^2} \Big/ \left(N-m\right)} \sim F\left(m, N-m\right),$$

то есть $\frac{1}{ms^2}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^T \mathbf{F}^T\mathbf{F}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \sim F\left(m, N-m\right)$.

Статистка Фишера (3.4) используется для проверки гипотезы о значимости линейной регрессионной модели в целом относительно нулевой гипотезы $H_0 : \boldsymbol{\theta} = \mathbf{0}$ (то есть $H_0 : \theta_1 = \theta_2 = \theta_3 = \ldots = \theta_m = 0$). Регрессионная модель считается значимой $\left(\boldsymbol{\theta} \neq \mathbf{0}\right)$ на уровне значимости α, если рассчитанное значение статистики

$$F = \frac{1}{ms^2}\hat{\boldsymbol{\theta}}^T\mathbf{F}^T\mathbf{F}\hat{\boldsymbol{\theta}}$$

превышает критическое значение $F_{m,N-m}^{1-\alpha}$ – квантиля на уровне $\left(1-\alpha\right)$ распределения Фишера с m и $\left(N-m\right)$ степенями свободы.

3.3 МНК-оценка линейной векторной функции коэффициентов регрессии

Линейная векторная функция параметров регрессии задаётся как

$$\boldsymbol{\tau} = \mathbf{T}\boldsymbol{\theta},$$

где $\mathbf{T}$ – произвольная матрица полного ранга линейного преобразования параметров, имеющая размер $k \times m$, $1 \le k \le m$ ($\operatorname{rank}\mathbf{T} = k$).

Тогда НЛН-оценка вектора $\boldsymbol{\tau}$ может быть определена через МНК-оценку параметров $\hat{\boldsymbol{\theta}}$ (2.9):

$$\hat{\boldsymbol{\tau}} = \mathbf{T}\hat{\boldsymbol{\theta}} = \mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}. \qquad (3.5)$$

Оценка $\hat{\boldsymbol{\tau}}$ (3.5) является несмещённой $\left(\mathbf{M}\hat{\boldsymbol{\tau}}=\boldsymbol{\tau}\right)$, а её дисперсионная матрица (дисперсионная матрица ошибки оценивания) равна

$$\mathbf{D}\left(\hat{\boldsymbol{\tau}}-\boldsymbol{\tau}\right)=\mathbf{D}\hat{\boldsymbol{\tau}}=\sigma^2\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T.$$

Основное назначение матрицы $\mathbf{T}$ – выделять статистически значимые компоненты регрессионной модели при проверке гипотез. Например, при $m=5,\ k=3$ матрица

$$\mathbf{T}=\begin{bmatrix}1&0&0&0&0\\0&0&1&0&0\\0&0&0&1&0\end{bmatrix}$$

выделяет из первоначального набора из пяти параметров $\boldsymbol{\theta}=\left(\theta_1\ \theta_2\ \theta_3\ \theta_4\ \theta_5\right)^T$ набор из трёх параметров $\boldsymbol{\tau}=\left(\theta_1\ \theta_2\ \theta_4\right)^T$ для анализа.

Другое назначение – приведение гипотезы о проверке параметров к нулевой гипотезе. Например, если требуется проверить гипотезу о равенстве четырёх параметров регрессии $H_0:\theta_1=\theta_2=\theta_3=\theta_4$, то есть $\theta_1-\theta_2=0$, $\theta_1-\theta_3=0$, $\theta_1-\theta_4=0$, то эквивалентная нулевая гипотеза $H_0':\boldsymbol{\tau}=\mathbf{0}$ получается при использовании линейного преобразования параметров:

$$\boldsymbol{\tau}=\mathbf{T}\boldsymbol{\theta}=\begin{bmatrix}1&-1&0&0\\1&0&-1&0\\1&0&0&-1\end{bmatrix}\begin{pmatrix}\theta_1\\\theta_2\\\theta_3\\\theta_4\end{pmatrix}=\mathbf{0},\ \text{где}\ \mathbf{T}=\begin{bmatrix}1&-1&0&0\\1&0&-1&0\\1&0&0&-1\end{bmatrix}.$$

Ниже без доказательства приведены вероятностные распределения и основные свойства оценки $\hat{\boldsymbol{\tau}}$. Доказательства аналогичны приведённым выше.

- Оценка $\hat{\boldsymbol{\tau}}=\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}$ является НЛН-оценкой $\boldsymbol{\tau}$.

- Вектор оценки $\hat{\boldsymbol{\tau}}$ имеет многомерное нормальное распределение с математическим ожиданием $\mathbf{M}\hat{\boldsymbol{\tau}} = \boldsymbol{\tau}$ и дисперсионной матрицей $\mathbf{D}\hat{\boldsymbol{\tau}} = \sigma^2 \mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T$:

$$\hat{\boldsymbol{\tau}} \sim \mathrm{N}\left(\boldsymbol{\tau}, \sigma^2 \mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T\right).$$

- Остаточная сумма квадратов R_0^2 и вектор оценки $\hat{\boldsymbol{\tau}}$ вероятностно независимы.
- Статистика

$$F = \frac{1}{ks^2}\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\right)^T \left[\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T\right]^{-1}\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\right), \qquad (3.6)$$

имеет распределение Фишера с k и $\left(N-m\right)$ степенями свободы: $F \sim F\left(k, N-m\right)$, где k – число строк в матрице $\mathbf{T}$.

3.4 Однофакторный двухуровневый дисперсионный анализ

Для начала рассмотрим простейший случай дисперсионного анализа, относящийся к исследованию значимости влияния какого-либо одного фактора на выходной фактор. На этом примере рассмотрим основные понятия и закономерности дисперсионного анализа.

Однофакторный двухуровневый дисперсионный анализ при одинаковом числе наблюдений решает простейшую задачу исследования значимости различия между средними значениями.

Рассмотрим наблюдения двух независимых выборок $\left\{y_1^1, y_1^2, y_1^3, \ldots, y_1^n\right\}$ и $\left\{y_2^1, y_2^2, y_2^3, \ldots, y_2^n\right\}$ из двух нормальных генеральных совокупностей Y_1 и Y_2 с одинаковыми дисперсиями σ^2, но различными математическими ожиданиями a и b:

$$y_1^i = a + \xi_1^i,\ i = \overline{1,n},$$

$$y_2^i = b + \xi_2^i,\ i = \overline{1,n},$$

где ξ_1^i и ξ_2^i можно интерпретировать как шум наблюдения с нулевым математическим ожиданием и дисперсией σ^2. Эти уравнения можно представить как результат наблюдения в классической линейной регрессионной модели (2.6) $\mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi}$ при $N = 2n,\ m = 2,$

$$\boldsymbol{\theta} = \left(a\ b\right)^T, \qquad \mathbf{Y} = \left(y_1^1\ y_1^2\ y_1^3 \ldots y_1^n\ y_2^1\ y_2^2\ y_2^3 \ldots y_2^n\right)^T,$$

$$\boldsymbol{\xi} = \left(\xi_1^1\ \xi_1^2\ \xi_1^3 \ldots \xi_1^n\ \xi_2^1\ \xi_2^2\ \xi_2^3 \ldots \xi_2^n\right)^T,$$

$$\mathbf{F} = \begin{bmatrix} \mathtt{111\ldots1000\ldots0} \\ \mathtt{000\ldots0111\ldots1} \end{bmatrix}^T,\ \mathbf{D}\boldsymbol{\xi} = \sigma^2\mathbf{I}_{2n}.$$

Тогда

$$\mathbf{F}^T\mathbf{F} = n\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = n\mathbf{I}_2,\ \left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \frac{1}{n}\mathbf{I}_2,\ \mathbf{F}^T\mathbf{Y} = \left(\sum_{i=1}^n y_1^i \quad \sum_{i=1}^n y_2^i\right)^T,$$

$$\hat{\boldsymbol{\theta}} = \left(\hat{a}\ \hat{b}\right)^T = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y} = \left(\overline{y}_1\ \ \overline{y}_2\right)^T, \text{ то есть } \hat{a} = \overline{y}_1,\ \hat{b} = \overline{y}_2,$$

где обозначено: $\overline{y}_1 = \frac{1}{n}\sum_{i=1}^n y_1^i,\ \overline{y}_2 = \frac{1}{n}\sum_{i=1}^n y_2^i.$

Для проверки гипотезы $H_0 : a = b$ используется статистика Фишера (3.6). При линейном преобразовании вектора параметров

$$\mathbf{T} = \left[\mathtt{1\ -1}\right],\ k = 1$$

получаем, что гипотеза $H_0 : a = b$ эквивалентна гипотезе $H_0' : \boldsymbol{\tau} = \mathbf{0},$

$$\left[\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T\right]^{-1}=\frac{n}{2},$$

$$\hat{\boldsymbol{\tau}}=\mathbf{T}\hat{\boldsymbol{\theta}}=\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}=\overline{y}_1-\overline{y}_2.$$

Остаточная сумма квадратов ошибок:

$$R_0^2=\sum_{i=1}^{n}\left(y_1^i-\overline{y}_1\right)^2+\sum_{i=1}^{n}\left(y_2^i-\overline{y}_2\right)^2=$$

$$=\sum_{i=1}^{n}\left(y_1^i\right)^2-2\overline{y}_1\sum_{i=1}^{n}y_1^i+n\overline{y}_1^2+\sum_{i=1}^{n}\left(y_2^i\right)^2-2\overline{y}_2\sum_{i=1}^{n}y_2^i+n\overline{y}_2^2=$$

$$=n\left(\overline{y_1^2}-\overline{y}_1^2\right)+n\left(\overline{y_2^2}-\overline{y}_2^2\right).$$

Несмещённая оценка дисперсии шума наблюдения: $s^2=\dfrac{n\left(\overline{y_1^2}-\overline{y}_1^2\right)+n\left(\overline{y_2^2}-\overline{y}_2^2\right)}{2n-2}$.

Статистика Фишера (3.6):

$$F=\frac{1}{ks^2}\hat{\boldsymbol{\tau}}^T\left[\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T\right]^{-1}\hat{\boldsymbol{\tau}}=\frac{\dfrac{n}{2}\left(\overline{y}_1-\overline{y}_2\right)^2}{\dfrac{n\left(\overline{y_1^2}-\overline{y}_1^2\right)+n\left(\overline{y_2^2}-\overline{y}_2^2\right)}{2n-2}}= \tag{3.7}$$

$$=\frac{(n-1)\left(\overline{y}_1-\overline{y}_2\right)^2}{\left(\overline{y_1^2}-\overline{y}_1^2\right)+\left(\overline{y_2^2}-\overline{y}_2^2\right)}\sim F\left(1,2n-2\right).$$

Обозначим через $F_{1,2n-2}^{1-\alpha}$ квантиль на уровне $(1-\alpha)$ распределения Фишера с $(1,2n-2)$ степенями свободы. Если $F_{1,2n-2}>F_{1,2n-2}^{1-\alpha}$, то гипотезу H_0 о равенстве математических ожиданий отвергаем на уровне значимости α, то есть считаем, что различие математических ожиданий существенно, значимо ($a\neq b$). Если $F_{1,2n-2}\le F_{1,2n-2}^{1-\alpha}$, то считаем, что эксперименталь-

ные данные не противоречат гипотезе $H_0 : a = b$ (различие математических ожиданий несущественно и объясняется наличием шума наблюдения).

Найдём ***общую*** изменчивость результирующего фактора y (сумму квадратов отклонений значений результирующего фактора от его среднего значения $\overline{y}$ по всем наблюдениям), которая обозначается как $SS_{общ}$ (от английского ***S****um of* ***S****quares*). Иначе $SS_{общ}$ обозначается как R_1^2. В общем виде в рамках классической линейной регрессионной модели с $N = 2n$ наблюдениями:

$$SS_{общ} = R_1^2 = \sum_{i=1}^{N}\left(y^i - \overline{y}\right)^2 = \sum_{i=1}^{N}\left(y^i\right)^2 - 2\sum_{i=1}^{N} y^i \overline{y} + \sum_{i=1}^{N}\overline{y}^2 =$$

$$= N\overline{y^2} - 2N\overline{y}^2 + N\overline{y}^2 = N\overline{y^2} - N\overline{y}^2 = n\overline{y_1^2} + n\overline{y_2^2} - \frac{n}{2}\left(\overline{y}_1 + \overline{y}_2\right)^2.$$

Изменчивость результирующего фактора y в результате воздействия ***шума наблюдения*** (обозначается $SS_{ош}$ – сумма квадратов ошибок, или R_0^2) описывается известной нам остаточной суммой квадратов ошибок

$$SS_{ош} = R_0^2 = \sum_{i=1}^{n}\left(y_1^i - \overline{y}_1\right)^2 + \sum_{i=1}^{n}\left(y_2^i - \overline{y}_2\right)^2 =$$

$$= \sum_{i=1}^{n}\left(y_1^i\right)^2 - 2\overline{y}_1\sum_{i=1}^{n} y_1^i + \sum_{i=1}^{n}\overline{y}_1^2 + \sum_{i=1}^{n}\left(y_1^i\right)^2 - 2\overline{y}_1\sum_{i=1}^{n} y_1^i + \sum_{i=1}^{n}\overline{y}_1^2 =$$

$$= n\left(\overline{y_1^2} - \overline{y}_1^2\right) + n\left(\overline{y_2^2} - \overline{y}_2^2\right).$$

Изменчивость результирующего фактора y в результате различных уровней ***входного фактора*** (обозначается $SS_{факт}$ – сумма квадратов отклонений значений выходного фактора, со-

ответствующих различным значениям входного фактора, от общего среднего):

$$SS_{факт} = \sum_{i=1}^{n} \left(\overline{y}_1 - \overline{y} \right)^2 + \sum_{i=1}^{n} \left(\overline{y}_2 - \overline{y} \right)^2 .$$

Для нашего случая однофакторного двухуровневого дисперсионного анализа можно получить более простое выражение для $SS_{факт}$:

$$\begin{aligned} SS_{факт} &= n\left(\overline{y}_1 - \overline{y} \right)^2 + n\left(\overline{y}_2 - \overline{y} \right)^2 = \\ &= n\left(\overline{y}_1 - \frac{\overline{y}_1 + \overline{y}_2}{2} \right)^2 + n\left(\overline{y}_2 - \frac{\overline{y}_1 + \overline{y}_2}{2} \right)^2 = \\ &= n\left(\frac{\overline{y}_1 - \overline{y}_2}{2} \right)^2 + n\left(\frac{\overline{y}_2 - \overline{y}_1}{2} \right)^2 = \frac{n}{2}\left(\overline{y}_1 - \overline{y}_2 \right)^2 . \end{aligned}$$

Докажем «закон сохранения суммы квадратов»

$$SS_{общ} = SS_{факт} + SS_{ош} \qquad (3.8)$$

на примере нашего частного случая:

$$\begin{aligned} SS_{общ} - SS_{ош} &= \sum_{i=1}^{n} \left(y_1^i - \overline{y} \right)^2 + \sum_{i=1}^{n} \left(y_2^i - \overline{y} \right)^2 - \sum_{i=1}^{n} \left(y_1^i - \overline{y}_1 \right)^2 - \sum_{i=1}^{n} \left(y_2^i - \overline{y}_2 \right)^2 = \\ &= \cancel{\sum_{i=1}^{n} \left(y_1^i \right)^2} - 2\overline{y} \sum_{i=1}^{n} y_1^i + n\overline{y}^2 + \cancel{\sum_{i=1}^{n} \left(y_2^i \right)^2} - 2\overline{y} \sum_{i=1}^{n} y_2^i + n\overline{y}^2 - \\ &- \cancel{\sum_{i=1}^{n} \left(y_1^i \right)^2} + 2\overline{y}_1 \sum_{i=1}^{n} y_1^i - n\overline{y}_1^2 - \cancel{\sum_{i=1}^{n} \left(y_2^i \right)^2} + 2\overline{y}_2 \sum_{i=1}^{n} y_2^i - n\overline{y}_2^2 = \\ &= -2n\overline{y}\,\overline{y}_1 + n\overline{y}^2 - 2n\overline{y}\,\overline{y}_2 + n\overline{y}^2 + n\overline{y}_1^2 + n\overline{y}_2^2 = n\left(\overline{y}_1 - \overline{y} \right)^2 + n\left(\overline{y}_2 - \overline{y} \right)^2 = \\ &= SS_{факт}. \end{aligned}$$

На самом деле вычисление статистики Фишера на основе соотношения (3.6) является частным случаем гораздо более

глобального соотношения, которое будет рассмотрено ниже для других моделей наблюдения. Общий подход к проверке гипотезы о значимости влияния входных факторов на результат основывается на вычислении статистики Фишера в виде:

$$F_{K,N-m} = \frac{SS_{факт}/K}{SS_{ош}/(N-m)} \sim F(K, N-m), \qquad (3.9)$$

где N – общее число наблюдений,
m – число оцениваемых параметров,
K – число связей (ограничений), наложенных на входные факторы проверяемой гипотезой.

В нашем случае $N=2n$, $m=2$, $K=1$.

$$\frac{SS_{факт}/K}{SS_{ош}/(N-m)} = \frac{\frac{n}{2}(\bar{y}_1 - \bar{y}_2)^2 \Big/ 1}{\left(n\left(\overline{y_1^2} - \bar{y}_1^2\right) + n\left(\overline{y_2^2} - \bar{y}_2^2\right)\right)\Big/(2n-2)} =$$

$$= \frac{(n-1)(\bar{y}_1 - \bar{y}_2)^2}{\left(\overline{y_1^2} - \bar{y}_1^2\right) + \left(\overline{y_2^2} - \bar{y}_2^2\right)} \ .$$

Видим, что результат совпадает с (3.7). Важным является то, что расчёт статистики Фишера на основе общего подхода является более простым и наглядным.

3.5 Однофакторный многоуровневый дисперсионный анализ

В табл. 3.1 приведены данные для однофакторного многоуровнего дисперсионного анализа. Уравнения линейной регрессии в скалярном виде имеют вид:

$$y_i^k = \theta_i + \xi_i^k \ , \ i = \overline{1,m} \ , \ k = \overline{1,n} \ , \qquad (3.10)$$

где $\theta_1, \theta_2, \theta_3, \ldots, \theta_m$ – влияния значений входного фактора различных уровней на результат, m – число уровней входного фактора, n – число наблюдений на каждом уровне входного фактора (всего $N = mn$ наблюдений), ξ_i^k – некоррелированный гауссовский шум наблюдения с нулевым математическим ожиданием и дисперсией σ^2.

Таблица 3.1 – **Основная таблица дисперсионного анализа**

Номер уровня входного фактора	Значение входного фактора	Влияние уровня входного фактора на результат	Значения выходного фактора					Среднее значение
			1	2	3	…	n	
1	$\mathbf{x}_1$	θ_1	y_1^1	y_1^2	y_1^3	…	y_1^n	$\overline{y}_1$
2	$\mathbf{x}_2$	θ_2	y_2^1	y_2^2	y_2^3	…	y_2^n	$\overline{y}_2$
…	…	…	…	…	…	…	…	…
m	$\mathbf{x}_m$	θ_m	y_m^1	y_m^2	y_m^3	…	y_m^n	$\overline{y}_m$
Общее среднее: $\overline{y} = \frac{1}{mn}\sum_{k=1}^{m}\sum_{i=1}^{n} y_k^i$								$\overline{y}$

В традиционной матричной записи уравнения (3.10) имеют вид:

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\theta} + \boldsymbol{\xi},$$

где $\boldsymbol{\theta} = \left(\theta_1\ \theta_2\ \theta_3\ \ldots\ \theta_m\right)^T$ – вектор параметров,

$\mathbf{Y} = \left(y_1^1\ y_1^2 \ldots y_1^n\ \ y_2^1\, y_2^2 \ldots y_2^n\ \ \ldots\ \ y_m^1 \ldots y_m^n\right)^T$ – вектор наблюдений,

$\boldsymbol{\xi} = \left(\xi_1^1\ \xi_1^2 \ldots \xi_1^n\ \ \xi_2^1\ \xi_2^2 \ldots \xi_2^n\ \ \ldots\ \ \xi_m^1 \ldots \xi_m^n\right)^T$ – вектор шума наблюдения, который считается гауссовским с нулевым математическим ожиданием и дисперсионной матрицей $\mathbf{D}\boldsymbol{\xi} = \sigma^2 \mathbf{I}_N$, $N = mn$,

$$\mathbf{F}^T = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \ldots & \mathbf{1} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \ldots & \ldots & \ldots & \ldots & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{1} & \mathbf{1} & \ldots & \mathbf{1} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \ldots & \ldots & \ldots & \ldots & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{1} & \mathbf{1} & \ldots & \mathbf{1} & \ldots & \ldots & \ldots & \ldots & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \ldots & \ldots & \ldots & \ldots & \mathbf{1} & \mathbf{1} & \ldots & \mathbf{1} \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ \\ m \end{matrix}$$

$$\begin{matrix} 1 & 2 & \ldots & n & & & & & & & & & & & & & & & mn \end{matrix}$$

Тогда

$$\mathbf{F}^T\mathbf{F} = \begin{bmatrix} n & 0 & \ldots & 0 \\ 0 & n & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & n \end{bmatrix} = nI_m,$$

$$\left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \begin{bmatrix} 1/n & 0 & \ldots & 0 \\ 0 & 1/n & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1/n \end{bmatrix} = \frac{1}{n}\mathbf{I}_m,$$

$$\mathbf{F}^T\mathbf{Y} = \begin{pmatrix} \sum\limits_{i=1}^{n} y_1^i \\ \sum\limits_{i=1}^{n} y_2^i \\ \ldots \\ \sum\limits_{i=1}^{n} y_m^i \end{pmatrix} = n\begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \ldots \\ \overline{y}_m \end{pmatrix},$$

$$\widehat{\boldsymbol{\theta}} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y} = \frac{1}{n}\mathbf{I}_m n\begin{pmatrix}\overline{y}_1\\ \overline{y}_2\\ \dots\\ \overline{y}_m\end{pmatrix} = \begin{pmatrix}\overline{y}_1\\ \overline{y}_2\\ \dots\\ \overline{y}_m\end{pmatrix}.$$

МНК-оценки параметров: $\widehat{\theta}_k = \overline{y}_k = \frac{1}{n}\sum_{i=1}^{n} y_k^i,\ k = \overline{1,m}$.

Для проверки гипотезы $H_0 : \theta_1 = \theta_2 = \theta_3 = \ldots = \theta_m$ (нулевая гипотеза о незначимом влиянии различных уровней входного фактора на результат) воспользуемся статистикой Фишера (3.6): $F = \frac{1}{ks^2}\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\right)^T\left[\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T\right]^{-1}\left(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}\right) \sim F\left(k, N-m\right)$ при следующих значениях параметров: $\boldsymbol{\tau} = \mathbf{0}$, $k = m-1$. Матрицу $\mathbf{T}$ определим из условия гипотезы $\mathbf{T}\boldsymbol{\theta} = \mathbf{0}$:

$$\mathbf{T} = \begin{bmatrix} 1 & -1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 & \dots & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & \dots & 0 & 0 & -1 \end{bmatrix} \begin{matrix} 1 \\ 2 \\ 3 \\ \dots \\ m-2 \\ m-1 \end{matrix}.$$
$$\begin{matrix} 1 & 2 & 3 & & & & & m \end{matrix}$$

Явный вид статистики (3.6):

$$F_{m-1,m(n-1)} = \frac{mn\left(n-1\right)\sum_{k=1}^{m}\left(\overline{y}_k - \overline{y}\right)^2}{\left(m-1\right)\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i - \overline{y}_k\right)^2} \sim F\left(m-1, nm-m\right). \quad (3.11)$$

Доказательство проведём с использованием классической модели линейной регрессии при гауссовом шуме наблюдения, затем полученный результат сравним с *SS*-технологией Фишера.

1. Регрессионная модель.

$$\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T=\begin{bmatrix}1&-1&0&0&\dots&0&0&0\\1&0&-1&0&\dots&0&0&0\\1&0&0&-1&\dots&0&0&0\\\dots&\dots&\dots&\dots&\dots&\dots&\dots&\dots\\1&0&0&0&\dots&0&-1&0\\1&0&0&0&\dots&0&0&-1\end{bmatrix}_{(m-1)\times m}\cdot\frac{1}{n}\cdot\mathbf{I}_m\begin{bmatrix}1&1&1&1&\dots&1&1&1\\-1&0&0&0&\dots&0&0&0\\0&-1&0&0&\dots&0&0&0\\\dots&\dots&\dots&\dots&\dots&\dots&\dots&\dots\\0&0&0&0&\dots&0&-1&0\\0&0&0&0&\dots&0&0&-1\end{bmatrix}_{m\times(m-1)}=$$

$$=\frac{1}{n}\begin{bmatrix}2&1&1&1&\dots&1&1&1\\1&2&1&1&\dots&1&1&1\\1&1&2&1&\dots&1&1&1\\\dots&\dots&\dots&\dots&\dots&\dots&\dots&\dots\\1&1&1&1&\dots&1&2&1\\1&1&1&1&\dots&1&1&2\end{bmatrix}_{(m-1)\times(m-1)}=\frac{1}{n}\left(\mathbf{I}_{m-1}+\mathbf{E}_{m-1}\right),$$

где

$$\mathbf{E}_{m-1}=\begin{bmatrix}1&1&1&1&\dots&1&1&1\\1&1&1&1&\dots&1&1&1\\1&1&1&1&\dots&1&1&1\\\dots&\dots&\dots&\dots&\dots&\dots&\dots&\dots\\1&1&1&1&\dots&1&1&1\\1&1&1&1&\dots&1&1&1\end{bmatrix}_{(m-1)\times(m-1)},$$

$$\mathbf{I}_{m-1}=\begin{bmatrix}1&0&0&0&\dots&0&0&0\\0&1&0&0&\dots&0&0&0\\0&0&1&0&\dots&0&0&0\\\dots&\dots&\dots&\dots&\dots&\dots&\dots&\dots\\0&0&0&0&\dots&0&1&0\\0&0&0&0&\dots&0&0&1\end{bmatrix}_{(m-1)\times(m-1)}.$$

$\left[\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T\right]^{-1}=\left[\frac{1}{n}\left(\mathbf{I}_{m-1}+\mathbf{E}_{m-1}\right)\right]^{-1}=n\left(\mathbf{I}_{m-1}-\frac{1}{m}\mathbf{E}_{m-1}\right)$, так как

$$\left(\mathbf{I}_{m-1}+\mathbf{E}_{m-1}\right)\left(\mathbf{I}_{m-1}-\frac{1}{m}\mathbf{E}_{m-1}\right)=\mathbf{I}_{m-1}+\mathbf{E}_{m-1}-\frac{1}{m}\mathbf{E}_{m-1}-\frac{1}{m}\mathbf{E}_{m-1}^2=$$

$$=\mathbf{I}_{m-1}+\frac{m-1}{m}\mathbf{E}_{m-1}-\frac{m-1}{m}\mathbf{E}_{m-1}=\mathbf{I}_{m-1}.$$

(Здесь использовано матричное тождество для матриц из единиц: $\mathbf{E}_m\cdot\mathbf{E}_m=m\mathbf{E}_m$).

$$\hat{\boldsymbol{\tau}}=\mathbf{T}\hat{\boldsymbol{\theta}}=\begin{bmatrix}1&-1&0&0&\dots&0&0&0\\1&0&-1&0&\dots&0&0&0\\1&0&0&-1&\dots&0&0&0\\\dots&\dots&\dots&\dots&\dots&\dots&\dots&\dots\\1&0&0&0&\dots&0&-1&0\\1&0&0&0&\dots&0&0&-1\end{bmatrix}\begin{pmatrix}\hat{\theta}_1\\\hat{\theta}_2\\\hat{\theta}_3\\\dots\\\hat{\theta}_m\end{pmatrix}=\begin{bmatrix}\bar{y}_1-\bar{y}_2\\\bar{y}_1-\bar{y}_3\\\bar{y}_1-\bar{y}_4\\\cdots\\\bar{y}_1-\bar{y}_m\end{bmatrix},$$

$$\hat{\boldsymbol{\tau}}^T\left[\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T\right]^{-1}\hat{\boldsymbol{\tau}}=\hat{\boldsymbol{\tau}}^T n\left(\mathbf{I}_{m-1}-\frac{1}{m}\mathbf{E}_{m-1}\right)\hat{\boldsymbol{\tau}}=n\hat{\boldsymbol{\tau}}^T\hat{\boldsymbol{\tau}}-\frac{n}{m}\hat{\boldsymbol{\tau}}^T\mathbf{E}_{m-1}\hat{\boldsymbol{\tau}}=$$

$$=n\sum_{k=2}^{m}\left(\bar{y}_1-\bar{y}_k\right)^2-\frac{n}{m}\sum_{k=2}^{m}\sum_{i=2}^{m}\left(\bar{y}_1-\bar{y}_k\right)\left(\bar{y}_1-\bar{y}_i\right)=n\sum_{k=1}^{m}\left(\bar{y}_1-\bar{y}_k\right)^2-\frac{n}{m}\sum_{k=1}^{m}\sum_{i=1}^{m}\left(\bar{y}_1-\bar{y}_k\right),$$

$$\left(\bar{y}_1-\bar{y}_i\right)=\underline{nm\bar{y}_1^2}-\underline{\underline{2nm\bar{y}_1\bar{y}}}+n\sum_{k=1}^{m}\bar{y}_k^2-\underline{nm\bar{y}_1^2}+\underline{\underline{2nm\bar{y}_1\bar{y}}}-nm\bar{y}^2=n\sum_{k=1}^{m}\bar{y}_k^2-nm\bar{y}^2=$$

$$=n\sum_{k=1}^{m}\bar{y}_k^2-2nm\bar{y}^2+nm\bar{y}^2=n\sum_{k=1}^{m}\bar{y}_k^2-2n\bar{y}\sum_{k=1}^{m}\bar{y}_k+n\sum_{k=1}^{m}\bar{y}^2=n\sum_{k=1}^{m}\left(\bar{y}_k-\bar{y}\right)^2,$$

$$ks^2=\left(m-1\right)\frac{R_0^2}{N-m}=\frac{m-1}{N-m}\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i-\hat{\theta}_k\right)^2=\frac{m-1}{nm-m}\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i-\bar{y}_k\right)^2,$$

$$F=\frac{1}{ks^2}\hat{\boldsymbol{\tau}}^T\left[\mathbf{T}\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{T}^T\right]^{-1}\hat{\boldsymbol{\tau}}=\frac{n\sum_{k=1}^{m}\left(\bar{y}_k-\bar{y}\right)^2}{\frac{m-1}{nm-m}\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i-\bar{y}_k\right)^2}=$$

$$=\frac{mn\left(n-1\right)\sum_{k=1}^{m}\left(\bar{y}_k-\bar{y}\right)^2}{\left(m-1\right)\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i-\bar{y}_k\right)^2}.$$

2. SS-технология.

$$F = \frac{SS_{факт}/(m-1)}{SS_{ош}/(N-m)},$$

$$SS_{ош} = R_0^2 = \sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i - \bar{y}_k\right)^2,\ SS_{факт} = \sum_{k=1}^{m}\sum_{i=1}^{n}\left(\bar{y}_k - \bar{y}\right)^2 = n\sum_{k=1}^{m}\left(\bar{y}_k - \bar{y}\right)^2,$$

$$F = \frac{SS_{факт}/(m-1)}{SS_{ош}/(nm-m)} = \frac{n\sum_{k=1}^{m}\left(\bar{y}_k - \bar{y}\right)^2 \Big/ (m-1)}{\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i - \bar{y}_k\right)^2 \Big/ (nm-m)} =$$

$$= \frac{mn(n-1)\sum_{k=1}^{m}\left(\bar{y}_k - \bar{y}\right)^2}{(m-1)\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i - \bar{y}_k\right)^2} \sim F(m-1, nm-m).$$

В основе дисперсионного анализа лежит разделение дисперсии на части или компоненты. В обозначениях, принятых в дисперсионном анализе (*SS* – от английского ***S**um of **S**quares*):

$SS_{общ} = R_1^2$	компонента, описывающая общую дисперсию (изменчивость) выходного фактора;
$SS_{ош} = R_0^2$	компонента, описывающая влияние шума наблюдения на выходной фактор;
$SS_{факт} = R_1^2 - R_0^2$	компонента, описывающая влияние входного фактора (входных факторов) на выходной фактор.

Доказательство зависимости между указанными компонентами

$$SS_{общ} = SS_{факт} + SS_{ош}$$

аналогично (3.8) («закон сохранения суммы квадратов»).

3.6 Двухфакторный дисперсионный анализ при произвольном числе наблюдений

В этом случае уравнения линейной регрессии в скалярном виде имеют вид:

$$y_{ij}^k = \theta_{ij} + \xi_{ij}^k,\ i = \overline{1, m_1},\ j = \overline{1, m_2},\ k = \overline{1, n_{ij}},$$

где θ_{ij} – совместное влияние значения первого входного фактора с номером уровня i и второго входного фактора с номером уровня j на результат; m_1 – число уровней первого входного фактора; m_2 – число уровней второго входного фактора; n_{ij} – число наблюдений выходного фактора при значениях первого входного фактора с номером уровня i и второго входного фактора с номером уровня j (всего $m = m_1 m_2$ неизвестных параметров и $N = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij}$ наблюдений); y_{ij}^k – значение выходного фактора в k-м наблюдении при сочетании воздействия первого входного фактора с номером уровня i и второго входного фактора с номером уровня j; ξ_{ij}^k – некоррелированный гауссовский шум наблюдения с нулевым математическим ожиданием и дисперсией σ^2. Будем считать, что число наблюдений n_{ij} при каждом сочетании уровней двух входных факторов не является одинаковым (полагаем, что $\forall(i, j)\ n_{ij} \neq 0$).

В табл. 3.2 обозначено: $n_i = \sum_{j=1}^{m_2} n_{ij}$, $n_j = \sum_{i=1}^{m_1} n_{ij}$, $n = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij}$.

Таблица 3.2 – **Таблица средних для двухфакторного многоуровнего дисперсионного анализа при неодинаковом числе наблюдений сочетаний двух факторов**

Уровни первого фактора	*Уровни второго фактора*								*Общие средние*	
	1		*2*		*3*		…	m_2		
1	$\overline{y}_{11}$	n_{11}	$\overline{y}_{12}$	n_{12}	$\overline{y}_{13}$	n_{13}	…	$\overline{y}_{1m_2}$ n_{1m_2}	$\overline{y}_1$	n_1
2	$\overline{y}_{21}$	n_{21}	$\overline{y}_{22}$	n_{22}	$\overline{y}_{23}$	n_{23}	…	$\overline{y}_{2m_2}$ n_{2m_2}	$\overline{y}_2$	n_2
…	…	…	…	…	…	…	…	… …	…	…
m_1	$\overline{y}_{m_1 1}$	$n_{m_1 1}$	$\overline{y}_{m_1 2}$	$n_{m_1 2}$	$\overline{y}_{m_1 3}$	$n_{m_1 3}$	…	$\overline{y}_{m_1 m_2}$ $n_{m_1 m_2}$	$\overline{y}_{m_1}$	n_{m_1}
Общие средние	$\overline{y}_1$	n_1	$\overline{y}_2$	n_2	$\overline{y}_3$	n_3	…	$\overline{y}_{m_2}$ n_{m_2}	$\overline{y}$	n

Параметры статистики Фишера:

$N = n = \sum_{i=1}^{m_1} \sum_{j=1}^{m_2} n_{ij}$ – общее число наблюдений;

$m = m_1 m_2$ – число оцениваемых параметров

$\theta_{11}, \theta_{12}, \ldots, \theta_{1m_2}, \theta_{21}, \theta_{22}, \ldots, \theta_{2m_2}, \ldots, \theta_{m_1 1}, \theta_{m_1 2}, \ldots, \theta_{m_1 m_2}$;

$K = m - 1 = m_1 m_2 - 1$ – число связей, накладываемых проверяемой гипотезой

$H_0 : \theta_{11} = \theta_{12} = \ldots = \theta_{1m_2} = \theta_{21} = \theta_{22} = \ldots = \theta_{2m_2} = \ldots = \theta_{m_1 1} = \theta_{m_1 2} = \ldots = \theta_{m_1 m_2}$.

Средние значения:

$\overline{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ij}^k$ – среднее значение наблюдаемого (выходного, результирующего) фактора при воздействии

первого фактора с номером уровня i, $i=\overline{1,m_1}$ и второго фактора с номером уровня j, $j=\overline{1,m_2}$;

$\overline{y}_i=\frac{1}{n_i}\sum_{j=1}^{m_2}n_{ij}\overline{y}_{ij}$ – среднее значение наблюдаемого фактора при воздействии первого фактора с номером уровня i, $i=\overline{1,m_1}$;

$\overline{y}_j=\frac{1}{n_j}\sum_{i=1}^{m_1}n_{ij}\overline{y}_{ij}$ – среднее значение наблюдаемого фактора при воздействии второго фактора с номером уровня j, $j=\overline{1,m_2}$;

$\overline{y}=\frac{1}{n}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{n_{ij}}y_{ij}^k=\frac{1}{n}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}n_{ij}\overline{y}_{ij}=\frac{1}{n}\sum_{j=1}^{m_2}n_j\overline{y}_j=\frac{1}{n}\sum_{i=1}^{m_1}n_i\overline{y}_i$ – глобальное среднее значение наблюдаемого фактора;

$SS_{общ}=\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{n_{ij}}\left(y_{ij}^k-\overline{y}\right)^2$ – общая изменчивость выходного фактора.

Заметим, что МНК-оценка неизвестных параметров в рамках используемой модели наблюдения всегда равна среднему значению выходного фактора при заданном сочетании входных факторов:

$$\hat{\theta}_{ij}=\overline{y}_{ij},\ i=\overline{1,m_1},\ j=\overline{1,m_2}.$$

Поэтому сумма квадратов ошибок наблюдения равна

$$SS_{ош}=R_0^2=\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{n_{ij}}\left(y_{ij}^k-\hat{\theta}_{ij}\right)^2=\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\sum_{k=1}^{n_{ij}}\left(y_{ij}^k-\overline{y}_{ij}\right)^2.$$

Статистика Фишера:

$$F = \frac{\left(SS_{общ} - SS_{ош}\right)/\left(m-1\right)}{SS_{ош}/\left(N-m\right)} \sim F\left(m-1, N-m\right).$$

3.7 Общий подход к дисперсионному анализу

Повторим ещё раз *SS*-технологию Фишера применительно к самому общему случаю дисперсионного анализа для произвольного числа входных факторов. Изменчивость (вариабельность) наблюдаемого (результирующего) фактора определяется: 1) изменчивостью шума наблюдения; 2) изменчивостью значений входных факторов. Проверка гипотезы о значимости влияния входных факторов на результат основывается на вычислении статистики Фишера:

$$F = \frac{SS_{факт}/K}{SS_{ош}/\left(N-m\right)} = \frac{\left(SS_{общ} - SS_{ош}\right)/K}{SS_{ош}/\left(N-m\right)} \sim F\left(K, N-m\right). \quad (3.12)$$

$SS_{общ}$ описывает общую изменчивость (дисперсию, вариабельность) выходного фактора и вычисляется как сумма квадратов отклонений наблюдаемого выходного фактора от его среднего значения:

$$SS_{общ} = R_1^2 = \sum_{\mathbf{i}} \left(y^{\mathbf{i}} - \bar{y}\right)^2, \ \bar{y} = \frac{1}{N}\sum_{\mathbf{i}} y^{\mathbf{i}}, \ \mathbf{i} = \left\{\left(i_1, i_2, i_3 \ldots\right)\right\},$$

где $\bar{y}$ – среднее значение выходного фактора (по всем наблюдениям $y^{\mathbf{i}}$); N – общее число наблюдений (опытов); $\mathbf{i}$ – значение вектора индексов, определяющее все используемые сочетания значений входных факторов и номера опытов для каждого сочетания значений входных факторов.

Компонента $SS_{ош}$ описывает влияние шума наблюдения и других случайных факторов на выходной фактор и рассчитывается как остаточная сумма квадратов значений невязки:

$$SS_{ош} = R_0^2 = \sum_{\mathbf{k}} \sum_{\mathbf{i}} \left(y_{\mathbf{k}}^{\mathbf{i}} - \overline{y}_{\mathbf{k}} \right)^2 ,$$

где $\overline{y}_{\mathbf{k}} = \dfrac{1}{n_{\mathbf{k}}} \sum_{\mathbf{i}} y_{\mathbf{k}}^{\mathbf{i}}$ – среднее значение выходного фактора в группе $\mathbf{k}$; $\mathbf{k} = \left\{ \left(k_1, k_2, k_3 \ldots \right) \right\}$ – вектор индексов групп входных факторов.

Компонента $SS_{факт}$ описывает влияние значений исследуемых входных факторов на результат наблюдения (выходной фактор):

$$SS_{факт} = \sum_{\mathbf{k}} n_{\mathbf{k}} \left(\overline{y}_{\mathbf{k}} - \overline{y} \right)^2 .$$

Здесь $n_{\mathbf{k}}$ – число наблюдений в группе $\mathbf{k}$, $\overline{y}_{\mathbf{k}}$ – среднее значение выходного фактора в группе $\mathbf{k}$. Эту факторную компоненту общей изменчивости можно также определить как

$$SS_{факт} = SS_{общ} - SS_{ош} .$$

Через K обозначено число связей (ограничений), наложенных на входные факторы проверяемой гипотезой.

Все рассмотренные выше модели дисперсионного анализа могут быть сведены к общей модели дисперсионного анализа. В табл. 3.3 показано, как это может быть выполнено. Результаты использования общей модели (3.12) совпадают с ранее полученными выражениями для расчёта статистики Фишера.

Таблица 3.3 – **Частные модели дисперсионного анализа**

$$F = \frac{SS_{факт}/K}{SS_{ош}/(N-m)} = \frac{(SS_{общ} - SS_{ош})/K}{SS_{ош}/(N-m)} \sim F(K, N-m)$$

Модель дисперсионного анализа	Число опытов, N	Число параметров, m	Связи входных факторов, накладываемые проверяемой гипотезой; число связей K	Сумма квадратов ошибок $SS_{ош}$	Общая изменчивость $SS_{общ}$
Однофакторная двухуровневая (п.п.3.2, 3.4)	$N = 2n$	$m = 2$	$a = b,\ K = 1$	$n\sum_{i=1}^{n}\left(y_1^i - \overline{y}_1\right)^2 + n\sum_{i=1}^{n}\left(y_2^i - \overline{y}_2\right)^2$	$\sum_{i=1}^{n}\left(y_1^i - \overline{y}\right)^2 + \sum_{i=1}^{n}\left(y_2^i - \overline{y}\right)^2$
Однофакторная многоуровневая (п.3.5)	$N = mn$	m	$\theta_1 = \theta_2 = \theta_3 = \ldots = \theta_m,\ K = m-1$	$\sum_{i=1}^{m}\sum_{k=1}^{n}\left(y_i^k - \overline{y}_i\right)^2$	$\sum_{i=1}^{m}\sum_{k=1}^{n}\left(y_i^k - \overline{y}\right)^2$
Двухфакторная многоуровневая (п.3.6)	$N = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} n_{ij}$	$m = m_1 m_2$	$\theta_{11} = \theta_{12} = \ldots = \theta_{1m_2} = \theta_{21} = \theta_{22} = \ldots = \theta_{2m_2} = \ldots = \theta_{m_1 1} = \theta_{m_1 2} = \ldots = \theta_{m_1 m_2},\ K = m_1 m_2 - 1$	$\sum_{k=1}^{n}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\left(y_{ij}^k - \overline{y}_{ij}\right)^2$	$\sum_{k=1}^{n}\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}\left(y_{ij}^k - \overline{y}\right)^2$
Общая (п.3.7)	$N = \sum_{\mathbf{k}} n_{\mathbf{k}}$	m	$K = m-1$	$\sum_{\mathbf{k}}\sum_{\mathbf{i}}\left(y_{\mathbf{k}}^{\mathbf{i}} - \overline{y}_{\mathbf{k}}\right)^2$	$\sum_{\mathbf{k}} n_{\mathbf{k}}\left(\overline{y}_{\mathbf{k}} - \overline{y}\right)^2$

Для проведения дисперсионного анализа не всегда нужно записывать регрессионную модель наблюдения и строить МНК-оценки параметров. Достаточно использовать общий подход, определяемый уравнением

3.8 Пример двухфакторного дисперсионного анализа

Исследование потребления топлива автомобилем. Исследователь хочет выяснить, оказывают ли тип потребляемого бензина и тип автомобиля влияние на расход топлива. Для этого будут использованы два типа бензина – обычный и высокооктановый, и для каждой группы будут использованы два типа автомобилей – с двумя ведущими колесами и с четырьмя. Для каждой группы будут использованы по два автомобиля, всего восемь. (Американский галлон ≈ 3,785 литра). Исходные данные заданы табл. 3.4.

Таблица 3.4 – **Исходные данные для исследования потребления топлива автомобилем**

Топливо [«92», «98»]	*Привод* [«2WD», «4WD»]	*Расход топлива* [миль/галлон]
92	2WD	**26.7**
92	2WD	**25.2**
92	4WD	**28.6**
92	4WD	**29.3**
98	2WD	**32.3**
98	2WD	**32.8**
98	4WD	**26.1**
98	4WD	**24.2**

Проверяем гипотезу $\mathbf{H_0}$, заключающуюся в том, что тип топлива и тип трансмиссии автомобиля не оказывают эффекта

на потребление бензина. Альтернативной является гипотеза $\mathbf{H_1}$, заключающаяся в том, что тип топлива **или** тип трансмиссии автомобиля оказывают влияние на потребление бензина. Таблица средних представлена в табл. 3.5.

Таблица 3.5 – **Таблица средних для двухфакторного дисперсионного анализа при одинаковом числе наблюдений сочетаний двух факторов**

Уровни первого фактора	*Уровни второго фактора*		*Общие средние*
	2WD	4WD	
«92»	25.95	28.95	27.45
«98»	32.55	25.15	28.85
Общие средние	29.25	27.05	**28.15**

Общая изменчивость (вариабельность) выходного фактора:

$$SS_{общ} = R_1^2 = \sum_{\mathbf{i}} \left(y^{\mathbf{i}} - \bar{y} \right)^2 = 70.98.$$

Остаточная сумма квадратов, характеризующая влияние шума наблюдения:

$$SS_{ош} = R_0^2 = \sum_{\mathbf{k}} \sum_{\mathbf{i}} \left(y_{\mathbf{k}}^{\mathbf{i}} - \bar{y}_{\mathbf{k}} \right)^2 =$$

$$= \left(26.7 - 25.95\right)^2 + \left(25.2 - 25.95\right)^2 + \ldots + \left(24.2 - 27.05\right)^2 = \mathbf{3.3}.$$

Изменчивость выходного фактора из-за влияния значений входных факторов:

$$SS_{факт} = \sum_{\mathbf{k}} n_{\mathbf{k}} \left(\bar{y}_{\mathbf{k}} - \bar{y} \right)^2 =$$

$$= 2\left(25.95 - 28.15\right)^2 + 2\left(28.95 - 28.15\right)^2 + 2\left(32.55 - 28.15\right)^2 +$$

$$+2\left(27.05 - 28.15\right)^2 = \mathbf{67.68}.$$

Или так: $SS_{факт} = SS_{общ} - SS_{ош} = 70.98 - 3.3 = \mathbf{67.68}$;

$$F = \frac{SS_{факт}/K}{SS_{ош}/(N-m)} = \frac{67.68/3}{3.3/(8-4)} = \mathbf{27.34545455} \sim$$
$$\sim F(K, N-m) = F(3,4).$$

K – число связей (ограничений), наложенных на входные факторы проверяемой гипотезой.

```
N = 8; m = 4; K = 3.
```

Проведём исследование гипотезы **H₀** при различных значениях уровня значимости α (табл. 36).

Таблица 3.6 – **Анализ результатов дисперсионного анализа**

Уровень значимости α	Квантиль распределения Фишера $F_{3,4}^{1-\alpha}$	Сравнение F и $F_{3,4}^{1-\alpha}$	Анализ выполнения гипотезы **H₀**
0.05	6.591382	$F > F_{3,4}^{0.95}$	Гипотеза **H₀** не подтверждается. Тип топлива **или** тип трансмиссии автомобиля оказывают влияние на потребление бензина. Вероятность ошибки равна 0.05
0.01	16.69437	$F > F_{3,4}^{0.99}$	Гипотеза **H₀** не подтверждается. Тип топлива **или** тип трансмиссии автомобиля оказывают влияние на потребление бензина. Вероятность ошибки равна 0.01

Окончание табл. 3.6

0.005	24.25912	$F > F_{3,4}^{0.995}$	Гипотеза **H_0** не подтверждается. Тип топлива **или** тип трансмиссии автомобиля оказывают влияние на потребление бензина. Вероятность ошибки равна 0.005
0.001	56.17719	$F < F_{3,4}^{0.999}$	Данные не противоречат гипотезе **H_0**. Гипотеза **H_0** принимается (тип топлива и тип трансмиссии автомобиля не оказывают эффекта на потребление бензина). Вероятность ошибки не установлена
0.003989453	**27.34545455**	$\mathbf{F = F_{3,4}^{0.996010547}}$	Граничное значение уровня значимости. Гипотеза **H_0** не подтверждается. Тип топлива **или** тип трансмиссии автомобиля оказывают влияние на потребление бензина. Вероятность ошибки равна 0.004

4 МЕТОДЫ И АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ ДАННЫХ

4.1 Задача кластеризации

Кластер (англ. *cluster*) переводится как «*группа», «рой», «пачка», «скопление», «сгусток», «связка», «гроздь»*. Употребляется этот термин в экономике, компьютерной инженерии, музыке, астрономии и означает множество объектов функционально схожих между собой и собранных в одну связку. Применительно к (интеллектуальному) анализу данных под кластерным анализом (*Data clustering*) понимается задача разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами, так чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Классификация, являясь наиболее простой задачей *Data Mining*, относится к стратегии «обучение с учителем», для ее решения обучающая выборка должна содержать значения как входных переменных, так и выходных (целевых) переменных. *Кластеризация*, напротив, является задачей *Data Mining*, относящейся к стратегии «обучение без учителя», т.е. не требует наличия значения целевых переменных в обучающей выборке. Таким образом, синонимами термина **«кластеризация»** являются *«автоматическая классификация», «обучение без учителя»* и *«таксономия»* (*taxonomy*).

Таксоно́мия (от древнегреческого *τάξις – строй, порядок и νόμος – закон*) – учение о принципах и практике *классификации* и

систематизации. Термин «таксономия» впервые был предложен в 1813 году Огюстеном Декандолем, занимавшимся классификацией растений, и изначально применялся только в биологии. Позже этот термин стал использоваться для обозначения общей теории классификации и систематизации в других областях знаний. Классическим примером таксономии на основе сходства является биноминальная номенклатура (деление на роды / виды живых существ), предложенная Карлом Линнеем в середине XVIII века. В современном представлении биологическая иерархия имеет около 30 уровней, 7 из них считаются основными: царство, тип, класс, отряд, семейство, род, вид. Аналогичные систематизации строятся во многих областях знания, чтобы упорядочить информацию о большом количестве объектов.

Кластеризация предназначена для разбиения совокупности объектов на однородные группы (кластеры или классы). Обычно данные представляют собой выборки точек в признаковом пространстве. Наибольшее применение кластеризация первоначально получила в таких науках как биология, антропология, психология. Кластерный анализ полезен, когда нужно классифицировать большое количество информации. Так, в медицине используется кластеризация заболеваний, лечения заболеваний или их симптомов, а также таксономия пациентов, препаратов и т.д. В археологии устанавливаются таксономии каменных сооружений и древних объектов и т.д. В маркетинге это может быть задача сегментации конкурентов и потребителей. В менеджменте примером задачи кластеризации будет разбиение персонала на различные группы, классификация потребителей и поставщиков, анализ поведения клиента-потребителя, выявление схожих производственных ситуаций, при которых возникает брак. В медицине – классификация симптомов. В социологии задача кластеризации – разбиение респондентов на однородные группы.

Алгоритмы кластеризации являются в большой степени *эвристическими*. Эвристический алгоритм – это алгоритм решения задачи, правильность которого для всех возможных случаев не доказана, но про который известно, что он даёт достаточно хорошее решение в большинстве случаев. В действительности может быть даже известно, что эвристический алгоритм формально неверен. Его всё равно можно применять, если при этом он даёт неверный результат только в отдельных, достаточно редких и хорошо выделяемых случаях или же даёт неточный, но всё же приемлемый результат. Проще говоря, эвристика – это не полностью математически обоснованный (или даже «не совсем корректный»), но при этом практически полезный алгоритм. Важно понимать, что эвристика, в отличие от корректного алгоритма решения задачи, обладает следующими особенностями:

- Она не гарантирует нахождение лучшего решения.
- Она не гарантирует нахождение решения, даже если оно заведомо существует (возможен «пропуск цели»).
- Она может дать неверное решение в некоторых случаях.

Формальная постановка задачи кластеризации

Пусть $\mathbf{X} = \left\{ x^1,\ x^2,\ x^3,\ \ldots,\ x^N \right\}$ – множество N объектов, заданных в n-мерном векторном признаковом пространстве: $x^k = \left(x_1^k\ \ x_2^k\ \ x_3^k\ \ldots\ x_n^k \right)^T$, $k = \overline{1, N}$,

$\mathbf{Y} = \{1,\ 2,\ 3,\ \ldots\}$ – множество номеров (имён, меток) кластеров.

Задана функция расстояния между объектами $\boldsymbol{\rho}\left(x, x' \right)$. Наиболее часто используется евклидова метрика:

$$\boldsymbol{\rho}_2 \left(x, x' \right) = \sqrt{\sum_{m=1}^{n} \left(x_m - x'_m \right)^2} .$$

Используются также

- манхэттенское расстояние (расстояние городских кварталов)

$$\boldsymbol{\rho}_1(x, x') = \sum_{j=1}^{n} \left| x_j - x'_j \right|;$$

- максимальное различие по координатам

$$\boldsymbol{\rho}_\infty(x, x') = \max_{j \in [1;n]} \left| x_j - x'_j \right|.$$

Все эти расстояния являются частным случаем расстояния Минковского:

$$\boldsymbol{\rho}_p(x, y) = \left(\sum_{j=1}^{n} \left| x_j - y_j \right|^p \right)^{1/p}.$$

Имеется конечная выборка объектов $\mathbf{X} = \left\{ x^1, x^2, \ldots, x^N \right\}$. Требуется разбить эту выборку на $\boldsymbol{K}$ непересекающихся подмножеств S_k, $k = \overline{1, K}$; $\mathbf{X} = \bigcup_{k=1}^{K} S_k$, называемых кластерами, так чтобы каждый кластер состоял из объектов, близких по метрике $\boldsymbol{\rho}$, а объекты разных кластеров существенно отличались. При этом каждому объекту $x^i \in \mathbf{X}$ приписывается номер кластера $y_i \in \mathbf{Y}$. Алгоритм кластеризации – это функция $\mathbf{X} \to \mathbf{Y}$, которая любому объекту $x \in \mathbf{X}$ ставит в соответствие номер кластера $y \in \mathbf{Y}$. Множество $\mathbf{Y}$ в некоторых случаях известно заранее, однако чаще ставится задача определить оптимальное число кластеров с точки зрения того или иного критерия качества кластеризации.

Кластеризация (обучение без учителя) отличается от классификации (обучения с учителем) тем, что метки исходных объектов y_i изначально не заданы и даже может быть неизвестно само множество $\mathbf{Y}$, т.е. число объектов в кластере.

Центром кластера S_k (центроидом) называется геометрический центр точек кластера k в евклидовом пространстве:

$X_k = \frac{1}{|S_k|} \sum_{x^i \in S_k} x^i$, где $|S_k|$ – число точек в кластере k, $k = 1, 2, 3, \ldots, K$; K – число кластеров.

Дисперсия кластера – это мера рассеяния точек в пространстве относительно центра кластера:

$$D_k = \frac{1}{|S_k|} \sum_{x^i \in S_k} \boldsymbol{\rho}^2 \left(x^i, X_k \right).$$

Радиус кластера – это тоже мера рассеяния точек в пространстве относительно центра кластера – максимальное расстояние до центра кластера:

$$R_k = \max_{x^i \in S_k} \boldsymbol{\rho} \left(x^i, X_k \right).$$

Цели кластеризации могут быть различными в зависимости от особенностей конкретной прикладной задачи:

- Понять структуру множества объектов $\mathbf{X}$, разбив его на группы схожих объектов. Упростить дальнейшую обработку данных и принятия решений, работая с каждым кластером по отдельности (стратегия «разделяй и властвуй»).
- Сократить объём хранимых данных в случае сверхбольшой выборки $\mathbf{X}$, оставив по одному наиболее типичному представителю от каждого кластера.
- Выделить нетипичные объекты, которые не подходят ни к одному из кластеров. Эту задачу называют одноклассовой классификацией, обнаружением нетипичности или новизны (novelty detection).

В первом случае число кластеров стараются сделать поменьше. Во втором случае важнее обеспечить высокую степень сходства объектов внутри каждого кластера, а кластеров может быть сколько угодно. В третьем случае наибольший интерес представляют отдельные объекты, не вписывающиеся ни в один из кластеров.

Во всех этих случаях может применяться иерархическая кластеризация, когда крупные кластеры дробятся на более мелкие, те, в свою очередь, дробятся ещё мельче и т. д. Такие задачи называются задачами таксономии (taxonomy). Результатом таксономии является не простое разбиение множества объектов на кластеры, а древообразная иерархическая структура. Вместо номера кластера объект характеризуется перечислением всех кластеров, которым он принадлежит: от крупного к мелкому. Таксономии строятся во многих областях знания, чтобы упорядочить информацию о большом количестве объектов. Мы будем рассматривать алгоритмы иерархической кластеризации, позволяющие автоматизировать процесс построения таксономий.

Решение задачи кластеризации принципиально неоднозначно по следующим причинам:

1. Не существует однозначно наилучшего критерия качества кластеризации. Известен целый ряд эвристических критериев, а также ряд алгоритмов, не имеющих чётко выраженного критерия, но осуществляющих достаточно разумную кластеризацию «по построению». Все они могут давать разные результаты. Следовательно, для определения качества кластеризации требуется эксперт предметной области, который бы мог оценить осмысленность выделения кластеров.

2. Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием. Это справедливо только для методов дискриминации, так

как в методах кластеризации выделение кластеров идёт за счёт формализованного подхода на основе мер близости.

3. Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом. Но стоит отметить, что есть ряд рекомендаций к выбору мер близости для различных задач.

4.2 Метод k-средних (k-means)

К настоящему времени разработано более сотни различных алгоритмов кластеризации, в результате применения которых получаются неодинаковые результаты, что объясняется особенностью работы того или иного алгоритма, ориентированного на решение *конкретной* задачи.

Наиболее популярным является метод *k-средних*. Название «*k-means*» было впервые использовано James MacQueen в 1967 году. Этот алгоритм относится к классу эвристических EM-алгоритмов (англ. *Expectation-maximization*), используемых в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей. Каждая итерация алгоритма состоит из двух шагов. На E-шаге (*expectation*) вычисляется ожидаемое значение функции правдоподобия. На M-шаге (*maximization*) вычисляется оценка максимального правдоподобия, увеличивая ожидаемое правдоподобие, вычисляемое на E-шаге. Затем это значение используется для E-шага на следующей итерации. Алгоритм выполняется до сходимости.

Основная идея алгоритма *k-средних* заключается в минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров, то есть

$$J = \sum_{k=1}^{K} \sum_{x_i \in S_k} \boldsymbol{\rho}\left(x^i, X_k\right) \to \min_{S_1, S_2, \ldots, S_K},$$

где K – число кластеров (считается известным); S_k – полученные кластеры; X_k – центр масс (центроид) векторов из кластера S_k.

Алгоритм заключается в том, что на каждой итерации заново вычисляется центр масс X'_k для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике:

$$x^i \in S'_k, \text{ если } \boldsymbol{\rho}\left(x^i, X'_k\right) = \min_{k=\overline{1,K}}.$$

На рис. 4.1 показан пример работы алгоритма k-means.

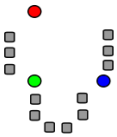
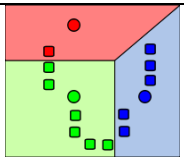
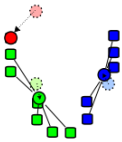
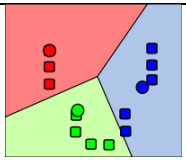
Шаг 1		Начальное расположение 10 объектов и случайно выбранные центры кластеров (красная, зелёная и синяя точки)
Шаг 2		Точки, отнесённые к кластерам по критерию близости к центру кластера по евклидовому расстоянию
Шаг 3		Вычисление новых центров кластеров
Шаг 4		Переход к шагу 2 с новым расположением центров кластеров или остановка, если нет изменений в кластерах

Рис. 4.1. Пример алгоритма k-means для трёх кластеров (https://ru.wikipedia.org/wiki/K-means)

Алгоритм завершается, когда на какой-то итерации не происходит изменения центра масс кластеров. Число итераций конечно, так как конечно множество возможных разбиений конечного множества векторов, а зацикливания не происходит, так как J не увеличивается. Доказано (*David Arthur*, *Sergei Vassilvitskii*; 2006), что на некоторых классах множеств вычислительная сложность алгоритма (число итераций) равна $2^{\Omega(\sqrt{n})}$, где $\Omega(\cdot)$ – полиномиальная сложность.

Алгоритм кластеризация k-средних

Начало

Сформировать начальное приближение центров всех ***K*** *кластеров, взяв наиболее удалённые друг от друга объекты выборки* $X_k^{(0)},\ k \in Y = \{1,2,3,\ldots,K\}$.

Repeat

Отнести каждый объект к ближайшему центру, разбивая пространство признаков и формируя новые кластеры $S_k^{(j)},\ k \in \{1,2,3,\ldots,K\}$*:*

$$\forall x_i,\ i = \overline{1,N}:\ k = \arg\min_{k \in \overline{1,K}} \boldsymbol{\rho}\left(x^i, X_k^{(j-1)}\right), \quad S_k^{(j)} = S_k^{(j)} \bigcup x^i.$$

Вычислить новое положение центров:

$$X_k^{(j)} = \frac{1}{\left|S_k^{(j)}\right|} \sum_{x_i \in S_k^{(j)}} x_i.$$

Перейти к следующей итерации:

$$j := j+1$$

Until

Пока состав кластеров $S_k^{(j)}$ не перестанет изменяться.

Недостатки алгоритма *k-means*:

- Не гарантируется достижение глобального минимума суммарного квадратичного отклонения J, а только одного из локальных минимумов.
- Результат зависит от начального выбора центров кластеров $\left\{X_k^{(0)}\right\}$, их оптимальный выбор неизвестен.
- Число кластеров K надо знать заранее.

Пример 1. Дан набор из 8 точек на плоскости:

	A	B	C	D	E	F	G	H
	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x_1	**1**	**4**	**5**	**6**	**1**	**5**	**1**	**2**
x_2	**3**	**3**	**3**	**3**	**2**	**2**	**1**	**1**

Выполнить кластеризацию точек по алгоритмау *k-means*: число кластеров равно 2, метрика – манхеттенская.

K=2. Произвольные начальные центры кластеров: $X_1^{(0)} = \begin{pmatrix} 5 \\ 2 \end{pmatrix}$; $X_2^{(0)} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$.

Начальные расстояния до центров кластеров и отнесение в кластер:

	A	B	C	D	E	F	G	H
	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x_1	**1**	**4**	**5**	**6**	**1**	**5**	**1**	**2**
x_2	**3**	**3**	**3**	**3**	**2**	**2**	**1**	**1**
$\boldsymbol{\rho_1}\left(x^i, \begin{pmatrix} 5 \\ 2 \end{pmatrix}\right)$	5	2	1	2	4	0	5	4
$\boldsymbol{\rho_1}\left(x^i, \begin{pmatrix} 4 \\ 3 \end{pmatrix}\right)$	3	0	1	2	4	2	5	4
Кластер	**2**	**2**	**1**	**1**	**1**	**1**	**1**	**1**

При одинаковых расстояниях до центров кластеров принадлежность кластеру вектора признаков назначается произвольно (для определённости выбираем первый кластер).

Центры кластеров на следующем шаге:

$$X_1^{(1)} = \begin{pmatrix} 10/3 \\ 2 \end{pmatrix}; \quad X_2^{(1)} = \begin{pmatrix} 5/2 \\ 3 \end{pmatrix}.$$

Новые расстояния до центров и отнесение в кластер:

	A	B	C	D	E	F	G	H
	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x_1	**1**	**4**	**5**	**6**	**1**	**5**	**1**	**2**
x_2	**3**	**3**	**3**	**3**	**2**	**2**	**1**	**1**
$\boldsymbol{\rho_1}\left(x^i, \begin{pmatrix} 10/3 \\ 2 \end{pmatrix}\right)$	10/3	5/3	8/3	11/3	7/3	5/3	10/3	7/3
$\boldsymbol{\rho_1}\left(x^i, \begin{pmatrix} 5/2 \\ 3 \end{pmatrix}\right)$	3/2	3/2	5/2	7/2	5/2	7/2	7/2	5/2
Кластер	**2**	**2**	**2**	**2**	**1**	**1**	**1**	**1**

Центры кластеров на следующем шаге:

$$X_1^{(2)} = \begin{pmatrix} 9/4 \\ 3/2 \end{pmatrix}; \quad X_2^{(2)} = \begin{pmatrix} 4 \\ 3 \end{pmatrix}.$$

Новые расстояния до центров и отнесение в кластер:

	A	B	C	D	E	F	G	H
	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x_1	**1**	**4**	**5**	**6**	**1**	**5**	**1**	**2**
x_2	**3**	**3**	**3**	**3**	**2**	**2**	**1**	**1**
$\boldsymbol{\rho_1}\left(x^i, \begin{pmatrix} 9/4 \\ 3/2 \end{pmatrix}\right)$	11/4	13/4	17/4	21/4	7/4	13/4	7/4	3/4
$\boldsymbol{\rho_1}\left(x^i, \begin{pmatrix} 4 \\ 3 \end{pmatrix}\right)$	3	0	1	2	4	2	5	4
Кластер	**1**	**2**	**2**	**2**	**1**	**2**	**1**	**1**

Центры кластеров на следующем шаге:

$$X_1^{(3)} = \begin{pmatrix} 5/4 \\ 7/4 \end{pmatrix}; \quad X_2^{(3)} = \begin{pmatrix} 5 \\ 11/4 \end{pmatrix}.$$

Новые расстояния до центров и отнесение в кластер:

	A	B	C	D	E	F	G	H
	x^1	x^2	x^3	x^4	x^5	x^6	x^7	x^8
x_1	**1**	**4**	**5**	**6**	**1**	**5**	**1**	**2**
x_2	**3**	**3**	**3**	**3**	**2**	**2**	**1**	**1**
$\boldsymbol{\rho_1}\left(x^i, \begin{pmatrix} 5/4 \\ 7/4 \end{pmatrix}\right)$	3/2	4	5	6	1/2	4	1	3/2
$\boldsymbol{\rho_1}\left(x^i, \begin{pmatrix} 5 \\ 11/4 \end{pmatrix}\right)$	17/4	5/4	1/4	5/4	19/4	3/4	23/4	19/4
Кластер	**1**	**2**	**2**	**2**	**1**	**2**	**1**	**1**

Видим, что состав кастеров на последней итерации не изменился. Поэтому итерации заканчиваются.

Ответ: $S_1 = \{A, E, G, H\}; \quad S_2 = \{B, C, D, F\}$.

Алгоритм кластеризации k-means++

Это расширение направлено на оптимальный, а не случайный **выбор начальных значений** центров кластеров (*David Arthur, Sergei Vassilvitskii*; 2007). Описание алгоритма:

1 Выбрать первый центроид случайным образом (среди всех точек).

2 Для каждой точки найти значение квадрата расстояния до ближайшего центроида (из тех, которые уже выбраны) dx^2.

3 Выбрать из этих точек следующий центроид так, чтобы вероятность выбора точки была пропорциональна вычисленному для неё квадрату расстояния. Это можно сделать следу-

ющим образом. На шаге 2 нужно параллельно с расчётом dx^2 подсчитывать сумму $S = \sum dx^2$. После накопления суммы взять случайное значение *R* из раномерно распределённого на отрезке $[0; S]$, по которому остаётся только определить, какой точке это соответствует. Для этого нужно снова начать подсчитывать сумму $S = \sum dx^2$ до тех пор, пока сумма не превысит *R*. Текущая точка берётся в качестве центроида.

4 При выборе каждого следующего центроида специально следить за тем, чтобы он не совпал с одной из уже выбранных в качестве центроидов точек, не нужно, так как вероятность повторного выбора некоторой точки равна 0.

5 Повторять шаги 2 и 3 до тех пор, пока не будут найдены все необходимые центроиды.

Далее выполняется основной алгоритм *k-means*.

4.3 Алгоритм кластеризации Isodata (ИСОМАД)

ИСОМАД (Итеративный самоорганизующийся метод анализа данных, **Isodata** – Iterative Self-Organizing Data Analysis Techniques) аналогичен методу ***k*-means**, однако обладает более широким набором параметров и вспомогательных эвристических процедур. Алгоритм эвристический, результат работы во многом зависит от исходных данных и заданных начальных значениях параметров.

Исходные данные: набор векторов $\{x_1, x_2, x_3, \ldots, x_N\}$. Параметры алгоритма: K – необходимое число кластеров; Q_N – параметр, с которым сравнивается количество выборочных об-

разов, вошедших в кластер; Q_S – параметр, характеризующий среднеквадратичное отклонение; Q_C – параметр, характеризующий компактность; L – максимальное количество пар центров кластеров, которые можно объединить; I – допустимое число циклов итерации.

Подробное описание алгоритма ***Isodata*** приведено в книге [1].

4.4 Иерархическая кластеризация

Алгоритмы **иерархической кластеризации** называются также графовыми алгоритмами кластеризации. **Дендрограмма** – дерево (граф без циклов), построенное по матрице мер близости. Дендрограмма позволяет изобразить взаимные связи между объектами из заданного множества. Для создания дендрограммы требуется матрица сходства, которая определяет уровень сходства между любой парой объектов.

Для **агломеративных методов** (*AGglomerative NESting*, ***AGNES***) – методы слияния – проводится рекурсивное поочерёдное слияние близких кластеров. При слиянии отдельные точки-объекты также считаются кластерами. На каждом шаге выбирается для слияния пара наиболее схожих кластеров. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Дивизивные методы (*DIvisive ANAlysis*, ***DIANA***) – методы деления – являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры и в результате образуется последовательность расщепляющих групп.

Различные алгоритмы отличаются выбором метрики измерения расстояний (см. п.4.1) и критерия схожести кластеров (приведены ниже).

Методы измерения расстояний между кластерами

Метод ближнего соседа. Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах.

Метод наиболее удаленных соседей. Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. «наиболее удаленными соседями»).

Метод Варда (Ward). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения. В отличие от других методов кластерного анализа для оценки расстояний между кластерами здесь используются методы дисперсионного анализа. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов.

Метод невзвешенного попарного среднего. В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них.

Метод взвешенного попарного среднего. Этот метод похож на метод невзвешенного попарного среднего, разница состоит лишь в том, что здесь в качестве весового коэффициента используется размер кластера (число объектов, содержащихся в кластере).

Невзвешенный центроидный метод. В качестве расстояния между двумя кластерами в этом методе берется расстояние между их геометрическими центрами.

Взвешенный центроидный метод. Для учета разницы между размерами кластеров (числе объектов в них) используются веса.

Пример. Необходимо провести кластеризацию пяти предприятий, каждое из которых характеризуется тремя переменными: x_1 – среднегодовая стоимость основных производственных фондов, млрд. руб.; x_2 – материальные затраты на 1 руб. произведенной продукции; x_3 – объем произведенной продукции, млрд. руб. Значения переменных приведены в табл. 4.1.

Таблица 4.1 – **Финансовые показатели производственных предприятий**

Номер предприятия	x_1	x_2	x_3
1	120,0	94,0	164,0
2	85,.0	75,2	92,0
3	145,0	81,0	120,0
4	78,0	76,8	86,0
5	70,0	75,9	104,0
Среднее значение, $\overline{x}_k$	99,6	80,6	113,2
Среднее квадратическое отклонение, s_k	28,4	10,9	27,9

Нормировка исходных данных:

$$z_k^i = \frac{x_k^i - \overline{x}_k}{s_k},$$

$$Z = \begin{pmatrix} 0.718 & 1.229 & 1.821 \\ -0.514 & -2.238 & -0.760 \\ 1.514 & 0.037 & 0.244 \\ -0.760 & -0.349 & -0.975 \\ -1{,}042 & -0.431 & -0.330 \end{pmatrix}.$$

Классификацию проведем при помощи иерархического **агломеративного** метода. Для построения матрицы расстояний воспользуемся квадратом евклидового расстояния. Тогда, например, квадрат расстояния между первым и вторым объектами:

$$d^2(1,2) = \left(0.718 - \left(-0.514\right)\right)^2 + \left(1.229 - \left(-2.238\right)\right)^2 +$$

$$+\left(1.821 - \left(-0.760\right)\right)^2 = 20.20,$$

квадрат расстояния между первым и третьим объектами:

$$d^2(1,3) = \left(0.718 - 1.514\right)^2 + \left(1.229 - 0.037\right)^2 +$$

$$+\left(1.821 - 0.244\right)^2 = 4.54$$ и т.д.

В результате получаем первоначальную матрицу расстояний, характеризующую расстояния между отдельными объектами, каждый из которых изначально является отдельным кластером:

$D_0^2 =$

	1	2	3	4	5
1	0	20.20	4.54	12.49	10.48
2		0	10.30	3.68	3.73
3			0	6.81	7.08
4				0	**0.50**
5					0

Как видно по элементам матрицы D_0^2, наиболее близкими являются объекты 4 и 5: $d^2(4,5)=0.50$. Объединим их в один кластер и присвоим ему метку 45. Пересчитаем расстояния всех оставшихся объектов (кластеров) до кластера 45. В матрице D_1^2 расстояния между кластерами определяются по алгоритму **«дальнего соседа»**.

Расстояние между кластером 1 и кластером 45:
$d^2(1,45) = \max(d^2(1,4), d^2(1,5)) = \max(12.49, 10.48) = 12.49$.

Расстояние между кластером 2 и кластером 45:
$d^2(2,45) = \max(d^2(2,4), d^2(2,5)) = \max(3.68, 3.73) = 3.73$.

Расстояние между кластером 3 и кластером 45:
$d^2(3,45) = \max(d^2(3,4), d^2(3,5)) = \max(6.81, 7.08) = 7.08$.

Получим новую матрицу расстояний:

$D_1^2 =$

	1	2	3	45
1	0	20.20	4.54	12.49
2		0	10.30	**3.73**
3		10.30	0	7.08
45				0

В матрице D_1^2 опять находим самые близкие кластеры. Это будут кластеры 2 и 45, $d^2(2,45)=3.73$. Следовательно, на этом шаге объединяем кластеры 2 и 45; получим новый кластер, состоящий из кластеров 2 и 45 (объекты 2,4,5). Присвоим ему метку 245. Теперь имеем три кластера: 1,245,3. Пересчитываем расстояния $d^2(1,245)$, $d^2(3,245)$ и получаем матрицу D_2:

$d^2(1,245) = \max(d^2(1,2), d^2(1,45)) =$
$= \max(20.2, 12.49) = 20.2,$
$d^2(3,245) = \max(d^2(3,2), d^2(3,45)) =$
$= \max(10.3, 7.08) = 10.3.$

$$D_2^2 = \begin{array}{c|ccc} & 1 & 245 & 3 \\ 1 & 0 & 20.20 & \mathbf{4.54} \\ 245 & & 0 & 10.30 \\ 3 & & & 0 \end{array}$$

На следующем шаге объединяем кластеры 1 и 3 ($d^2(1,3)=4.54$) в один кластер и присваиваем ему номер 13. Теперь имеем только два кластера: 13, 245.

$d^2(13,245) = \max(d^2(1,245), d^2(3,245)) =$
$= \max(20.20, 10.3) = 20.20$

И, наконец, на последнем шаге объединяем кластеры 13 и 245 на квадрате расстояния 20.20.

$$D_3^2 = \begin{array}{c|cc} & 13 & 245 \\ 13 & 0 & \mathbf{20.20} \\ 245 & & 0 \end{array}$$

Представим результаты классификации в виде дендрограммы (рис. 4.2). Дендрограмма свидетельствует о том, что кластер S_2 более однороден по составу входящих объектов, так как в нем объединение происходит при меньших расстояниях, чем в кластере S_1.

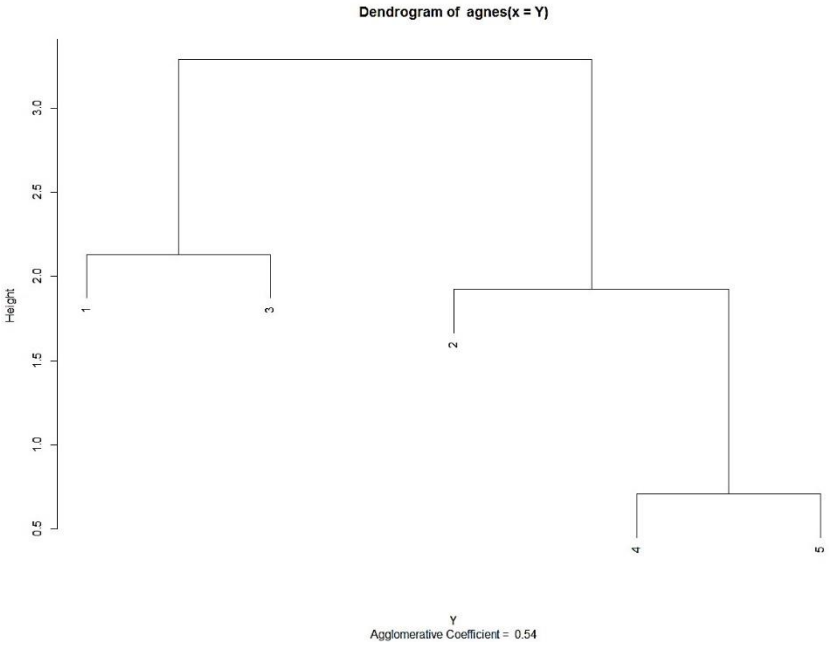


Рис. 4.2. Дендрограмма кластеризации пяти производственных предприятий

5 РАСПОЗНАВАНИЕ ОБРАЗОВ

Задачей распознавания образов является отнесение наблюдаемого объекта (вектора признаков) $\mathbf{x}$ к одному из классов $\{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3, \boldsymbol{\omega}_4, \ldots\}$. То есть требуется разбить всё множество $\mathbf{X}$ возможных значений признаков объектов на непересекающиеся области $G_1, G_2, G_3, G_4, \ldots$ из условия минимизации вероятности ошибки. Будем считать, что условная вероятность принадлежности объекта каждому из классов (априорная вероятность) известна или её можно определить/измерить на основе известной модели наблюдения.

5.1 Статистическая проверка гипотез

Простейшей задачей распознавания образов является задача статистической проверки гипотез из математической статистики. Рассмотрим **байесовское решающее правило**, минимизирующее вероятность ошибки решения, для случая двух альтернатив. Рассмотрим модель наблюдения, в которой признаки объекта являются непрерывными случайными величинами, которые описываются безусловной и условной плотностями вероятностей. Пусть x – вектор наблюдений признаков (объект); задача заключается в том, чтобы определить его принадлежность классу $\boldsymbol{\omega}_1$ (гипотеза H_1) или $\boldsymbol{\omega}_2$ (гипотеза H_2). Решающее правило основано на максимизации **апостериорной**

вероятности $\mathbf{Pr}\{\boldsymbol{\omega}_i|x\}$, вычисляемой по формуле Байеса (То́мас Байе́с (англ. *Reverend Thomas Bayes*) – английский математик XVIII века), которая для случая непрерывных признаков имеет вид:

$$\mathbf{Pr}\{\boldsymbol{\omega}_i|x\} = \frac{f(x|\boldsymbol{\omega}_i)\mathbf{p}_i}{f(x)},\ i = \overline{1,2}, \qquad (5.1)$$

$$\boldsymbol{\omega}_i = \arg\max_{\boldsymbol{\omega}_i} \mathbf{Pr}\{\boldsymbol{\omega}_i|x\} = \arg\max_{\boldsymbol{\omega}_i} \frac{f(x|\boldsymbol{\omega}_i)\mathbf{p}_i}{f(x)} =$$

$$= \arg\max_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} \left\{ \frac{f(x|\boldsymbol{\omega}_1)\mathbf{p}_1}{f(x)}, \frac{f(x|\boldsymbol{\omega}_2)\mathbf{p}_2}{f(x)} \right\},$$

где $\mathbf{p}_1 = \mathbf{Pr}\{\boldsymbol{\omega}_1\}$, $\mathbf{p}_2 = \mathbf{Pr}\{\boldsymbol{\omega}_2\}$ – **априорные** вероятности классов объектов (гипотез);

$f(x) = f(x|\boldsymbol{\omega}_1)\mathbf{p}_1 + f(x|\boldsymbol{\omega}_2)\mathbf{p}_2$ – безусловная плотность вероятности, $\mathbf{p}_1 + \mathbf{p}_2 = 1$;

$f(x|\boldsymbol{\omega}_1)$, $f(x|\boldsymbol{\omega}_2)$ – условные плотности вероятности.

Так как решающее правило (5.1) не зависит от значения знаменателя $f(x)$, его можно записать в виде

$$f(x|\boldsymbol{\omega}_1)\mathbf{p}_1 > f(x|\boldsymbol{\omega}_2)\mathbf{p}_2 \rightarrow x \in \boldsymbol{\omega}_1,$$

$$f(x|\boldsymbol{\omega}_1)\mathbf{p}_1 < f(x|\boldsymbol{\omega}_2)\mathbf{p}_2 \rightarrow x \in \boldsymbol{\omega}_2.$$

С использованием отношения правдоподобия $l(x)$ решающее правило принимает вид:

$$l(x)=\frac{f\left(x|\boldsymbol{\omega}_1\right)}{f\left(x|\boldsymbol{\omega}_2\right)}>h \;\rightarrow\; x\in\boldsymbol{\omega}_1,$$
$$l(x)=\frac{f\left(x|\boldsymbol{\omega}_1\right)}{f\left(x|\boldsymbol{\omega}_2\right)}<h \;\rightarrow\; x\in\boldsymbol{\omega}_2. \qquad (5.2)$$

где $h=\mathbf{p}_2/\mathbf{p}_1$ – пороговое значение.

Уравнение (5.2) называется **байесовским критерием**, минимизирующим ошибку решения. Вероятность ошибки классификации:

$$\varepsilon=\mathbf{Pr}\{ou\}=\mathbf{p_1}\mathbf{Pr}\{ou|\boldsymbol{\omega}_1\}+\mathbf{p}_2\mathbf{Pr}\{ou|\boldsymbol{\omega}_2\}=$$
$$=\mathbf{p_1}\mathbf{Pr}\{x\in G_2|\boldsymbol{\omega}_1\}+\mathbf{p}_2\mathbf{Pr}\{x\in G_1|\boldsymbol{\omega}_2\}=$$
$$=\mathbf{p_1}\int_{G_2} f\left(x|\boldsymbol{\omega}_1\right)dx+\mathbf{p}_2\int_{G_1} f\left(x|\boldsymbol{\omega}_2\right)dx=\mathbf{p}_1\varepsilon_1+\mathbf{p}_2\varepsilon_2.$$

Величины

$$\varepsilon_1=\int_{G_2} f\left(x|\boldsymbol{\omega}_1\right)dx \quad \text{и} \quad \varepsilon_2=\int_{G_1} f\left(x|\boldsymbol{\omega}_2\right)dx$$

называются вероятностями ошибок *первого рода* и *второго рода*.

В радиолокации эти ошибки называются вероятностью *пропуска цели* и вероятностью *ложной тревоги*.

В медицинской диагностике используются названия *чувствительность* $(1-\varepsilon_1)$ и *специфичность* $(1-\varepsilon_2)$.

Докажем, что байесовский классификатор обеспечивает минимальную вероятность ошибки. Воспользуемся тем фактом, что события $x\in G_1$ и $x\in G_2$ образуют полную группу:

$$\int_{G_2} f\left(x|\boldsymbol{\omega}_{\mathbf{i}}\right)dx+\int_{G_1} f\left(x|\boldsymbol{\omega}_{\mathbf{i}}\right)dx=1.$$

$$\varepsilon=\mathbf{p}_1\int_{G_2} f\left(x|\boldsymbol{\omega}_1\right)dx+\mathbf{p}_2\int_{G_1} f\left(x|\boldsymbol{\omega}_2\right)dx=$$

$$=\mathbf{p}_1\left(1-\int_{G_1} f\left(x|\boldsymbol{\omega}_1\right)dx\right)+\mathbf{p}_2\int_{G_1} f\left(x|\boldsymbol{\omega}_2\right)dx=$$

$$=\mathbf{p}_1+\int_{G_1}\left[\mathbf{p}_2 f\left(x|\boldsymbol{\omega}_2\right)-\mathbf{p}_1 f\left(x|\boldsymbol{\omega}_1\right)\right]dx\to\min_{G_1}.$$

Из последнего соотношения видно, что минимум вероятности ошибки достигается, когда в область G_1 включаются точки x, для которых значение подынтегральной функции отрицательно, то есть $\mathbf{p}_2 f\left(x|\boldsymbol{\omega}_2\right)-\mathbf{p}_1 f\left(x|\boldsymbol{\omega}_1\right)<0$ или $\dfrac{f\left(x|\boldsymbol{\omega}_1\right)}{f\left(x|\boldsymbol{\omega}_2\right)}>\dfrac{\mathbf{p}_2}{\mathbf{p}_1}\ \rightarrow\ x\in G_1$, что совпадает с правилом Байеса (5.2).

Пример 5.1. Построить байесовское решающее правило для двух классов независимых гауссовских наблюдений случайной величины X с плотностями вероятностей

$$f\left(x|\boldsymbol{\omega}_1\right)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-1)^2\right],$$

$$f\left(x|\boldsymbol{\omega}_2\right)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-3)^2\right]$$

и априорными вероятностями классов $\mathbf{p}_1=1/3,\ \mathbf{p}_2=2/3$.

Отношение правдоподобия:

$$l(x)=\frac{f\left(x|\boldsymbol{\omega}_1\right)}{f\left(x|\boldsymbol{\omega}_2\right)}=\frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(x-1)^2}{2}\right]}{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{(x-3)^2}{2}\right]}=$$

$$=\exp\left[\frac{(x-3)^2-(x-1)^2}{2}\right]=\exp(4-2x).$$

Получаем следующее решающее правило:

$\exp(4-2x)>\dfrac{\mathbf{p}_2}{\mathbf{p}_1}=2\to x\in\boldsymbol{\omega}_1$, то есть

если $x<2-\dfrac{\ln 2}{2}$, *то* $x\in\boldsymbol{\omega}_1$;

если $x>2-\dfrac{\ln 2}{2}$, *то* $x\in\boldsymbol{\omega}_2$.

Вероятность ошибки классификации:

$$\varepsilon=\mathbf{p}_1\mathbf{Pr}\left\{\textit{ошибка}\,|\boldsymbol{\omega}_1\right\}+\mathbf{p}_2\mathbf{Pr}\left\{\textit{ошибка}\,|\boldsymbol{\omega}_2\right\}=$$

$$=\mathbf{p}_1\mathbf{Pr}\left\{X\in G_2\,|\boldsymbol{\omega}_1\right\}+\mathbf{p}_2\mathbf{Pr}\left\{X\in G_1|\,\boldsymbol{\omega}_2\right\}=$$

$$=\frac{1}{3}\int_{2-\ln 2/2}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-1)^2\right]dx+\frac{2}{3}\int_{-\infty}^{2-\ln 2/2}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}(x-2)^2\right]dx=$$

$$=\frac{1}{3}\int_{1-\ln 2/2}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}x^2\right]dx+\frac{2}{3}\int_{-\infty}^{-\ln 2/2}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}x^2\right]dx=$$

$$=\frac{1}{3}\left(1-\Phi\left(1-\ln 2/2\right)\right)+\frac{2}{3}\Phi\left(-\ln 2/2\right)\approx 0.4135779.$$

Здесь $\Phi(z)=\dfrac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left(-\dfrac{x^2}{2}\right)dx$ – интегральная функция стандартного нормального распределения (интеграл вероятности, функция Лапласа).

При гауссовских наблюдениях x удобнее использовать логарифм отношения правдоподобия. Тогда решающее правило (5.2) для случая двух классов может быть записано в виде

$$\begin{aligned} -\ln l(x) = -\ln \frac{f(x|\boldsymbol{\omega}_1)}{f(x|\boldsymbol{\omega}_2)} < \ln \frac{\mathbf{p}_1}{\mathbf{p}_2} \quad \rightarrow \quad x \in \boldsymbol{\omega}_1, \\ -\ln l(x) = -\ln \frac{f(x|\boldsymbol{\omega}_1)}{f(x|\boldsymbol{\omega}_2)} > \ln \frac{\mathbf{p}_1}{\mathbf{p}_2} \quad \rightarrow \quad x \in \boldsymbol{\omega}_2. \end{aligned} \tag{5.3}$$

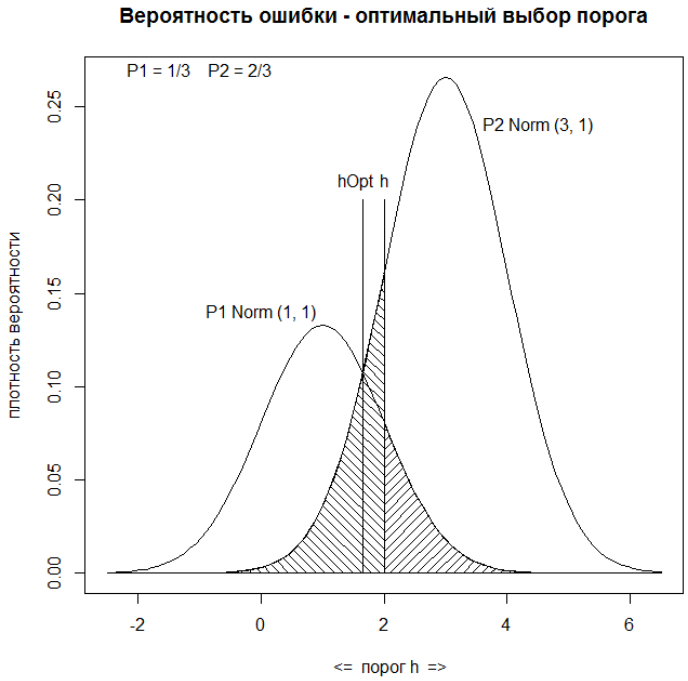


Рис. 5.1. Графическая иллюстрация оптимального выбора порога (к примеру 5.1)

Пример 5.2. Построить байесовское решающее правило (5.2) для двух классов независимых гауссовских наблюдений

двумерного случайного вектора x с математическими ожиданиями $\mathbf{a}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\mathbf{a}_2 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$, корреляционными матрицами $\mathbf{R}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$, $\mathbf{R}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ и априорными вероятностями классов $\mathbf{p}_1 = p$, $\mathbf{p}_2 = 1 - p$. Изобразить графически семейство разделяющих поверхностей для различных значений p.

$$f\left(\mathbf{x}|\boldsymbol{\omega}_1\right) = f\left(x_1, x_2 | \boldsymbol{\omega}_1\right) = \frac{1}{2\pi\sqrt{|\mathbf{R}_1|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{x}-\mathbf{a}_1\right)\right],$$

$$f\left(\mathbf{x}|\boldsymbol{\omega}_2\right) = f\left(x_1, x_2 | \boldsymbol{\omega}_2\right) = \frac{1}{2\pi\sqrt{|\mathbf{R}_2|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{x}-\mathbf{a}_2\right)\right].$$

Логарифм отношения правдоподобия:

$$-\ln l\left(\mathbf{x}\right) = -\ln \frac{f\left(\mathbf{x}|\boldsymbol{\omega}_1\right)}{f\left(\mathbf{x}|\boldsymbol{\omega}_2\right)} = -\ln \frac{\frac{1}{2\pi\sqrt{|\mathbf{R}_1|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{x}-\mathbf{a}_1\right)\right]}{\frac{1}{2\pi\sqrt{|\mathbf{R}_2|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{x}-\mathbf{a}_2\right)\right]} =$$

$$= -\ln \frac{|\mathbf{R}_2|}{|\mathbf{R}_1|} + \frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{x}-\mathbf{a}_1\right) - \frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{x}-\mathbf{a}_2\right) =$$

$$= \frac{1}{2}\begin{pmatrix} x_1 \\ x_2 - 1 \end{pmatrix}^T \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 - 1 \end{pmatrix} - \frac{1}{2}\begin{pmatrix} x_1 - 2 \\ x_2 \end{pmatrix}^T \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 - 2 \\ x_2 \end{pmatrix} =$$

$$= \frac{1}{2}x_1^2 + \frac{1}{4}\left(x_2 - 1\right)^2 - \frac{1}{4}\left(x_1 - 2\right)^2 - \frac{1}{2}x_2^2 = \frac{1}{4}\left(x_1 + 2\right)^2 - \frac{1}{4}\left(x_2 + 1\right)^2 - \frac{3}{4}.$$

Решающее правило:

$$\textit{Если}\quad -\ln l(\mathbf{x}) < -\ln\frac{1-p}{p},\quad \textit{то}\quad x\in\boldsymbol{\omega}_1;\quad \textit{иначе}\quad x\in\boldsymbol{\omega}_2.$$

Уравнение разделяющей поверхности:

$$(x_1+2)^2-(x_2+1)^2=3-4\ln\frac{1-p}{p}.$$

Это есть гипербола с центром в точке `(-2,-1)` и асимптотами $x_2=x_1+1$ и $x_2=-x_1-3$.

Центры классов (математические ожидания) обозначены на рис. 5.2 красной (класс 1) и синей (класс 2) точками. График разделяющей поверхности построен для p = 0.73 ($\mathbf{p}_1=0.73$, $\mathbf{p}_2=0.27$). Заметим, что точки, лежащие левее левой ветви гиперболы, будут оптимально классифицироваться в класс 2 (апостериорная вероятность выше для второго класса, хотя эти точки лежат ближе к центру первого класса).

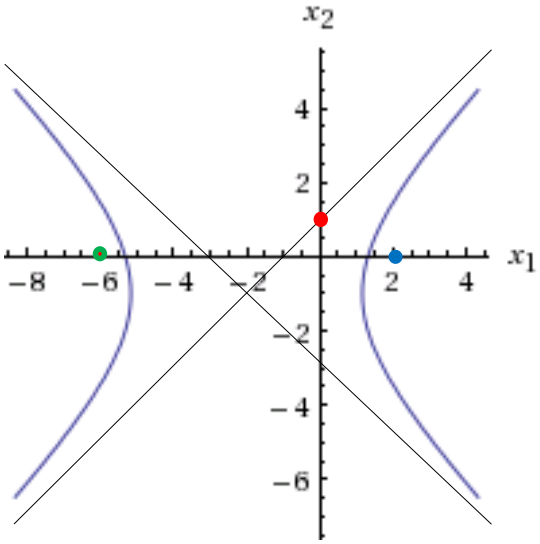


Рис. 5.2. Разделяющая поверхность двух классов гауссовых наблюдений с различными априорными вероятностями

Покажем, что для точки $\mathbf{x}=\begin{pmatrix}-6\\0\end{pmatrix}$ (зелёный цвет) выполняется неравенство $\mathbf{p}_2 f\left(\mathbf{x}|\boldsymbol{\omega}_2\right) > \mathbf{p}_1 f\left(\mathbf{x}|\boldsymbol{\omega}_1\right)$ (то есть точка классифицируется в класс 2):

$$\mathbf{p}_2 f\left(\mathbf{x}|\boldsymbol{\omega}_2\right) - \mathbf{p}_1 f\left(\mathbf{x}|\boldsymbol{\omega}_1\right) = \mathbf{p}_2 \frac{1}{2\pi\sqrt{|\mathbf{R}_2|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{x}-\mathbf{a}_2\right)\right] -$$

$$-\mathbf{p}_1 \frac{1}{2\pi\sqrt{|\mathbf{R}_1|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{x}-\mathbf{a}_1\right)\right].$$

$$\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{x}-\mathbf{a}_1\right) = \frac{1}{2}\begin{pmatrix}-6\\-1\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}1&0\\0&2\end{pmatrix}^{-1}\begin{pmatrix}-6\\-1\end{pmatrix} = 18.25,$$

$$\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{x}-\mathbf{a}_2\right) = \frac{1}{2}\begin{pmatrix}-8\\0\end{pmatrix}^{\mathrm{T}}\begin{pmatrix}2&0\\0&1\end{pmatrix}^{-1}\begin{pmatrix}-8\\0\end{pmatrix} = 16,$$

$$p_2 f\left(\mathbf{x}|\boldsymbol{\omega}_2\right) - p_1 f\left(\mathbf{x}|\boldsymbol{\omega}_1\right) = \frac{1}{2\pi\sqrt{2}} 0.73 e^{-16} - \frac{1}{2\pi\sqrt{2}} 0.27 e^{-18.25} =$$

$$= \frac{1}{2\pi\sqrt{2}}\left(8.215\cdot 10^{-8} - 3.203\cdot 10^{-9}\right) > 0.$$

Пример 5.3. Найти разделяющие поверхности трёх классов **I, II, III** гауссовских наблюдений двумерного случайного вектора X с математическими ожиданиями $\mathbf{a}_1=\begin{pmatrix}1\\0\end{pmatrix}$, $\mathbf{a}_2=\begin{pmatrix}0\\1\end{pmatrix}$, $\mathbf{a}_3=\begin{pmatrix}1\\1\end{pmatrix}$, некоррелированными компонентами $\mathbf{R}_1=\mathbf{R}_2=\mathbf{R}_3=\mathbf{R}=\begin{pmatrix}1&0\\0&1\end{pmatrix}$ и априорными вероятностями классов $\mathbf{p_1}=1/2$, $\mathbf{p_2}=1/3$, $\mathbf{p}_3=1/6$. Плотность вероятности каждого класса

$$f\left(\mathbf{x}|\boldsymbol{\omega}_i\right) = f\left(x_1, x_2|\boldsymbol{\omega}_i\right) = \frac{1}{2\pi\sqrt{|\mathbf{R}|}} \exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_i\right)^T \mathbf{R}^{-1}\left(\mathbf{x}-\mathbf{a}_i\right)\right],$$

$i=1,2,3.$

Логарифм функции правдоподобия для разделения объектов первого и второго классов:

$$-\ln\frac{f\left(\mathbf{x}|\boldsymbol{\omega}_1\right)}{f\left(\mathbf{x}|\boldsymbol{\omega}_2\right)}=-\ln\frac{\frac{1}{2\pi\sqrt{|\mathbf{R}|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T\mathbf{R}^{-1}\left(\mathbf{x}-\mathbf{a}_1\right)\right]}{\frac{1}{2\pi\sqrt{|\mathbf{R}|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T\mathbf{R}^{-1}\left(\mathbf{x}-\mathbf{a}_2\right)\right]}=$$

$$=\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T\mathbf{R}^{-1}\left(\mathbf{x}-\mathbf{a}_1\right)-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T\mathbf{R}^{-1}\left(\mathbf{x}-\mathbf{a}_2\right)=$$

$$=\frac{1}{2}\left\|\mathbf{x}-\mathbf{a}_1\right\|^2-\frac{1}{2}\left\|\mathbf{x}-\mathbf{a}_2\right\|^2=\frac{1}{2}\left[\left(x_1-1\right)^2+x_2^2-x_1^2-\left(x_2-1\right)^2\right]=x_2-x_1.$$

Разделяющая поверхность первого и второго классов находится из уравнения

$$-\ln\frac{f\left(\mathbf{x}|\boldsymbol{\omega}_1\right)}{f\left(\mathbf{x}|\boldsymbol{\omega}_2\right)}=\ln\frac{\mathbf{p}_1}{\mathbf{p}_2}\quad\rightarrow\quad x_2-x_1=\ln\frac{3}{2}\quad\rightarrow\quad x_2=x_1+\ln\frac{3}{2}.$$

Логарифм функции правдоподобия для разделения объектов первого и третьего классов:

$$-\ln\frac{f\left(\mathbf{x}|\boldsymbol{\omega}_1\right)}{f\left(\mathbf{x}|\boldsymbol{\omega}_3\right)}=\frac{1}{2}\left\|\mathbf{x}-\mathbf{a}_1\right\|^2-\frac{1}{2}\left\|\mathbf{x}-\mathbf{a}_3\right\|^2=$$

$$=\frac{1}{2}\left[\left(x_1-1\right)^2+x_2^2-\left(x_1-1\right)^2-\left(x_2-1\right)^2\right]=x_2-\frac{1}{2}.$$

Разделяющая поверхность первого и третьего классов находится из уравнения

$$-\ln\frac{f\left(\mathbf{x}|\boldsymbol{\omega}_1\right)}{f\left(\mathbf{x}|\boldsymbol{\omega}_3\right)}=\ln\frac{\mathbf{p}_1}{\mathbf{p}_3}\quad\rightarrow\quad x_2-\frac{1}{2}=\ln 3\quad\rightarrow\quad x_2=\frac{1}{2}+\ln 3.$$

Логарифм функции правдоподобия для разделения объектов второго и третьего классов:

$$-\ln\frac{f\left(\mathbf{x}|\boldsymbol{\omega}_2\right)}{f\left(\mathbf{x}|\boldsymbol{\omega}_3\right)}=\frac{1}{2}\left\|\mathbf{x}-\mathbf{a}_2\right\|^2-\frac{1}{2}\left\|\mathbf{x}-\mathbf{a}_3\right\|^2=$$

$$=\frac{1}{2}\left[x_1^2+\left(x_2-1\right)^2-\left(x_1-1\right)^2-\left(x_2-1\right)^2\right]=x_1-\frac{1}{2}.$$

Разделяющая поверхность второго и третьего классов находится из уравнения

$$-\ln\frac{f\left(\mathbf{x}|\boldsymbol{\omega}_2\right)}{f\left(\mathbf{x}|\boldsymbol{\omega}_3\right)}=\ln\frac{\mathbf{p}_2}{\mathbf{p}_3}\quad\rightarrow\quad x_1-\frac{1}{2}=\ln 2\quad\rightarrow\quad x_1=\frac{1}{2}+\ln 2.$$

На рис. 5.3 цветными точками обозначены центры (математические ожидания) классов. Видим, что из-за неодинаковой априорной вероятности классов происходит смещение области решений в сторону класса с низкой априорной вероятностью.

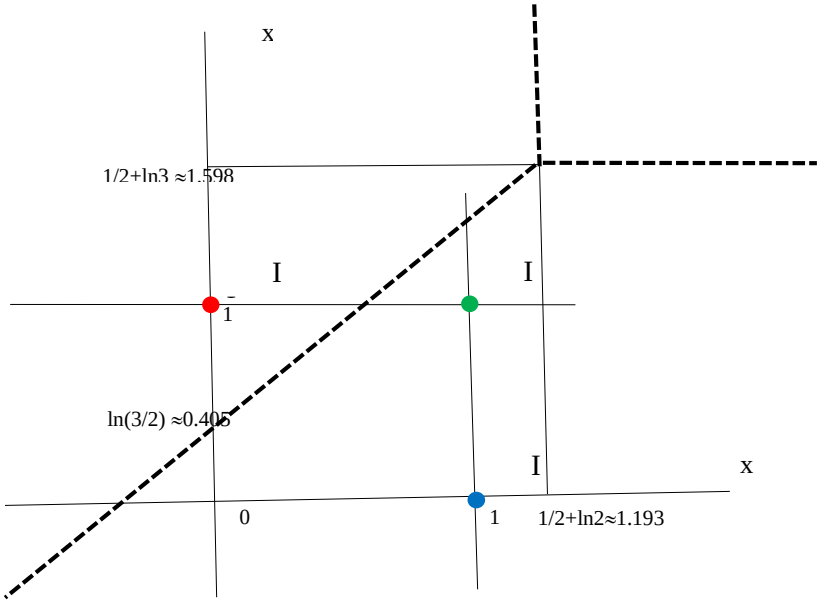


Рис. 5.3. Разделяющие прямые

Например, математическое ожидание третьего класса – точка $\mathbf{a}_3$ – будет оптимально (с минимальной вероятностью ошибки) классифицирована как принадлежащая первому классу, так как

$$\mathbf{p_1} f\left(\mathbf{x}|\boldsymbol{\omega}_1\right)\Big|_{\mathbf{x}=\mathbf{a_3}} > \mathbf{p_3} f\left(\mathbf{x}|\boldsymbol{\omega}_3\right)\Big|_{\mathbf{x}=\mathbf{a_3}} \Leftrightarrow \frac{1}{2}\frac{1}{2\pi\sqrt{|\mathbf{R}|}}\exp\left[-\frac{1}{2}\|\mathbf{a_3}-\mathbf{a_1}\|^2\right] >$$

$$> \frac{1}{6}\frac{1}{2\pi\sqrt{|\mathbf{R}|}}e^0 \Leftrightarrow \frac{1}{\sqrt{e}} > \frac{1}{3}.$$

5.2 Минимизация функции риска

Четыре типа штрафов: c_{ij}, $i=1,2,\ j=1,2$ – штраф за решение $x\in\boldsymbol{\omega}_\mathbf{j}$, если в действительности $x\in\boldsymbol{\omega}_\mathbf{i}$. Риск $\mathbf{r}$ – математическое ожидание штрафа:

$$\mathbf{r} = c_{11}\mathbf{p_1}\mathbf{Pr}\left\{x\in G_1|\boldsymbol{\omega}_1\right\} + c_{21}\mathbf{p_2}\mathbf{Pr}\left\{x\in G_1|\boldsymbol{\omega}_2\right\} + c_{12}\mathbf{p_1}\mathbf{Pr}\left\{x\in G_2|\boldsymbol{\omega}_1\right\} +$$
$$+c_{22}\mathbf{p_2}\mathbf{Pr}\left\{x\in G_2|\boldsymbol{\omega}_2\right\}.$$

Риск можно минимизировать за счёт оптимального выбора областей принятия решений G_1, G_2:

$$\begin{aligned}
&\mathbf{r} = c_{11}\mathbf{p_1}\int_{G_1} f\left(x|\boldsymbol{\omega}_1\right)dx + c_{21}\mathbf{p_2}\int_{G_1} f\left(x|\boldsymbol{\omega}_2\right)dx + c_{12}\mathbf{p_1}\int_{G_2} f\left(x|\boldsymbol{\omega}_1\right)dx + c_{22}\mathbf{p_2}\int_{G_2} f\left(x|\boldsymbol{\omega}_2\right)dx = \\
&= c_{11}\mathbf{p_1}\int_{G_1} f\left(x|\boldsymbol{\omega}_1\right)dx + c_{21}\mathbf{p_2}\int_{G_1} f\left(x|\boldsymbol{\omega}_2\right)dx + c_{12}\mathbf{p_1}\left(1-\int_{G1} f\left(x|\boldsymbol{\omega}_1\right)dx\right) + c_{22}\mathbf{p_2}\left(1-\int_{G_1} f\left(x|\boldsymbol{\omega}_2\right)dx\right) = \\
&= c_{12}\mathbf{p_1} + c_{22}\mathbf{p_2} + \int_{G_1}\left[-\left(c_{12}-c_{11}\right)\mathbf{p_1} f\left(x|\boldsymbol{\omega}_1\right) + \left(c_{21}-c_{22}\right)\mathbf{p_2} f\left(x|\boldsymbol{\omega}_2\right)\right]dx \to \min_{G_1}.
\end{aligned} \quad (5.4)$$

При выводе соотношения (5.4) использовался тот факт, что события $x\in G_1$ и $x\in G_2$ образуют полную группу: $\int_{G_2} f\left(x|\boldsymbol{\omega}_\mathbf{i}\right)dx + \int_{G_1} f\left(x|\boldsymbol{\omega}_\mathbf{i}\right)dx = 1$.

Для случая $c_{21} > c_{22}$, $c_{12} > c_{11}$, минимизируя подынтегральное выражение в (5.4) в каждой точке x, получаем по аналогии с (5.2) **байесовский критерий, минимизирующий риск:**

$$l(x) = \frac{f(x|\boldsymbol{\omega_1})}{f(x|\boldsymbol{\omega_2})} > \frac{(c_{21} - c_{22})\mathbf{p_2}}{(c_{12} - c_{11})\mathbf{p_1}} \rightarrow x \in \boldsymbol{\omega_1},$$
$$l(x) = \frac{f(x|\boldsymbol{\omega_1})}{f(x|\boldsymbol{\omega_2})} \leq \frac{(c_{21} - c_{22})\mathbf{p_2}}{(c_{12} - c_{11})\mathbf{p_1}} \rightarrow x \in \boldsymbol{\omega}_2. \quad (5.5)$$

Сравнив (5.2) и (5.5), видим, что назначение штрафов эквивалентно изменению отношения априорных вероятностей (порога функции правдоподобия). В случае симметричной функции штрафа ($c_{21} - c_{22} = c_{12} - c_{11}$) получаем байесовский критерий (5.2), минимизирующий вероятность ошибки.

5.3 Критерий Неймана-Пирсона

Одним из существенных недостатков байесовского правила обнаружения сигналов является большое количество априорной информации о потерях и вероятностях состоянии объекта, которая должна быть в распоряжении наблюдателя. Иногда указать априорные вероятности наличия цели и потери за счет ложной тревоги или пропуска цели оказывается весьма затруднительным. Поэтому в подобных задачах вместо байесовского критерия обычно используется критерий Неймана-Пирсона. Согласно этому критерию выбирается такое правило обнаружения, которое обеспечивает минимальную величину вероятности пропуска сигнала, то есть минимальную вероятность ошибки первого рода ε_1 (максимальную вероятность правильного обнаружения), при условии, что вероятность ложной тре-

воги (вероятность ошибки второго рода ε_2) не превышает заданной величины.

Минимизируется вероятность ошибки ε_1 при условии, что вероятность ошибки ε_2 равна заданной величине (например, ε_0). Для определения решающего правила минимизируем выражение

$$\mathbf{r} = \varepsilon_1 + \lambda\left(\varepsilon_2 - \varepsilon_0\right),$$

где λ – множитель Лагранжа.

$$\mathbf{r} = \varepsilon_1 + \lambda\left(\varepsilon_2 - \varepsilon_0\right) = \int\limits_{G_2} f\left(\mathbf{x}\middle|\boldsymbol{\omega}_1\right) d\mathbf{x} + \lambda\left(\int\limits_{G_1} f\left(\mathbf{x}\middle|\boldsymbol{\omega}_2\right) d\mathbf{x} - \varepsilon_0\right) =$$

$$= \left(1 - \lambda\varepsilon_0\right) + \int\limits_{G_1}\left[\lambda f\left(\mathbf{x}\middle|\boldsymbol{\omega}_2\right) - f\left(\mathbf{x}\middle|\boldsymbol{\omega}_1\right)\right] d\mathbf{x} \to \min_{G_1}.$$

Решающее правило:

$$\begin{aligned} l\left(\mathbf{x}\right) &= \frac{f\left(\mathbf{x}\middle|\boldsymbol{\omega}_1\right)}{f\left(\mathbf{x}\middle|\boldsymbol{\omega}_2\right)} > \lambda \quad \to \quad \mathbf{x} \in \boldsymbol{\omega}_1, \\ l\left(\mathbf{x}\right) &= \frac{f\left(\mathbf{x}\middle|\boldsymbol{\omega}_1\right)}{f\left(\mathbf{x}\middle|\boldsymbol{\omega}_2\right)} \le \lambda \quad \to \quad \mathbf{x} \in \boldsymbol{\omega}_2. \end{aligned} \tag{5.6}$$

Из сравнения (5.6) и (5.2) можно заключить, что оптимальное, в смысле критерия Неймана-Пирсона, правило обнаружения отличается от байесовского лишь величиной порогового уровня, с которым производится сравнение отношения правдоподобия. Порог λ находится из уравнения $\dfrac{\partial \mathbf{r}}{\partial \lambda} = 0$, то есть $\int\limits_{G_1(\lambda)} f\left(\mathbf{x}\middle|\boldsymbol{\omega}_2\right) d\mathbf{x} = \varepsilon_0 = \varepsilon_2$.

Пример 5.4. Построить решающее правило Неймана-Пирсона (5.6) для двух классов независимых гауссовских наблюдений случайной величины X с условной плотностью вероятностей

$$f\left(x|\boldsymbol{\omega}_1\right)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(x-1\right)^2\right],$$

$$f\left(x|\boldsymbol{\omega}_2\right)=\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(x-3\right)^2\right].$$

Отношение правдоподобия:

$$l\left(x\right)=\frac{f_X\left(x|\boldsymbol{\omega_1}\right)}{f_X\left(x|\boldsymbol{\omega_2}\right)}=\frac{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(x-1\right)^2\right]}{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(x-3\right)^2\right]}=$$

$$=\exp\left[\frac{1}{2}\left(x-3\right)^2-\frac{1}{2}\left(x-1\right)^2\right]=\exp\left(4-2x\right).$$

Решающее правило:

$$x<h\rightarrow x\in\boldsymbol{\omega}_1.$$

Порог h находится из уравнения, фиксирующего вероятность ошибки второго рода:

$$\int_{-\infty}^{h}f\left(x|\boldsymbol{\omega}_2\right)dx=\varepsilon_2,\qquad\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{h}\exp\left[-\frac{1}{2}\left(x-3\right)^2\right]dx=\varepsilon_2,$$

$$\Phi\left(h-3\right)=\varepsilon_2,\ h=\Phi^{-1}\left(\varepsilon_2\right)+3.$$

Здесь, как и раньше, через $\Phi\left(z\right)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z}\exp\left(-\frac{x^2}{2}\right)dx$ обозначена интегральная функция стандартного нормального рас-

пределения (интеграл вероятности, функция Лапласа). Тогда вероятность ошибки первого рода:

$$\varepsilon_1 = \int_h^{+\infty} f\left(x|\boldsymbol{\omega}_1\right)dx = \int_{\Phi^{-1}(\varepsilon_2)+3}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-1)^2\right]dx =$$

$$= \int_{\Phi^{-1}(\varepsilon_2)+2}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)du = 1-\Phi\left(\Phi^{-1}(\varepsilon_2)+2\right),$$

то есть: $\varepsilon_1 = 1-\Phi\left(\Phi^{-1}(\varepsilon_2)+2\right)$. (5.7)

Уравнение (5.7) задаёт зависимость вероятности ошибки первого рода от установленного значения вероятности ошибки второго рода. В табл. 5.1 приведены значения связанных по этому уравнению вероятностей ошибок первого и второго рода.

Таблица 5.1 – **Соотношение между вероятностями ошибок первого и второго рода**

ε_2	0.010	0.050	0.100	0.500	0.900	0.950
h	0.674	1.355	1.718	3.000	4.282	4.645
ε_1	0.628	0.361	0.236	0.023	0.001	0.000

На рис. 5.1 приведена графическая иллюстрация соотношения между вероятностями ошибок первого и второго рода.

Пример 5.5. Построить решающее правило Неймана-Пирсона (5.5) для двух классов независимых гауссовских наблюдений двумерного случайного вектора X с математическими ожиданиями $\mathbf{a}_1 = \begin{pmatrix}-1 & 0\end{pmatrix}^T$, $\mathbf{a}_2 = \begin{pmatrix}1 & 0\end{pmatrix}^T$, корреляционными матрицами $\mathbf{R}_1 = \mathbf{R}_2 = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$.

Логарифм отношения правдоподобия:

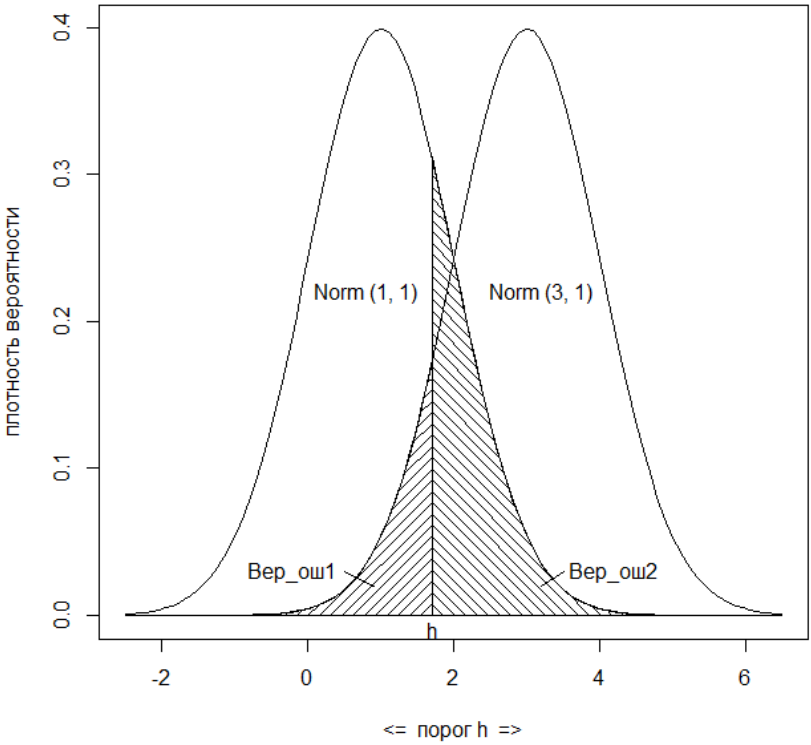


Рис. 5.4. Решающее правило Неймана–Пирсона.
Соотношение между вероятностями ошибок первого и второго рода

$$-\ln l(\mathbf{x}) = -\ln \frac{f(\mathbf{x}|\boldsymbol{\omega}_1)}{f(\mathbf{x}|\boldsymbol{\omega}_2)} = -\ln \frac{\frac{1}{2\pi\sqrt{|\mathbf{R}_1|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{a}_1)^T \mathbf{R}_1^{-1}(\mathbf{x}-\mathbf{a}_1)\right]}{\frac{1}{2\pi\sqrt{|\mathbf{R}_2|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{a}_2)^T \mathbf{R}_2^{-1}(\mathbf{x}-\mathbf{a}_2)\right]} =$$

$$= -\frac{1}{2}\begin{pmatrix} x_1 - 1 \\ x_2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x_1 - 1 \\ x_2 \end{pmatrix} + \frac{1}{2}\begin{pmatrix} x_1 + 1 \\ x_2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} x_1 + 1 \\ x_2 \end{pmatrix} = 2x_1 < \ln\lambda \rightarrow \mathbf{x} \in \boldsymbol{\omega}_1.$$

Порог λ находится из уравнения $\int\limits_{2x_1 > \ln\lambda} f(\mathbf{x}|\boldsymbol{\omega}_1)\,d\mathbf{x} = \varepsilon_2$, т.е.

$$\varepsilon_2 = \frac{1}{2\pi}\int\limits_{+\infty}^{+\infty}\left(\int\limits_{\frac{\ln\lambda}{2}}^{+\infty}\exp\left[-\frac{1}{2}\left(\left(x_1+1\right)^2+x_2^2\right)\right]dx_1\right)dx_2 =$$

$$= \frac{1}{\sqrt{2\pi}}\int\limits_{+\infty}^{+\infty}\left(\frac{1}{\sqrt{2\pi}}\int\limits_{\frac{\ln\lambda}{2}+1}^{+\infty}\exp\left[-\frac{1}{2}x_1^2\right]dx_1\right)\exp\left(-\frac{1}{2}x_2^2\right)dx_2 = 1-\Phi\left(\frac{\ln\lambda}{2}+1\right).$$

Вероятность ошибки первого рода находится из уравнения

$$\varepsilon_1 = \int\limits_{2x_1<\ln\lambda} f_X\left(\mathbf{x}\middle|\boldsymbol{\omega}_2\right)d\mathbf{x}, \text{ т.е.}$$

$$\varepsilon_1 = \frac{1}{2\pi}\int\limits_{+\infty}^{+\infty}\left(\int\limits_{-\infty}^{\frac{\ln\lambda}{2}}\exp\left[-\frac{1}{2}\left(\left(x_1-1\right)^2+x_2^2\right)\right]dx_1\right) =$$

$$= \frac{1}{\sqrt{2\pi}}\int\limits_{+\infty}^{+\infty}\left(\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{\frac{\ln\lambda}{2}-1}\exp\left(-\frac{1}{2}x_1^2\right)dx_1\right)\exp\left(-\frac{1}{2}x_2^2\right)dx_2 = \Phi\left(\frac{\ln\lambda}{2}-1\right).$$

Из последних двух соотношений $\varepsilon_1 = \Phi\left(\frac{\ln\lambda}{2}-1\right)$, $\varepsilon_2 = 1-\Phi\left(\frac{\ln\lambda}{2}+1\right)$ получаем явную зависимость $\varepsilon_1 = \varepsilon_1\left(\varepsilon_2\right)$:

$$\varepsilon_1 = \Phi\left(\Phi^{-1}\left(1-\varepsilon_2\right)-2\right).$$

В табл. 5.2 приведён ряд рассчитанных значений вероятностей ошибок первого и второго рода для некоторых значений порога λ.

Таблица 5.2 – **Соотношение между вероятностями ошибок первого и второго рода**

λ	8	4	2	1	1/2	1/4	1/8
ε_1	0.516	0.379	0.257	0.159	0.090	0.045	0.021
ε_2	0.021	0.045	0.090	0.159	0.257	0.379	0.516

5.4 ROC-кривая

В предыдущих разделах мы видели, что вероятности ошибок первого и второго рода изменяются в зависимости от выбранного порога, но жестко связаны друг с другом. Их зависимость традиционно представляется в виде так называемой *ROC-кривой* (*Receiver Operating Characteristic*, рабочая характеристика приёмника).

Эта кривая позволяет оценить качество бинарной классификации, отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак (*True Positive Rate, TPR*, называемой **чувствительностью** (*Sensitivity*) алгоритма классификации, в наших обозначениях $TPR = 1 - \varepsilon_1$) и долей объектов, не несущих признака, ошибочно классифицированных как несущих признак (*False Positive Rate, FPR*, величина (*1–FPR*) называется **специфичностью** (*Specificity*) алгоритма классификации, в наших обозначениях $FPR = \varepsilon_2$) при варьировании порога решающего правила.

На рис. 5.5 приведена ROC-кривая для решающего правило Неймана-Пирсона для двух классов независимых гауссовских наблюдений двумерного случайного вектора (см. пример 5.4, уравнение (5.7)): $TPR = \Phi\left(\Phi^{-1}\left(FPR\right) - 2\right)$.

Площадь под ROC-кривой (*AUC – Area Under Curve*) $AUC = \int_0^1 TPR\left(FPR\right) d\,FPR$ позволяет оценить качество используемого алгоритма классификации при варьировании порога решающего правила.

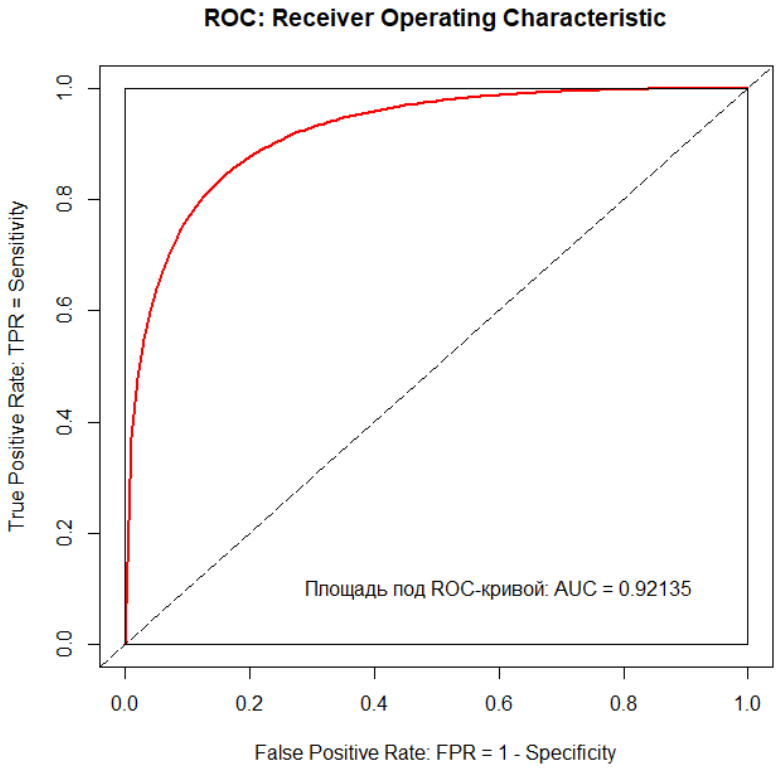


Рис. 5.5. ROC-кривая для ошибок классификации из примера 5.4

5.5 Наивный байесовский классификатор

Наивный байесовский классификатор основан на наивном предположении, что объекты описываются *независимыми признаками*.

Будем считать, что n вероятностно независимых признаков (не обязательно числовых) $(x_1, x_2, \ldots, x_n)$ представлены вектором $\mathbf{x}$. Вероятностная независимость означает, что многомерное распределение вероятностей равно произведению одномерных распределений:

$$P(\mathbf{x}) = P(x_1)P(x_2)\ldots P(x_n).$$

В случае условных вероятностей, опуская обозначения признаков, получаем:

$$P(\mathbf{x}|\boldsymbol{\omega}_i) = P(x_1 \mid \boldsymbol{\omega}_i)P(x_2 \mid \boldsymbol{\omega}_i)\ldots P(x_n \mid \boldsymbol{\omega}_i), i = \overline{1,m} - \textit{номер класса.}$$

Запишем формулу Байеса **для апостериорной вероятности** класса в виде

$$P(\boldsymbol{\omega}_i|\mathbf{x}) = \frac{P(\mathbf{x}|\boldsymbol{\omega}_i)P(\boldsymbol{\omega}_i)}{P(\mathbf{x})} = \frac{P(\mathbf{x}|\boldsymbol{\omega}_i)P(\boldsymbol{\omega}_i)}{P(\mathbf{x}|\boldsymbol{\omega}_1)P(\boldsymbol{\omega}_1)+\ldots+P(\mathbf{x}|\boldsymbol{\omega}_m)P(\boldsymbol{\omega}_m)}, \quad i = \overline{1,m},$$

где $P(\boldsymbol{\omega}_i)$ – **априорные** вероятности классов объектов;

$P(\mathbf{x}|\boldsymbol{\omega}_i)$ – условные распределения вероятности признаков для различных классов объектов;

$P(\mathbf{x}) = P(\boldsymbol{\omega}_1)P(\mathbf{x}|\boldsymbol{\omega}_1) + \ldots + P(\boldsymbol{\omega}_m)P(\mathbf{x}|\boldsymbol{\omega}_m)$ – распределение вероятностей признаков (формула полной вероятности).

Так как в знаменателе формулы Байеса стоит величина вероятности, не зависящая от класса объекта $\mathbf{x}$, то класс объекта, соответствующий максимуму апостериорной вероятности, определяется числителем формулы Байеса:

$$\boldsymbol{\omega}_* = \arg\max_{\boldsymbol{\omega}_1,\ldots,\boldsymbol{\omega}_m} \left\{ P\{\boldsymbol{\omega}_1\} P(\mathbf{x}|\boldsymbol{\omega}_1), \ldots, P\{\boldsymbol{\omega}_m\} P(\mathbf{x}|\boldsymbol{\omega}_m) \right\}.$$

Для случая двух классов получаем:

$$\boldsymbol{\omega}_* = \arg\max_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} \left\{ P(\boldsymbol{\omega}_1) P(\mathbf{x}|\boldsymbol{\omega}_1), P(\boldsymbol{\omega}_2) P(\mathbf{x}|\boldsymbol{\omega}_2) \right\}.$$

Пример 5.6. **Байесовская фильтрация спама**. Обозначим события: S – рассматриваемое сообщение является спамом (*Spam,* класс $\boldsymbol{\omega}_1$), H – сообщение не является спамом (*non-spam* или *ham,* класс $\boldsymbol{\omega}_2$). Недавние статистические исследования показали, что на сегодняшний день вероятность любого сообщения быть спамом составляет по меньшей мере 80%: $P\{S\} = 0.8$, $P\{\bar{S}\} = 0.2$. Однако большинство байесовских программ обнаружения спама делают предположение об отсутствии априорных предпочтений у сообщения быть спамом, а не «ham», и полагает, что у обеих альтернатив есть равные вероятности 50%: $P\{S\} = P\{S\} = 0.5$ (ещё одно наивное предположение). О фильтрах спама, которые используют эту гипотезу, говорят как о фильтрах «без предубеждений». Это означает, что у них нет никакого предубеждения относительно входящей электронной почты. Это предположение позволяет упростить общую формулу (считаем, что $P\{S\} = P\{\bar{S}\} = 0.5$):

$$P\{S|\mathbf{x}\} = \frac{P(\mathbf{x}|S) P\{S\}}{P(\mathbf{x}|S) P\{S\} + P(\mathbf{x}|\bar{S}) P\{\bar{S}\}}, \text{ то есть}$$

$$\mathbf{Pr}\{S|x_1, x_2, \ldots, x_n\} = \frac{P(x_1|S)P(x_2|S)\ldots P(x_n|S)}{P(x_1|S)P(x_2|S)\ldots P(x_n|S) + P(x_1|\bar{S})P(x_2|\bar{S})\ldots P(x_n|\bar{S})}.$$

Здесь $x_1, x_2, \ldots, x_n$ – слова в проверяемом на спам сообщении; $P(x_k|S)$ – оценка вероятность наличия слова x_k в спамовом сообщении, определяемой относительной частотой сооб-

щений, содержащих слово x_k в сообщениях, идентифицированных как спам во время фазы обучения; $P\left(x_k \middle| \overline{S}\right)$ – оценка вероятности наличия слова x_k в неспамовом сообщении, определяемой относительной частотой сообщений, содержащих слово x_k в сообщениях, идентифицированных как неспам во время фазы обучения.

$$P\{S|\mathbf{x}\} = \frac{\frac{P(S|x_1)}{P(S)} \cdot \frac{P(S|x_2)}{P(S)} \cdot \ldots \cdot \frac{P(S|x_n)}{P(S)}}{\frac{P(S|x_1)}{P(S)} \cdot \frac{P(S|x_2)}{P(S)} \cdot \ldots \cdot \frac{P(S|x_n)}{P(S)} + \frac{P(\overline{S}|x_1)}{P(\overline{S})} \cdot \frac{P(\overline{S}|x_2)}{P(\overline{S})} \cdot \ldots \cdot \frac{P(\overline{S}|x_n)}{P(\overline{S})}} =$$

$$= \frac{p_1 p_2 \ldots p_n}{p_1 p_2 \ldots p_n + (1-p_1)(1-p_2)\ldots(1-p_n)},$$

где $p_k = \mathbf{P}\left(S \middle| x_k\right)$ – условная вероятность того, что сообщение является спамом при условии, что оно содержит *k*-е слово ($k = \overline{1,n}$) (тогда $P\left(\overline{S} \middle| x_k\right) = 1 - p_k$). То есть p_1 – условная вероятность того, что сообщение S – спам при условии, что оно содержит слово x_1; p_2 – условная вероятность того, что сообщение S – спам при условии, что оно содержит слово x_2 и т.д.

6 ЛИНЕЙНЫЕ КЛАССИФИКАТОРЫ

Байесовский классификатор, рассмотренный в предыдущем разделе, основан на вычислении отношения правдоподобия и является оптимальным в смысле минимизации риска или вероятности ошибки. Более простые методы классификации основаны на задании математического вида классификатора с последующим определением его оптимальных параметров. Наиболее простым является *линейный* или *кусочно-линейный* классификатор. Следует, однако, иметь в виду, что никакой линейный классификатор по качеству работы не превосходит классификатор, основанный на вычислении отношения правдоподобия.

6.1 Байесовский линейный классификатор

Для случайного вектора признаков $\mathbf{x} \in R^n$ при двух классах гауссовских распределений с параметрами $(\mathbf{M}_1, R_1)$ и $(\mathbf{M}_2, R_2)$, **с одинаковыми ковариационными матрицами** $R_1 = R_2 = R$ и с заданными априорными вероятностями классов $\mathbf{p}_1 = \mathbf{Pr}\{\boldsymbol{\omega}_1\}$ и $\mathbf{p}_2 = \mathbf{Pr}\{\boldsymbol{\omega}_2\}$ и условными плотностями распределений $f_1(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\omega}_1)$, $f_2(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\omega}_2)$ плотность вероятности смеси распределений имеет вид:

$$f(\mathbf{x}) = \mathbf{p}_1 f_1(\mathbf{x}) + \mathbf{p}_2 f_2(\mathbf{x}) = \mathbf{p}_1 \frac{1}{(2\pi)^{n/2}|R|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{M}_1)^T R^{-1}(\mathbf{x}-\mathbf{M}_1)\right] +$$

$$+\mathbf{p}_2 \frac{1}{(2\pi)^{n/2}|R|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{M}_2)^T R^{-1}(\mathbf{x}-\mathbf{M}_2)\right].$$

Байесовское решающее правило имеет линейный относительно вектора наблюдений $\mathbf{x}$ вид:

$$-\ln l(\mathbf{x}) = -\ln \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} = \frac{1}{2}(\mathbf{x}-\mathbf{M}_1)^T R^{-1}(\mathbf{x}-\mathbf{M}_1) -$$

$$-\frac{1}{2}(\mathbf{x}-\mathbf{M}_2)^T R^{-1}(\mathbf{x}-\mathbf{M}_2) \; < \; \ln\frac{\mathbf{p}_1}{\mathbf{p}_2} \; \rightarrow \; \boldsymbol{\omega}_1,$$

то есть

$$(\mathbf{M}_2-\mathbf{M}_1)^T R^{-1}\mathbf{x} + \frac{1}{2}\left(\mathbf{M}_1^T R^{-1}\mathbf{M}_1 - \mathbf{M}_2^T R^{-1}\mathbf{M}_2\right) < \ln(\mathbf{p}_1/\mathbf{p}_2) \rightarrow \boldsymbol{\omega}_1.$$

Уравнение разделяющей поверхности:

$$(\mathbf{M}_2-\mathbf{M}_1)^T R^{-1}\mathbf{x} = \ln(\mathbf{p}_1/\mathbf{p}_2) - \frac{1}{2}\left(\mathbf{M}_1^T R^{-1}\mathbf{M}_1 - \mathbf{M}_2^T R^{-1}\mathbf{M}_2\right).$$

Это есть гиперплоскость в пространстве $\mathbf{R}^n$.

Если компоненты вектора наблюдений X являются некоррелированными ($R = \sigma^2 I_n$, σ^2 – дисперсия белого шума наблюдения), то из

$$\frac{1}{2}(\mathbf{x}-\mathbf{M}_1)^T \sigma^{-2} I_n (\mathbf{x}-\mathbf{M}_1) - \frac{1}{2}(\mathbf{x}-\mathbf{M}_2)^T \sigma^{-2} I_n (\mathbf{x}-\mathbf{M}_2) < \ln(\mathbf{p}_1/\mathbf{p}_2) \rightarrow \boldsymbol{\omega}_1$$

получаем:

$$\|\mathbf{x}-\mathbf{M}_1\|^2 - \|\mathbf{x}-\mathbf{M}_2\|^2 < 2\sigma^2 \ln(\mathbf{p}_1/\mathbf{p}_2) \rightarrow \boldsymbol{\omega}_1,$$

где $\|\cdot\|$ – евклидова норма вектора.

Уравнение разделяющей поверхности:

$$\|\mathbf{x}-\mathbf{M}_1\|^2-\|\mathbf{x}-\mathbf{M}_2\|^2 = 2\sigma^2\ln(\mathbf{p}_1/\mathbf{p}_2).$$

Геометрическая интерпретация. Это есть точка $x=(M_1+M_2)/2+\sigma^2\ln(\mathbf{p}_1/\mathbf{p}_2)$ в случае n=1, прямая в случае n=2, плоскость в случае n=3 или гиперплоскость при n>3, перпендикулярная отрезку, соединяющему точки $\mathbf{M}_1$ и $\mathbf{M}_2$. При $\mathbf{p}_1=\mathbf{p}_2=\mathbf{1/2}$ разделяющая поверхность проходит через середину этого отрезка.

В случае, когда наблюдения не являются некоррелированными, т.е. когда $R_X\neq\sigma^2 I_n$, также можно использовать принцип минимизации расстояния. Для этого проведём декорреляцию вектора наблюдений с помощью некоторого линейного преобразования $A: Y=AX$:

$$R_Y=\mathbf{M}Y_0Y_0^T=\mathbf{M}AX_0X_0^TA^T=A\mathbf{M}\left(X_0X_0^T\right)A^T=AR_XA^T=I_n.$$

Так как ковариационная матрица является положительно определённой, то матрица линейного преобразования A всегда существует и невырождена, т.е. преобразование A является обратимым. Поэтому можно классифицировать гауссовский случайный вектор Y с условными математическими ожиданиями $\mathbf{M}_1^Y=A\mathbf{M}_1$ и $\mathbf{M}_2^Y=A\mathbf{M}_2$ и ковариационной матрицей $R_Y=I_n$, сравнивая расстояния $\|Y-\mathbf{M}_1^Y\|^2$ и $\|Y-\mathbf{M}_2^Y\|^2$.

6.2 Линейная разделяющая функция, минимизирующая вероятность ошибки

Во многих прикладных задачах предположение о равенстве ковариационных функций при наблюдении объектов различных классов не выполняется. Рассмотрим задачу построения линейного классификатора, который определяется *линейной разделяющей функцией*

$$h(x) = V^T x + v_0 : \begin{cases} < 0 \to x \in \boldsymbol{\omega}_1, \\ \geq 0 \to x \in \boldsymbol{\omega}_2. \end{cases}$$

Задача синтеза классификатора заключается в определении коэффициентов $V^T = \left[v_1 \ \ldots \ v_n \right]$ и v_0, оптимальных по некоторому критерию. При нормальных распределениях с одинаковыми ковариационными матрицами линейная разделяющая функция $h(x)$ совпадает с отношением правдоподобия.

Построим линейную разделяющую функцию, минимизирующую вероятность ошибки классификации (как мы видели выше, в общем случае разделяющая функция не является линейной). Даже если наблюдения не подчиняются нормальному закону, то в силу центральной предельной теоремы величина $h(X)$ имеет распределение, близкое к нормальному при $n \to \infty$. Условные математическое ожидание и дисперсия скалярной величины $h(X)$ для первого и второго классов объектов равны:

$$m_i = \mathbf{M}\left(h(X) \mid \boldsymbol{\omega}_i \right) = \mathbf{M}\left(V^T X \mid \boldsymbol{\omega}_i \right) + v_0 = V^T \mathbf{M}\left(X \mid \boldsymbol{\omega}_i \right) + v_0 = \\ = V^T \mathbf{M}_i + v_0, \quad i = \overline{1,2},$$

$$\sigma_i^2 = \mathbf{D}\left\{h(X) \mid \boldsymbol{\omega}_i\right\} = \mathbf{D}\left\{V^T X \mid \boldsymbol{\omega}_i\right\} =$$

$$= \mathbf{M}\left\{V^T \left(X - m_i\right)\left(X - m_i\right)^T V \mid \boldsymbol{\omega}_i\right\} = V^T R_i V, \quad i = \overline{1,2},$$

где $\mathbf{M}_i$ и R_i – условное математическое ожидание и условная корреляционная матрица (при условии принадлежности классу i) вектора признаков X.

Вероятность ошибки:

$$\boldsymbol{\varepsilon} = \mathbf{p}_1 \int_0^{+\infty} \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{\left(x - m_1\right)^2}{2\sigma_1^2}\right) dx +$$

$$+ \mathbf{p}_2 \int_{-\infty}^{0} \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{\left(x - m_2\right)^2}{2\sigma_2^2}\right) dx \to \min_{v_0, v_1, \ldots, v_n}.$$

Здесь $m_1 = m_1\left(v_0, v_1, \ldots, v_n\right)$, $\sigma_1 = \sigma_1\left(v_0, v_1, \ldots, v_n\right)$,

$m_2 = m_2\left(v_0, v_1, \ldots, v_n\right)$, $\sigma_2 = \sigma_2\left(v_0, v_1, \ldots, v_n\right)$.

В общем случае эта оптимизационная задача может быть решена численно с использованием специально разработанных итерационных процедур (см., например, [2, с. 105-106]).

Пример 6.1. Построить линейный классификатор для двух классов независимых гауссовских наблюдений двумерного случайного вектора $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ с математическими ожиданиями $\mathbf{a}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ и $\mathbf{a}_2 = \begin{pmatrix} 0 \\ 4 \end{pmatrix}$, корреляционными матрицами $\mathbf{R}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{R}_2 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$ и равными априорными вероятностями классов: $\mathbf{p}_1 = \mathbf{p}_2 = 1/2$. Сравнить вероятность ошибки классификации построенного линейного классификатора с вероятностью ошибки байесовской классификации.

Линейный классификатор.

Линейная разделяющая функция: $h(x) = v_0 + v_1 x_1 + v_2 x_2$. Из соображений симметрии следует, что $v_1 = 0$, то есть $h(x) = v_0 + v_2 x_2$, т.е. уравнение разделяющей поверхности: $x_2 = -\dfrac{v_0}{v_2}$.

Так как параметры v_0, v_2 являются детерминированными (неслучайными) величинами, то случайная величина $X = v_0 + v_2 x_2$ имеет нормальное распределение для объектов $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ как первого, так и второго классов. Условные математические ожидания и дисперсии случайной величина X:

$$m_1 = \mathbf{M}(\mathrm{X} \mid \boldsymbol{\omega}_1) = \mathbf{M}(v_0 + v_2 x_2 \mid \boldsymbol{\omega}_1) = v_0,$$

$$m_2 = \mathbf{M}(X \mid \boldsymbol{\omega}_2) = \mathbf{M}(v_0 + v_2 x_2 \mid \boldsymbol{\omega}_2) = v_0 + 4v_2,$$

$$\sigma_1^2 = \mathbf{D}(X \mid \boldsymbol{\omega}_1) = \mathbf{D}(v_0 + v_2 x_2 \mid \boldsymbol{\omega}_1) = v_2^2,$$

$$\sigma_2^2 = \mathbf{D}(X \mid \boldsymbol{\omega}_2) = \mathbf{D}(v_0 + v_2 x_2 \mid \boldsymbol{\omega}_2) = v_2^2.$$

Уравнению разделяющей поверхности $h(x) = v_0 + v_2 x_2 = 0$ соответствует значение случайной величины $X = 0$.

Вероятность ошибочной классификации:

$$\boldsymbol{\varepsilon} = \mathbf{p}_1\boldsymbol{\varepsilon}_1 + \mathbf{p}_2\boldsymbol{\varepsilon}_2 = \frac{1}{2}\int_0^{+\infty} f_X(x \mid \boldsymbol{\omega}_1)\,dx + \frac{1}{2}\int_{-\infty}^{0} f_X(x \mid \boldsymbol{\omega}_2)\,dx =$$

$$= \frac{1}{2}\int_0^{+\infty} \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left(-\frac{(x-m_1)^2}{2\sigma_1^2}\right)dx + \frac{1}{2}\int_{-\infty}^{0} \frac{1}{\sigma_2\sqrt{2\pi}} \exp\left(-\frac{(x-m_2)^2}{2\sigma_2^2}\right)dx =$$

$$= \frac{1}{2}\int_0^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-v_0)^2}{2v_2^2}\right)\frac{dx}{v_2} + \frac{1}{2}\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-v_0-4v_2)^2}{2v_2^2}\right)\frac{dx}{v_2} =$$

$$= \left\{u = \frac{x}{v_2},\ du = \frac{dx}{v_2},\ z = \frac{v_0}{v_2}\right\} =$$

$$= \frac{1}{2}\int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u-z)^2\right) du + \frac{1}{2}\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(u-z-4)^2\right) du =$$

$$= \frac{1}{2\sqrt{2\pi}}\int_{-z}^{+\infty} \exp\left(-\frac{u^2}{2}\right) du + \frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{-z-4} \exp\left(-\frac{u^2}{2}\right) du.$$

Приравниваем нулю производную ε по z и ищем вероятность ошибки в точке минимума:

$$\frac{\partial \varepsilon}{\partial z} = -\left(-\frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)\right) + \left(-\frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(z+4)^2}{2}\right)\right) = 0 \;\rightarrow$$

$$\rightarrow \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) = \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{(z+4)^2}{2}\right) \rightarrow$$

$$\rightarrow \quad (z+4)^2 = z^2 \quad \rightarrow \quad z = -2 \quad \rightarrow \quad \frac{v_0}{v_2} = -2;$$

$$\varepsilon = \frac{1}{2\sqrt{2\pi}}\int_{2}^{+\infty} \exp\left(-\frac{x^2}{2}\right) dx + \frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{-2} \exp\left(-\frac{x^2}{2}\right) dx =$$

$$= \frac{1}{2}\left(1-\Phi(2)\right) + \frac{1}{2}\Phi(-2) = \Phi(-2) = 0.02275013,$$

где $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du$ – интеграл вероятности (интегральная функция стандартного нормального распределения).

Использованы также соотношения:

$$\frac{1}{\sqrt{2\pi}}\int_{x}^{+\infty} \exp\left(-\frac{u^2}{2}\right) du = 1-\Phi(x), \qquad \Phi(-x) = 1-\Phi(x).$$

Вероятность ошибочной классификации: ε = **0.02275013**.

Уравнение разделяющей поверхности: $x_2 = 2$.

Области принятия решений: $G_1 = \{(x_1, x_2):\ x_2 < 2\}$, $G_2 = \{(x_1, x_2):\ x_2 > 2\}$.

Оптимальный байесовский классификатор

Теперь для сравнения построим уравнение оптимальной разделяющей поверхности при байесовской классификации и рассчитаем вероятность ошибки.

$$\mathbf{p}_1 f\left(\mathbf{x}|\boldsymbol{\omega}_1\right) = \mathbf{p}_2 f\left(\mathbf{x}|\boldsymbol{\omega}_2\right), \quad \mathbf{p}_1 = \mathbf{p}_2 = 1/2,$$

$$\frac{1}{2}\frac{1}{\sqrt{|\mathbf{R}_1|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{x}-\mathbf{a}_1\right)\right] = \frac{1}{2}\frac{1}{\sqrt{|\mathbf{R}_2|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{x}-\mathbf{a}_2\right)\right],$$

$$-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{x}-\mathbf{a}_1\right) + \frac{1}{2}\left(\mathbf{x}-\mathbf{a}_2\right)^T \mathbf{R}_2^{-1}\left(\mathbf{x}-\mathbf{a}_2\right) = \ln\frac{1}{2},$$

$$-\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} x_1 \\ x_2 - 4 \end{pmatrix}^T \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 - 4 \end{pmatrix} = -2\ln 2,$$

$$-x_1^2 - x_2^2 + \frac{1}{4}x_1^2 + \left(x_2 - 4\right)^2 = -2\ln 2, \qquad -\frac{3}{4}x_1^2 - 8x_2 + 16 = -2\ln 2.$$

Уравнение разделяющей поверхности в виде $x_2 = x_2\left(x_1\right)$:

$$x_2 = -\frac{3x_1^2}{32} + 2 + \frac{\ln 2}{4}$$ (парабола).

Области принятия решений:

$$G_1 = \left\{\left(x_1, x_2\right): \; x_2 < -\frac{3x_1^2}{32} + 2 + \frac{\ln 2}{4}\right\},$$

$$G_2 = \left\{\left(x_1, x_2\right): \; x_2 > -\frac{3x_1^2}{32} + 2 + \frac{\ln 2}{4}\right\}.$$

Вероятность ошибочной классификации:

$$\boldsymbol{\varepsilon} = \mathbf{p}_1 \mathbf{Pr}\left\{\textit{ошибка}|\boldsymbol{\omega}_1\right\} + \mathbf{p}_2 \mathbf{Pr}\left\{\textit{ошибка}|\boldsymbol{\omega}_2\right\} =$$

$$= \frac{1}{2}\left(\mathbf{Pr}\left\{X \in G_2|\boldsymbol{\omega}_1\right\} + \mathbf{Pr}\left\{X \in G_1|\boldsymbol{\omega}_2\right\}\right) = \frac{1}{2}\left(\boldsymbol{\varepsilon}_1 + \boldsymbol{\varepsilon}_2\right),$$

Рассчитаем численно вероятности ошибок первого и второго рода:

$$\boldsymbol{\varepsilon}_1 = \iint\limits_{G_2} f_1\left(\mathbf{x}\right) d\mathbf{x} = \iint\limits_{G_2} \frac{1}{2\pi\sqrt{|\mathbf{R}_1|}}\exp\left[-\frac{1}{2}\left(\mathbf{x}-\mathbf{a}_1\right)^T \mathbf{R}_1^{-1}\left(\mathbf{x}-\mathbf{a}_1\right)\right] d\mathbf{x} =$$

$$= \iint\limits_{x_2 > -\frac{3x_1^2}{32} + 2 + \frac{\ln 2}{4}} \frac{1}{2\pi}\exp\left[-\frac{1}{2}\left(x_1^2 + x_2^2\right)\right] dx_1 dx_2 = 0.01980355.$$

$$\varepsilon_2 = \iint\limits_{G_1} f_2(\mathbf{x})\,d\mathbf{x} = \iint\limits_{G_1} \frac{1}{2\pi\sqrt{|\mathbf{R}_2|}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{a}_2)^T\,\mathbf{R}_2^{-1}(\mathbf{x}-\mathbf{a}_2)\right] d\mathbf{x} =$$

$$= \iint\limits_{G_1} \frac{1}{2\pi\sqrt{4}} \exp\left[-\frac{1}{2}\begin{pmatrix} x_1 \\ x_2-4 \end{pmatrix}^T \begin{pmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2-4 \end{pmatrix}\right] dx_1 dx_2 =$$

$$= \iint\limits_{x_2 < -\frac{3x_1^2}{32}+2+\frac{\ln 2}{4}} \frac{1}{2\cdot 2\pi} \exp\left[-\frac{1}{2}\left(\frac{x_1^2}{4}+(x_2-4)^2\right)\right] dx_1 dx_2 = 0.02026693.$$

Общая вероятность ошибки **ε = 0.02003524** (приблизительно на 12% меньше, чем у линейного классификатора).

```
    # R version 3.4.2 (2017-09-28) + pracma   2017 (c) MacroSoft
    # Линейное и байесовское разделение двух гауссовых классов с
равными
    # априорными вероятностями и некоррелированными компонентами:
    # N1: m=(0;0), D=(1,0; 0,1), Pr1=1/2
    # N2: m=(0;4), D=(4,0; 0,1), Pr2=1/2

    require(pracma)

    e0 <- pnorm(-2)  # Ошибка линейного классификатора e0

    f1 <- function(x1, x2) # Ошибка первого рода байесовского клас-
сификатора e1
     ifelse(x2  >  -3*x1*x1/32  +  2  +  log(2)/4,  exp(-x1^2/2  -
x2^2/2)/(2*pi), 0)
    e1 <- dblquad(f1, -Inf, Inf, -Inf, Inf, dim=1)

    f2 <- function(x1, x2) # Ошибка второго рода байесовского клас-
сификатора e2
     ifelse(x2 > -3*x1*x1/32 + 2 + log(2)/4, 0, exp(-x1^2/8 - (x2-
4)^2/2)/(2*pi))/2
    e2 <- dblquad(f2, -Inf, Inf, -Inf, Inf, dim=1)

    e <- .5*(e1+e2) # Общая ошибка байесовского классификатора e

    fprintf ("\nОшибка линейного классификатора e0 = %10.8f\n", e0)
    fprintf ("\nОшибка первого рода  e1 = %10.8f", e1)
    fprintf ("\nОшибка второго рода  e2 = %10.8f", e2)
    fprintf ("\nОбщая ошибка   e = %10.8f", e)
```

Рис. 6.1. Фрагмент R-программы для вычисления вероятностей ошибок

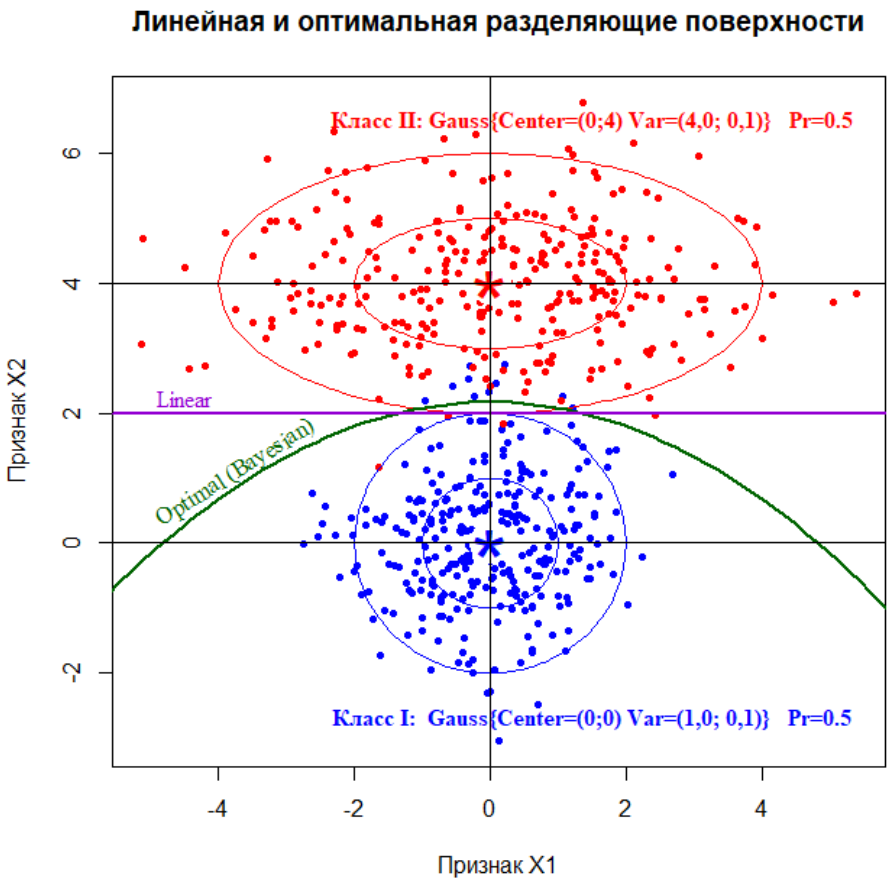


Рис. 6.2. Разделяющие поверхности для оптимальной байесовской и для оптимальной линейной классификации

Пример 6.2. Построить линейный классификатор для двух классов независимых гауссовских наблюдений двумерного случайного вектора $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ с математическими ожиданиями $\mathbf{a}_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}$ и $\mathbf{a}_2 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}$, корреляционными матрицами $\mathbf{R}_1 = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$,

$\mathbf{R}_2 = \begin{pmatrix} 0.2 & -0.1 \\ -0.1 & 0.2 \end{pmatrix}$ и одинаковыми априорными вероятностями классов: $\mathbf{p}_1 = \mathbf{p}_2 = 0.5$.

Линейная разделяющая функция: $h(x) = v_0 + v_1 x_1 + v_2 x_2$. Так как параметры v_0, v_1, v_2 являются детерминированными (неслучайными) величинами, то случайная величина $X = v_0 + v_1 x_1 + v_2 x_2$ имеет нормальное распределение для объектов $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ как первого, так и второго классов. Условные математические ожидания и дисперсии случайной величина X:

$$m_1 = \mathbf{M}(\mathrm{X} \mid \boldsymbol{\omega}_1) = \mathbf{M}(v_0 + v_1 x_1 + v_2 x_2 \mid \boldsymbol{\omega}_1) =$$

$$= v_0 + v_1 \mathbf{M}(x_1 \mid \boldsymbol{\omega}_1) + v_2 \mathbf{M}(x_2 \mid \boldsymbol{\omega}_1) = v_0 + 2v_1,$$

$$m_2 = \mathbf{M}(\mathrm{X} \mid \boldsymbol{\omega}_2) = \mathbf{M}(v_0 + v_1 x_1 + v_2 x_2 \mid \boldsymbol{\omega}_2) =$$

$$= v_0 + v_1 \mathbf{M}(x_1 \mid \boldsymbol{\omega}_2) + v_2 \mathbf{M}(x_2 \mid \boldsymbol{\omega}_2) = v_0 + 2v_2,$$

$$\sigma_1^2 = \mathbf{D}(X \mid \boldsymbol{\omega}_1) = \mathbf{D}(v_0 + v_1 x_1 + v_2 x_2 \mid \boldsymbol{\omega}_1) =$$

$$= v_1^2 \mathbf{D}(x_1 \mid \boldsymbol{\omega}_1) + v_2^2 \mathbf{D}(x_2 \mid \boldsymbol{\omega}_1) + 2v_1 v_2 \mathbf{R}(x_1, x_2 \mid \boldsymbol{\omega}_1) = 0.2v_1^2 + 0.2v_2^2 + 0.2v_1 v_2 = 0.2(v_1^2 + v_2^2 + v_1 v_2),$$

$$\sigma_2^2 = \mathbf{D}(X \mid \boldsymbol{\omega}_2) = \mathbf{D}(v_0 + v_1 x_1 + v_2 x_2 \mid \boldsymbol{\omega}_2) =$$

$$= v_1^2 \mathbf{D}(x_1 \mid \boldsymbol{\omega}_2) + v_2^2 \mathbf{D}(x_2 \mid \boldsymbol{\omega}_2) + 2v_1 v_2 \mathbf{R}(x_1, x_2 \mid \boldsymbol{\omega}_2) =$$

$$= 0.2v_1^2 + 0.2v_2^2 - 0.2v_1 v_2 = 0.2(v_1^2 + v_2^2 - v_1 v_2).$$

Область принятия решений в пользу первого класса $G_1 = \{(x_1, x_2):\ v_0 + v_1 x_1 + v_2 x_2 < 0\}$.

Область принятия решений в пользу второго класса $G_2 = \{(x_1, x_2):\ v_0 + v_1 x_1 + v_2 x_2 > 0\}$.

Вероятность ошибочной классификации:

$$\varepsilon = \mathbf{p}_1\varepsilon_1 + \mathbf{p}_2\varepsilon_2 = \frac{1}{2}\int_0^{+\infty} f_X\left(x \mid \boldsymbol{\omega}_1\right)dx + \frac{1}{2}\int_{-\infty}^{0} f_X\left(x \mid \boldsymbol{\omega}_2\right)dx =$$

$$= \frac{1}{2}\int_0^{+\infty} \frac{1}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{\left(x-m_1\right)^2}{2\sigma_1^2}\right)dx + \frac{1}{2}\int_{-\infty}^{0} \frac{1}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{\left(x-m_2\right)^2}{2\sigma_2^2}\right)dx =$$

$$= \frac{1}{2}\int_0^{+\infty} \frac{1}{\sqrt{2\pi 0.2\left(v_1^2+v_2^2+v_1v_2\right)}}\exp\left(-\frac{\left(x-v_0-2v_1\right)^2}{2\cdot 0.2\left(v_1^2+v_2^2+v_1v_2\right)}\right)dx +$$

$$\left\{ z = \frac{x-v_0-2v_1}{\sqrt{0.2\left(v_1^2+v_2^2+v_1v_2\right)}};\ \ dz = \frac{dx}{\sqrt{0.2\left(v_1^2+v_2^2+v_1v_2\right)}} \quad \left[\begin{array}{ll} x=+\infty & \Rightarrow\ z=+\infty \\ x=0 & \Rightarrow\ z=\dfrac{-v_0-2v_1}{\sqrt{0.2\left(v_1^2+v_2^2+v_1v_2\right)}} \end{array}\right] \right\}$$

$$+\frac{1}{2}\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi 0.2\left(v_1^2+v_2^2-v_1v_2\right)}}\exp\left(-\frac{\left(x-v_0-2v_2\right)^2}{2\cdot 0.2\left(v_1^2+v_2^2-v_1v_2\right)}\right)dx =$$

$$\left\{ z = \frac{x-v_0-2v_2}{\sqrt{0.2\left(v_1^2+v_2^2-v_1v_2\right)}};\ \ dz = \frac{dx}{\sqrt{0.2\left(v_1^2+v_2^2-v_1v_2\right)}} \quad \left[\begin{array}{ll} x=-\infty & \Rightarrow\ z=-\infty \\ x=0 & \Rightarrow\ z=\dfrac{-v_0-2v_2}{\sqrt{0.2\left(v_1^2+v_2^2-v_1v_2\right)}} \end{array}\right] \right\}$$

$$= \frac{1}{2}\int_{\frac{-v_0-2v_1}{\sqrt{v_1^2+v_2^2+v_1v_2}}}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)dx + \frac{1}{2}\int_{-\infty}^{\frac{-v_0-2v_2}{\sqrt{v_1^2+v_2^2-v_1v_2}}} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)dx =$$

$$= \frac{1}{2} - \frac{1}{2}\Phi\left(\frac{-v_0-2v_1}{\sqrt{0.2\left(v_1^2+v_2^2+v_1v_2\right)}}\right) + \frac{1}{2}\Phi\left(\frac{-v_0-2v_2}{\sqrt{0.2\left(v_1^2+v_2^2-v_1v_2\right)}}\right) \to \min_{v_0,v_1,v_2},$$

где $\Phi\left(x\right) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)du$ – интеграл вероятности (интегральная функция стандартного нормального распределения). Численная минимизация даёт результат: $v_0 = 0.453$, $v_1 = -0.927$, $v_2 = 0.927$. Разделяющая поверхность $0.453 - 0.927x_1 + 0.927x_2 = 0$, или $x_2 = x_1 - 0{,}489$. Минимальная вероятность ошибки равна 0.0005101382.

```
    # Filename: LinClass2.r R version 3.4.0 (2017-04-21) MacroSoft
(c) 2017
    # Вероятность ошибки линейного классификатора
    # p1 = 1/2; p2 = 1/2; M1 = (2 0); M2 = (0 2);
    # R1 = [.2, .1; .1, .2]; R2 = [.2, -.1; -.1, .2]

    require(mvtnorm)

    fr <- function(x) {
     v1 <- x[1]; v2 <- x[2]; v0 <- x[3]
     (1 - pnorm((-v0-2*v1)/sqrt(.2*(v1*v1+v2*v2+v1*v2)))
     + pnorm((-v0-2*v2)/sqrt(.2*(v1*v1+v2*v2-v1*v2))))/2
    }
    opt_res <- optim(c(-1, 1, 0), fr) # Filename: LinClass2.r R ver-
sion 3.4.0 (2017-04-21) MacroSoft (c) 2017
    # Вероятность ошибки линейного классификатора
    # p1 = 1/2; p2 = 1/2; M1 = (2 0); M2 = (0 2);
    # R1 = [.2, .1; .1, .2]; R2 = [.2, -.1; -.1, .2]

    require(mvtnorm)

    fr <- function(x) {
     v1 <- x[1]; v2 <- x[2]; v0 <- x[3]
     (1 - pnorm((-v0-2*v1)/sqrt(.2*(v1*v1+v2*v2+v1*v2)))
     + pnorm((-v0-2*v2)/sqrt(.2*(v1*v1+v2*v2-v1*v2))))/2
    }
    opt_res <- optim(c(-1, 1, 0), fr)
    print(opt_res$par); print(opt_res$value)
    windows(600, 600); n <- 200
    M1 <- c(2, 0); R1 <- .2*array(c(1, .5, .5, 1), dim=c(2,2))
    M2 <- c(0, 2); R2 <- .2*array(c(1, -.5, -.5, 1), dim=c(2,2))
    xy <- rbind(rmvnorm(n,M1,R1), rmvnorm(n,M2,R2))
    plot(xy, xlab='X1', ylab='X2')
    abline(h=0); abline(v=0); abline(a=-.489, b=1)
    text(2, 2.4, "x2 = x1 - 0.489")
    print(opt_res$par); print(opt_res$value)
    windows(600, 600); n <- 200
    M1 <- c(2, 0); R1 <- .2*array(c(1, .5, .5, 1), dim=c(2,2))
    M2 <- c(0, 2); R2 <- .2*array(c(1, -.5, -.5, 1), dim=c(2,2))
    xy <- rbind(rmvnorm(n,M1,R1), rmvnorm(n,M2,R2))
    plot(xy, xlab='X1', ylab='X2')
    abline(h=0); abline(v=0); abline(a=-.489, b=1)
    text(2, 2.4, "x2 = x1 - 0.489")
```

Рис. 6.3. Фрагмент R-программы для расчёта
и визуализации минимальной вероятности ошибки

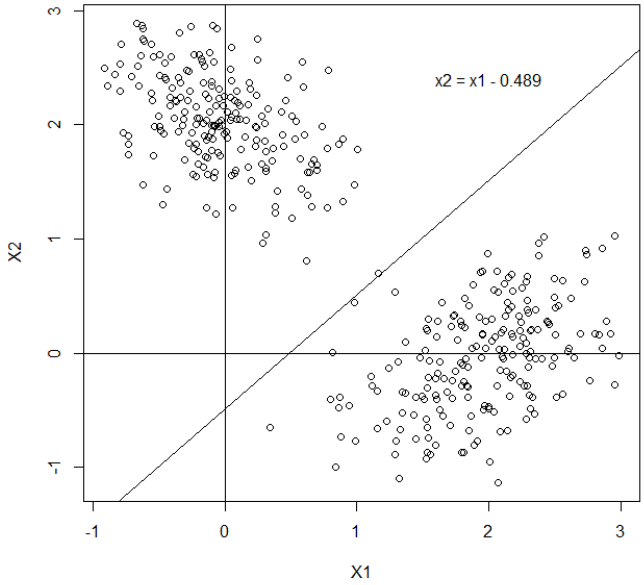


Рис. 6.4. Разделяющая поверхность
для оптимальной линейной классификации

7 ДИСКРИМИНАНТНЫЙ АНАЛИЗ

7.1 Матрицы рассеяния и критерий разделимости

Дискриминант (лат. discriminans (dis-criminantis)) переводится с латыни как различающий, разделяющий). Целью дискриминантного анализа является расчёт некоторого коэффициента, влияющего на эффективность классификации (вероятности ошибок первого и второго рода) при использовании того или иного набора признаков для классификации. Этот коэффициент не должен зависеть от вероятностного распределения признаков, от используемых методов классификации и от соответствующей вероятности ошибки классификации. Рассмотрим два класса объектов с *n*-мерными признаками X_1, X_2 и априорными вероятностями классов $p_1 = \mathbf{Pr}\{\boldsymbol{\omega}_1\}$, $p_2 = \mathbf{Pr}\{\boldsymbol{\omega}_2\}$.

В дискриминантном анализе критерии разделимости классов формулируются с использованием матриц рассеяния внутри классов и матриц рассеяния между классами [2].

Матрица рассеяния между классами S_1 показывает разброс объектов между классами. Эта матрица определяется несколькими способами. Мы используем самый простой:

$S_1 = (M_1 - M_2)(M_1 - M_2)^T$ – расстояние между центрами классов.

Другие методы определения межклассового разброса признаков:

$S_1' = \sum_{i=1}^{2} p_i \left[\left(M_i - M_0 \right) \left(M_i - M_0 \right)^T \right]$ – среднее расстояние до общего центра, где M_0 – математическое ожидание смеси: $M_0 = p_1 M_1 + p_2 M_2$;

$S_1'' = \mathbf{M}\left(X^{(1)} - X^{(2)} \right) \left(X^{(1)} - X^{(2)} \right)^T$ – математическое ожидание расстояния между объектами первого и второго классов.

Матрица рассеяния внутри классов S_2 показывает разброс признаков объектов относительно векторов математических ожиданий классов:

$$S_2 = p_1 R_1 + p_2 R_2 = p_1 \mathbf{M}\left[\left(X_1 - M_1 \right) \left(X_1 - M_1 \right)^T \right] +$$

$$+ p_2 \mathbf{M}\left[\left(X_2 - M_2 \right) \left(X_2 - M_2 \right)^T \right], \text{ где } M_1 = \mathbf{M} X_1, \; M_2 = \mathbf{M} X_2 \; .$$

Критерием разделимости классов должна быть функция от матриц рассеяния, определённых выше. Эта функция должна увеличиваться при увеличении рассеяния между классами S_1 и при уменьшении рассеяния внутри классов S_2. Таким условиям удовлетворяет множество критериев, из которых наиболее простым является

$$J_1 = \operatorname{tr}\left(S_2^{-1} S_1 \right).$$

Другими наиболее употребительными критериями разделимости являются следующие:

$$J_2 = \ln\left(\left| S_2^{-1} S_1 \right| \right) = \ln\left(|S_1| / |S_2| \right),$$

$$J_3 = \operatorname{tr} S_1 - \operatorname{tr} S_2,$$

$$J_4 = \operatorname{tr} S_1 / \operatorname{tr} S_2 .$$

7.2 Разделимость выборки

Допустим, что имеется выборка признаков (обучающая выборка) двух классов объектов (априорные вероятности классов p_1 и p_2, число объектов каждого класса N_1 и N_2)

$$X_1 = \begin{pmatrix} x_1^1 \\ x_1^2 \\ x_1^3 \\ \dots \\ x_1^{N_1} \end{pmatrix}, \quad X_2 = \begin{pmatrix} x_2^1 \\ x_2^2 \\ x_2^3 \\ \dots \\ x_2^{N_2} \end{pmatrix},$$

где $x_i^j = \left(x_{i1}^j \; x_{i2}^j \; \dots \; x_{in}^j \right)^T$ – n-мерный вектор признаков объектов i-го класса в j-м наблюдении. Требуется оценить критерий разделимости $\mathbf{J_1} = \mathrm{tr}\left(S_2^{-1} S_1 \right)$.

Векторы центров классов:

$$M_1 = \left(M_{11} \; M_{12} \; \dots \; M_{1n} \right)^T, \;\; M_{1k} = \frac{1}{N_1} \sum_{j=1}^{N_1} x_{1k}^j,$$

$$M_2 = \left(M_{21} \; M_{22} \; \dots \; M_{2n} \right)^T, \;\; M_{2k} = \frac{1}{N_2} \sum_{j=1}^{N_2} x_{2k}^j.$$

Матрица рассеяния между классами:

$$S_1 = \left(M_1 - M_2 \right) \left(M_1 - M_2 \right)^T.$$

Ковариационные матрицы признаков первого и второго классов:

$$R_1 = \left\{ \frac{1}{N_1} \sum_{j=1}^{N_1} \tilde{x}_{1k}^j \tilde{x}_{1l}^j \right\}_{n \times n}, \; R_2 = \left\{ \frac{1}{N_2} \sum_{j=1}^{N_2} \tilde{x}_{2k}^j \tilde{x}_{2l}^j \right\}_{n \times n},$$

где $\tilde{x}_{ik}^{j}$ – центрированные значения признаков: $\tilde{x}_{ik}^{j} = x_{ik}^{j} - M\left(x_{ik}\right)$.

Матрица рассеяния внутри классов:

$$S_2 = p_1 R_1 + p_2 R_2 = \frac{p_1}{N_1}\sum_{j=1}^{N_1} \tilde{x}_{1k}^{j}\tilde{x}_{1l}^{j} + \frac{p_2}{N_2}\sum_{j=1}^{N_2} \tilde{x}_{2k}^{j}\tilde{x}_{2l}^{j} .$$

7.3 Выбор признаков, максимизирующих критерий $\mathbf{J_1}$

7.3.1 Линейное преобразование случайных векторов

Пусть случайный вектор-столбец $X = \left(x_1 \;\; x_2 \;\; \ldots \;\; x_n\right)^T$ имеет математическое ожидание $\mathbf{M}X$ и ковариационную матрицу R_X. Тогда случайный вектор $Y = \left(y_1 \;\; y_2 \;\; \ldots \;\; y_m\right)^T$, являющийся линейным преобразованием вектора *X*,

$$Y = AX$$

имеет математическое ожидание

$$\mathbf{M}Y = A\mathbf{M}X$$

и ковариационную матрицу

$$R_Y = AR_X A^T,$$

так как

$$R_Y = \mathbf{M}\left[\left(Y - \mathbf{M}Y\right)\left(Y - \mathbf{M}Y\right)^T\right] = \mathbf{M}\left[\left(AX - A\mathbf{M}X\right)\left(AX - A\mathbf{M}X\right)^T\right] =$$
$$= \mathbf{M}\left[A\left(X - \mathbf{M}X\right)\left(X - \mathbf{M}X\right)^T A^T\right] = A\mathbf{M}\left[\left(X - \mathbf{M}X\right)\left(X - \mathbf{M}X\right)^T\right]A^T = AR_X A^T.$$

7.3.2 *Диагонализирующее и декоррелирующее преобразования*

Обозначим через $\lambda_1,\ldots,\lambda_n$ собственные значения, а через $U_1, U_2, \ldots, U_n$ – ортонормированные собственные векторы некоторой ковариационной матрицы R размера $n \times n$.

$$\lambda_i U_i = RU_i, \quad i = \overline{1,n}.$$

Так как ковариационная матрица R является симметричной и положительно определённой, то все её n собственных значений $\lambda_1,\ldots,\lambda_n$ являются вещественными и положительными. Справедливо матричное равенство:

$$U\Lambda = RU,$$

где $U = \left[U_1\, U_2 \ldots U_n\right]$ – ортонормированная матрица собственных векторов матрицы R, расположенных по столбцам; $\Lambda = diag\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right)$ – диагональная матрица собственных значений матрицы R. Из матричной алгебры известно, что $U^T U = I$, $UU^T = I$, $U^{-1} = U^T$, $R = U\Lambda U^T$, $U^T RU = \Lambda$, $\Lambda^{-1/2} U^T RU \Lambda^{-1/2} = I$.

Если в качестве матрицы A линейного преобразования случайного вектора X в случайный вектор Y

$$Y = AX$$

выбрать транспонированную матрицу собственных векторов

$$A = U^T,$$

то ковариационная матрица результирующего вектора $R_Y = AR_X A^T = U^T R_X U = \Lambda$, то есть компоненты вектора Y бу-

дут некоррелированными случайными величинами с дисперсиями $\lambda_1,\ldots,\lambda_n$ – собственными числами матрицы R_X. Такое преобразование A в этом разделе будем называть *диагонализирующим преобразованием*.

Если также потребовать, чтобы компоненты вектора Y обладали единичной дисперсией, в качестве матрицы линейного преобразования следует выбрать

$$A = \Lambda^{-1/2} U^T .$$

Тогда $R_Y = A R_X A^T = \Lambda^{-1/2} U^T R_X U \Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I$. Преобразование $A = \Lambda^{-1/2} U^T$ будем называть *декоррелирующим преобразованием*.

Декоррелированность декоррелированного случайного вектора сохраняется при *любом* ортонормированном преобразовании $A = \Psi^T$: $Z = \Psi^T Y$. То есть, если $R_Y = I$, то

$$R_Z = \Psi^T R_Y \Psi = \Psi^T I \Psi = \Psi^T \Psi = I .$$

Фактически, оба преобразования $A = U^T$ и $A = \Lambda^{-1/2} U^T$ являются одновременно и диагонализирующими, и декоррелирующими.

7.3.3 Одновременная диагонализация двух ковариационных матриц

Пусть R_{1X} и R_{2X} – две в общем случае различные ковариационные матрицы случайных векторов X_1 и X_2. Сначала применим декоррелирующее преобразование $Y_1 = \Lambda_{1X}^{-1/2} U_{1X}^T X_1$ к вектору X_1. Тогда

$$R_{Y1} = \Lambda_{X1}^{-1/2} U_{X1}^T R_{X1} U_{X1} \Lambda_{X1}^{-1/2} = I .$$

То же самое преобразование (согласованное с R_{X1}) $Y_2 = \Lambda_{X1}^{-1/2} U_{X1}^T X_2$ применим к вектору X_2. Получим:

$$R_{Y2} = \Lambda_{X1}^{-1/2} U_{X1}^T R_{X2} U_{X1} \Lambda_{X1}^{-1/2}.$$

В общем случае матрица R_{Y2} не будет диагональной.

Теперь применим ортонормированное преобразование $Z_2 = U_{Y2}^T Y_2$ для диагонализации матрицы R_{Y2} (U_{Y2}^T – ортонормированная матрица собственных векторов ковариационной матрицы R_{2Y} случайного вектора Y_2):

$$R_{Z2} = U_{Y2}^T R_{Y2} U_{Y2} = \Lambda_{Y2}.$$

И, наконец, применим последнее ортонормированное преобразование $Z_1 = U_{X2}^T Y_1$ к декоррелированному вектору Y_1 (при этом её декоррелированность сохраняется, что доказано выше):

$$R_{1Z} = U_{2X}^T R_{1Y} U_{2X} = U_{2X}^T I U_{2X} = U_{2X}^T U_{2X} = I.$$

Таким образом, искомая матрица преобразования, одновременно декоррелирующая X_1 и X_2, равна:

$$A = U_{Y2}^T \Lambda_{X1}^{-1/2} U_{X1}^T: \; Z_1 = \left(U_{Y2}^T \Lambda_{X1}^{-1/2} U_{X1}^T\right) X_1, \; Z_2 = \left(U_{Y2}^T \Lambda_{X1}^{-1/2} U_{X1}^T\right) X_2.$$

$$R_{Z1} = U_{Y2}^T \Lambda_{X1}^{-1/2} \left(U_{X1}^T R_{X1} U_{X1}\right) \Lambda_{X1}^{-1/2} U_{Y2} = U_{Y2}^T \Lambda_{X1}^{-1/2} \left(\Lambda_{X1}\right) \Lambda_{X1}^{-1/2} U_{Y2} =$$
$$= U_{Y2}^T U_{Y2} = I.$$

$$R_{Z2} = U_{Y2}^T \left(\Lambda_{X1}^{-1/2} U_{X1}^T R_{X2} U_{X1} \Lambda_{X1}^{-1/2}\right) U_{Y2} = U_{Y2}^T \left(R_{Y2}\right) U_{Y2} = \Lambda_{Y2}$$

Полученный алгоритм поясняет следующая схема (рис. 7.1).

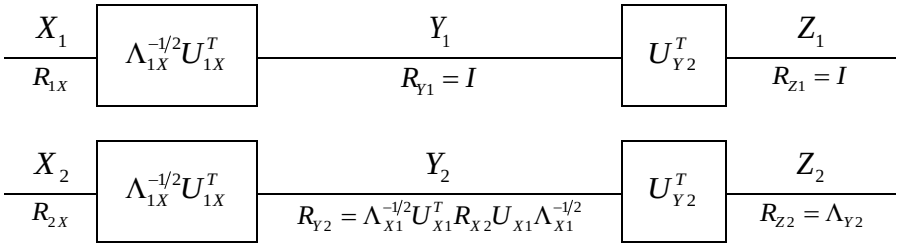


Рис. 7.1. Схема одновременной диагонализации
двух ковариационных матриц

Докажем, что полученная матрица преобразования A является транспонированной матрицей собственных векторов матрицы $R_{X1}^{-1}R_{X2}$.

Собственные числа λ матрицы R_{Y2} определяются из уравнения $\left|R_{Y2}-\lambda I\right|=0$. Подставляя формулы декоррелирующего преобразования, приведённые выше, и учитывая, что матрица декоррелирующего преобразования $\Lambda_{X1}^{-1/2}U_{X1}^T$ и ковариационная матрица R_{X1} не вырождены, получаем:

$$\left|\Lambda_{X1}^{-1/2}U_{X1}^T R_{X2} U_{X1}\Lambda_{X1}^{-1/2}-\lambda\Lambda_{X1}^{-1/2}U_{X1}^T R_{X1} U_{X1}\Lambda_{X1}^{-1/2}\right|=0,$$
$$\left|\left(\Lambda_{X1}^{-1/2}U_{X1}^T\right)\left(R_{X2}-\lambda R_{X1}\right)\left(U_{X1}\Lambda_{X1}^{-1/2}\right)\right|=0,\ \left|R_{X2}-\lambda R_{X1}\right|=0,\quad \left|R_{X1}^{-1}R_{X2}-\lambda I\right|=0.$$

Это означает, что собственные числа λ_i матрицы $R_{X1}^{-1}R_{X2}$ совпадают с собственными числами матрицы R_{Y2}. Подставив в уравнение $U_{Y2}^T R_{Y2} U_{Y2}=\Lambda$ значение $R_{Y2}=\Lambda_{X1}^{-1/2}U_{X1}^T R_{X2} U_{X1}\Lambda_{X1}^{-1/2}$, получаем:

$$U_{Y2}^T\Lambda_{X1}^{-1/2}U_{X1}^T R_{X2} U_{X1}\Lambda_{X1}^{-1/2}U_{Y2}=\Lambda_{Y2},$$
$$R_{2X}U_{1X}\Lambda_{1X}^{-1/2}U_{2Y}=\left(U_{2Y}^T\Lambda_{1X}^{-1/2}U_{1X}^T\right)^{-1}\Lambda_{Y2},$$
$$R_{1X}^{-1}R_{2X}U_{1X}\Lambda_{1X}^{-1/2}U_{2Y}=U_{1X}\Lambda_{1X}^{-1}U_{1X}^T\left(U_{2Y}^T\Lambda_{1X}^{-1/2}U_{1X}^T\right)^{-1}\Lambda_{Y2},$$
$$R_{1X}^{-1}R_{2X}\left(U_{1X}\Lambda_{1X}^{-1/2}U_{2Y}\right)=\left(U_{1X}\Lambda_{1X}^{-1/2}U_{2Y}\right)\Lambda_{Y2}.$$

Таким образом, матрица диагонализирующего преобразования A равна транспонированной матрице собственных векторов $\left(U_{1X}\Lambda_{1X}^{-1/2}U_{2Y}\right)^T$ матрицы $R_{1X}^{-1}R_{2X}$.

7.3.4 *Оптимизация сокращения размерности пространства признаков*

Критерий $J_1 = \mathrm{tr}\left(S_{X2}^{-1}S_{X1}\right)$ может быть максимизирован с использованием линейного преобразования пространства признаков:

$$Y = AX\ ,$$

где матрица преобразования A имеет размер $m \times n\ \left(m \le n\right)$, n – размерность исходного пространства признаков X, m – новая размерность пространства признаков.

Идея заключается в нахождении такой матрицы преобразования A, которая позволяет одновременно диагонализировать две матрицы рассеяния S_{X1} и S_{X2}. При любом выборе матриц рассеяния S_{X1} и S_{X2} в пространстве X соответствующие матрицы рассеяния в пространстве Y в силу линейности преобразования равны

$$S_{1Y} = AS_{1X}A^T,\ S_{2Y} = AS_{2X}A^T.$$

Пусть $\lambda_i, U_i,\ i = 1, 2, 3, \ldots, n$ – собственные значения и собственные векторы матрицы $S_{2X}^{-1}S_{1X}$, $\mu_j, V_j,\ j = 1, 2, 3, \ldots, m$ – собственные значения и собственные векторы матрицы $S_{2Y}^{-1}S_{1Y}$.

За счёт выбора матрицы преобразования A можно провести одновременную диагонализацию матриц S_{1X} и S_{2X}.

$$AS_{1X}A^T = \Lambda,\ AS_{2X}A^T = I_n.$$

При этом $J_{1X} = \operatorname{tr} S_{2X}^{-1} S_{1X} = \operatorname{tr} U \Lambda U^T = \operatorname{tr} U^T U \Lambda = \operatorname{tr} \Lambda = \sum_{i=1}^{n} \lambda_i$.

Аналогично: $J_{1Y} = \operatorname{tr} S_{2Y}^{-1} S_{1Y} = \sum_{i=1}^{m} \mu_i$.

Здесь использовано свойство: $\operatorname{tr} AB = \operatorname{tr} BA$ для любых матриц согласованного размера. Можно доказать [2, п.9.2.2], что критерий J_{1Y} при $m < n$ достигает максимального значения при выборе линейного преобразования $A = \left[U_1\ U_2\ \ldots\ U_m\right]^T$, где $U_1, U_2, \ldots, U_m$ – собственные векторы, соответствующие m максимальным собственным значениям матрицы $S_{2X}^{-1} S_{1X}$, при этом $\mu_i = \lambda_i$, $i = 1, 2, 3, \ldots, m$. Собственные значения считаются упорядоченными по убыванию: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_n$.

8 МЕТОД ОПОРНЫХ ВЕКТОРОВ (SVM)

Под методом опорных векторов (в англоязычной литературе *SVM – Support Vectors Machine – машина опорных векторов*) понимается набор похожих алгоритмов обучения с учителем, разработанных различными авторами в конце прошлого века [4] и использующихся для задач классификации и регрессионного анализа. Мы сосредоточимся на задаче классификации. Обозначения и формулировки в этом разделе приводятся в соответствии с лекциями Воронцова [3].

Формальная постановка задачи. Пусть X – пространство объектов (признаков); Y – множество классов; $X \to Y$ – функциональная зависимость, известная только на объектах обучающей выборки: $X^L = \left(x_i, y_i\right)_{i=1}^{L}$, $y_i = y\left(x_i\right)$; L – количество объектов в обучающей выборке. Требуется построить алгоритм $a: X \to Y$, аппроксимирующий целевую функциональную зависимость на всём пространстве X.

8.1 Линейно разделимая выборка

Рассмотрим классификацию на два класса объектов, описываемых n-мерными векторами: $X = \mathbb{R}^n \to Y = \{-1, +1\}$. Уравнение разделяющей гиперплоскости:

$$\sum_{j=1}^{n} w_j x^j = b, \qquad (8.1)$$

где $x = \left(x^1, x^2, \ldots, x^n\right)$ – признаки объекта x; $w = \left(w^1, w^2, \ldots, w^n\right) \in \mathbb{R}^n$, $b \in \mathbb{R}$ – параметры уравнения гиперплоскости.

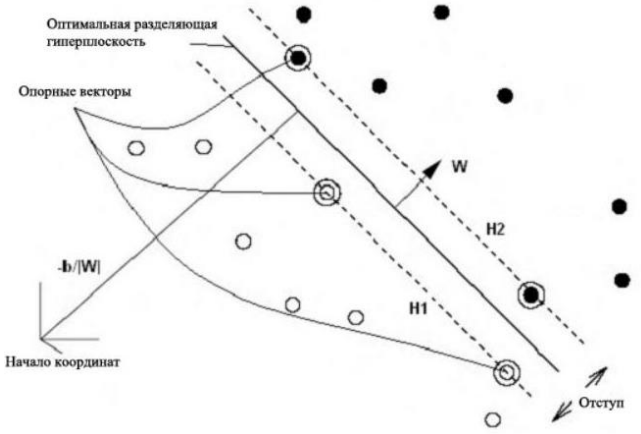


Рис. 8.1. Разделяющая гиперплоскость

В векторном виде уравнение разделяющей гиперплоскости: $\left(w, x\right) = b$, где через $\left(\cdot, \cdot\right)$ здесь и далее обозначается скалярное произведение векторов. Алгоритм классификации:

$$a\left(x\right) = \operatorname{sgn}\left(\left(w, x\right) - b\right) =$$

$$= \begin{cases} +1, \text{ } x \text{ находится с одной стороны} \\ \qquad \text{от гиперплоскости}, \\ -1, \text{ } x \text{ находится с другой стороны} \\ \qquad \text{от гиперплоскости}. \end{cases}$$

От значений коэффициентов уравнения гиперплоскости w, b зависит общее число ошибок классификации в обучающей выборке:

$$Q(w,b)=\sum_{i=1}^{L}\left[\,y_i\cdot\big((w,x_i)-b\big)<0\right],$$

где обозначено: $[\textit{условие}]=\begin{cases}1, \textit{ если условие истинно,}\\ 0, \textit{ если условие ложно.}\end{cases}$

Если $Q(w,b)=0$, то ошибки отсутствуют при некотором положении разделяющей гиперплоскости. Это положение не единственное (см. рис. 8.1). Поэтому для более уверенной классификации потребуем максимизации зазора (margine, отступа) между классами.

Заметим, что алгоритм классификации $a(x)=\operatorname{sgn}\big((w,x_i)-b\big)$ не изменится, если w и b умножить на одну и ту же положительную константу. Удобно выбрать эту константу так, чтобы на всех *граничных* объектах x_i обучающей выборки выполнялись условия

$$(w,x_i)-b=y_i.$$

Это можно сделать, так как при оптимальном положении разделяющей гиперплоскости все граничные объекты находятся от неё на одинаковом расстоянии. Остальные объекты находятся дальше. Таким образом, для всех $x_i\in X^L$:

$$\begin{bmatrix}(w,x_i)-b\le -1, & \textit{если } y_i=-1;\\ (w,x_i)-b\ge +1, & \textit{если } y_i=+1,\end{bmatrix} \text{ или } y_i\cdot\big((w,x_i)-b\big)=1,\ i=\overline{1,L}. \quad (8.2)$$

Условие $-1<(w,x)-b<+1$ задаёт полосу, разделяющую классы. Ни одна из точек обучающей выборки не может лежать внутри этой полосы. Границами полосы служат две параллельные гиперплоскости с направляющим вектором w. Точки, ближайшие к разделяющей гиперплоскости, лежат в точности на границах полосы. При этом сама разделяющая гиперплоскость проходит ровно по середине полосы. *Ширина разделяющей по-*

лосы должна быть максимальной. Пусть x_+ и x_- – две произвольные точки классов $+1$ и -1, лежащие на границе полосы. Тогда ширина полосы есть

$$\left((x_+ - x_-), \frac{w}{\|w\|}\right) = \frac{(w, x_+) - (w, x_+)}{\|w\|} = \frac{(b+1) - (b-1)}{\|w\|} = \frac{2}{\|w\|} \to \min_w.$$

Ширина полосы максимальна, когда норма вектора w минимальна. Требуется найти такие значения параметров w и b, при которых норма вектора w минимальна при условии (8.1). Это задача квадратичного программирования, которая заключается в минимизации квадратичной формы при линейных ограничениях-неравенствах вида (8.1) относительно $(n+1)$ переменных $w^1, w^2, \ldots, w^n, b$:

$$\begin{cases} (w, w) \to \min\limits_{w,b}; \\ y_i \cdot \big((w, x_i) - b\big) \geq 1, \ i = \overline{1, L}. \end{cases} \tag{8.3}$$

В скалярном виде:

$$\begin{cases} w_1^2 + w_2^2 + \ldots + w_n^2 \to \min\limits_{w_1, \ldots, w_n, b}; \\ y_i \cdot \left(w_1 x_i^1 + w_2 x_i^2 + \ldots + w_n x_i^n - b\right) \geq 1, \quad i = \overline{1, L}. \end{cases}$$

По условию Куна–Таккера в теории нелинейного программирования эта задача эквивалентна задаче поиска *седловой точки* функции Лагранжа:

$$\begin{cases} \mathbf{L}(w, b; \lambda) = \dfrac{1}{2}(w, w) - \sum\limits_{i=1}^{L} \lambda_i \left(y_i \cdot \big((w, x_i) - b\big) - 1\right) \to \min\limits_{w,b} \max\limits_{\lambda}; \\ \lambda_i \geq 0, \ i = 1, 2, \ldots, L; \\ \lambda_i = 0 \ \ \textit{или} \ \ (w, x_i) - b = y_i, \ \ i = 1, 2, \ldots, L, \end{cases}$$

где $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ – вектор двойственных переменных.

Условие $\lambda_i = 0$ или $(w, x_i) - b = y_i$, $i = 1,2,\ldots,L$, называется *условием дополняющей нежёсткости*.

Необходимым условием седловой точки является равенство нулю производных Лагранжиана. Из уравнений $\frac{\partial \mathbf{L}(w,b;\lambda)}{\partial w} = 0$ и $\frac{\partial \mathbf{L}(w,b;\lambda)}{\partial b} = 0$ следует:

$$w = \sum_{i=1}^{L} \lambda_i y_i x_i , \tag{8.4}$$

$$\sum_{i=1}^{L} \lambda_i y_i = 0 . \tag{8.5}$$

Из (8.4) следует, что искомый вектор весов w является линейной комбинацией векторов обучающей выборки, причём только тех, для которых $\lambda_i \neq 0$. Согласно условию дополняющей нежёсткости на этих векторах x_i ограничения-неравенства обращаются в равенства: $\langle w, x_i \rangle - b = y_i$, следовательно эти векторы находятся на границе разделяющей полосы. Все остальные векторы отстоят дальше от границы, для них $\lambda_i = 0$ и они не участвуют в суммировании (8.4).

Определение. Если $\lambda_i > 0$ и $(w, x_i) - b = y_i$, то объект обучающей выборки x_i называется опорным вектором (*support vector*).

Подставляя (8.4) и (8.5) обратно в Лагранжиан, получим эквивалентную задачу квадратичного программирования, содержащую только двойственные переменные:

$$\begin{cases} -\mathbf{L}(\lambda) = \dfrac{1}{2}\sum\limits_{i=1}^{L}\sum\limits_{j}^{L}\lambda_i\lambda_j y_i y_j \left(x_i, x_j\right) - \sum\limits_{i=1}^{L}\lambda_i \to \min\limits_{\lambda}; \\ \lambda_i \geq 0,\ i = 1, 2, \ldots, L; \\ \sum\limits_{i=1}^{L}\lambda_i y_i = 0. \end{cases} \tag{8.6}$$

Полученная задача (8.6) имеет единственное решение из-за выпуклости функционала и выпуклости ограничений. Вектор w вычисляется из (8.4). Для любого опорного вектора x_i класса y_i порог $b = (w, x_i) - y_i$. На практике для повышения помехоустойчивости следует брать в качестве b среднее по всем опорным векторам, а ещё лучше медиану:

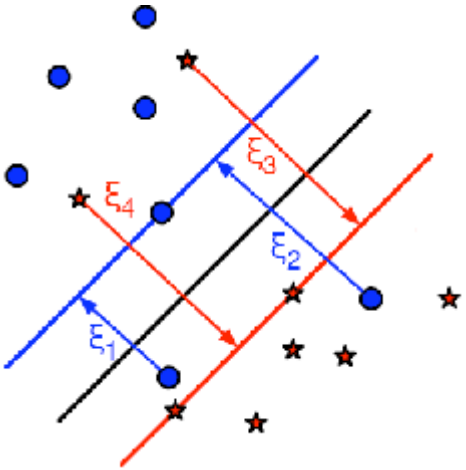
$$b = \operatorname{med}\left\{(w, x_i) - y_i :\ \lambda_i > 0,\ i = \overline{1, L}\right\}. \tag{8.7}$$

В итоге искомый алгоритм классификации может быть записан в следующем виде:

$$a(x) = \operatorname{sgn}\left(\sum_{i=1}^{L}\lambda_i y_i \langle x_i, x\rangle - b\right). \tag{8.8}$$

8.2 Линейно неразделимая выборка

Чтобы обобщить SVM на случай линейной неразделимости, позволим алгоритму допускать ошибки на обучающих объектах, но при этом постараемся, чтобы ошибок было поменьше. Введём набор дополнительных переменных $\xi_i > 0$, характеризующих величину ошибки на объектах ξ_i, $i = \overline{1, L}$. Будем исходить из исходной задачи минимизации квадратичной формы (8.3)



$$\begin{cases} (w,w) \to \min\limits_{w,b}; \\ y_i \cdot \big((w,x_i) - b\big) \geq 1, \;\; i = \overline{1,L}. \end{cases}$$

Смягчим в ней ограничения-неравенства, и одновременно введём в минимизируемый функционал штраф за суммарную ошибку:

$$\begin{cases} \dfrac{1}{2}(w,w) + C\sum\limits_{i=1}^{L} \xi_i \to \min\limits_{w,b,\xi}; \\ y_i \cdot \big((w,x_i) - b\big) \geq 1 - \xi_i, \;\; i = \overline{1,L}; \\ \xi_i \geq 0. \end{cases} \qquad (8.9)$$

Положительная константа C является управляющим параметром метода и позволяет находить компромисс между максимизацией разделяющей полосы и минимизацией суммарной ошибки. Вернёмся к задаче (8.9) и запишем её функцию Лагранжа:

$$\mathbf{L}(w,b,\xi;\lambda,\eta) = \frac{1}{2}(w,w) - \sum_{i=1}^{L} \lambda_i \big(y_i \cdot \big((w,x_i) - b\big) - 1\big) - \sum_{i=1}^{L} \xi_i \cdot (\lambda_i + \eta_i - C), \quad (8.10)$$

где $\eta = (\eta_1, \eta_2, \ldots, \eta_L)$ – вектор переменных, двойственных переменным $\xi = (\xi_1, \xi_2, \ldots, \xi_L)$. Как и в прошлый раз, условия Куна–Таккера сводят задачу к поиску седловой точки функции Лагранжа:

$$\begin{cases} \mathbf{L}\left(w,b,\xi;\lambda,\eta\right) \to \min\limits_{w,b,\xi}\max\limits_{\lambda,\eta}; \\ \xi_i \ge 0,\ \lambda_i \ge 0,\ \eta_i \ge 0,\ \ i=\overline{1,L}; \\ \lambda_i = 0\ \ \textit{или}\ \ y_i \cdot \left(\left(w,x_i\right)-b\right) = 1-\xi_i,\ \ i=\overline{1,L}, \\ \eta_i = 0\ \ \textit{или}\ \ \xi_i = 0,\ \ i=\overline{1,L}. \end{cases}$$

В последних двух строках записаны условия *дополняющей нежёсткости*. Необходимым условием седловой точки является равенство нулю производных Лагранжиана. Из уравнений $\frac{\partial \mathbf{L}}{\partial w} = \frac{\partial \mathbf{L}}{\partial b} = \frac{\partial \mathbf{L}}{\partial \xi_i} = 0$ получаются три полезных соотношения:

$$w = \sum_{i=1}^{L} \lambda_i y_i x_i\ , \tag{8.11}$$

$$\sum_{i=1}^{L} \lambda_i y_i = 0\,, \tag{8.12}$$

$$\eta_i + \lambda_i = C,\ \ i=\overline{1,L}. \tag{8.13}$$

В силу соотношения (8.13) в Лагранжиане обнуляются все члены, содержащие переменные ξ_i и η_i и он принимает тот же вид, что и в случае линейной разделимости. Параметры разделяющей поверхности w и b, согласно формулам (8.11) и (8.12), также выражаются только через двойственные переменные λ_i. Таким образом, задача снова сводится к квадратичному программированию относительно двойственных переменных λ_i. Единственное отличие от линейно разделимого случая состоит в появлении ограничения сверху $\lambda_i \le C$:

$$\begin{cases} -\mathbf{L}(\lambda) = \dfrac{1}{2}\sum\limits_{i=1}^{L}\sum\limits_{j}^{L}\lambda_i\lambda_j y_i y_j \left(x_i, x_j\right) - \sum\limits_{i=1}^{L}\lambda_i \to \min\limits_{\lambda}; \\ 0 \le \lambda_i \le C, \ i = \overline{1, L}; \\ \sum\limits_{i=1}^{L}\lambda_i y_i = 0. \end{cases} \tag{8.14}$$

На практике для построения SVM решают именно эту задачу, а не (8.6), так как гарантировать линейную разделимость выборки в общем случае не представляется возможным. Этот вариант алгоритма называют *SVM с мягким зазором* (*soft-margin SVM*), тогда как в линейно разделимом случае говорят об *SVM с жёстким зазором* (*hard-margin SVM*).

Для алгоритма классификации сохраняется формула (8.8) с той лишь разницей, что теперь ненулевыми λ_i обладают не только опорные объекты, но и объекты-нарушители. В определённом смысле это недостаток *SVM*, поскольку нарушителями часто оказываются шумовые выбросы, а построенное на них решающее правило, по сути дела, опирается на шум.

Константу C обычно выбирают по критерию скользящего контроля. Это трудоёмкий способ, так как задачу приходится решать заново при каждом значении C. Если есть основания полагать, что выборка почти линейно разделима и лишь объекты-выбросы классифицируются неверно, то можно применить фильтрацию выбросов. Сначала задача решается при некотором C и из выборки удаляется небольшая доля объектов, имеющих наибольшую величину ошибки ξ_i. После этого задача решается заново по усечённой выборке. Возможно, придётся проделать несколько таких итераций, пока оставшиеся объекты не окажутся линейно разделимыми.

8.3 Ядра и спрямляющие пространства

Другой подход к решению проблемы линейной неразделимости основан на переходе от исходного пространства признаковых описаний объектов X к новому пространству H с помощью некоторого преобразования $\psi: X \to H$. Если пространство H имеет достаточно высокую размерность, то можно надеяться, что в нём выборка окажется линейно разделимой. Пространство H называют *спрямляющим*. Если предположить, что признаковыми описаниями объектов являются векторы $\psi(x_i)$, а не векторы x_i, то построение *SVM* проводится точно так же, как и ранее. Единственное отличие состоит в том, что скалярное произведение (x, x') в пространстве X всюду заменяется скалярным произведением $(\psi(x), \psi(x'))$ в пространстве H. Отсюда вытекает естественное требование: пространство H должно быть наделено скалярным произведением, в частности, подойдёт любое евклидово, а в общем случае и гильбертово, пространство.

> **Определение**. Функция $K: X \times X \to \mathbb{R}$ *называется ядром (kernel function)*, если она представима в виде $K(x, x') = (\psi(x), \psi(x'))$ при некотором отображении $\psi: X \to H$, где H – пространство со скалярным произведением.

Постановка задачи (8.14), и сам алгоритм классификации (8.8) зависят только от скалярных произведений объектов, но не от самих признаковых описаний. Это означает, что скалярное произведение $\langle x, x' \rangle$ можно формально заменить ядром

$K(x,x')$. Поскольку ядро в общем случае нелинейно, такая замена приводит к существенному расширению множества реализуемых алгоритмов $a: X \to Y$. Более того, можно вообще не строить спрямляющее пространство H в явном виде и вместо подбора отображения $\psi: X \to H$ заниматься непосредственно подбором ядра. Можно пойти ещё дальше и вовсе отказаться от признаковых описаний объектов. Во многих практических задачах объекты изначально задаются информацией об их попарном взаимоотношении, например, отношении сходства. Если эта информация допускает представление в виде двуместной функции $K(x,x')$, удовлетворяющей аксиомам скалярного произведения, то задача может решаться методом *SVM*. Для такого подхода недавно был придуман термин *беспризнаковое распознавание* (*featureless recognition*), хотя многие давно известные метрические алгоритмы классификации (kNN, RBF и др.) также не требуют задания признаковых описаний.

Теорема Мерсера. Функция $K(x,x')$ является ядром тогда и только тогда, когда она симметрична, $K(x,x') = K(x',x)$ и неотрицательно определена:

$\int_X \int_X K(x,x') g(x) g(x') dx dx' \geq 0$ для любой функции $g: X \to \mathbb{R}$.

Конструктивные способы построения ядер. Следующие правила порождения позволяют строить ядра в практических задачах.

1. Произвольное скалярное произведение $K(x,x') = (x,x')$ является ядром.
2. Константа $K(x,x') = 1$ является ядром.

3. Произведение ядер $K(x,x') = K_1(x,x')K_2(x,x')$ является ядром.
4. Для любой функции $\psi : X \to \mathrm{R}$ произведение $K(x,x') = \psi(x)\psi(x')$ является ядром.
5. Линейная комбинация ядер с неотрицательными коэффициентами $K(x,x') = \alpha_1 K_1(x,x') + \alpha_2 K_2(x,x')$ является ядром.
6. Композиция произвольной функции $\phi : X \to X$ и произвольного ядра K_0 является ядром: $K(x,x') = K_0(\phi(x), \phi(x'))$.
7. Если $s : X \times X \to \mathrm{R}$ – произвольная симметричная интегрируемая функция, то $K(x,x') = \int_X s(x,z)s(x',z)dz$ является ядром.
8. Функция вида $K(x,x') = k(x-x')$ является ядром тогда и только тогда, когда Фурье-образ $F[k](\omega) = (2\pi)^{n/2} \int_X e^{-i(\omega,x)} k(x)dx$ неотрицателен.
9. Композиция произвольного ядра K_0 и произвольной функции $f : \mathrm{R} \to \mathrm{R}$, представимой в виде сходящегося степенного ряда с неотрицательными коэффициентами $K(x,x') = f(K_0(x,x'))$, является ядром. В частности, функции $f(z) = e^z$ и $f(z) = \frac{1}{1-z}$ от ядра являются ядрами.

9 ЗАДАЧИ ДЛЯ САМОСТОЯТЕЛЬНОГО РЕШЕНИЯ

1. Доказать, что при одинаковом или противоположном ранжировании факторов X и Y коэффициент ранговой корреляции Кендалла $\tau_{XY} = \pm 1$.

2. Построить МНК-оценки параметров a, b, c, d несимметричной билинейной функции двух переменных $z(x, y) = axy + bx + cy + d$ по наблюдениям в девяти точках:

$z_1 = z(1;1)$, $z_2 = z(1;0)$, $z_3 = z(1;-1)$,
$z_4 = z(0;1)$, $z_5 = z(0;0)$, $z_6 = z(0;-1)$,
$z_7 = z(-1;1)$, $z_8 = z(-1;0)$, $z_9 = z(-1;-1)$.

3. Доказать, что при удвоении числа наблюдений в каждой точке увеличивается точность оценивания параметров линейной регрессии. Рассмотреть пример МНК-оценивания параметров a, b линейной зависимости $y = ax + b$ по независимым наблюдениям в трёх точках $\{-1, 0, 1\}$ и по двойным независимым наблюдениям в трёх точках $\{-1, -1, 0, 0, 1, 1\}$. Как сильно уменьшается при этом погрешность оценивания параметров?

4. Построить МНК-оценки параметров a, b нелинейной функции $y = ax^b$ по трём наблюдениям y_1, y_2, y_3 в точках $x_1 = 1, x_2 = 2, x_3 = 3$.

Указание. Применить логарифмирование исходного уравнения.

5. Найти несмещённую оценку дисперсии σ^2 шума наблюдения ξ в стандартной классической модели линейной

регрессии $y = ax + b + \xi$ по наблюдениям y_1, y_2, y_3, y_4, y_5 в точках $x_1 = -2,\ x_2 = -1,\ x_3 = 0,\ x_4 = 1,\ x_5 = 2$.

6. Производится измерение веса предмета на цифровых весах, имеющих случайную погрешностью ξ (среднее значение погрешености равно нулю, дисперсия равна σ^2). Определить, как изменяется погрешность измерения веса предмета при увеличении числа независимых измерений n.

7. Измерения веса двух предметов производятся на чашечных весах со стрелкой. Стрелка показывает разность весов предметов, располагающихся на двух чашках, с некоторой случайной погрешностью ξ, имеющей нулевое среднее и дисперсию σ^2. Требуется определить вес каждого предмета с максимальной точностью. Допускается провести четыре независимых измерения.

Подсказка. Рассмотреть случаи, когда предметы взвешиваются отдельно по одному и когда два предмета помещаются на различные чашки весов.

8. Сформулировать задачу классической линейной регрессии для определения весов $\theta_1, \theta_2, \theta_3$ трёх предметов при взвешивании на чашечных весах со стрелкой. Построить матрицу эксперимента F и рассчитать погрешности определения весов предметов, если известно, что независимые измерения производятся с некоторой случайной погрешностью ξ, имеющей нулевое среднее и дисперсию σ^2. Рассчитать дисперсию ошибки оценивания веса каждого предмета: 1) при независимом взвешивании каждого из трёх предметов; 2) при размещении одновременно всех трёх предметов на весах в разных комбинациях. Общее число взвешиваний в каждом случае равно 6.

Подсказка. На чашечных весах определяется разность весов предметов, находящихся в двух чашках. На каждую чашку

весов может помещаться несколько предметов (или ни одного). Эксперимент определяется значениями базовой переменной

$$f_i = \begin{cases} +1, \textit{ если } i^{\text{й}} \textit{ предмет находится на левой чашке весов,} \\ -1, \textit{ если } i^{\text{й}} \textit{ предмет находится на правой чашке весов,} \\ 0, \textit{ если } i^{\text{й}} \textit{ предмет не участвует во взвешивании.} \end{cases}$$

(1)

$$\mathrm{F} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \qquad \mathrm{F}^T\mathrm{F} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} = 2\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{D}\hat{\boldsymbol{\theta}} = \sigma^2\left(\mathrm{F}^{\mathrm{T}}\mathrm{F}\right)^{-1} = \frac{\sigma^2}{2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

(2)

$$\mathrm{F} = \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \ \mathrm{F}^T\mathrm{F} = \begin{pmatrix} 6 & 2 & 2 \\ 2 & 6 & -2 \\ 2 & -2 & 6 \end{pmatrix} = 2\begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{pmatrix},$$

$$\mathbf{D}\hat{\boldsymbol{\theta}} = \sigma^2\left(\mathrm{F}^{\mathrm{T}}\mathrm{F}\right)^{-1} = \frac{\sigma^2}{2}\begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & -1 \\ 1 & -1 & 3 \end{pmatrix}^{-1} = \frac{\sigma^2}{32}\begin{pmatrix} 8 & ? & ? \\ ? & 8 & ? \\ ? & ? & 8 \end{pmatrix} = \frac{\sigma^2}{4}\begin{pmatrix} 1 & ? & ? \\ ? & 1 & ? \\ ? & ? & 1 \end{pmatrix}.$$

9. Доказать, что $\overline{\left(x-\overline{x}\right)^2}=\overline{x^2}-\overline{x}^2$, где $x^1,x^2,x^3,\ldots,x^N$ – произвольный набор чисел, $\overline{(\cdot)}$ – операция вычисления среднего арифметического.

10. Доказать общее соотношение для *SS*-технологии Фишера

$$SS_{общ}=SS_{факт}+SS_{ош}$$

на следующем примере:

> Производится многократное измерение некоторого параметра с помощью *m* различных приборов. Каждым прибором выполняется *n* независимых измерений (всего *N=mn* измерений). Результаты измерний: $y_1^1,y_1^2,\ldots,y_1^n$, $y_2^1,y_2^2,\ldots,y_2^n,y_m^1,y_m^2,\ldots,y_m^n$. Средние значения измерений для каждого прибора: $\overline{y}_k=\sum_{i=1}^{n}y_k^i,\ k=\overline{1;m}$.

Подсказка:

Общая вариабельность наблюдаемой переменной:

$SS_{общ}=R_1^2=\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i-\overline{y}\right)^2$, $\overline{y}=\sum_{k=1}^{m}\sum_{i=1}^{n}y_k^i=\sum_{k=1}^{m}\sum_{i=1}^{n}\overline{y}_k$ – общее среднее.

Изменчивость наблюдаемой переменной за счёт случайного шума наблюдения (остаточная сумма квадратов ошибок):

$$SS_{ош}=R_0^2=\sum_{k=1}^{m}\sum_{i=1}^{n}\left(y_k^i-\overline{y}_k\right)^2.$$

Изменчивость наблюдаемой переменной за счёт изменения уровней входного фактора:

$$SS_{факт}=\sum_{k=1}^{m}\sum_{i=1}^{n}\left(\overline{y}_k-\overline{y}\right)^2.$$

11. Дан набор из 8 точек на плоскости:

	A	B	C	D	E	F	G	H
X	**1**	**3**	**4**	**5**	**1**	**4**	**1**	**2**
Y	**3**	**3**	**3**	**3**	**2**	**2**	**1**	**1**

Выполнить кластеризацию точек по двум алгоритмам: 1) *k-means* (число кластеров равно 2, метрика – манхеттенская); 2) иерархическая *агломеративная* кластеризация (расстояние между кластерами измеряется по наиболее удалённому объекту, метрика – манхеттенская). Нарисовать дендрограмму для иерархической кластеризации.

12. Построить байесовский классификатор для двух классов наблюдений по одному признаку, распределённому по равномерному (первый класс) и гауссовскому (второй класс) законам с одинаковыми математическими ожиданиями и одинаковыми дисперсиями, если априорные вероятности классов одинаковы. Найти вероятность ошибочной классификации.

13. Найти разделяющие поверхности четырёх классов **I, II, III, IV** гауссовских наблюдений двумерного случайного вектора X с математическими ожиданиями $\mathbf{a}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mathbf{a}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\mathbf{a}_3 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, $\mathbf{a}_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, некоррелированными компонентами: $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}_3 = \mathbf{R}_4 = \mathbf{R} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ и априорными вероятностями классов $\mathbf{p}_1 = 1/2$, $\mathbf{p}_2 = 1/4$, $\mathbf{p}_3 = 1/8$, $\mathbf{p}_4 = 1/8$.

14. Построить решающее правило для трёх признаков для трёх классов независимых гауссовских наблюдений трёхмерного случайного вектора X с математическими ожиданиями $\mathbf{a}_1 = (1\ 0\ 0)^T$, $\mathbf{a}_2 = (0\ 1\ 0)^T$, $\mathbf{a}_3 = (0\ 0\ 1)^T$ и одинаковыми

корреляционными матрицами $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, если априорные вероятности классов равны **1/2**, **1/3**, **1/6**.

15. Медицинский диагностический прибор анализирует заболевание на основе некоторого признака *X*, обеспечивая **чувствительность** диагностики по критерию Неймана–Пирсона на уровне 0.9. Найти уровень **специфичности**, если решение принимается исходя из значения признака *X*, распределённого по нормальному закону с параметрами (0;1) при отсутствии заболевания или с параметрами (1;1) при наличии заболевания.

16. Построить решающее правило Неймана–Пирсона для разделения двух классов объектов по признаку *x*, распределённому по закону Лапласа с плотностью вероятности $f_1(x) = e^{-|x+1|}$, $x \in (-\infty, \infty)$ (для первого класса) и $f_2(x) = e^{-|x-1|}$, $x \in (-\infty, \infty)$ (для второго класса). Установить связь между вероятностями ошибок первого и второго рода, считая, что пороговое значение $h \in [-1;1]$.

17. Производится обнаружение сигнала s на фоне шума ξ:

$$x = \begin{cases} s + \xi, \textit{ если сигнал присутствует;} \\ \xi, \textit{ если наблюдается только шум.} \end{cases}$$

Построить решающее правило Неймана–Пирсона, если шум является гауссовским с нулевым математическим ожиданием и дисперсией σ^2 и задана вероятность ложной тревоги ε. Найти вероятность пропуска цели.

18. Приведите простейший пример, показывающий, что при оптимальной (байесовской) классификации центр (матема-

тическое ожидание) одного класса может оптимально классифицироваться в другой класс.

19. Построить линейный классификатор для двух классов независимых гауссовских наблюдений двумерного случайного вектора $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ с математическими ожиданиями $\mathbf{a}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ и $\mathbf{a}_2 = \begin{pmatrix} 0 \\ 4 \end{pmatrix}$, корреляционными матрицами $\mathbf{R}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\mathbf{R}_2 = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix}$ и равными априорными вероятностями классов: $\mathbf{p}_1 = \mathbf{p}_2 = 1/2$. Сравнить вероятность ошибки классификации построенного линейного классификатора с вероятностью ошибки байесовской классификации.

20. Построить оптимальный линейный классификатор для двух классов объектов по двум признакам $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, распределённых по нормальному закону с математическими ожиданиями $\mathbf{m}_1 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$, $\mathbf{m}_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$, корреляционными матрицами $\mathbf{R}_1 = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}, \mathbf{R}_2 = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}$ и одинаковыми априорными вероятностями классов. Найти вероятность ошибочной классификации.

21. Построить оптимальный линейный классификатор для трёх классов объектов по двум признакам $\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, распределённых по нормальному закону с математическими ожиданиями $\mathbf{m}_1 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}$, $\mathbf{m}_2 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$, $\mathbf{m}_3 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$, равными корреляционными

матрицами $\mathbf{R}_1 = \mathbf{R}_2 = \mathbf{R}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ и априорными вероятностями классов **1/2**, **1/3**, **1/6**. Найти вероятность ошибочной классификации.

22. Какое расстояние должно быть между центрами двух классов гауссовых наблюдений с дисперсией σ^2 для обеспечения вероятности ошибки классификации не более 1%?

23. Привести графический пример с минимальным количеством объектов с линейно разделимой выборкой, для которого линейное разделение даёт безошибочную классификацию на два класса, а метод классификации по ближайшему соседу даёт одну ошибку.

24. Сравнить с использованием критерия дискриминантного анализа $\mathbf{J}_1$ эффективность классификации с использованием признаков задачи **19** и задачи **20**.

25. Случайный вектор X имеет ковариационную матрицу $R_X = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$. Найти линейное преобразование $A: Y = AX$, диагонализирующее ковариационную матрицу результирующего случайного вектора Y: $R_Y = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

26. Какое максимальное значение может принимать вероятность ошибки при оптимальной байесовской классификации объектов на два класса?

27. Построить МНК-оценки параметров a, b уравнения плоскости $z = ax + by$, проходящей через точку (0,0,0), по наблюдениям в N точках с известными координатами:
$z_1 = z(x_1, y_1)$, $z_2 = z(x_2, y_2)$, $z_3 = z(x_3, y_3)$, …, $z_N = z(x_N, y_N)$.

Список литературы и интернет-ресурсы

1. Кендалл, М. Дж. Теория распределений / М.Дж. Кендалл, А. Стьюарт. – М.: Наука, 1966.
2. Ермаков, С.М. Математическая теория оптимального эксперимента: учеб. пособие / С.М. Ермаков, А.А. Жиглявский. – М.: Наука, 1987.
3. Шеффе, Г. Дисперсионный анализ / Г. Шеффе. – М.: Наука, 1980.
4. Ту, Дж. Принципы распознавания образов: пер. с англ. / Дж. Ту, Р. Гонсалес. – М.: Мир, 1978.
5. Фукунага, К. Введение в статистическую теорию распознавания образов: пер. с англ. / К. Фукунага. – М.: Наука, 1979.
6. Воронцов, К.В. Лекции по методу опорных векторов / К.В. Воронцов. – М.: Наука, 2007.
7. Vapnik, V.N. Statistical Learning Theory / V.N. Vapnik. – М.: Наука, 1998.

Видеолекции

1. Support Vector Machines [Yaser Abu-Mostafa; 2012]. http://www.youtube.com/watch?v=eHsErlPJWUU&list=PLo2GY1dZ5AGmsdzWundK-ZeniCqfiqTxh&index=3
2. Kernel Methods [*Yaser Abu-Mostafa; 2012*]. http://www.youtube.com/watch?v=XUj5JbQihlU&index=2&list=PLo2GY1dZ5AGmsdzWundK-ZeniCqfiqTxh
3. How SVM (Support Vector Machine) algorithm works [*Thales Sehn Körting; 2014*]http://www.youtube.com/watch?v=1NxnPkZM9bc
4. Support Vector Machine & In-depth Convex Analysis [*Sanjeev Sharma*] http://www.youtube.com/watch?v=eh3sM4-3heo&index=1&list=PLo2GY1dZ5AGmsdzWundK-ZeniCqfiqTxh

Приложение А
Статистические функции

А.1 Квантиль распределения

Будем называть *квантилем распределения* непрерывной случайной величины функцию, обратную к функции распределения. Если $f_\xi(x)$ и $F_\xi(x)$ – соответственно плотность вероятности и функция распределения некоторой случайной величины ξ, то

$$F_\xi(x_p) = \int_{-\infty}^{x_p} f_\xi(u)\,du = p, \qquad \text{(А.1)}$$

где $x_p = F_\xi^{-1}(p)$ – квантиль распределения этой случайной величины на уровне p, $0 < p < 1$.

Ниже в таблицах используются следующие обозначения:
df - число степеней свободы; p - вероятность; q - квантиль.

А.2 Нормальное распределение

Плотность вероятности нормальной случайной величины с математическим ожиданием a и дисперсией σ^2:

$$f_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-a)^2}{2\sigma^2}\right]. \qquad \text{(А.2)}$$

Функция нормального распределения (см. табл. А.1 и А.2):

$$F_\xi(x)=\frac{1}{\sigma\sqrt{2\pi}}\int\limits_{-\infty}^{x}e^{-\frac{(u-a)^2}{2\sigma^2}}du=\Phi\left(\frac{x-a}{\sigma}\right)=\frac{1}{2}+\Phi_0\left(\frac{x-a}{\sigma}\right)=$$

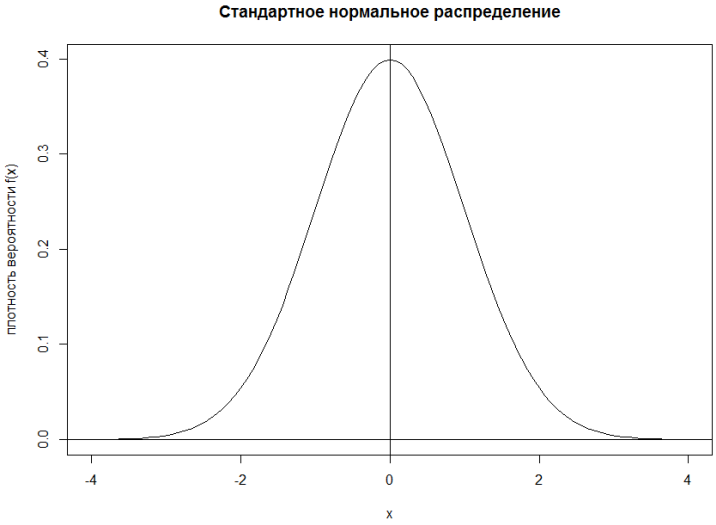
$$=\frac{1}{2}+\frac{1}{2}erf\left(\frac{x-a}{\sigma\sqrt{2}}\right),$$

где $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}e^{-u^2/2}du$ – интеграл вероятности,

$\Phi(x)=\frac{1}{2}+\Phi_0(x)$;

$\Phi_0(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{0}^{x}e^{-u^2/2}du$ – функция Лапласа, $\Phi_0(x)=\frac{1}{2}erf\left(\frac{x}{\sqrt{2}}\right)$;

$erf(x)=\frac{2}{\sqrt{\pi}}\int\limits_{0}^{x}e^{-u^2}du$ – функция ошибок, $erf(x)=2\Phi_0\left(x\sqrt{2}\right)$.



Плотность вероятности нормального случайного вектора $\boldsymbol{\xi}=\left(\xi_1\ \xi_2\ \xi_3\ \ldots\ \xi_n\right)^T\sim \mathrm{N}\left(\mathbf{a};\mathbf{D}_\xi\right)$:

$$f_{\xi}(x_1, x_2, \ldots, x_n) = (2\pi)^{-n/2} \left(\det \mathbf{D}_{\xi}\right)^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{a})^T \mathbf{D}_{\xi}^{-1}(\mathbf{x}-\mathbf{a})\right], \quad (A.3)$$

где $\mathbf{a} = \mathbf{M}\xi$ – вектор математических ожиданий, $\mathbf{D}_{\xi} = \mathbf{M}(\xi - \mathbf{a})(\xi - \mathbf{a})^T$ – дисперсионная (корреляционная) матрица, $\mathbf{x} = (x_1\ x_2\ x_3\ \ldots\ x_n)^T$.

Таблица А.1 – **p-квантили стандартного нормального распределения**

```
p = 0.999 0.998 0.995 0.990 0.980 0.950 0.900 0.800
q = 3.0902323 2.8781617 2.5758293 2.3263479 2.0537489
1.6448536 1.2815516 0.8416212
```

Таблица А.2 – **Интеграл вероятности**
(функция стандартного нормального распределения)[z F(z)]

```
0.000 0.50000 0.050 0.51994 0.100 0.53983 0.150 0.55962 0.200 0.57926 0.250 0.59871 0.300 0.61791 0.350
0.63683 0.400 0.65542 0.450 0.67364
0.500 0.69146 0.550 0.70884 0.600 0.72575 0.650 0.74215 0.700 0.75804 0.750 0.77337 0.800 0.78814 0.850
0.80234 0.900 0.81594 0.950 0.82894
1.000 0.84134 1.050 0.85314 1.100 0.86433 1.150 0.87493 1.200 0.88493 1.250 0.89435 1.300 0.90320 1.350
0.91149 1.400 0.91924 1.450 0.92647
1.500 0.93319 1.550 0.93943 1.600 0.94520 1.650 0.95053 1.700 0.95543 1.750 0.95994 1.800 0.96407 1.850
0.96784 1.900 0.97128 1.950 0.97441
2.000 0.97725 2.050 0.97982 2.100 0.98214 2.150 0.98422 2.200 0.98610 2.250 0.98778 2.300 0.98928 2.350
0.99061 2.400 0.99180 2.450 0.99286
2.500 0.99379 2.520 0.99413 2.540 0.99446 2.560 0.99477 2.580 0.99506 2.600 0.99534 2.620 0.99560 2.640
0.99585 2.660 0.99609 2.680 0.99632
2.700 0.99653 2.720 0.99674 2.740 0.99693 2.760 0.99711 2.780 0.99728 2.800 0.99744 2.820 0.99760 2.840
0.99774 2.860 0.99788 2.880 0.99801
2.900 0.99813 2.910 0.99819 2.920 0.99825 2.930 0.99831 2.940 0.99836 2.950 0.99841 2.960 0.99846 2.970
0.99851 2.980 0.99856 2.990 0.99861
2.990 0.99861 2.991 0.99861 2.992 0.99861 2.993 0.99862 2.994 0.99862 2.995 0.99863 2.996 0.99863 2.997
0.99864 2.998 0.99864 2.999 0.99865
3.000 0.99865 3.050 0.99886 3.100 0.99903 3.150 0.99918 3.200 0.99931 3.250 0.99942 3.300 0.99952 3.350
0.99960 3.400 0.99966 3.450 0.99972
3.500 0.99977 3.550 0.99981 3.600 0.99984 3.650 0.99987 3.700 0.99989 3.750 0.99991 3.800 0.99993 3.850
0.99994 3.900 0.99995 3.950 0.99996
4.000 0.99997 4.050 0.99997 4.100 0.99998 4.150 0.99998 4.200 0.99999 4.250 0.99999 4.300 0.99999 4.350
0.99999 4.400 0.99999 4.450 1.00000
4.500 1.00000 4.600 1.00000 4.700 1.00000 4.800 1.00000 4.900 1.00000 5.000 1.00000 5.100 1.00000 5.200
1.00000 5.300 1.00000 5.400 1.00000
```

А.3 Распределение хи-квадрат

Если независимые случайные величины $\xi_1, \xi_2, \ldots, \xi_k$ имеют нормальное распределение с нулевым математическим

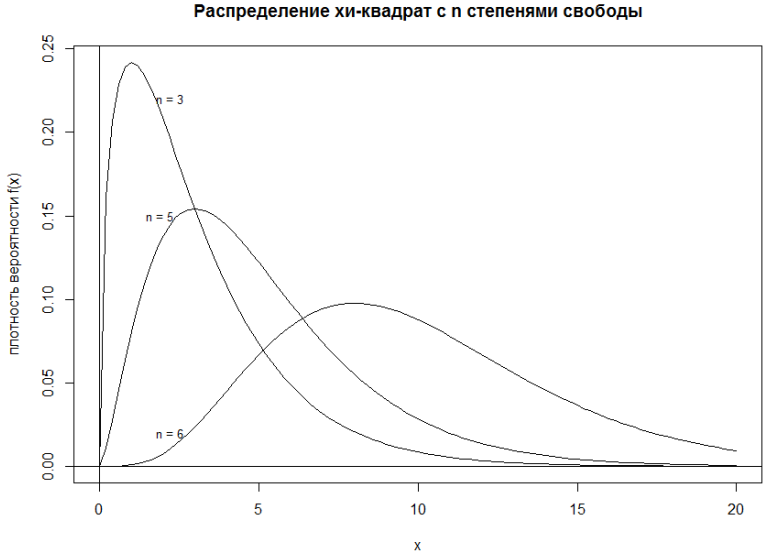
ожиданием и единичной дисперсией, то случайная величина $\chi_k^2 = \sum_{i=1}^{k} \xi_i^2$ имеет χ^2-распределение с k степенями свободы:

$$\sum_{i=1}^{k} \xi_i^2 \sim \chi^2(k). \qquad \text{(A.4)}$$

Если $\boldsymbol{\xi} = (\xi_1\ \xi_2\ \xi_3\ \ldots\ \xi_k)^T \sim \mathrm{N}(\mathbf{0}; \mathbf{D}_\xi)$, то

$$\boldsymbol{\xi}^T \mathbf{D}_\xi^{-1} \boldsymbol{\xi} \sim \chi^2(k). \qquad \text{(A.5)}$$

Если $\boldsymbol{\xi} = (\xi_1\ \xi_2\ \xi_3\ \ldots\ \xi_k)^T \sim \mathrm{N}(\mathbf{0}; \mathbf{I}_k)$, а $k \times k$-матрица $\mathbf{A}$ является симметричной и идемпотентной: $\mathbf{A}^T = \mathbf{A}$, $\mathbf{A}^2 = \mathbf{A}$, то $\boldsymbol{\xi}^T \mathbf{A} \boldsymbol{\xi} \sim \chi^2(\mathrm{tr}\,\mathbf{A})$.



Плотность вероятности распределения хи-квадрат с k степенями свободы:

$$f_{\chi_k^2}(x)=\begin{cases}\dfrac{(1/2)^{k/2}}{\Gamma(k/2)}x^{k/2-1}e^{-x/2},\ x\ge 0,\\ 0,\ x<0,\end{cases} \qquad \text{(A.6)}$$

где $\Gamma(s)=\int_0^\infty t^{s-1}e^{-t}dt$ – гамма-функция.

Числовые характеристики распределения хи-квадрат:

$$\mathbf{M}\chi_k^2=k,\ \mathbf{D}\chi_k^2=2k.$$

Таблица А.3 – **p-квантили хи-квадрат распределения с df степенями свободы**

```
p = 0.999 0.998 0.995 0.990 0.980 0.950 0.900 0.800
df = 1 q = 10.8275662 9.5495357 7.8794386 6.6348966 5.4118944 3.8414588
2.7055435 1.6423744
df = 2 q = 13.8155106 12.4292162 10.5966347 9.2103404 7.8240460 5.9914645
4.6051702 3.2188758
df = 3 q = 16.2662362 14.7955171 12.8381565 11.3448667 9.8374093 7.8147279
6.2513886 4.6416277
df = 4 q = 18.4668270 16.9237582 14.8602590 13.2767041 11.6678434 9.4877290
7.7794403 5.9886167
df = 5 q = 20.5150057 18.9073774 16.7496023 15.0862725 13.3882226 11.0704977
9.2363569 7.2892761
df = 6 q = 22.4577445 20.7911677 18.5475842 16.8118938 15.0332078 12.5915872
10.6446407 8.5580597
df = 7 q = 24.3218863 22.6006709 20.2777399 18.4753069 16.6224219 14.0671404
12.0170366 9.8032499
df = 8 q = 26.1244816 24.3520814 21.9549550 20.0902350 18.1682308 15.5073131
13.3615661 11.0300914
df = 9 q = 27.8771649 26.0564333 23.5893508 21.6659943 19.6790161 16.9189776
14.6836566 12.2421455
df = 10 q = 29.5882984 27.7216472 25.1881796 23.2092512 21.1607675 18.3070381
15.9871792 13.4419576
df = 11 q = 31.2641336 29.3536376 26.7568489 24.7249703 22.6179408 19.6751376
17.2750085 14.6314205
df = 12 q = 32.9094904 30.9569605 28.2995188 26.2169673 24.0539567 21.0260698
18.5493478 15.8119862
df = 13 q = 34.5281790 32.5352145 29.8194712 27.6882496 25.4715091 22.3620325
19.8119293 16.9847970
df = 14 q = 36.1232737 34.0913010 31.3193496 29.1412377 26.8727646 23.6847913
21.0641442 18.1507706
df = 15 q = 37.6972982 35.6276001 32.8013206 30.5779142 28.2594963 24.9957901
22.3071296 19.3106571
df = 16 q = 39.2523548 37.1460934 34.2671865 31.9999269 29.6331773 26.2962276
23.5418289 20.4650793
```

```
df = 17 q = 40.7902167 38.6484515 35.7184657 33.4086636 30.9950472 27.5871116
24.7690353 21.6145605
df = 18 q = 42.3123963 40.1360984 37.1564515 34.8053057 32.3461609 28.8692994
25.9894231 22.7595458
df = 19 q = 43.8201960 41.6102599 38.5822566 36.1908691 33.6874251 30.1435272
27.2035710 23.9004172
df = 20 q = 45.3147466 43.0720001 39.9968463 37.5662348 35.0196255 31.4104328
28.4119806 25.0375056
df = 22 q = 48.2679423 45.9618287 42.7956550 40.2893604 37.6594993 33.9244385
30.8132823 27.3014540
df = 24 q = 51.1785978 48.8118021 45.5585119 42.9798201 40.2703610 36.4150285
33.1962443 29.5533152
df = 26 q = 54.0519624 51.6268519 48.2898823 45.6416827 42.8558348 38.8851387
35.5631713 31.7946101
df = 28 q = 56.8922854 54.4109680 50.9933763 48.2782358 45.4188474 41.3371382
37.9159225 34.0265651
df = 30 q = 59.7030643 57.1674331 53.6719619 50.8921813 47.9618028 43.7729718
40.2560237 36.2501868
df = 32 q = 62.4872191 59.8989864 56.3281150 53.4857718 50.4867045 46.1942595
42.5847451 38.4663128
df = 34 q = 65.2472175 62.6079425 58.9639259 56.0609087 52.9952429 48.6023674
44.9031575 40.6756494
df = 36 q = 67.9851676 65.2962777 61.5811791 58.6192145 55.4888599 50.9984602
47.2121739 42.8787986
df = 38 q = 70.7028874 67.9656958 64.1814124 61.1620868 57.9687973 53.3835406
49.5125798 45.0762782
df = 40 q = 73.4019575 70.6176778 66.7659618 63.6907398 60.4361336 55.7584793
51.8050572 47.2685377
df = 45 q = 80.0767320 77.1794917 73.1660608 69.9568321 66.5552662 61.6562334
57.5053047 52.7288148
df = 50 q = 86.6608152 83.6565574 79.4899785 76.1538912 72.6132524 67.5048065
63.1671210 58.1637966
df = 55 q = 93.1675328 90.0613479 85.7489516 82.2921168 78.6191417 73.3114930
68.7962142 63.5772438
df = 60 q = 99.6072331 96.4035454 91.9516982 88.3794189 84.5799493 79.0819445
74.3970057 68.9720687
```

А.4 Распределение Стьюдента

Если независимые случайные величины $\xi_0, \xi_1, \xi_2, \ldots, \xi_k$ имеют нормальное распределение с нулевым математическим ожиданием и единичной дисперсией, то случайная величина

$$t_k = \xi_0 \left(\frac{1}{k} \sum_{i=1}^{k} \xi_i^2 \right)^{-1/2} \sim t(k) \qquad \text{(А.7)}$$

имеет t-распределение Стьюдента с k степенями свободы.

Другое определение. Если случайные величины $\xi_0 \sim \mathrm{N}(0;1)$, $\chi_k^2 \sim \chi^2(k)$ независимы, то

$$\xi_0 \Big/ \sqrt{\chi_k^2 / k} \sim t(k) . \qquad \text{(A.8)}$$

Плотность вероятности t-распределения Стьюдента:

$$f_{t_k}(x) = \frac{\Gamma\left((k+1)/2\right)}{\sqrt{\pi k}\,\Gamma\left(k/2\right)} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}} .$$

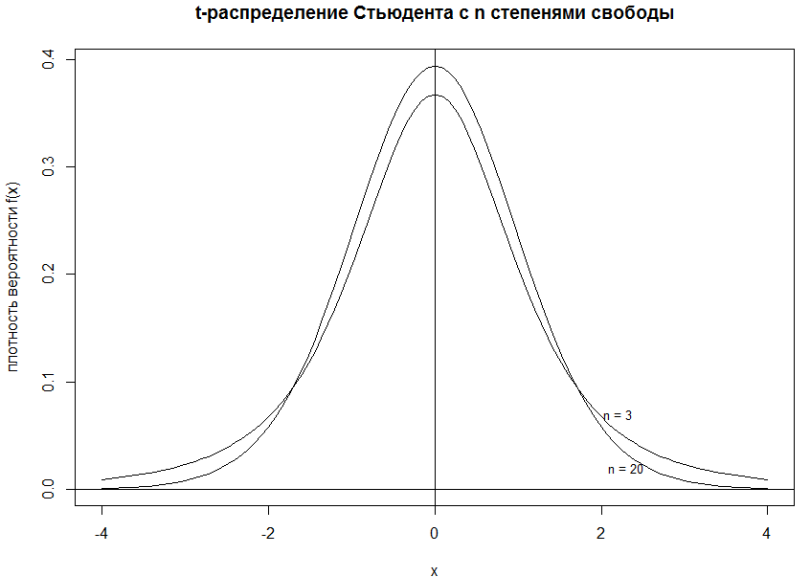


Таблица А.4 – **p-квантили t-распределения Стьюдента с df степенями свободы**

```
p = 0.999 0.998 0.995 0.990 0.980 0.950 0.900 0.800
df = 1 q = 318.3088390 159.1528487 63.6567412 31.8205160 15.8945448 6.3137515
3.0776835 1.3763819
df = 2 q = 22.3271248 15.7639146 9.9248432 6.9645567 4.8487322 2.9199856
1.8856181 1.0606602
df = 3 q = 10.2145319 8.0526131 5.8409093 4.5407029 3.4819088 2.3533634
1.6377444 0.9784723
```

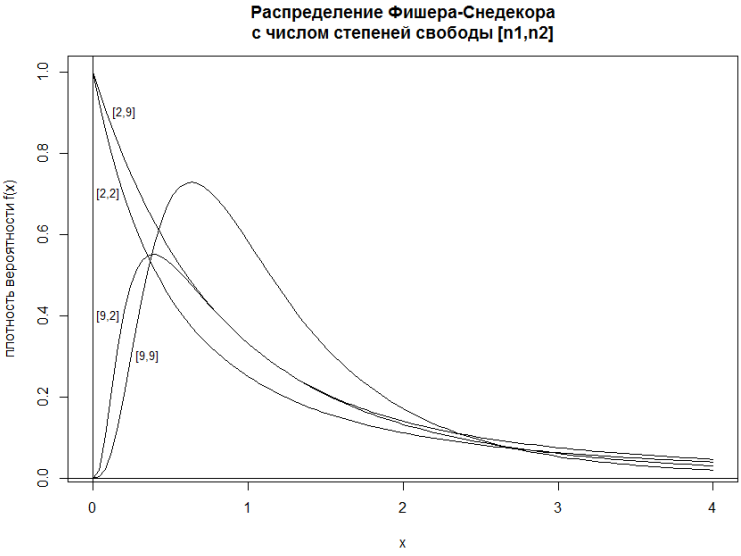
df = 4 q = 7.1731822 5.9513728 4.6040949 3.7469474 2.9985279 2.1318468
1.5332063 0.9409646
df = 5 q = 5.8934295 5.0302444 4.0321430 3.3649300 2.7565085 2.0150484
1.4758840 0.9195438
df = 6 q = 5.2076262 4.5241279 3.7074280 3.1426684 2.6122418 1.9431803
1.4397557 0.9057033
df = 7 q = 4.7852896 4.2071246 3.4994833 2.9979516 2.5167524 1.8945786
1.4149239 0.8960296
df = 8 q = 4.5007909 3.9909504 3.3553873 2.8964594 2.4489850 1.8595480
1.3968153 0.8888895
df = 9 q = 4.2968057 3.8345103 3.2498355 2.8214379 2.3984410 1.8331129
1.3830287 0.8834039
df = 10 q = 4.1437005 3.7162355 3.1692727 2.7637695 2.3593146 1.8124611
1.3721836 0.8790578
df = 11 q = 4.0247010 3.6237693 3.1058065 2.7180792 2.3281398 1.7958848
1.3634303 0.8755300
df = 12 q = 3.9296333 3.5495437 3.0545396 2.6809980 2.3027217 1.7822876
1.3562173 0.8726093
df = 13 q = 3.8519824 3.4886730 3.0122758 2.6503088 2.2816036 1.7709334
1.3501713 0.8701515
df = 14 q = 3.7873902 3.4378666 2.9768427 2.6244941 2.2637813 1.7613101
1.3450304 0.8680548
df = 15 q = 3.7328344 3.3948290 2.9467129 2.6024803 2.2485403 1.7530504
1.3406056 0.8662450
df = 16 q = 3.6861548 3.3579113 2.9207816 2.5834872 2.2353584 1.7458837
1.3367572 0.8646670
df = 17 q = 3.6457674 3.3258988 2.8982305 2.5669340 2.2238453 1.7396067
1.3333794 0.8632790
df = 18 q = 3.6104849 3.2978778 2.8784405 2.5523796 2.2137033 1.7340636
1.3303909 0.8620487
df = 19 q = 3.5794001 3.2731475 2.8609346 2.5394832 2.2047014 1.7291328
1.3277282 0.8609506
df = 20 q = 3.5518083 3.2511618 2.8453397 2.5279770 2.1966577 1.7247182
1.3253407 0.8599644
df = 22 q = 3.5049920 3.2137824 2.8187561 2.5083246 2.1828926 1.7171444
1.3212367 0.8582661
df = 24 q = 3.4667773 3.1831995 2.7969395 2.4921595 2.1715447 1.7108821
1.3178359 0.8568555
df = 26 q = 3.4349972 3.1577165 2.7787145 2.4786298 2.1620289 1.7056179
1.3149719 0.8556652
df = 28 q = 3.4081552 3.1361574 2.7632625 2.4671401 2.1539349 1.7011309
1.3125268 0.8546475
df = 30 q = 3.3851849 3.1176818 2.7499957 2.4572615 2.1469663 1.6972609
1.3104150 0.8537673
df = 32 q = 3.3653059 3.1016728 2.7384815 2.4486776 2.1409037 1.6938887
1.3085728 0.8529985
df = 34 q = 3.3479343 3.0876678 2.7283944 2.4411496 2.1355813 1.6909243
1.3069516 0.8523212
df = 36 q = 3.3326243 3.0753130 2.7194846 2.4344941 2.1308714 1.6882977
1.3055139 0.8517200
df = 38 q = 3.3190297 3.0643332 2.7115576 2.4285676 2.1266740 1.6859545
1.3042302 0.8511828
df = 40 q = 3.3068777 3.0545110 2.7044593 2.4232568 2.1229098 1.6838510
1.3030771 0.8506998

```
df = 45 q = 3.2814798 3.0339594 2.6895850 2.4121159 2.1150048 1.6794274
1.3006493 0.8496819
df = 50 q = 3.2614091 3.0176960 2.6777933 2.4032719 2.1087213 1.6759050
1.2987137 0.8488692
df = 55 q = 3.2451491 3.0045060 2.6682160 2.3960811 2.1036068 1.6730340
1.2971343 0.8482054
df = 60 q = 3.2317091 2.9935936 2.6602830 2.3901195 2.0993628 1.6706489
1.2958211 0.8476530
```

А.5 Распределение Фишера

Если независимые случайные величины χ_k^2 и χ_l^2 имеют χ^2-распределение соответственно с k и l степенями свободы, то случайная величина $F_{k,l} = \dfrac{\chi_k^2/k}{\chi_l^2/l}$ имеет распределение Фишера (*F*-распределение, распределение Фишера–Снедекора) с k и l степенями свободы:

$$\frac{\chi_k^2/k}{\chi_l^2/l} \sim F(k,l).$$



Если случайная величина t_k имеет распределение Стьюдента с k степенями свободы, то случайная величина t_k^2 имеет распределение Фишера с 1 и k степенями свободы:

$$t_k^2 \sim F(1,k). \qquad \text{(A.9)}$$

Отсюда следует соотношение между квантилем t-распределения Стьюдента и квантилем распределения Фишера:

$$t_k^{1-\alpha/2} = \sqrt{F_{1,k}^{1-\alpha}}\,. \qquad \text{(A.10)}$$

Плотность вероятности распределения Фишера:

$$f_{F_{k,l}}(x) = \begin{cases} \dfrac{\Gamma\left((k+l)/2\right)}{\Gamma\left(k/2\right)\Gamma\left(l/2\right)}\left(\dfrac{k}{l}\right)^{k/2} x^{k/2-1}\left(1+\dfrac{k}{l}x\right)^{-(k+l)/2}, & x \ge 0, \\ 0, & x < 0. \end{cases} \qquad \text{(A.11)}$$

Таблица А.5 – **p-квантили F-распределения Фишера с числом степеней свободы [n1,n2]**

```
p = 0.900
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 39.86346 8.526316 5.538319 4.544771 4.060420 3.775950 3.589428
3.457919 3.360303
[2,] 49.50000 9.000000 5.462383 4.324555 3.779716 3.463304 3.257442
3.113118 3.006452
[3,] 53.59324 9.161790 5.390773 4.190860 3.619477 3.288762 3.074072
2.923796 2.812863
[4,] 55.83296 9.243416 5.342644 4.107250 3.520196 3.180763 2.960534
2.806426 2.692680
[5,] 57.24008 9.292626 5.309157 4.050579 3.452982 3.107512 2.883344
2.726447 2.610613
[6,] 58.20442 9.325530 5.284732 4.009749 3.404507 3.054551 2.827392
2.668335 2.550855
[7,] 58.90595 9.349081 5.266195 3.978966 3.367899 3.014457 2.784930
2.624135 2.505313
[8,] 59.43898 9.366770 5.251671 3.954940 3.339276 2.983036 2.751580
2.589349 2.469406
```

```
[9,] 59.85759 9.380544 5.239996 3.935671 3.316281 2.957741 2.724678
2.561238 2.440340
p = 0.950
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 161.4476 18.51282 10.127964 7.708647 6.607891 5.987378 5.591448
5.317655 5.117355
[2,] 199.5000 19.00000 9.552094 6.944272 5.786135 5.143253 4.737414
4.458970 4.256495
[3,] 215.7073 19.16429 9.276628 6.591382 5.409451 4.757063 4.346831
4.066181 3.862548
[4,] 224.5832 19.24679 9.117182 6.388233 5.192168 4.533677 4.120312
3.837853 3.633089
[5,] 230.1619 19.29641 9.013455 6.256057 5.050329 4.387374 3.971523
3.687499 3.481659
[6,] 233.9860 19.32953 8.940645 6.163132 4.950288 4.283866 3.865969
3.580580 3.373754
[7,] 236.7684 19.35322 8.886743 6.094211 4.875872 4.206658 3.787044
3.500464 3.292746
[8,] 238.8827 19.37099 8.845238 6.041044 4.818320 4.146804 3.725725
3.438101 3.229583
[9,] 240.5433 19.38483 8.812300 5.998779 4.772466 4.099016 3.676675
3.388130 3.178893

p = 0.980
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 1012.545 48.50505 20.61798 14.03961 11.322754 9.876365 8.987714
8.389477 7.960512
[2,] 1249.500 49.00000 18.85813 12.14214 9.454406 8.052094 7.202569
6.636592 6.233995
[3,] 1350.505 49.16573 18.10970 11.34354 8.670194 7.286980 6.453946
5.901384 5.509665
[4,] 1405.833 49.24874 17.69377 10.89942 8.233030 6.859440 6.034727
5.488919 5.102656
[5,] 1440.612 49.29859 17.42876 10.61574 7.952932 6.584743 5.764727
5.222715 4.839502
[6,] 1464.455 49.33184 17.24510 10.41859 7.757703 6.392778 5.575612
5.035888 4.654497
[7,] 1481.803 49.35560 17.11028 10.27352 7.613672 6.250824 5.435476
4.897197 4.516938
[8,] 1494.986 49.37342 17.00711 10.16225 7.502956 6.141480 5.327335
4.789995 4.410456
[9,] 1505.341 49.38729 16.92561 10.07418 7.415156 6.054611 5.241280
4.704564 4.325487

   p = 0.990
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,]  4052.181  98.50251  34.11622  21.19769  16.25818  13.745023
12.246383 11.258624 10.561431
```

```
[2,] 4999.500 99.00000 30.81652 18.00000 13.27393 10.924767 9.546578
8.649111 8.021517
[3,] 5403.352 99.16620 29.45670 16.69437 12.05995 9.779538 8.451285
7.590992 6.991917
[4,] 5624.583 99.24937 28.70990 15.97702 11.39193 9.148301 7.846645
7.006077 6.422085
[5,] 5763.650 99.29930 28.23708 15.52186 10.96702 8.745895 7.460435
6.631825 6.056941
[6,] 5858.986 99.33259 27.91066 15.20686 10.67225 8.466125 7.191405
6.370681 5.801770
[7,] 5928.356 99.35637 27.67170 14.97576 10.45551 8.259995 6.992833
6.177624 5.612865
[8,] 5981.070 99.37421 27.48918 14.79889 10.28931 8.101651 6.840049
6.028870 5.467123
[9,] 6022.473 99.38809 27.34521 14.65913 10.15776 7.976121 6.718752
5.910619 5.351129

    p = 0.995
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
[1,] 16210.72 198.5013 55.55196 31.33277 22.78478 18.63500 16.235558
14.688199 13.613609
[2,] 19999.50 199.0000 49.79928 26.28427 18.31383 14.54411 12.403957
11.042412 10.106714
[3,] 21614.74 199.1664 47.46723 24.25912 16.52977 12.91660 10.882447
9.596475 8.717055
[4,] 22499.58 199.2497 46.19462 23.15450 15.55606 12.02753 10.050491
8.805130 7.955885
[5,] 23055.80 199.2996 45.39165 22.45643 14.93961 11.46370 9.522059
8.301799 7.471158
[6,] 23437.11 199.3330 44.83847 21.97458 14.51326 11.07304 9.155336
7.951992 7.133850
[7,] 23714.57 199.3568 44.43410 21.62169 14.20045 10.78592 8.885389
7.694143 6.884908
[8,] 23925.41 199.3746 44.12557 21.35198 13.96096 10.56576 8.678115
7.495906 6.693300
[9,] 24091.00 199.3885 43.88240 21.13908 13.77165 10.39149 8.513823
7.338595 6.541090
```

Приложение Б
Сведения из теории матриц

Б.1 Спектральное разложение симметричной матрицы

Рассмотрим произвольную *симметричную* $n \times n$-матрицу $\mathbf{A}$; $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ – её собственные числа.

Матрица $\mathbf{A}$ может быть представлена в виде:

$$\mathbf{A} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T = \sum_{i=1}^{n} \lambda_i P_i P_i^T , \qquad (Б.1)$$

где $P_1, P_2, P_3, \ldots, P_n$ – ортонормированные собственные векторы-столбцы матрицы $\mathbf{A}$, $\mathbf{P} = \left[P_1 \; P_2 \; P_3 \; \ldots \; P_n \right]$ – ортогональная $n \times n$-матрица собственных векторов матрицы $\mathbf{A}$ ($\mathbf{P}^{-1} = \mathbf{P}^T$, $\mathbf{P}\mathbf{P}^T = \mathbf{P}^T\mathbf{P} = \mathbf{I}_n$, $\mathbf{I}_n$ – единичная $n \times n$-матрица), $\mathbf{\Lambda} = diag\left(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n\right)$ – диагональная матрица собственных чисел матрицы $\mathbf{A}$.

Определитель матрицы $\mathbf{A}$:

$$\det \mathbf{A} = \prod_{i=1}^{n} \lambda_i . \qquad (Б.2)$$

Если $n \times n$-матрица $\mathbf{A}$ является положительно (неотрицательно) определённой, то

$$\left[\mathbf{A}\right]_{ii} > 0 \;\; \left(\left[\mathbf{A}\right]_{ii} \geq 0\right), \; \lambda_i > 0 \;\; \left(\lambda_i \geq 0\right), \; i = \overline{1,n} . \qquad (Б.3)$$

Б.2 Свойства операции *trace* (след)

Во всех рассмотренных ниже свойствах считается, что фигурирующие в формулах матрицы имеют согласованные размеры.

Если $\mathbf{A}$ и $\mathbf{B}$ – произвольные прямоугольные матрицы, то

$$\operatorname{tr}\mathbf{AB} = \operatorname{tr}\mathbf{BA}. \qquad \text{(Б.4)}$$

Если $\mathbf{P}$ – ортогональная матрица ($\mathbf{PP}^T = \mathbf{P}^T\mathbf{P} = \mathbf{I}_n$), то

$$\operatorname{tr}\mathbf{PAP}^T = \operatorname{tr}\mathbf{A}. \qquad \text{(Б.5)}$$

Если $\mathbf{A}$ – симметричная матрица, то

$$\operatorname{tr}\mathbf{A}^k = \sum_{i=1}^{n} \lambda_i^k, \ k = 0, \pm 1, \pm 2, \pm 3, \ldots, \qquad \text{(Б.6)}$$

где $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ – собственные числа матрицы $\mathbf{A}$.

Б.3 Линейное преобразование случайного вектора

Если ξ – случайный вектор, а $\mathbf{c}$ и $\mathbf{B}$ – произвольные детерминированные вектор и матрица соответствующих размеров, то

$$\mathbf{M}\left(\mathbf{c}+\mathbf{B}\xi\right) = \mathbf{c}+\mathbf{B}\mathbf{M}\xi, \ \mathbf{D}\left(\mathbf{c}+\mathbf{B}\xi\right) = \mathbf{B}\mathbf{D}\xi\mathbf{B}^T. \qquad \text{(Б.7)}$$

Если $\xi \sim \mathrm{N}\left(\mathbf{a}, \mathbf{D}_\xi\right)$, $\eta = \mathbf{c} + \mathbf{B}\xi$, то $\eta \sim \mathrm{N}\left(\mathbf{a}+\mathbf{c}, \mathbf{B}\mathbf{D}\xi\mathbf{B}^T\right)$.

Приложение В
Десять основных алгоритмов анализа данных

На IEEE International Conference on Data Mining (ICDM) in December 2006 выделены десять основных алгоритмов анализа данных (**The top 10 data mining algorithms** – табл. В.1).

Таблица В.1 – **Десять основных алгоритмов анализа данных**

ID3	**ID3** (**Iterative Dichotomiser 3**) is an algorithm invented by Ross Quinlan used to generate a decision tree from a dataset
C4.5	Алгоритм для **построения** деревьев решений, разработанный в 1993 Джоном Квинланом (вариант машинного обучения с учителем). C4.5 является усовершенствованной версией алгоритма ID3 того же автора. В частности, в новую версию были добавлены отсечение ветвей, возможность работы с числовыми атрибутами, а также возможность построения дерева из неполной обучающей выборки, в которой отсутствуют значения некоторых атрибутов
***k*-Means**	**Метод k-средних** (быстрый кластерный анализ) – наиболее популярный метод кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузомhttps://ru.wikipedia.org/wiki/K-means_-_cite_note-1 и Стюартом Ллойдом. Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров
ИСОМАД (Isodata)	**ИСОМАД** (Итеративный самоорганизующийся метод анализа данных, **Isodata** – Iterative Self-Organizing Data Analysis Techniques) аналогичен методу ***k*-Means**, однако обладает более широким набором параметров и вспомогательных эвристических процедур
SVM	**Метод опорных векторов** (*SVM – support vector machine*) – набор схожих алгоритмов **обучения с учителем**, использующихся для задач классификации и регрессионного анализа. Принадлежит к семейству линейных классификаторов, может также рассматриваться как специальный случай регуляризации по Тихонову. Особым свойством метода опорных

Продолжение табл. В1

	векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора, поэтому метод также известен как метод классификатора с максимальным зазором. Основная идея метода – перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей наши классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора
Apriori	**Apriori** – алгоритм поиска ассоциативных правил. Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis
EM	**EM-алгоритм** (*Expectation-maximization algorithm*) – алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей в случае, когда модель зависит от некоторых скрытых переменных. Каждая итерация алгоритма состоит из двух шагов. На ***E***-шаге (**expectation**) вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. На ***M***-шаге (**maximization**) вычисляется оценка максимального правдоподобия, таким образом увеличивается ожидаемое правдоподобие, вычисляемое на ***E***-шаге. Затем это значение используется для ***E***-шага на следующей итерации. Алгоритм выполняется до сходимости. Часто EM-алгоритм используют для разделения смеси гауссиан

Продолжение табл. В1

PageRank	**PageRank** – один из алгоритмов **ссылочного ранжирования**. Алгоритм применяется к коллекции документов, связанных гиперссылками (таких, как веб-страницы из всемирной паутины), и назначает каждому из них некоторое численное значение, измеряющее его «важность» или «авторитетность» среди остальных документов. Вообще говоря, алгоритм может применяться не только к веб-страницам, но и к любому набору объектов, связанных между собой взаимными ссылками, то есть к любому графу
AdaBoost	**AdaBoost** (сокращение от *Adaptive Boosting*) – **алгоритм усиления классификаторов** путем объединения их в комитет, предложенный Йоавом Фройндом и Робертом Шапире. Этот алгоритм может использоваться в сочетании с несколькими алгоритмами классификации для улучшения их эффективности. AdaBoost является адаптивным в том смысле, что каждый следующий комитет классификаторов строится по объектам, неверно классифицированным предыдущими комитетами. AdaBoost чувствителен к шуму в данных и выбросам. Однако он менее подвержен переобучению по сравнению с другими алгоритмами машинного обучения
***k*NN**	Метод ***k*** ближайших соседей (англ. ***k****-nearest neighbor algorithm*, ***k***NN) — метод **автоматической** классификации объектов. Основным принципом **метода ближайших соседей** является то, что объект присваивается тому классу, который является наиболее распространённым среди соседей данного элемента. Соседи берутся исходя из множества объектов, классы которых уже известны, и, исходя из ключевого для данного метода значения k высчитывается, какой класс наиболее многочислен среди них. Каждый объект имеет конечное количество атрибутов (размерностей). Предполагается, что существует определенный набор объектов с уже имеющейся классификацией
Naive Bayes	**Наивный байесовский классификатор** – простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими (наивными) предположениями о независимости. В зависимости от точной природы вероятностной модели наивные байесовские классификаторы могут обучаться очень эффективно. Во многих практи-

	ческих приложениях для оценки параметров для наивных байесовых моделей используют метод максимального правдоподобия; другими словами, можно работать с наивной байесовской моделью не веря в байесовскую вероятность и не используя байесовские методы. Несмотря на наивный вид и, несомненно, очень упрощенные условия, наивные байесовские классификаторы часто работают намного лучше во многих сложных жизненных ситуациях. Достоинством наивного байесовского классификатора является малое количество данных для обучения, необходимых для оценки параметров, требуемых для классификации
CART	**Classification and Regression Trees** (Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone, 1984) is a binary recursive partitioning procedure capable of processing continuous and nominal attributes both as targets and predictors. Data are handled in their raw form; no binning is required or recommended. Trees are grown to a maximal size without the use of a stopping rule and then pruned back (essentially split by split) to the root via cost-complexity pruning. The next split to be pruned is the one contributing least to the overall performance of the tree on training data (and more than one split may be removed at a time)

Учебное издание

Храмов Александр Григорьевич

МЕТОДЫ И АЛГОРИТМЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Учебное пособие

Редактор Н.С. Куприянова
Компьютерная верстка Л.Р. Дмитриенко

Подписано в печать 22.07.2019. Формат 60x84 1/16.
Бумага офсетная. Печ. л. 11,0.
Тираж 25 экз. Заказ . Арт. – 27(Р1У)/2019.

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»
(САМАРСКИЙ УНИВЕРСИТЕТ)
443086 Самара, Московское шоссе, 34.

Изд-во Самарского университета.
443086 Самара, Московское шоссе, 34.