МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
УНИВЕРСИТЕТ имени академика С.П. КОРОЛЕВА»
(Самарский университет)

М. А. ПОРУЧИКОВ

АНАЛИЗ 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 студентов, обучающихся по программе высшего образования по направлению подготовки 38.03.05 Бизнес-информатика

САМАРА
Издательство Самарского университета
2016

УДК 33(075)
ББК 65.050я7
П 602

Рецензенты: д-р экон. наук, проф. Д. Ю. Иванов,
д-р техн. наук, проф. Н. Н. Васин

Поручиков, Михаил Алексеевич

П 602 **Анализ данных:** учеб. пособие / *М.А. Поручиков.* – Самара: Изд-во Самарского университета, 2016. – 88 с.

ISBN 978-5-7883-1085-5

Приведены общие сведения о месте и роли анализа данных в современной системе знаний. Рассмотрены основы регрессионного анализа, классификации, кластерного анализа, быстродействия систем анализа данных. Приведены вопросы для самоконтроля, задачи для самостоятельного решения. Представлены указания по выполнению лабораторных работ.

Пособие предназначено для студентов, изучающих дисциплину «Анализ данных» по направлению подготовки 38.03.05 Бизнес-информатика.

Разработано на кафедре математических методов в экономике.

УДК 33(075)
ББК 65.050я7

ISBN 978-5-7883-1085-5

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ 4
ОСНОВЫ АНАЛИЗА ДАННЫХ 5
Роль анализа данных в современном мире 5
Научные исследования 6
Программное обеспечение 7
Построение системы анализа данных 7
Вопросы для самоконтроля 7
Практические задания 8
СБОР И ПОДГОТОВКА ДАННЫХ 9
Общие сведения 9
Источники данных 10
Сбор данных 10
Подготовка данных 13
Вопросы для самоконтроля 14
Лабораторная работа «Сбор и подготовка данных» 15
РЕГРЕССИОННЫЙ АНАЛИЗ 20
Общие сведения 20
Аналитическое решение 22
Численное решение 24
Выбор функции гипотезы 27
Вопросы для самоконтроля 32
Лабораторная работа «Регрессионный анализ» 33
КЛАССИФИКАЦИЯ ДАННЫХ 37
Общие сведения 37
Бинарная классификация 38
Качество классификации 42
Множественная классификация 44
Вопросы для самоконтроля 54
Лабораторная работа «Бинарная классификация» 55
Лабораторная работа «Множественная классификация» 57
КЛАСТЕРНЫЙ АНАЛИЗ 61
Общие сведения 61
Метод к-средних 61
Вопросы для самоконтроля 69
Лабораторная работа «Кластерный анализ» 70
БЫСТРОДЕЙСТВИЕ СИСТЕМ АНАЛИЗА ДАННЫХ 75
Общие сведения 75
Вычислительная сложность 75
Вопросы для самоконтроля 78
Задачи 79
Лабораторная работа «Быстродействие систем анализа данных» 80
ЗАКЛЮЧЕНИЕ 84
БИБЛИОГРАФИЧЕСКИЙ СПИСОК 85

ВВЕДЕНИЕ

Данное учебное пособие предназначено для студентов, обучающихся по программе высшего образования по направлению 38.03.05 «Бизнес-информатика» и направлено на формирование следующих компетенций согласно соответствующему федеральному государственному образовательному стандарту [1]:

- способность к самоорганизации и самообразованию;
- способность работать с компьютером как средством управления информацией, работать с информацией из различных источников, в том числе в глобальных компьютерных сетях;
- способность использовать основные методы естественно-научных дисциплин в профессиональной деятельности для теоретического и экспериментального исследования;
- способность использовать соответствующий математический аппарат и инструментальные средства для обработки, анализа и систематизации информации по теме исследования;
- умение готовить научно-технические отчеты, презентации, научные публикации по результатам выполненных исследований.

Для успешного освоения материала, представленного в учебном пособии, необходимо владение основами линейной алгебры и математического анализа, а также базовыми навыками работы с электронными таблицами *Microsoft Excel* и оформления документов в текстовом редакторе (например, *Microsoft Word* или *OpenOffice Write*).

Пособие построено по модульному принципу. Каждый модуль включает теоретический материал, вопросы для самоконтроля и задания для самостоятельного решения, лабораторные работы. Выполнение лабораторных работ предполагается с помощью специализированного программного обеспечения, размещенного в курсе «Анализ данных» системы дистанционного обучения (СДО) Самарского университета [2].

ОСНОВЫ АНАЛИЗА ДАННЫХ

Роль анализа данных в современном мире

Современный этап развития человечества характеризуется экспоненциальным ростом количества накопленной информации. Согласно исследованию [3], к 2007 году человечество имело возможность хранения информации объемом $2.9 \cdot 10^{20}$ байт. Большой объем данных порождают научные эксперименты. Так, в апреле 2016 года в открытый доступ поступили 300 Тбайт экспериментальных данных, полученных на большом адронном коллайдере [4]. Функционирование многих технических систем также сопровождается сбором большого количества данных. Например, самолет Боинг-787 генерирует около 500 Гбайт данных за один полет [5]. Однако для выделения из накопленных данных полезной информации требуется определенная обработка этих данных.

Также существует тенденция к переложению функции принятия решений – изначально функции человека – на так называемые экспертные системы (специализированные информационные системы). Экспертные системы позволяют повысить скорость и точность принятия решений. Как правило, функционирование экспертных систем связано с анализом большого объема данных (рис. 1).

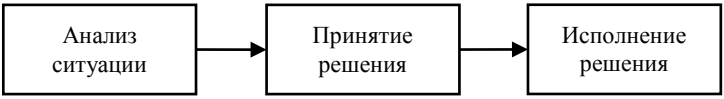


Рис. 1. Процесс управления

В целом анализ данных можно определить как процесс поиска скрытых закономерностей и генерации новых знаний. К основным задачам анализа данных можно отнести прогнозирование, классификацию, поиск схожих черт, выдачу рекомендаций, выявление отклонений. Анализ данных – междисциплинарная область знаний, находящаяся на стыке математики, теории алгоритмов и информационных технологий.

В англоязычных источниках для обозначения сферы анализа данных используется термины Data Mining и Machine Learning

(машинное обучение). Согласно энциклопедии Британника, машинное обучение является дисциплиной направления «искусственный интеллект» (Artificial Intelligence), в свою очередь принадлежащего к области компьютерных наук (Computer Science) [6].

Необходимость анализа больших объемов накопленных данных привела к созданию специализированных подразделений во многих компаниях. Некоторые компании, например Яндекс, реализуют собственные образовательные проекты в этой области [7].

Научные исследования

В сфере анализа данных ведутся активные научные исследования. Анализ публикаций, индексированных в реферативной базе данных SCOPUS, показывает устойчивый рост количества научных работ (табл. 1).

Таблица 1. Количество публикаций в базе SCOPUS

Ключевые слова	Периоды				
	1991-1995	1996-2000	2001-2005	2006-2010	2011-2015
Data Analysis	221072	333879	536742	906164	1008666
Machine Learning	1889	2958	9301	26388	43548
Expert System	13073	14372	19861	34466	41352
Data Mining	1376	4662	18185	42347	59646
Big Data	1321	2554	5110	10544	28220
Deep Learning	263	596	1221	3377	7410

В последнее время особый интерес в сфере анализа данных вызывают такие направления исследований, как «большие данные» (Big Data) и «глубокое обучение» (Deep Learning).

Передовые разработки в сфере искусственного интеллекта поражают воображение. В 1997 году компьютер DeepBlue впервые выиграл матч из шести партий у чемпиона мира по шахматам [8]. В рамках проекта DeepQA разрабатывается система искусственного интеллекта, позволяющая воспринимать вопросы на естественных языках [9]. Ведется разработка беспилотных автомобилей, ядром системы управления которых является система искусственного

интеллекта. Лидером в этой области можно считать компанию Google с проектом Google Self-Driving Car Project [10].

Программное обеспечение

В основе систем анализа данных лежит программное обеспечение. При проектировании систем анализа данных могут быть использованы следующие подходы:

- использование «коробочного» программного обеспечения общего назначения (например *Microsoft Excel*);
- использование программного обеспечения, ориентированного на математические задачи (например *Matlab*, *Octave*, *R*);
- разработка специализированного программного обеспечения с использованием готовых библиотек, включающих наборы специальных функций обработки данных.

При разработке специализированного ПО рекомендуется использовать готовые библиотеки функций обработки данных. Так, для нейросетевого анализа можно применить библиотеку *FANN*, имеющую версии для языков программирования *C#, C++, Java, Python, R, Matlab* [11], а для решения задач обработки изображений – библиотеку *OpenCV*, имеющую версии для языков *Python, Java, Ruby, Matlab* и др. [12].

Построение системы анализа данных

Можно предложить следующий общий алгоритм построения системы анализа данных:

1 Постановка задачи.
2 Определение источников данных.
3 Выбор метода и алгоритма обработки данных.
4 Выбор аппаратной платформы.
5 Выбор или разработка программного обеспечения.
6 Верификация построенной системы.

Отметим, что шаги 3 - 5 тесно связаны друг с другом: например, изменение аппаратной платформы может повлечь необходимость повторной разработки программного обеспечения.

Вопросы для самоконтроля

1 Дайте определение понятия «анализ данных».
2 Перечислите основные задачи анализа данных.

3 Приведите примеры применения методов анализа данных.

4 Приведите пример актуального направления в области анализа данных.

5 Приведите алгоритм построения системы анализа данных.

Практические задания

1 Найдите в сети Интернет два сайта, на которых используются системы прогнозирования.

2 Найдите в сети Интернет два сайта, на которых используются рекомендательные системы.

3 Пользуясь системой SCOPUS, проанализируйте динамику количества публикаций за пять лет по направлениям Deep Learning, Big Data, Recommender Systems, Social Network Analysis.

4 Пользуясь системой SCOPUS, найдите пять публикаций с наибольшей цитируемостью публикаций за последние десять лет по направлениям Deep Learning, Big Data, Recommender Systems, Social Network Analysis.

5 Пользуясь системами SCOPUS, Web of Science, E-library (РИНЦ), выявите нескольких ведущих ученых в сфере анализа данных.

СБОР И ПОДГОТОВКА ДАННЫХ

Общие сведения

Анализ данных включает три основных этапа (рис. 2).

Рис. 2. Этапы анализа данных

Данные по виду можно подразделить на числовые и категориальные.

Числовые данные (Numerical Data) – это данные, характеризующие состояние какого-либо параметра изучаемого объекта. Наиболее часто такие данные бывают представлены вещественными числами. Примерами числовых данных являются заработная плата, население страны, артериальное давление, температура воздуха.

Категориальные данные (Categorical Data) – это данные, образующие признак принадлежности к какой-либо группе. Примерами категориальных данных являются экзаменационная оценка, цвет автомобиля, уровень образования человека.

В фрагменте набора данных по маркетинговой кампании в банке [13] поля Age и Balance являются числовыми, а поля Job, Marital, Education и Housing – категориальными (табл. 2).

Таблица 2. Анкетные данные клиентов банка

Age	Job	Marital	Education	Balance	Housing
58	management	married	tertiary	2143	yes
44	technician	single	secondary	29	yes
33	entrepreneur	married	secondary	2	yes
47	blue-collar	married	unknown	1506	yes
33	unknown	single	unknown	1	no
35	management	married	tertiary	231	yes
28	management	single	tertiary	447	yes
42	entrepreneur	divorced	tertiary	2	yes
58	retired	married	primary	121	yes
43	technician	single	secondary	593	yes

Источники данных

В настоящее время в открытом доступе есть большое количество баз данных, содержащих самые разнообразные сведения. Так, самым большим источником данных по разнообразным показателям стран мира в целом можно считать базу данных Всемирного банка [14], содержащую годовые значения 331 показателя стран мира за период с 1960 по 2014 годы в форматах HTML, XLS и XML.

По состоянию на 23 декабря 2015 года самым большим источником открытых данных по Российской Федерации является «Портал открытых данных Российской Федерации» [15], содержащий более 4,1 тыс. наборов данных. Предполагается, что предоставление свободного доступа к отдельным данным может способствовать повышению качества государственного, регионального и муниципального управления. Принцип открытости получил отдельное название – «открытые данные» (Open Data). В Российской Федерации концепция открытых данных упоминается в Федеральном законе «Об информации, информационных технологиях и о защите информации» [16].

Также большой объем открытых статистических данных содержится в банке данных Федеральной службы государственной статистики [17].

Сбор данных

Сбор данных – процесс формирования структурированного набора данных в цифровой форме. В некоторых случаях процесс сбора данных может включать также этап оцифровки.

Как правило, оцифрованные данные бывают представлены в виде:

- электронных таблиц в форматах XLS либо ODS;
- текстовых файлов в формате CSV;
- веб-страниц в формате HTML;
- файлов в формате XML;
- базы данных с доступом по технологии JSON либо через специализированный интерфейс (API).

Автоматизированный сбор данных

В случаях, когда источники данных структурированы и представлены в сети Интернет, возможна реализация

автоматизированного сбора данных. Программное обеспечение *Microsoft Excel* имеет специальное средство для сбора данных, в том числе из сети Интернет.

Рассмотрим пример реализации автоматизированного сбора данных на примере онлайн-табло аэропорта Домодедово (рис. 3).

Номер рейса	Пункт назначения	Дата и время плановые	Фактическое время		Статус рейса	
IB 3143	Мадрид(Барахас)	30 дек 07:00	30 дек 07:25		Отправлен	
S7 4063 (совмещен с IB 3143)	Мадрид(Барахас)	30 дек 07:00	30 дек 07:25		Отправлен	
7R 169	Пенза(Терновка)	30 дек 07:05	30 дек 07:14		Отправлен	
LH 1451	Франкфурт-На-Майне	30 дек 07:05	30 дек 07:27		Отправлен	
S7 153	Кишинев	30 дек 07:05	30 дек 08:45		Отправлен	
9U 9153 (совмещен с S7 153)	Кишинев	30 дек 07:05	30 дек 08:45		Отправлен	
U6 2997	Нукус	30 дек 07:20	30 дек 07:45		Отправлен	
U6 575	Симферополь (Центральный)	30 дек 07:20	30 дек 07:20		Отменен	

Рис. 3. Фрагмент онлайн-табло вылета аэропорта Домодедово

Для получения данных необходимо выполнить следующие шаги:

1) запустить программу *Microsoft Excel*;

2) перейти пункт главного меню «Данные»;

3) выбрать пункт «Из Веба» в подменю «Получить внешние данные» (рис. 4);

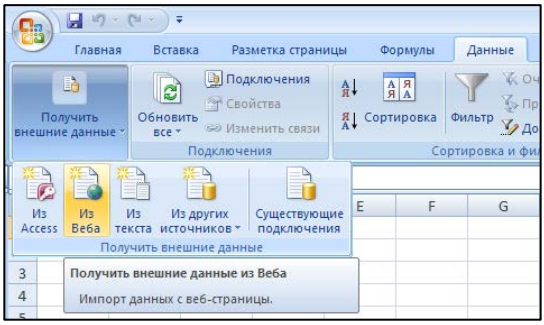


Рис. 4. Импорт данных. Шаг 1

4) в открывшемся окне «Создание веб-запроса»в поле «Адрес» набрать адрес интернет-страницы, содержащей искомые данные, и нажать кнопку «Пуск»;

5) на открывшейся странице с помощью зеленого маркера выделить таблицу, содержащую искомые данные (рис. 5).

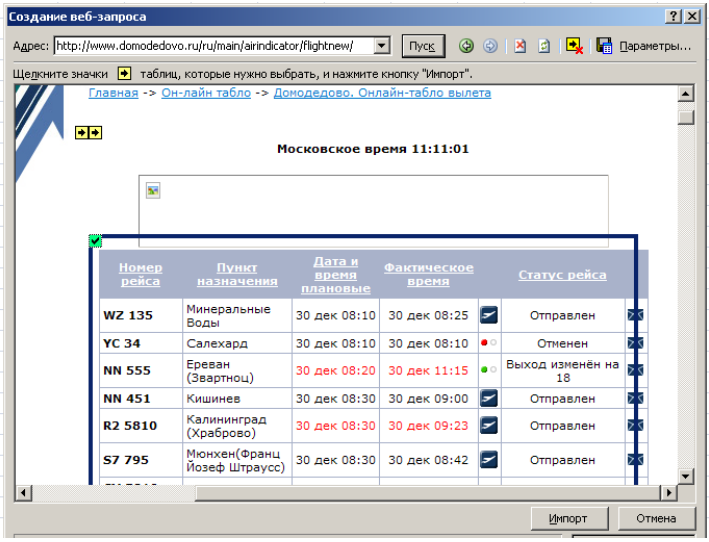
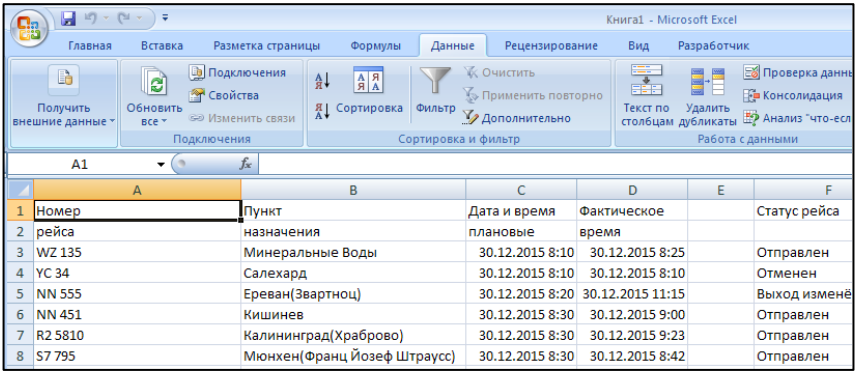


Рис. 5. Импорт данных. Шаг 2

В результате выполненных действий искомые данные будут импортированы на активный лист документа Excel (рис. 6).



	A	B	C	D	E	F
1	Номер	Пункт	Дата и время	Фактическое		Статус рейса
2	рейса	назначения	плановые	время		
3	WZ 135	Минеральные Воды	30.12.2015 8:10	30.12.2015 8:25		Отправлен
4	YC 34	Салехард	30.12.2015 8:10	30.12.2015 8:10		Отменен
5	NN 555	Ереван(Звартноц)	30.12.2015 8:20	30.12.2015 11:15		Выход изменё
6	NN 451	Кишинев	30.12.2015 8:30	30.12.2015 9:00		Отправлен
7	R2 5810	Калининград(Храброво)	30.12.2015 8:30	30.12.2015 9:23		Отправлен
8	S7 795	Мюнхен(Франц Йозеф Штраусс)	30.12.2015 8:30	30.12.2015 8:42		Отправлен

Рис. 6. Результат импорта данных

Аналогичным образом может быть построена система сбора любых данных, представленных в сети Интернет.

Подготовка данных

Для использования в системах анализа данные должны быть представлены в определенном, как правило, табличном виде. Однако зачастую наборы данных имеют следующие особенности:

- отличную от табличной форму представления;
- пропуски отдельных данных;
- некорректные значения;
- большие числовые значения;
- текстовые данные.

Перечисленные особенности могут либо привести к затруднениям в процессе дальнейшей обработки данных, либо сделать её невозможной.

Для устранения отмеченных несоответствий могут быть применены следующие операции:

- структурирование – приведение данных к табличному (матричному) виду;
- отбор – исключение записей с отсутствующими или некорректными значениями;
- нормализация – приведение числовых значений к определенному диапазону, например к диапазону 0...1;
- кодирование – это представление категориальных данных в числовой форме. Например, при бинарной классификации один из классов можно представить числом «0», а другой класс – числом «1». При множественной классификации система кодирования несколько усложняется: создается несколько числовых полей по количеству классов в выборке данных, каждый класс кодируется проставлением числа «1» в соответствующем поле.

Рассмотрим пример подготовки данных. Пусть имеется выборка анкетных данных клиентов банка (табл. 3).

Таблица 3. Анкетные данные клиентов банка

№	Age	Marital	Balance	Housing
1	47	married	1506	yes
2	33	single	1	no
3	35	married	high	yes
4	28	single	447	yes
5	42	divorced	2	yes
6	58		121	yes
7	43	single	593	yes

Для приведения этой выборки данных в «правильный» формат необходимо выполнить следующие операции:

1) исключить записи №3 и №6 как имеющие отсутствующие или некорректные значения;

2) нормализовать числовые значения в столбцах *Age* и *Balance*;

3) закодировать категориальные данные в столбцах *Marital* и *Housing*.

После выполнения этих операций набор данных примет следующий вид (табл. 4).

Таблица 4. Обработанная выборка данных

№	Age	Marital1	Marital2	Marital3	Balance	Housing
1	1,000	1	0	0	1,000	1
2	0,263	0	1	0	0,000	0
4	0,000	0	1	0	0,296	1
5	0,737	0	0	1	0,001	1
7	0,789	0	1	0	0,393	1

Вопросы для самоконтроля

1 Приведите примеры непрерывных данных.

2 Приведите примеры категориальных данных.

3 Дайте определения понятию «источник данных».

4 Приведите способы классификации источников данных.

5 Охарактеризуйте понятие «открытые данные».

6 Приведите примеры источников открытых данных.

7 Перечислите основные форматы хранения данных.

8 Приведите алгоритм построения системы сбора данных на основе программного обеспечения Microsoft Excel.

9 Обоснуйте необходимость подготовки данных.

10 Охарактеризуйте операцию форматирования данных.

11 Приведите пример форматирования данных.

12 Охарактеризуйте операцию отбора данных.

13 Приведите пример отбора данных

14 Охарактеризуйте операцию нормализации данных.

15 Приведите пример нормализации данных.

16 Охарактеризуйте операцию кодирования данных.

17 Приведите пример кодирования данных.

Лабораторная работа «Сбор и подготовка данных»

Общие сведения

Целями работы являются:

- ознакомление со структурой источников открытых данных, изучение способов хранения и представления данных;
- приобретение навыка построения системы сбора данных.

Задачи:

1 Исследование наборов данных, представленных на портале открытых данных data.gov.ru.

2 Исследование наборов данных, представленных на портале data.worldbank.org.

3 Построение автоматизированной системы сбора данных.

В качестве инструментального средства используется программное обеспечение *Microsoft Excel*.

Варианты задания

Таблица 5. Задания по части 1 «Исследование портала data.gov.ru»

Вариант	Тематика	Адрес в сети Интернет
1	Государство	http://data.gov.ru/rubriki/gosudarstvo
2	Экономика	http://data.gov.ru/rubriki/ekonomika
3	Образование	http://data.gov.ru/rubriki/education
4	Здоровье	http://data.gov.ru/rubriki/zdorove
5	Экология	http://data.gov.ru/rubrics/ecology
6	Транспорт	http://data.gov.ru/rubriki/transport
7	Культура	http://data.gov.ru/rubrics/culture
8	Спорт	http://data.gov.ru/rubrics/sport
9	Строительство	http://data.gov.ru/rubriki/stroitelstvo
10	Досуг и отдых	http://data.gov.ru/rubrics/leisure-and-entertainment
11	Торговля	http://data.gov.ru/rubriki/torgovlya
12	Туризм	http://data.gov.ru/rubrics/tourism
13	Электроника	http://data.gov.ru/rubrics/electronics
14	Картография	http://data.gov.ru/rubrics/cartography
15	Безопасность	http://data.gov.ru/rubriki/bezopasnost
16	Метеоданные	http://data.gov.ru/rubrics/weather

Таблица 6. Задания по части 2 «Исследование портала data.worldbank.org»

Вариант	Тематика
1	Agriculture & Rural Development
2	Aid Effectiveness
3	Climate Change
4	Economy & Growth
5	Education
6	Energy & Mining
7	Environment
8	External Debt
9	Financial Sector
10	Gender
11	Health
12	Infrastructure
13	Poverty
14	Private Sector
15	Public Sector
16	Science & Technology
17	Social Development
18	Social Protection & Labor
19	Trade
20	Urban Development

Таблица 7. Задания по части 2 «Автоматизированный сбор данных»

Вариант	Данные
1	Онлайн-табло какого-либо аэропорта/вокзала
2	Котировки акций / валют / драгоценных металлов / полезных ископаемых на какой-либо бирже
3	По предложению студента

Порядок выполнения

1 Исследование наборов данных на портале data.gov.ru:

1.1 Выберите вариант задания (табл. 5).

1.2 Найдите произвольный набор данных на портале data.gov.ru по тематике, указанной в выбранном варианте задания. Набор должен быть представлен в формате *csv* и кодировке *Windows*.

1.3 Загрузите на компьютер найденный набор данных и его паспорт.

1.4 Проведите анализ набора данных: определите количество записей и полей в наборе данных.

2 Исследование наборов данных на портале data.worldbank.org:

2.1 Выберите вариант задания (табл. 6).

2.2 Найдите произвольный набор данных на портале data.worldbank.org по тематике, указанной в выбранном варианте задания.

2.3 Загрузите на компьютер найденный набор данных в формате *XLS*.

2.4 На основе набора данных подготовьте выборку, содержащую значения показателя за все годы для трёх произвольно выбранных стран мира.

2.5 На основе подготовленной выборки постройте график, иллюстрирующий изменение показателя со временем для трёх стран мира.

2.6 Сохраните файл.

3 Построение системы автоматизированной системы сбора данных:

3.1 Выберите вариант задания (табл. 7).

3.2 Найдите интернет-сайт, содержащий указанные в задании данные.

3.3 Запустите *Microsoft Excel*.

3.4 Выберите пункт «Из Веба» в меню «Данные».

3.5 В адресной строке появившегося окна «Создание веб-запроса» наберите адрес найденной ранее веб-страницы.

3.6 Выберите таблицу, содержащую искомые данные.

3.7 Нажмите кнопку «Импорт».

3.8 В появившемся окне «Импорт данных» нажмите кнопку «Свойства».

3.9 В появившемся окне «Свойства внешнего диапазона» задайте параметр «Период обновления», равный 1 минуте, параметр «Обновление при открытии файла» - «Да».

3.10 Нажмите кнопку «ОК».

3.11 В окне «Импорт данных» нажмите кнопку «ОК».

3.12 Сохраните файл.

4 Отчет о работе:

4.1 Составьте отчет о работе.

4.2 Преобразуйте отчет в формат PDF.

4.3 Запакуйте отчет (PDF) и все использованные и созданные в работе файлы в архив формата ZIP.

4.4 Прикрепите архив в раздел «Отчет по лабораторной работе №1 (сбор и подготовка данных)» курса «Анализ данных» СДО университета [2].

Содержание отчета

Отчет должен содержать:

1 Титульный лист: наименование работы, вариант задания, ФИО студента, номер учебной группы, дата выполнения работы.

2 Реферат.

3 Оглавление.

4 Часть 1 «Исследование наборов данных на портале data.gov.ru»:

4.1 Задание.

4.2 Копия экрана с набором данных, открытом в *Microsoft Excel*.

4.3 Описание набора данных согласно нижеприведенной форме (табл. 8).

Таблица 8. Форма описания набора данных

Показатель	Значение
Наименование	
Ссылка	
Формат	
Количество записей	
Количество полей	
в т.ч. числовых	
в т.ч. текстовых	

5 Часть 2 «Исследование наборов данных на портале data.worldbank.org»:

5.1 Задание.

5.2 Копия экрана с набором данных, открытом в *Microsoft Excel*.

5.3 График изменения показателя со временем по трем произвольно выбранным странам мира.

6 Часть 3 «Построение автоматизированной системы сбора данных»:

6.1 Задание.

6.2 Копия экрана с интернет-сайтом, содержащим данные.

6.3 Копия экрана *Microsoft Excel* после импорта данных.

7 Список использованных источников:

7.1 Источники данных.

7.2 Нормативные документы.

Все представленные в отчете таблицы и рисунки должны иметь пояснения. Отчет должен быть оформлен в соответствии с действующими стандартами университета [18, 19].

РЕГРЕССИОННЫЙ АНАЛИЗ

Общие сведения

Предположим, что есть задача определения стоимости некоторой квартиры. Очевидно, что в общем случае стоимость квартир зависит от многих факторов: площади, географического расположения, этажа и т.п. Зная характер этой зависимости, можно оценить (предсказать) стоимость любой квартиры.

Подобные системы появились на сайтах агентств недвижимости (рис. 7).

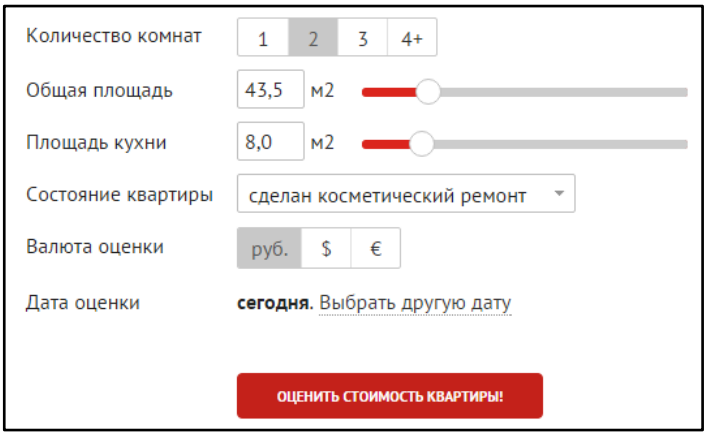


Рис. 7. Прогнозирование цены на сайте http://www.irn.ru/price

Предсказание значения зависимой переменной с помощью независимой переменной (независимых переменных) является задачей регрессионного анализа.

Регрессия относится к типу задач обучения с учителем (Supervised Learning в терминах Machine Learning). Предполагается, что имеется некоторая выборка данных, в которой представлены несколько объектов с известными свойствами.

Решение задачи предсказания включает два этапа: поиск характера зависимости и собственно предсказание (рис. 8).

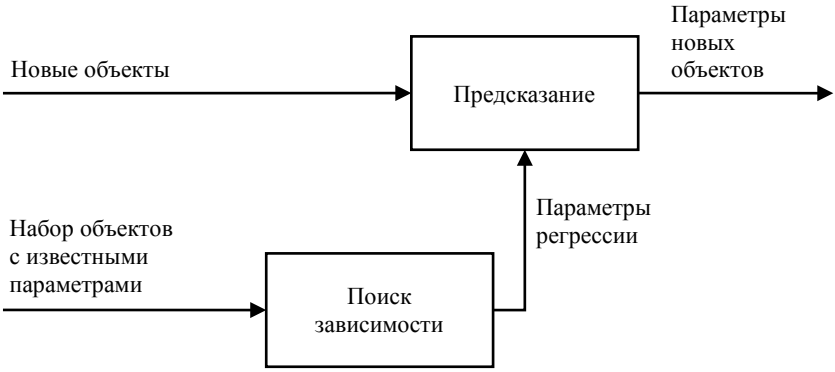


Рис. 8. Схема применения регрессии

Наиболее часто используется линейная функция гипотезы

$$h(x) = \theta_0 \cdot x_0 + \theta_1 \cdot x_1 + \ldots + \theta_m \cdot x_m = \sum_{j=0}^{m} \theta_j \cdot x_j \,. \qquad (1)$$

С учетом того, что наборы значений θ и x по сути являются векторами, выражение (1) для удобства записывают в виде произведения векторов:

$$h(x) = x \cdot \theta \,. \qquad (2)$$

В зависимости от характера функции гипотезы регрессию подразделяют на линейную и нелинейную. В зависимости от числа независимых переменных регрессию подразделяют на парную и множественную.

Примером парной линейной регрессии является задача выявления зависимости стоимости квартир от их площади (табл. 9, рис. 9).

Таблица 9. Характеристики квартир

Площадь, кв. м	Стоимость, млн. руб.
34	1,3
40	2,9
59	3,0
85	6,5

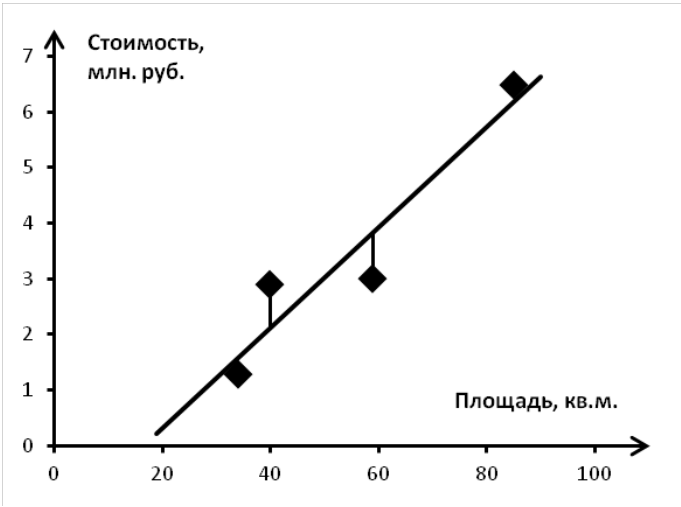


Рис. 9. Регрессия с помощью линейной функции

Подбор параметров регрессионной функции обычно осуществляется по критерию минимума суммы квадратов отклонений:

$$CF = \sum_{i=1}^{n} \left[h(x_i) - y_i \right]^2 \to \min . \qquad (3)$$

При этом выражение $\left[h(x_i) - y_i \right]^2$ называется функцией штрафа (cost function, CF; либо loss function, LF).

В формулировке (3) задача нахождения параметров регрессионной функции является оптимизационной. Существует два основных подхода к решению задачи регрессии в постановке (1): аналитический и численный. Следует отметить, что решения регрессионной задачи, полученные разными методами, могут различаться.

Аналитическое решение

Известно аналитическое решение задачи линейной регрессии в постановке (1):

$$\theta = (X^T X)^{-1} X^T y , \qquad (4)$$

где X – матрица, содержащая значения независимых переменных,

y – вектор, содержащий значений зависимых переменных.

Для вышеприведенного набора данных (табл. 9) матрица X и вектор y примут вид

$$X = \begin{bmatrix} 1 & 34 \\ 1 & 40 \\ 1 & 59 \\ 1 & 85 \end{bmatrix}, \; y = \begin{bmatrix} 1{,}3 \\ 2{,}9 \\ 3{,}0 \\ 6{,}5 \end{bmatrix}. \qquad (5)$$

При исходных данных (3) выражение (2) дает результат

$$\theta \approx \begin{bmatrix} -1{,}506 \\ 0{,}090 \end{bmatrix}.$$

Для вычисления выражений вида (2) удобно использовать специализированное математическое программное обеспечение, например *Matlab*, *Octave*. Однако широко распространенное ПО *Microsoft Excel* также имеет инструменты для решения подобных задач. Так, для умножения матриц используется функция МУМНОЖ, для транспонирования матриц – функция ТРАНСП, а для нахождения обратной матрицы – МОБР (рис. 10, рис. 11).

X		y	X^T			
1	34	1,3	1	1	1	1
1	40	2,9	34	40	59	85
1	59	3,0				
1	85	6,5				

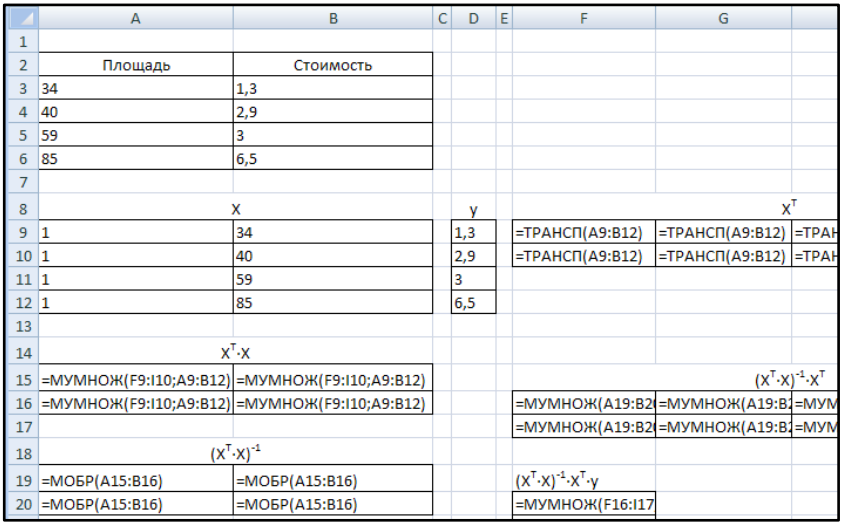
$X^T \cdot X$	
4	218
218	13462

$(X^T \cdot X)^{-1} \cdot X^T$			
0,957	0,750	0,095	-0,801
-0,013	-0,009	0,003	0,019

$(X^T \cdot X)^{-1}$	
2,128716	-0,0344719
-0,03447	0,0006325

$(X^T \cdot X)^{-1} \cdot X^T \cdot y$
-1,506
0,090

Рис. 10. Вычисления в Microsoft Excel (режим значений)



	A	B	C	D	E	F	G	
1								
2	Площадь	Стоимость						
3	34	1,3						
4	40	2,9						
5	59	3						
6	85	6,5						
7								
8	X			y			X^T	
9	1	34		1,3		=ТРАНСП(A9:B12)	=ТРАНСП(A9:B12)	=ТРАН
10	1	40		2,9		=ТРАНСП(A9:B12)	=ТРАНСП(A9:B12)	=ТРАН
11	1	59		3				
12	1	85		6,5				
13								
14	$X^T \cdot X$							
15	=МУМНОЖ(F9:I10;A9:B12)	=МУМНОЖ(F9:I10;A9:B12)					$(X^T \cdot X)^{-1} \cdot X^T$	
16	=МУМНОЖ(F9:I10;A9:B12)	=МУМНОЖ(F9:I10;A9:B12)				=МУМНОЖ(A19:B2	=МУМНОЖ(A19:B1	=МУМ
17						=МУМНОЖ(A19:B2	=МУМНОЖ(A19:B1	=МУМ
18	$(X^T \cdot X)^{-1}$							
19	=МОБР(A15:B16)	=МОБР(A15:B16)				$(X^T \cdot X)^{-1} \cdot X^T \cdot y$		
20	=МОБР(A15:B16)	=МОБР(A15:B16)				=МУМНОЖ(F16:I17		

Рис. 11. Вычисления в Microsoft Excel (режим формул)

Данный метод характеризуется следующими особенностями:

1 Относительно низкая устойчивость к отдельным сочетаниям данных. Так, дублирование какой-либо строки в наборе данных приведет к сбою в вычислениях при операции нахождения обратной матрицы.

2 Большая вычислительная сложность. Относительно большие наборы данных, содержащие порядка тысячи и более строк, будут обрабатываться относительно медленно.

3 Чувствительность к большим значениям. Для наборов данных, в отдельных столбцах которых содержатся большие значения, может потребоваться предварительная нормализация.

Численное решение

Для линейной регрессии задача в формулировке (1) имеет единственное решение, что позволяет без каких-либо оговорок применять численные методы. Например, можно использовать метод Ньютона либо метод сопряженных градиентов. Оба этих метода представлены в инструменте «Поиск решения» ПО *Microsoft Excel*.

Численное решение регрессионной задачи включает следующие шаги:

1) подготовку данных;

2) задание функции гипотезы, в том числе начальных значений её параметров;

3) задание целевой функции;

4) решение оптимизационной задачи каким-либо численным методом.

Рассмотрим численное решение задачи регрессии на основе данных о стоимости квартир (табл. 9) с помощью программного обеспечения *Microsoft Excel*.

Для удобства запишем выражение для функции гипотезы в следующей форме:

$$h(x) = a_0 + a_1 \cdot x\,. \tag{6}$$

Также запишем формулировку оптимизационной задачи:

$$CF = \sum_{i=1}^{4} \left[(a_0 + a_1 \cdot x_1) - y_i\right]^2 \to \min\,. \tag{7}$$

Зададим функцию гипотезы и начальные значения коэффициентов функции гипотезы, зададим функцию штрафа (рис. 12).

	A	B	C	D	E
1					
2	x	y	h(x)	h(x)-y	$[h(x)-y]^2$
3	34	1,3	4	2,7	7,3
4	40	2,9	4	1,1	1,2
5	59	3	4	1,0	1,0
6	85	6,5	4	-2,5	6,3
7				∑=	15,8
8					
9					
10	a0=	4			
11	a1=	0			

	A	B	C	D	E
1					
2	x	y	h(x)	h(x)-y	$[h(x)-y]^2$
3	34	1,3	=$B$10+$B$11*A3	=C3-B3	=D3^2
4	40	2,9	=$B$10+$B$11*A4	=C4-B4	=D4^2
5	59	3	=$B$10+$B$11*A5	=C5-B5	=D5^2
6	85	6,5	=$B$10+$B$11*A6	=C6-B6	=D6^2
7				∑=	=СУММ(E3:E6)
8					
9					
10	a0=	4			
11	a1=	0			

Рис. 12. Подготовка к численному решению

В настройках инструмента «Поиск решения» зададим целевую ячейку, содержащую выражение для функции штрафа, и изменяемые ячейки, содержащие значения коэффициентов функции гипотезы (рис. 13).

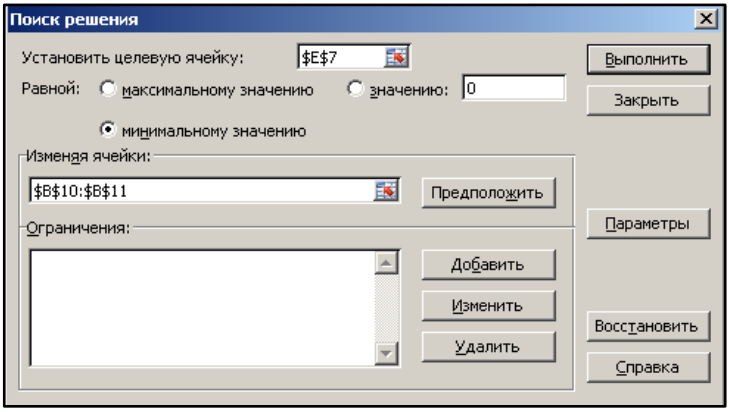


Рис. 13. Параметры поиска решения

В результате решения задачи (7) с помощью инструмента «Поиск решения» получим значения коэффициентов функции гипотезы $a_0 \approx -1{,}5062$, $a_1 \approx 0{,}0905$.

График функции гипотезы представляет собой прямую линию (рис. 14).

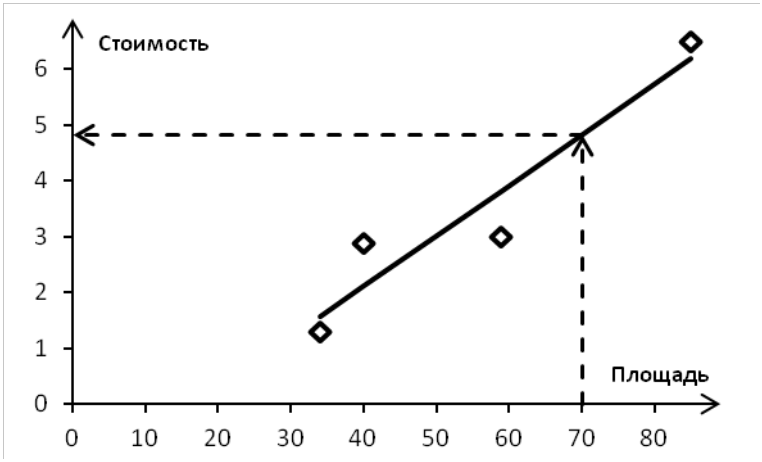


Рис. 14. Прогноз по графику функции гипотезы

Прогнозирование стоимости квартиры осуществляется с помощью подстановки площади квартиры и найденных коэффициентов

в выражение (6). Например, для квартиры площадью 70 кв. м прогнозная стоимость составит $-1{,}5062 + 0{,}0905 \cdot 70 \approx 4{,}83$ млн. руб. (рис. 14).

Выбор функции гипотезы

Одной из важных задач регрессионного анализа является задача выбора функции гипотезы. В случае парной регрессии выбор функции гипотезы можно осуществлять визуально по соответствующему графику. В случае множественной регрессии этот подход неприменим.

Предположим, что имеются данные о стоимости квартир (табл. 10).

Таблица 10. Стоимость квартир

Площадь, кв. м	Цена, млн. руб.
18	2,0
30	2,0
42	3,0
50	5,0
80	9,0

Рассмотрим два варианта решения задачи регрессии с применением линейной функции гипотезы и функции гипотезы, представляющей собой полином четвёртой степени. Опуская подробности решения этой задачи, приведем результаты (табл. 11, рис. 15).

Таблица 11. Параметры решений для различных функций гипотезы

Функция гипотезы	R^2	Функция штрафа
Линейная	0,935	2,271
Полином 4-й степени	1,000	0

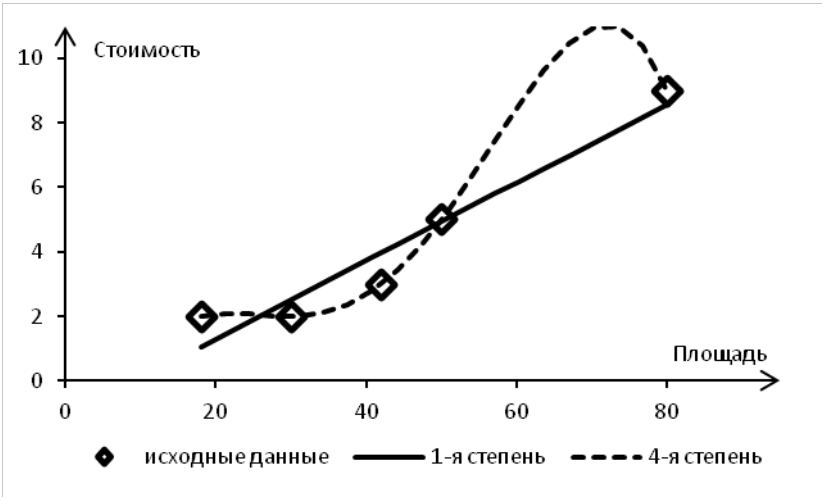


Рис. 15. Регрессия при разных функциях гипотезы

С точки зрения минимизации функции штрафа из представленных функций следует выбрать полином четвертой степени. С другой стороны, очевидно, что такая функция уже не вполне адекватно отражает тенденцию роста стоимости квартир с возрастанием их площади.

В терминологии Machine Learning ситуация, иллюстрируемая сплошной линией (рис. 15), соответствующей линейной функции гипотезы, обозначается термином underfitting (недообученность). В этом случае общая тенденция уже проявляется, но функция прогноза недостаточно хорошо аппроксимирует имеющийся набор данных.

Ситуация, иллюстрируемая пунктирной линией (рис. 15), соответствующей полиномиальной функции регрессии, обозначается термином «переобученность» (overfitting). Эта ситуация может быть описана следующим образом: аппроксимация очень хорошо либо идеально описывает выборку данных, но способность к обобщению потеряна.

Существуют разные способы выбора функции регрессии. Один из способов предполагает выполнение следующих шагов:

1 Разделение случайным образом исходной выборки данных на две части: обучающую, содержащую от 70 до 80% исходных данных, и проверочную, содержащую от 20 до 30% исходных данных.

2 Задание нескольких функций гипотезы.

3 Выполнение для каждой из функций гипотезы подбора параметров функции по обучающей выборке (минимизация функции штрафа по обучающей выборке) и вычисления функции штрафа по тестовой выборке.

4 Выбор функции гипотезы по критерию минимальной функции штрафа по тестовой выборке.

Рассмотрим пример выбора функции гипотезы на примере данных о площади и стоимости квартир (табл. 12, рис. 16).

Таблица 12. Стоимость квартир

Площадь ,кв. м	Стоимость, млн. руб.
30	2,8
100	7,0
46	4,9
69	6,5
84	6,7
77	7,2
54	5,9
84	7,4
66	6,0
93	6,7
33	1,9
65	6,9
44	3,5
54	5,3
61	6,0
67	6,1
89	7,8
62	5,6
41	3,6
92	8,4
70	7,4
45	5,1
35	3,7
68	6,7
65	5,5

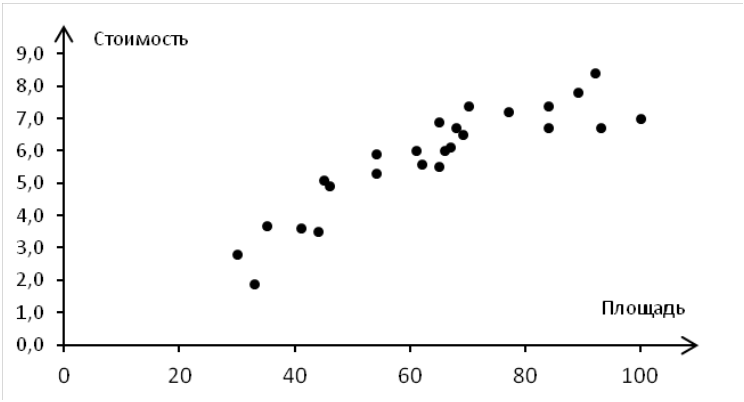


Рис. 16. Исходная выборка данных по стоимости квартир

Разделим исходную выборку данных на обучающую, содержащую 20 (80%) записей из исходной выборки (табл. 13), и проверочную, содержащую 5 (20%) записей из сходной выборки (табл. 14).

Таблица 13. Обучающая выборка

Площадь ,кв. м.	Стоимость, млн. руб.
30	2,8
100	7,0
46	4,9
69	6,5
84	6,7
77	7,2
54	5,9
84	7,4
66	6,0
93	6,7
33	1,9
65	6,9
44	3,5
54	5,3
61	6,0
67	6,1
89	7,8
62	5,6
41	3,6

Таблица 14. Проверочная выборка

Площадь, кв. м.	Стоимость, млн. руб.
70	7,4
45	5,1
35	3,7
68	6,7
65	5,5

Графическая интерпретация разделения исходной выборки на две приведена ниже (рис. 17).

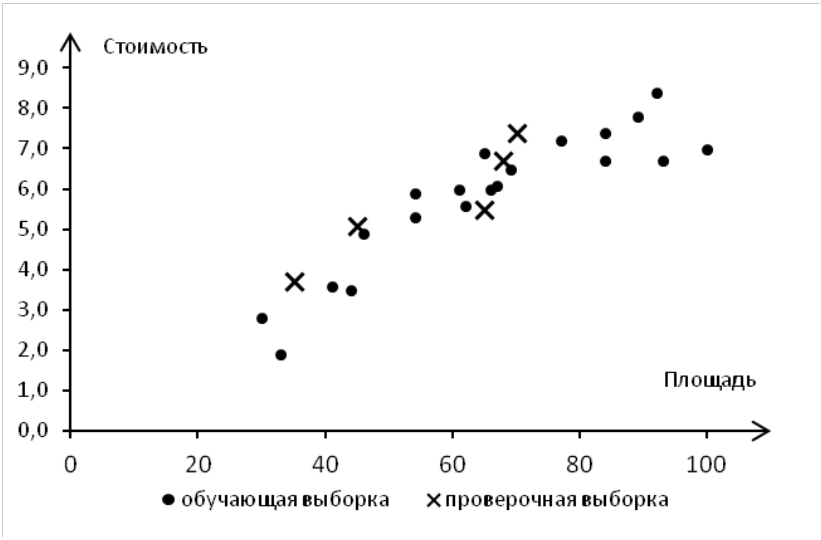


Рис. 17. Обучающая и проверочная выборки

Решение задачи регрессии приводит к следующим показателям (рис. 18).

Таким образом, исходя из критерия минимума функции штрафа по проверочной выборке, можно сделать вывод о том, что наиболее подходящей в данном случае является квадратичная функция гипотезы.

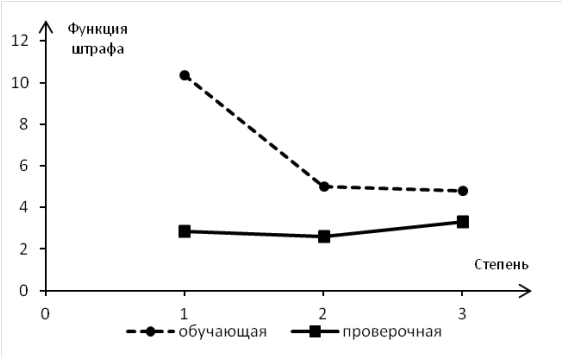


Рис. 18. Зависимость ошибки от степени функции регрессии

Вопросы для самоконтроля

1 Сформулируйте понятие регрессионного анализа.

2 Приведите и охарактеризуйте виды регрессии.

3 Приведите примеры практических задач, требующих применения регрессионного анализа.

4 Перечислите несколько факторов, от которых зависит стоимость: квартиры, автомобиля, авиабилета.

5 Перечислите способы решения задачи регрессии.

6 Дайте определение парной регрессии.

7 Дайте определение множественной регрессии.

8 Приведите порядок решения регрессионной задачи аналитическим методом.

9 Охарактеризуйте особенности решения регрессионной задачи аналитическим методом.

10 Приведите порядок решения регрессионной задачи численными методами.

11 Охарактеризуйте особенности решения регрессионной задачи численными методами.

12 Охарактеризуйте эффекты недообученности и переобученности.

13 Приведите алгоритм подбора функции регрессии.

Лабораторная работа «Регрессионный анализ»

Общие сведения

Целью работы является приобретение навыка регрессионного анализа.

В качестве инструментального средства используется программное обеспечение *Microsoft Excel*.

Исходные данные

Таблица 15. Варианты задания по регрессионному анализу

Вариант	Сфера	Данные	Зависимость
1	Пассажирские авиаперевозки	Дальность и время перелета между разными городами	Время от дальности
2	Пассажирские авиаперевозки	Дальность и стоимость перелета между разными городами экономическим классом	Стоимость от дальности
3	Пассажирские авиаперевозки	Дальность и стоимость перелета между разными городами бизнес-классом	Стоимость от дальности
4	Пассажирские железнодорожные перевозки	Дальность и время поездки между разными городами	Время от дальности
5	Пассажирские железнодорожные перевозки	Дальность и стоимость поездки между разными городами в купе	Стоимость от дальности
6	Пассажирские железнодорожные перевозки	Дальность и стоимость поездки между разными городами в плацкарте	Стоимость от дальности
7	Рынок недвижимости	Площадь и стоимость квартир на первичном рынке	Стоимость от площади
8	Рынок недвижимости	Площадь и стоимость квартир на вторичном рынке	Стоимость от площади
9	Рынок автотранспорта	Стоимость и пробег автомобилей какой-либо марки на вторичном рынке	Стоимость от пробега
10	Рынок автотранспорта	Стоимость и возраст автомобилей какой-либо марки на вторичном рынке	Стоимость от возраста
11	Рынок автотранспорта	Возраст и пробег автомобилей какой-либо марки на вторичном рынке	Пробег от возраста
12	Мировая экономика	Продолжительность жизни и доходы на душу населения стран мира	Продолжительность жизни от доходов

Таблица 16. Проверочные данные

Вариант	Данные
1	Время полета на 500, 1000 и 3000 км
2	Стоимость перелета на 500, 1000 и 3000 км
3	Стоимость перелета на 500, 1000 и 3000 км
4	Время поездки на 400, 800 и 2000 км
5	Стоимость поездки на 400, 800 и 2000 км
6	Стоимость поездки на 400, 800 и 2000 км
7	Стоимость для площади 30, 50, 100 кв.м.
8	Стоимость для площади 30, 50, 100 кв.м.
9	Стоимость для пробега 20 тыс., 50 тыс., 150 тыс. км.
10	Стоимость для возраста 2 года, 5 лет, 10 лет
11	Пробег для возраста 2 года, 5 лет, 10 лет
12	Продолжительность жизни для доходов 5, 20, 50 тыс. $

Порядок выполнения

1 Подготовка к работе:

1.1 Выберите вариант задания (табл. 15).

1.2 Найдите источник данных согласно заданию.

1.3 Запустите *Microsoft Excel*.

1.4 Создайте лист «Исходные данные» в документе *Excel*.

1.5 Подготовьте и разместите в листе «Исходные данные» выборку данных согласно выбранному варианту задания. Выборка должна содержать не менее 15 записей.

2 Построение линейной регрессии аналитическим методом:

2.1 Создайте лист «Аналитическое решение».

2.2 Скопируйте выборку данных с листа «Исходные данные» на лист «Аналитическое решение».

2.3 Выполните поиск параметров функции регрессии с помощью нормального уравнения.

2.4 Постройте на одном графике исходные данные и график функции регрессии.

2.5 Создайте прогноз. В качестве аргумента используйте проверочные данные (табл. 16).

3 Построение линейной регрессии численным методом:

3.1 Создайте лист «Численное решение».

3.2 Скопируйте выборку данных с листа «Исходные данные» на лист «Численное решение».

3.3 Выполните поиск параметров функции регрессии с помощью инструмента «Поиск решения» ПО *Microsoft Excel*.

3.4 Постройте на одном графике исходные данные и график функции регрессии.

3.5 Создайте прогноз. В качестве аргумента используйте проверочные данные (табл. 16).

4 Сравнительный анализ:

4.1 Сравните коэффициенты уравнения регрессии, полученные обоими методами.

4.2 Сравните прогнозы, полученные обоими методами.

5 Подбор функции регрессии.

5.1 Разделите исходную выборку на две части: обучающую и проверочную.

5.2 Постройте регрессию по обучающей части выборки для линейной, квадратичной и кубической функций.

5.3 Изобразите на одном графике исходные данные и графики трёх функций регрессии.

5.4 Изобразите на одном графике зависимость функции штрафа для обучающей выборки и функции штрафа для проверочной выборки от степени полинома функции гипотезы.

5.5 Выберите наилучшую функцию регрессии.

6 Отчет о работе:

6.1 Оформите отчет согласно требованиям, приведенным ниже.

6.2 Сохраните отчет в формате PDF.

6.3 Заархивируйте отчет и файлы Excel, использованные в работе.

6.4 Прикрепите архив в раздел «Отчет по лабораторной работе №2 «Регрессионный анализ» курса «Анализ данных» СДО университета [2].

Требования к отчету

Отчет должен содержать:

1 Титульный лист: наименование работы, вариант задания, ФИО студента, номер учебной группы, дата выполнения работы.

2 Реферат.

3 Оглавление.

4 Задание.

5 Описание выполненной работы:

5.1 Решение задачи регрессии аналитическим методом.

5.2 Решение задачи регрессии численными методами.
5.3 Подбор оптимальной функции регрессии.
6 Полученные результаты.
7 Анализ результатов.
8 Список использованных источников:
8.1 Источники данных.
8.2 Нормативные документы.
9 Приложения.

Отчет должен быть оформлен в соответствии с действующими стандартами университета [18, 19].

КЛАССИФИКАЦИЯ ДАННЫХ

Общие сведения

Классификация – это процесс определения принадлежности объектов к определенным классам.

Существует много практических задач классификации. В промышленности при оценке качества продукции возникает задача подразделения изделий на годные и бракованные. В банковском секторе при выдаче кредитов возникает задача подразделения заемщиков на кредитоспособных и некредитоспособных. В медицине при оценке состояния здоровья возникает задача постановки диагноза.

Как и регрессия, классификация относится к типу задач обучения с учителем (Supervised Learning в терминах Machine Learning). Предполагается, что имеется некоторая выборка данных, в которой представлены объекты нескольких классов. При этом выборка содержит как свойства объектов, так и признак принадлежности объекта к какому-либо классу.

Применение классификации производится в два этапа. На первом этапе выполняется обучение классификатора на некотором наборе данных, а на втором этапе – непосредственная классификация новых объектов (рис. 19).

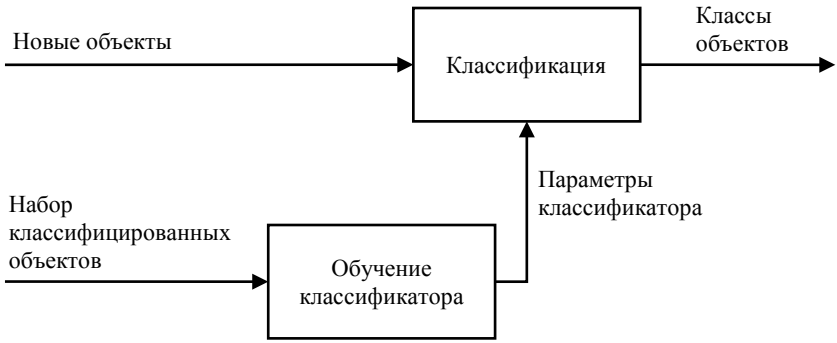


Рис. 19. Схема применения классификации

Различают бинарную и множественную классификацию. Бинарная классификация предполагает наличие двух классов, множественная – трех и более классов.

Классификация выполняется с помощью специальных методов (алгоритмов). Известно большое количество алгоритмов классификации. Так, в работе [20] проведены исследования 179 алгоритмов.

Бинарная классификация

Задачей бинарной классификации является определение принадлежности некоего объекта к одному из двух возможных классов. Например:

- является ли сообщение электронной почты «нормальным» или представляет собой спам;
- здоров или болен пациент;
- является ли заемщик банка надежным или ненадежным;
- качественная или бракованная деталь.

Наиболее известными методами бинарной классификации являются:

- логистическая регрессия (Logistic Regression);
- «наивный» байесовский классификатор (Naive Bayes Classifier);
- метод опорных векторов (Support Vector Machine, SVM);
- нейронная сеть (Neural Network).

Логистическая регрессия

Логистическая регрессия – один из методов бинарной классификации данных.

Алгоритм применения логистической регрессии:

1 Подготовка обучающей выборки – кодирование классов числами.

2 Задание функций штрафа.

3 Задание целевой функции.

4 Задание начальных значений коэффициентам функции.

5 Численное решение:

$$z = x \cdot \theta; \tag{8}$$

$$h(x_j) = \frac{1}{1+e^{-z}}; \tag{9}$$

$$CF(h_j, y_j) = -(1-y_j)\cdot\ln(1-h_j) - y_j\cdot\ln(h_j). \tag{10}$$

В ряде случаев использование численных методов может приводить к ошибкам вычислений, поэтому иногда удобнее использовать формулу (10) в другом варианте:

$$CF(h_j, y_j) = \begin{cases} -\ln(1-h_j), & y_j = 0 \\ -\ln(h_j), & y_j = 1 \end{cases}. \tag{11}$$

Оптимизационная задача по-прежнему формулируется как задача минимзации функции штрафа:

$$CF = \sum_j CF(h_j, y_j) \to \min. \tag{12}$$

Рассмотрим численное решение задачи логистической регрессии с помощью программного обеспечения Microsoft Excel:

1 В соответствии с предложенным выше алгоритмом представим исходные данные и расчетные формулы (рис. 20; 21).

	A	B	C	D	E	F	G
1	X0	X1	X2	y	z	h(x)	cost(h,y)
2	1	1	6	0	2,000	0,881	2,127
3	1	3	4	0	2,000	0,881	2,127
4	1	2	2	1	-1,000	0,269	1,313
5	1	3	3	1	1,000	0,731	0,313
6	1	4	3	1	2,000	0,881	0,127
7							6,007
8							
9		theta0	-5				
10		theta1	1				
11		theta2	1				

Рис. 20. Логистическая регрессия в Excel (режим значений)

	A	B	C	D	E	F	G
1	X0	X1	X2	y	z	h(x)	cost(h,y)
2	1	1	6	0	=МУМНОЖ(A2:C2;$C$9:$C$11)	=1/(1+EXP(-E2))	=-LN(1-F2)
3	1	3	4	0	=МУМНОЖ(A3:C3;$C$9:$C$11)	=1/(1+EXP(-E3))	=-LN(1-F3)
4	1	2	2	1	=МУМНОЖ(A4:C4;$C$9:$C$11)	=1/(1+EXP(-E4))	=-LN(F4)
5	1	3	3	1	=МУМНОЖ(A5:C5;$C$9:$C$11)	=1/(1+EXP(-E5))	=-LN(F5)
6	1	4	3	1	=МУМНОЖ(A6:C6;$C$9:$C$11)	=1/(1+EXP(-E6))	=-LN(F6)
7							=СУММ(G2:G6)

Рис. 21. Логистическая регрессия в Excel (режим формул)

2 Выполним численное решение с помощью инструмента «Поиск решения» (рис. 22).

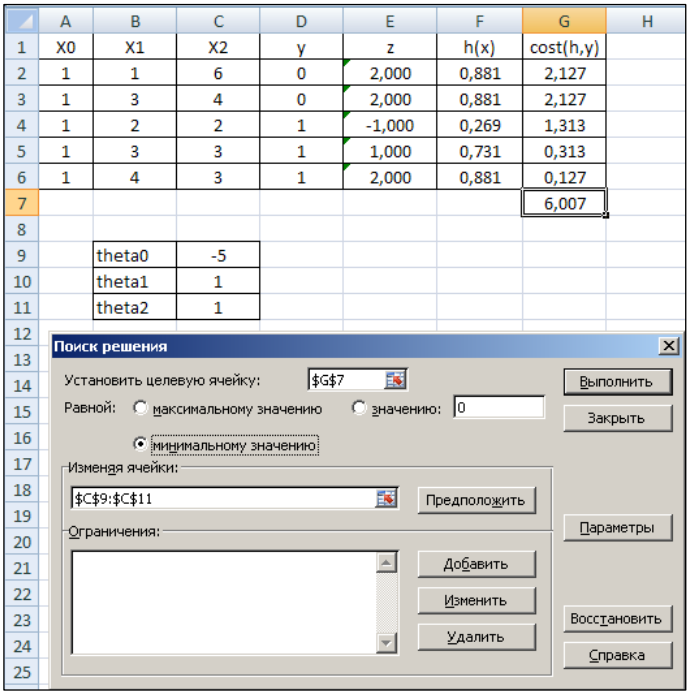


Рис. 22. Параметры поиска решения

В результате численного решения будут определены параметры функции линейного разделения. Визуальная проверка показывает корректность разделения двух классов (рис. 23).

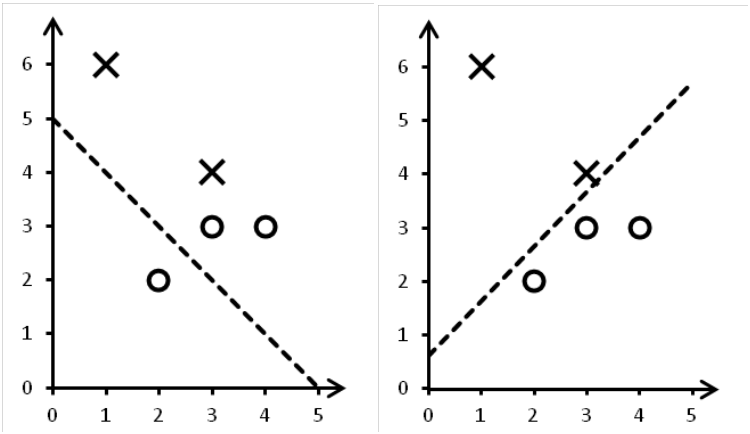


Рис. 23. Визуальное представление классов

Зачастую в реальных задачах бинарной классификации данные не могут быть разделены на два класса линейной функцией гипотезы (рис. 24).

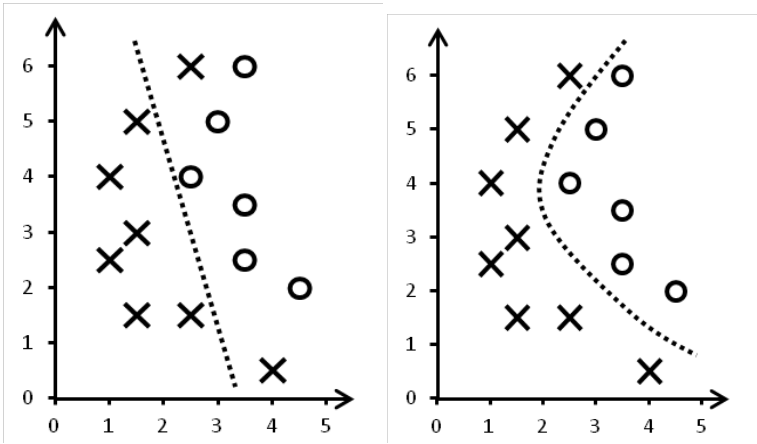


Рис. 24. Проблема линейной разделимости

Возможны следующие способы решения этой проблемы:

- применение нелинейной функции гипотезы;
- принципиальная замена логистической регрессии другим методом, например, нейросетевым классификатором.

Качество классификации

Очевидно, что при бинарной классификации возможны четыре сочетания реального класса каждого из объектов выборки данных и предположения алгоритма о классе объекта (рис. 25).

Правильно классифицированные алгоритмом объекты относятся либо к группе «true positives», либо к группе «true negatives». Неправильно классифицированные алгоритмом объекты относятся либо к группе «false positives», либо к группе «false negatives».

		Реальность	
		+	-
Предположение алгоритма	+	True positives (TP)	False positives (FP) Ошибка I рода
	-	False negatives (FN) Ошибка II рода	true negatives (TN)

Рис. 25. Сочетания при бинарной классификации

Реальные алгоритмы допускают ошибки классификации двух видов: ошибки I рода и ошибки II рода. Ошибки классификации объектов могут привести к последующим неправильным решениям и нежелательным последствиям (рис. 26).

		Реальность	
		Нормальное письмо	Письмо с вирусом
Предположение алгоритма	Нормальное письмо	Письмо пропущено в почтовый ящик	Письмо пропущено в почтовый ящик. Последствие: заражение компьютера вирусом
	Письмо с вирусом	Письмо отброшено. Последствие: пользователь не получит важную информацию	Письмо отброшено

Рис. 26. Последствия ошибок классификации

Существует несколько методов оценки качества классификации. Одним из методов является оценка с помощью F-критерия, выполняемая в четыре этапа:

1 Подсчет количества каждого сочетания случаев.

2 Расчет точности (precision)

$$P = \frac{TP}{TP + FP}. \tag{13}$$

3 Расчет чувствительности (recall)

$$R = \frac{TP}{TP + FN}. \tag{14}$$

4 Расчет F-критерия

$$F = \frac{2 \cdot P \cdot R}{P + R}. \tag{15}$$

Предположим, что в электронный почтовый ящик пришло 10 сообщений, часть из которых является нормальными, а часть – спамом (табл. 17).

Таблица 17. Сообщения электронной почты

№	Вид сообщения	«Мнение» антивируса
1	письмо	письмо
2	спам	письмо
3	письмо	спам
4	спам	письмо
5	письмо	спам
6	письмо	письмо
7	спам	спам
8	письмо	письмо
9	письмо	спам
10	письмо	письмо

Рассчитаем количество всех четырех сочетаний (табл. 18).

Таблица 18. Сочетания классификации

		Реальность	
		письмо	Спам
«Мнение» антивируса	письмо	4	2
	спам	3	1

В соответствии с формулами (13) - (15)

$$P = \frac{4}{4+2} \approx 0{,}667;\ R = \frac{4}{4+3} \approx 0{,}571;\ F = \frac{2 \cdot 0{,}667 \cdot 0{,}571}{0{,}667 + 0{,}571} \approx 0{,}615.$$

Для идеального алгоритма, не совершающего ошибок, $F = 0$.

Для проверки качества классификатора можно использовать репозиторий открытых наборов данных [21].

Множественная классификация

Задачей множественной классификации является определение принадлежности некоего объекта к одному из нескольких (трех или более) возможных классов, например постановка диагноза пациенту.

Наиболее известными методами множественной классификации являются:

- метод «один против всех» (One vs All);
- нейронная сеть (Neural Network).

Искусственная нейронная сеть (ИНС) – математическая модель нервной системы живого организма. Было обнаружено, что свойства ИНС позволяют использовать их для решения широкого круга прикладных задач, в том числе задач классификации.

Исторически первой была искусственная нейронная сеть под названием «перцептрон Розенблатта» (1957).

В общем случае ИНС имеет несколько входов и выходов. На входы подаются некоторые значения (сигналы). Результатом работы нейронной сети являются значения (сигналы) на её выходе (рис. 27).

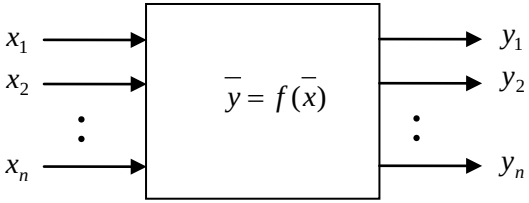


Рис. 27. Модель нейронной сети

Таким образом, ИНС можно рассматривать как векторную функцию векторного аргумента:

$$\bar{y} = h(\bar{x}). \qquad (16)$$

Нейронная сеть состоит из элементов – нейронов, связанных друг с другом (рис. 28).

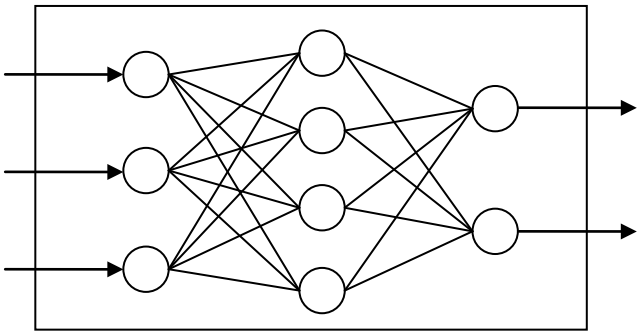


Рис. 28. Структура нейронной сети

Как правило, нейроны объединяются в группы, называемые слоями. Различают три вида слоёв: входной, выходной и скрытый. Так, выше изображена нейронная сеть, содержащая 3 нейрона во входном слое, 4 нейрона в скрытом слое и 2 нейрона во входном слое.

Нейрон является базовым составляющим элементом нейронной сети. В общем случае нейрон имеет несколько входов и один выход (рис.29).

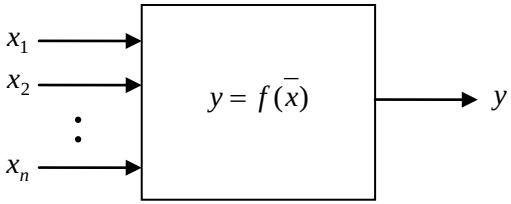


Рис. 29. Модель нейрона

Нейрон можно рассматривать как скалярную функцию векторного аргумента:

$$y = f(\bar{x}). \tag{17}$$

Предполагается, что каждому входу нейрона соответствует некоторый весовой коэффициент (рис. 30).

Значения на входе нейрона можно представить в виде вектора

$$\bar{x} = \{x_1, x_2, \ldots, x_n\}, \tag{18}$$

а весовые коэффициенты – в виде вектора

$$\bar{w} = \{w_1, w_2, \ldots, w_n\}. \tag{19}$$

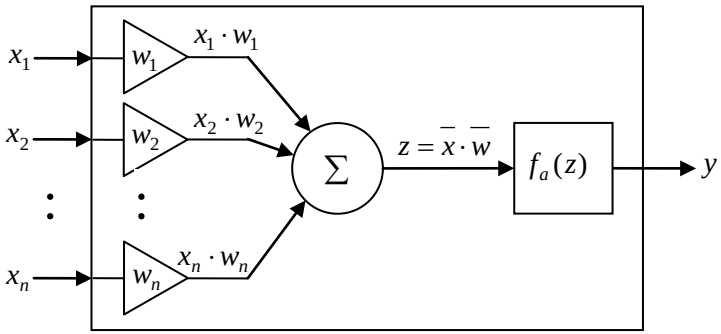


Рис. 30. Структура нейрона

Вычисление значения на выходе нейрона осуществляется в два этапа. На первом этапе рассчитывается взвешенная сумма

$$z = x_1 \cdot w_1 + x_2 \cdot w_2 + \ldots + x_n \cdot w_n = \sum_{i=1}^{n} x_i \cdot w_i = \overline{x} \cdot \overline{w}. \qquad (20)$$

На втором этапе рассчитывается значение функции активации $f_a(z)$. Наиболее часто применяется логистическая (сигмоидная) функция активации

$$f_a(z) = \frac{1}{1+e^{-z}}. \qquad (21)$$

Свойства функции нейронной сети определяются:

- структурой нейронной сети, то есть характером взаимосвязей между нейронами;
- свойствами нейронов: их весовыми коэффициентами и функциями активации.

Как и логистическая регрессия, нейронная сеть приобретает свои свойства в результате так называемого «обучения». Обучение ИНС – процесс подстройки весовых коэффициентов нейронов ИНС. Обучение производится на так называемой «обучающей выборке», представляющей собой набор «вопросов» и соответствующих «правильных ответов».

Качество обучения определяется степенью соответствия ответов сети («гипотез») «правильным ответам». Показателем качества обучения является значение функции штрафа, определяемой взвешенной суммой квадратов отклонений:

$$CF_j = \frac{1}{n}\sum_{i=1}^{n}(h(x_i^{(j)}) - y_i^{(j)})^2 ; \qquad (22)$$

$$CF = \frac{1}{m}\sum_{j=1}^{m} CF_j . \qquad (23)$$

В процессе обучения весовые коэффициенты нейронов ИНС изменяются согласно определенным правилам. Обучение производится шагами (эпохами). На одном шаге (в течение одной

эпохи) происходит одно обновление коэффициентов *W*. Обучение заканчивается в момент, когда значение функции штрафа достигает заданного пользователем порога. Также обучение может быть остановлено, если был превышен заданный лимит числа шагов.

Обучение сети производится с помощью специальных алгоритмов. В основе большинства алгоритмов лежат градиентные методы обучения. Исторически первым был так называемый «алгоритм обратного распространения ошибки» (error backpropagation). В дальнейшем были предложены еще несколько алгоритмов, наиболее известными из которых являются QPROP и RPROP.

В ходе обучения возможно проявление двух нежелательных эффектов: эффекта недообученности и эффекта переобученности.

Эффект недообученности

Эффект недообученности, как в регрессионном анализе, проявляется в виде недостаточного качества классификации объектов из обучающей выборки. Графически это иллюстрируется как приближение функции штрафа к некоему постоянному значению (рис. 31).

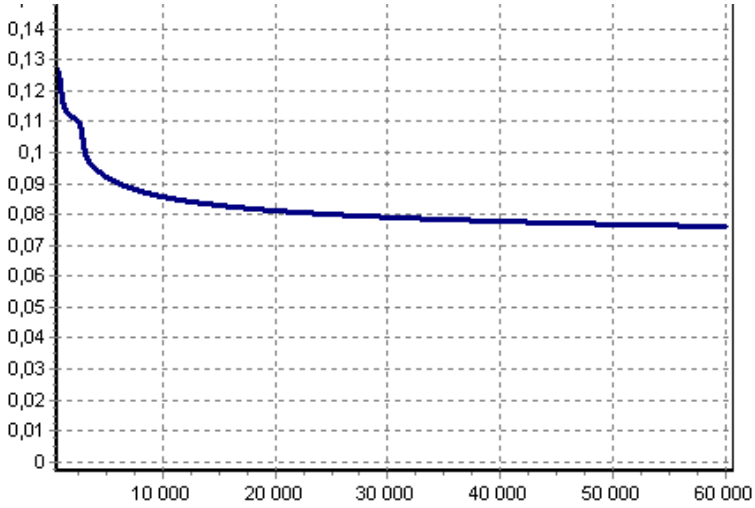


Рис. 31. Функция штрафа при недообученности

Для избежания эффекта недообученности можно использовать следующие способы:

1) увеличение числа нейронов в скрытом слое ИНС;

2) увеличение числа скрытых слоев.

Эффект переобученности

Можно выделить три признака переобучения:

1) относительно быстрое убывание функции штрафа в процессе обучения;

2) нулевое или близкое к нулю значение функции штрафа;

3) абсолютно точная при предъявлении объектов из обучающей выборки.

Одним из признаков переобученности является нулевое значение функции штрафа после обучения ИНС (рис. 32).

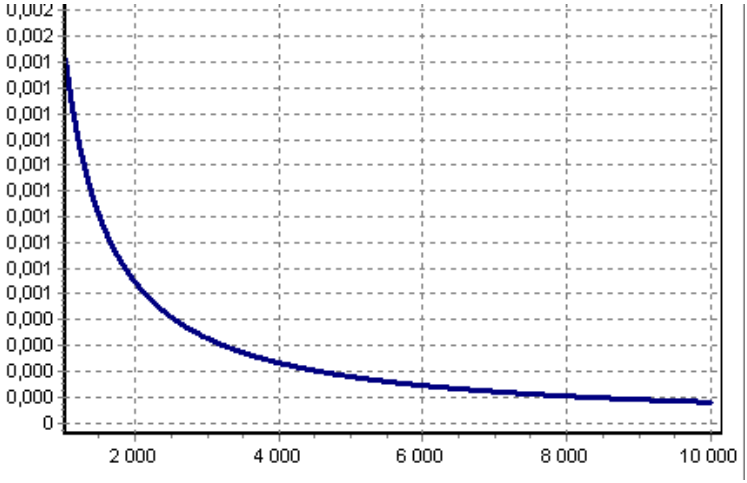


Рис. 32. Функция штрафа при переобучении

Переобучение приводит к потере классификатором способности к обобщению. Для избежания эффекта переобученности можно использовать следующие способы:

1) уменьшение числа нейронов в скрытом слое ИНС;

2) уменьшение числа скрытых слоев.

Программное обеспечение image_recognition

Данное программное обеспечение предназначено для классификации визуальных образов. В основе ПО лежит трехслойная нейронная сеть, размеры слоев которой задаются пользователем. Предполагается, что на вход сети подается монохромное изображение, при этом на выходе сети вычисляется вероятность принадлежности изображения к тому или иному классу (образу).

Интерфейс программы содержит следующие элементы (рис. 33):

- панель «Создание сети»;
- панель «Описание классов»;
- панель «Обучение сети»;
- панель «Проверка сети».

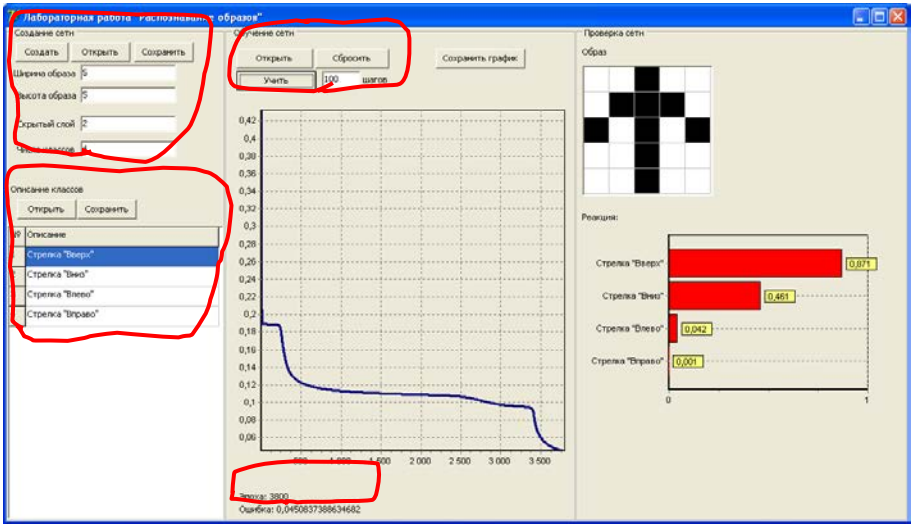


Рис. 33. Интерфейс программы *image_recognition*

Панель «Создание сети»

Кнопка «Создать» создает трехслойную НС заданной структуры. При этом размер входного слоя сети определяется произведением параметров «Ширина образа» и «Высота образа», размер скрытого слоя определяется значением параметра «Скрытый слой», размер выходного слоя определяется значением параметра «Число классов».

Кнопка «Открыть» позволяет загрузить в программу описание ранее созданной нейронной сети из файла с расширением *.net.

Кнопка «Сохранить» позволяет сохранить нейронную сеть в файл с расширением *.net.

Панель «Описание классов»

Кнопка «Открыть» позволяет загрузить в программу ранее созданное описание классов из файла с расширением *.txt.

Кнопка «Сохранить» позволяет сохранить описание классов в файл с расширением *.txt.

Таблица предназначена для текстового описания классов.

Панель «Обучение сети»

Кнопка «Открыть» предназначена для загрузки файла с обучающей выборкой.

Кнопка «Учить» предназначена для обучения сети в течение нескольких шагов, при этом число шагов определяется параметром «Число шагов».

Кнопка «Сбросить» задает произвольные значения весовым коэффициентам нейронов сети.

Кнопка «Сохранить график» сохраняет в файл график кривой обучения.

График кривой обучения показывает изменение ошибки в процессе обучения сети.

Параметр «Эпоха» показывает число сделанных шагов обучения.

Параметр «Ошибка» показывает текущую ошибку.

Панель «Проверка сети»

Изображение «Образ» предназначено для создания проверочного образа, подающегося на вход сети. Размеры образа определяются параметрами «Ширина образа» и «Высота образа», заданными ранее в ходе создания сети. Образ создается путем кликов по соответствующим элементам образа.

Диаграмма «Реакция» отображает выход сети.

Кнопка «Сохранить» сохраняет изображение образа и диаграмму с реакцией в соответствующие файлы.

Пример применения

Рассмотрим способ применения ПО *image_recognition* на примере задачи классификации визуальных образов стрелок (рис. 34).

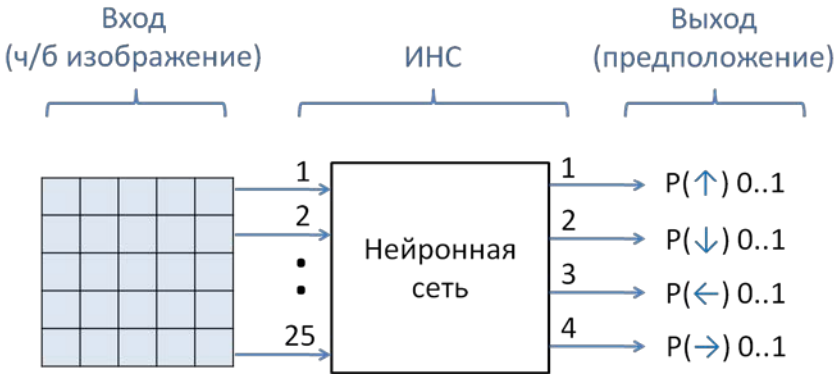


Рис. 34. Входные и выходные данные

Предположим, что входными образами являются изображения размером 5x5 элементов. В качестве обучающей выборки предлагается использовать образы четырех стрелок (рис. 35).

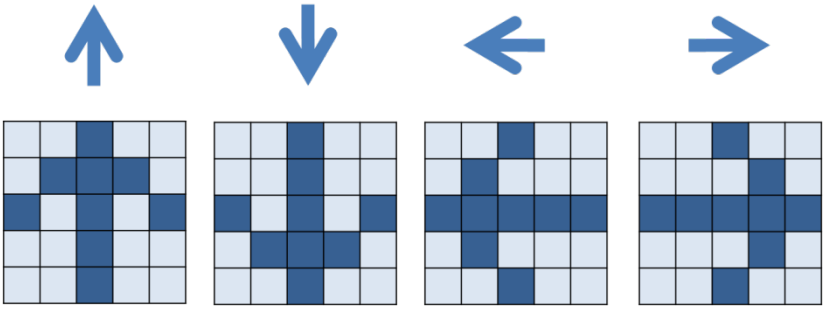


Рис. 35. Образы обучающей выборки

Для использования классификатора каждый известный образ должен быть закодирован численными значениями (рис. 36).

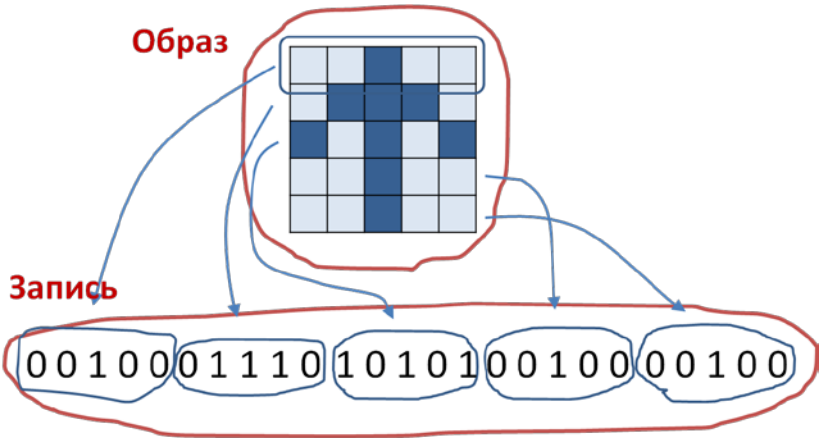


Рис. 36. Численное представление образа

Обучающая выборка для программы *image_recognition* представлена в текстовом файле в специальном формате (рис. 37).

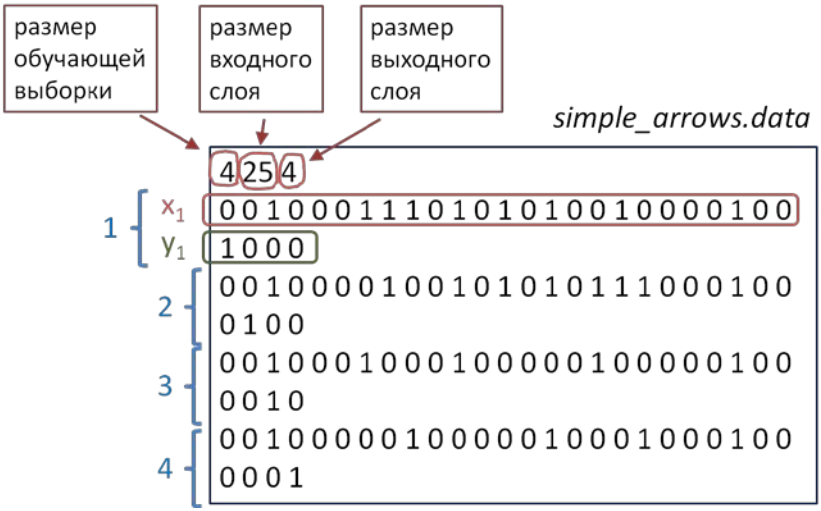


Рис. 37. Обучающая выборка

Вопросы для самоконтроля

1 Охарактеризуйте понятие классификации данных.

2 Виды классификации данных.

3 Перечислите методы классификации данных.

4 Приведите пример практического применения классификации.

5 Приведите алгоритм применения логистической регрессии.

6 Запишите функции штрафа при логистической регрессии.

7 Поясните суть проблемы линейного разделения классов.

8 Дайте определение ошибки классификации первого рода.

9 Дайте определение ошибки классификации второго рода.

10 Приведите примеры ошибок классификации и возможных последствий.

11 Приведите алгоритм оценки качества классификации по F1-критерию.

12 Запишите формулу расчета точности (precision).

13 Запишите формулу расчета чувствительности (recall).

14 Запишите формулу расчета F-критерия.

15 Охарактеризуйте понятие «искусственная нейронная сеть».

16 Дайте определение понятию «нейрон».

17 Охарактеризуйте эффекты обучения искусственной нейронной сети.

18 Изобразите кривую обучения, иллюстрирующую эффект недообученности ИНС.

19 Изобразите кривую обучения, иллюстрирующую эффект переобученности ИНС.

Лабораторная работа «Бинарная классификация»

Общие сведения

Целью работы является приобретение навыка бинарной классификации данных на основе логистической регрессии.

В качестве инструментального средства используется программное обеспечение *Microsoft Excel*.

Задание

Вариант 1. При проверке медицинской диагностической системы, основанной на бинарном классификаторе, получены следующие результаты (табл. 19).

Таблица 19. Экспериментальная проверка диагностической системы

№	Состояние пациента	Предположение классификатора
1	здоров	болен
2	болен	болен
3	здоров	здоров
4	болен	здоров
5	болен	болен
6	здоров	здоров
7	здоров	болен
8	болен	здоров
9	здоров	здоров
10	болен	болен

Вариант 2. При испытании антивируса, основанного на бинарном классификаторе, получены следующие результаты (табл. 20).

Таблица 20. Экспериментальная проверка антивируса

№	Наличие вируса	Предположение классификатора
1	есть	есть
2	нет	нет
3	нет	нет
4	есть	есть
5	нет	нет
6	есть	нет
7	есть	есть
8	нет	нет
9	нет	нет
10	нет	нет

Порядок выполнения

1 Подготовка:

1.1 Выберите вариант задание (см. с. 55).

1.2 Подготовьте выборку данных в ПО *Microsoft Excel*.

1.3 Постройте диаграмму, отображающую выборку данных.

2 Классификация:

2.1 Задайте целевую функцию.

2.2 Определите коэффициенты функции гипотезы с помощью инструмента «Поиск решения».

2.3 Рассчитайте значения точности, чувствительности, F-критерия.

3 Сделайте вывод об эффективности этого классификатора.

4 Отчет о работе:

4.1 Составьте отчет о работе.

4.2 Преобразуйте отчет в формат PDF.

4.3 Запакуйте отчет (PDF) и файл с данными (XLS) в один архив формата ZIP.

4.4 Прикрепите архив в раздел «Отчет по лабораторной работе №3 (бинарная классификация)» курса «Анализ данных» СДО университета [2].

Содержание отчета

Отчет должен содержать:

1 Титульный лист: наименование работы, вариант задания, ФИО студента, номер учебной группы, дата выполнения работы.

2 Реферат.

3 Оглавление.

4 Задание.

5 Описание выполненной работы.

6 Полученные результаты.

7 Анализ результатов.

8 Список использованных источников:

8.1 Источники данных.

8.2 Нормативные документы.

9 Приложения.

Отчет должен быть оформлен в соответствии с действующими стандартами университета [18, 19].

Лабораторная работа «Множественная классификация»

Общие сведения

Целью работы является приобретение навыка множественной классификации данных.

Задачи:

1 Подготовка обучающей выборки.

2 Обучение классификатора.

3 Проверка классификатора.

В качестве инструментального средства используется программное обеспечение *image_recognition*. Описание данного программного обеспечения приведено выше (с. 50).

Исходные данные

Таблица 21. Образы

Вариант	Образы
1	Арабские цифры
2	Римские цифры
3	Заглавные буквы кириллицы
4	Строчные буквы кириллицы
5	Заглавные буквы латиницы
6	Строчные буквы латиницы
7	Заглавные греческие буквы
8	Строчные греческие буквы
9	Математические символы
10	Смайлики
11	Дорожные знаки
12	Иконки социальных сетей
13	Логотипы автомобилей

Число распознаваемых образов – не менее пяти.

Порядок выполнения

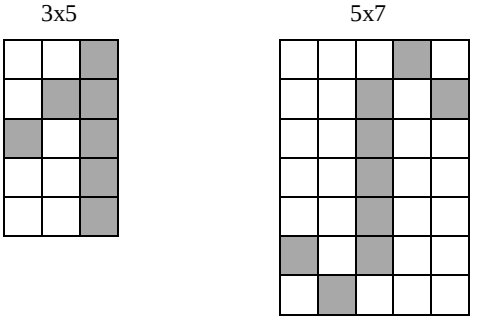
1 Подготовка.

1.1 Загрузите архивный файл *image_recognition.zip*, содержащий программное обеспечение, с сайта курса.

1.2 Распакуйте всё содержимое архивного файла в какую-либо папку.

1.3 Выберите вариант задания (табл. 21).

1.4 Выберите размер образа, то есть его высоту и ширину. Примеры образов приведены ниже (рис. 38).



а) цифра «1» б) символ интеграла

Рис. 38. Примеры образов

1.5 Выберите количество образов.

1.6 Создайте файл с обучающей выборкой.

2 Создание нейронной сети:

2.1 Запустите программу *image_recognition.exe*.

2.2 Задайте размеры изображения.

2.3 Задайте число классов.

2.4 Задайте число нейронов в скрытом слое, равное единице.

2.5 Обучите сеть.

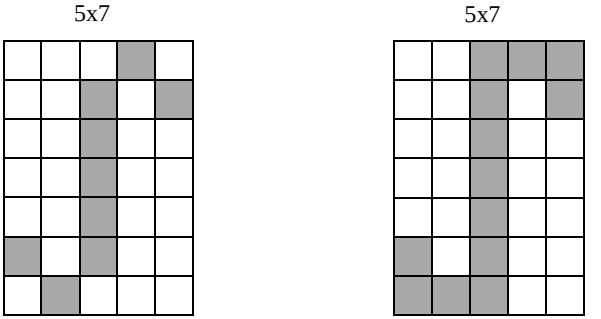
2.6 Проверьте распознавание всех известных образов.

2.7 Изменяя размер скрытого слоя, подберите минимальный размер скрытого слоя сети, при котором сеть уверенно распознает образы из обучающей выборки.

3 Проверка:

3.1 Проверьте распознавание всех известных образов.

3.2 Проверьте распознавание неизвестных образов. Например, если обучающая выборка содержит символ интеграла (рис. 39, а), то можно проверить неизвестный, но похожий образ (рис. 39, б).



а) известный образ б) неизвестный образ

Рис. 39. Примеры образов

3.3 Проведите анализ полученных результатов.

4 Продемонстрируйте преподавателю полученные результаты. При наличии замечаний проведите повторные эксперименты.

5 Отчет по работе:

5.1 Составьте отчет.

5.2 Преобразуйте отчет в формат PDF.

5.3 Создайте архив в формате ZIP, содержащий 1) отчет (PDF); 2) файл с обучающей выборкой (*.data); 3) файл со структурой сети (*.net); 4) файл с описанием классов (*.txt); 5) файл с описанием размера изображения (*.size) в один архив.

5.4 Прикрепите архив в раздел «Отчет по лабораторной работе №4» (множественная классификация) курса «Анализ данных» [2].

5.5 При наличии замечаний от преподавателя скорректируйте отчет.

Требования к отчету

Отчет должен содержать:

1 Титульный лист: наименование работы, вариант задания, ФИО студента, номер учебной группы, дата выполнения работы.

2 Реферат.

3 Оглавление.

4 Задание.

5 Описание образов:

5.1 Размер образа и число классов.

5.2 Изображения образов.

5.3 Обучающая выборка.

6 Описание нейронной сети:

6.1 Структура сети (число нейронов в слоях).

6.2 Число шагов обучения.

6.3 Достигнутое значение функции штрафа.

6.4 График функции штрафа (изменение значения функции штрафа в процессе обучения сети).

7 Результаты:

7.1 Реакция сети на все известные (т.е. имеющиеся в обучающей выборке) образы.

7.2 Реакция сети на неизвестные образы.

8 Анализ результатов.

9 Список использованных источников:

9.1 Источники данных.

9.2 Нормативные документы.

10 Приложения.

Отчет должен быть оформлен в соответствии с действующими стандартами университета [18, 19].

КЛАСТЕРНЫЙ АНАЛИЗ

Общие сведения

Кластерный анализ – (кластеризация) выявление групп (кластеров) объектов в выборке данных.

В отличие от регрессии и классификации, кластеризация относится к типу задач обучения без учителя (Unsupervised Learning в терминах Machine Learning). В отличие от классификации, в кластерном анализе не используется выборка ранее классифицированных объектов. Принятие решения о принадлежности объекта к той или иной группе принимается на основе свойств объектов (рис. 40).

Рис. 40. Схема кластерного анализа

Выборка данных в общем случае представляет собой таблицу (табл. 22).

Таблица 22. Шаблон набора данных

Наименование объекта	Свойство 1	Свойство 2	…	Свойство M
Объект 1				
Объект 2				
…				
Объект N				

Метод к-средних

Существует большое число методов кластерного анализа [22, 23], а наиболее известным является метод (алгоритм) к-средних.

Принцип: расчет средневзвешенного расстояния в нормированном эвклидовом пространстве свойств объектов (рис. 41).

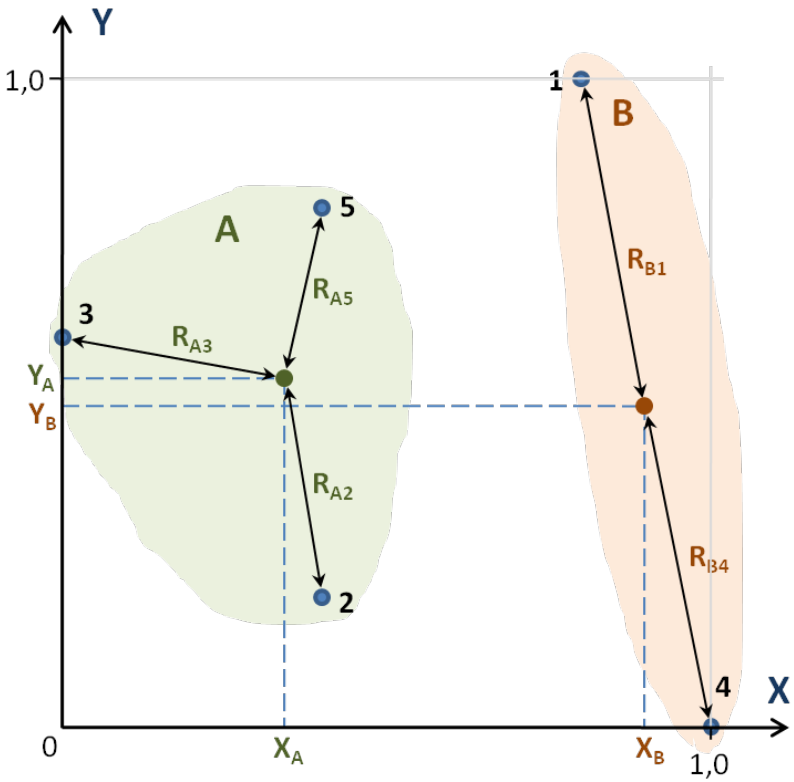


Рис. 41. Метод к-средних

Метод к-средних представляет собой следующую последовательность операций:

1 Пользователь задает количество кластеров.

2 Производится первоначальное случайное распределение объектов из выборки данных по кластерам.

3 Вычисляются коодинаты центров кластеров.

4 Вычисляются расстояния от каждого объекта до центров соответствущих кластеров.

5 Рассчитывается функция штрафа – сумма всех расстояний.

6 Каждый из объектов «прикрепляется» к тому кластеру, расстояние до центра которого наименьшее.

Шаги 3 – 6 повторяются до тех пор, пока не перестанут изменяться координаты центров кластеров.

Особенностью метода к-средних является разный результат выполнения алгоритма при повторном проведении кластерного анализа одной и той же выборки данных, поэтому рекомендуется многократный повтор кластеризации и выбор наилучшего результата.

Рассмотрим пример использования метода к-средних. Пусть имеется набор объектов, имеющих два свойства (табл. 23).

Таблица 23. Объекты и их свойства

Объект	Свойство 1	Свойство 2
1	10	7
2	12	5
3	35	2
4	45	4

Проведем нормализацию исходных данных, т.е. приведение их к диапазону 0..1 по каждому свойству (измерению) (табл. 24).

Таблица 24. Объекты и их нормализованные свойства

Объект	Свойство 1	Свойство 2
1	0,00	1,00
2	0,06	0,60
3	0,71	0,00
4	1,00	0,40

Последовательно применив алгоритм метода k-средних, получим следующие результаты (табл. 25).

Таблица 25. Варианты распределения объектов по кластерам

Шаг	Распределение по кластерам				R^2
	Объект 1	Объект 2	Объект 3	Объект 4	
1	1	1	1	2	0,822
2	1	1	2	1	0,817
3	1	2	1	1	1,037
4	2	1	1	1	0,654
5	1	2	2	1	1,076
6	1	1	2	2	0,202

Шаг 6 иллюстрирует следующая диаграмма (рис. 42). При этом найденное на этом шаге значение R^2 является минимальным, что свидетельствует о наилучшем распределении объектов по кластерам на этом шаге.

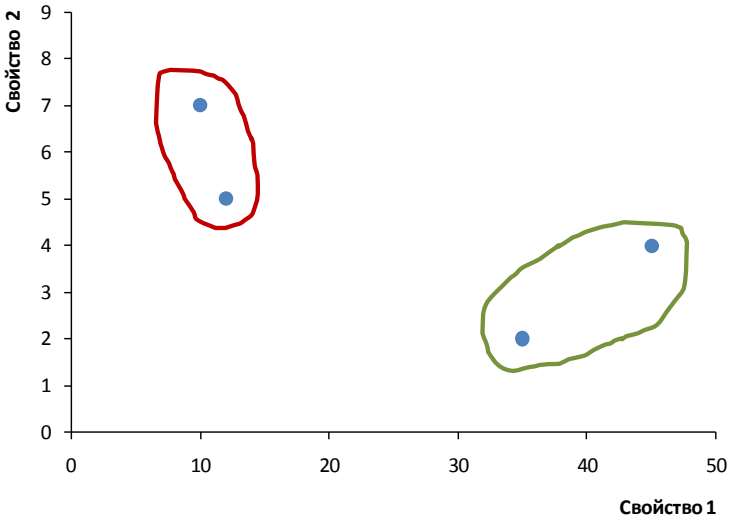


Рис. 42. Вариант распределения объектов по кластерам

Таким образом, найдено следующее оптимальное распределение объектов по кластерам (табл. 26).

Таблица 26. Распределение объектов по кластерам

Объект	Свойство 1	Свойство 2	Кластер
1	10	7	1
2	12	5	1
3	35	2	2
4	45	4	2

Метод к-средних не дает ответа на вопрос о количестве кластеров в выборке данных. Для определения количества кластеров можно воспользоваться так называемым методом локтя (Elbow Method). Метод локтя предполагает выполнение следующих шагов:

1 Выполняется кластеризация методом к-средних, при этом рассчитывается и записывается значение функции штрафа.

2 Строится график зависимости функции штрафа от заданного числа кластеров.

3 В качестве решения выбирается число кластеров, при котором происходит наибольший перегиб графика.

Программное обеспечение kmeans

Метод к-средних реализован в программном обеспечении *kmeans*. ПО *kmeans* предназначено для кластеризации набора объектов, имеющих два свойства и представлено в виде исполняемого файла для ОС Windows. Программное обеспечение имеет однооконный интерфейс (рис. 43).

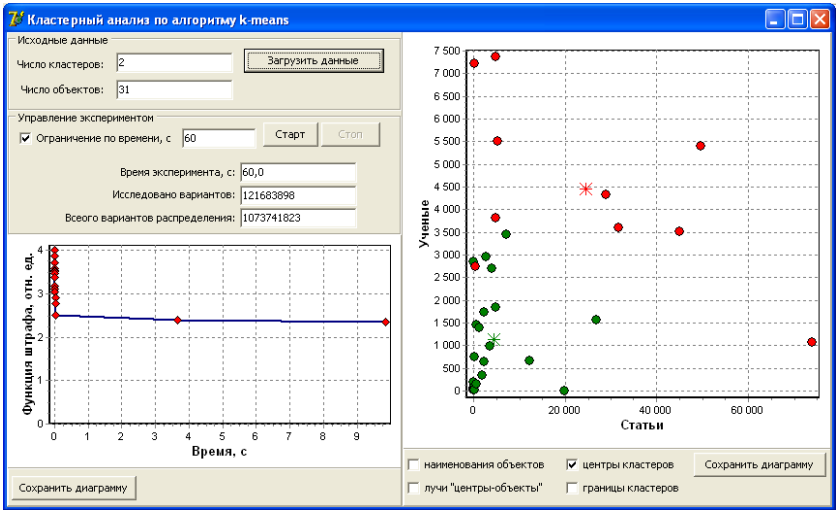


Рис. 43. Интерфейс программного обеспечения k-means

Основные функции управления реализованы в панелях «Исходные данные» и «Управление экспериментом» (рис. 44).

Панель «Исходные данные»:

- кнопка «Загрузить данные» предназначена для загрузки исходных данных для кластерного анализа;

- параметр «Число объектов» показывает количество объектов в загруженном файле данных;

- показатель «Число кластеров» предназначен для задания числа кластеров, по которым будет производиться разбивка объектов.

Панель «Управление экспериментом»:

- выключатель «Ограничение по времени»;

- кнопка «Старт» предназначена для запуска процедуры кластеризации. Кластеризация будет происходить, пока не перебраны все возможные варианты или не закончилось время, отпущенное на эксперимент;

- показатель «Время эксперимента» показывает длительность эксперимента;

показатель «Число вариантов» показывает сколько вариантов разбиения объектов по кластерам было исследовано на данный момент;

- показатель «Варианты распределения по кластерам» показывает число вариантов распределения объектов по кластерам.

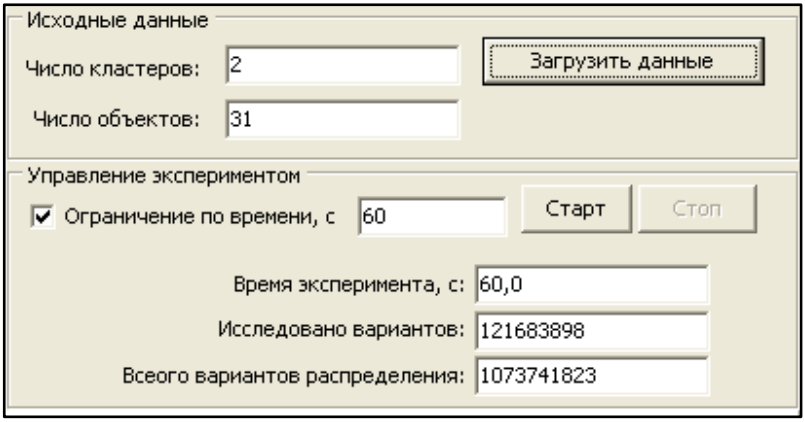


Рис. 44. Панели «Исходные данные» и «Управление экспериментом»

График функции штрафа (рис. 45) показывает изменение функции штрафа со временем в процессе кластеризации.

Диаграмма распределения по кластерам (рис. 46) показывает расположение объектов в декартовой системе координат и принадлежность объекта к тому или иному кластеру. Горизонтальной оси соответствует первый показатель из исходного

файла данных, вертикальной оси – второй. Отображение диаграммы регулируется следующими элементами управления:

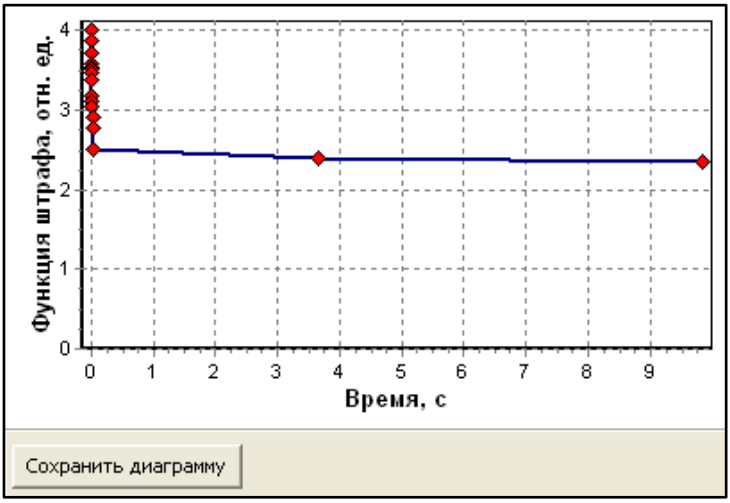


Рис. 45. График функции штрафа

- выключатель «Наименования объектов» позволяет подписать объекты на диаграмме;

- выключатель «Лучи “центры-объекты”» позволяет отобразить лучи от центра кластеров до всех объектов кластера;

- выключатель «Центры кластеров» позволяет отобразить условные центры кластеров;

- выключатель «Граница» позволяет отобразить условные границы кластеров. Граница представляет собой ломаную линию, объединяющую все объекты каждого кластера;

- кнопка «Сохранить диаграмму» записывает диаграмму в графический файл в формате PNG;

- кнопка «Сохранить кластеры» создает файл, содержащий номера кластеров, координаты их центров и количество объектов в кластерах.

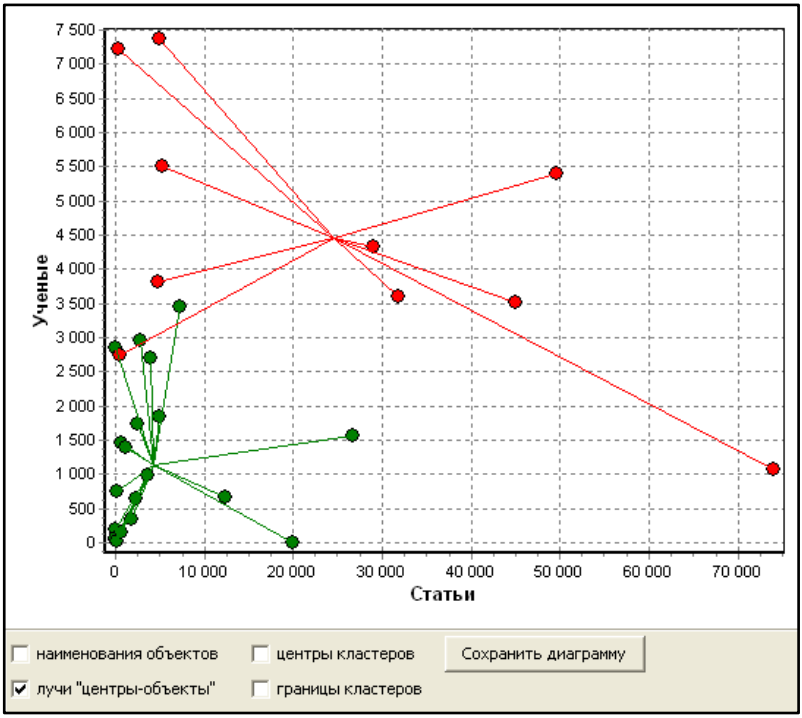


Рис. 46. Диаграмма распределения по кластерам

Анализ данных с использованием ПО *kmeans* проводится по следующей схеме (рис. 47).

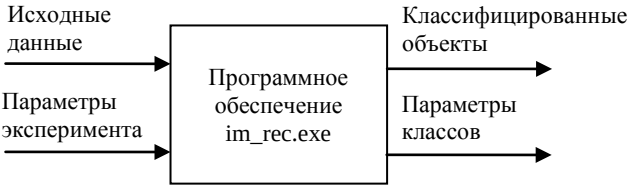


Рис. 47. Схема эксперимента

Исходные данные должны быть представлены в файле в формате CSV «Текст, разделенный». В файле должно быть три столбца:

1 Наименования объектов.

2 Значение первого свойства.

3 Значение второго свойства.

Первая строка файла должна содержать подписи свойств объектов. Подготовку исходных данных удобно производить в *Microsoft Excel* (рис. 48). После подготовки данных файл необходимо сохранить в формате CSV (разделители – запятые).

	A	B	C
1	Страна	Статьи	Ученые
2	ARG	3655,2	982,5
3	AUT	4832,2	3811,8
4	BEL	7217,6	3446,2
5	BIH	63,8	197,2
6	BRA	12306,3	658
7	BGR	735,4	1466,3
8	CAN	29016,9	4334,7

Рис. 48. Представление исходных данных в Microsoft Excel

Убедиться в корректности подготовленного файла можно, открыв его в блокноте Windows (рис. 49).

```
Страна;Статьи;Ученые
ARG;3655,2;982,5
AUT;4832,2;3811,8
BEL;7217,6;3446,2
BIH;63,8;197,2
BRA;12306,3;658,0
BGR;735,4;1466,3
CAN;29016,9;4334,7
CHL;1867,8;333,7
CHN;74019,2;1077,1
COL;608,4;156,1
```

Рис. 49. Представление исходных данных в блокноте Windows

Вопросы для самоконтроля

1 Дайте определение понятия «кластер».
2 Дайте определение понятия «кластеризация».
3 Охарактеризуйте два любых алгоритма кластеризации.
4 Назовите входные данные алгоритма к-средних.
5 Назовите выходные данные алгоритма к-средних.
6 Приведите последовательность шагов в алгоритме к-средних.
7 Приведите порядок кластерного анализа с помощью ПО *kmeans*.

Лабораторная работа «Кластерный анализ»

Общие сведения

Целью работы является приобретение навыка кластерного анализа на основе метода к-средних.

В качестве исходных данных используются статистические данные Всемирного банка. В качестве инструментального средства для проведения экспериментов используется программное обеспечение *kmeans*. Описание данного программного обеспечения приведено выше (см. с. 65).

Исходные данные

Таблица 27. Исходные данные для кластерного анализа

Вариант	Показатели	Год
1	1. Railways, goods transported (million ton-km). 2. Air transport, freight (million ton-km).	2008
2	1. Railways, goods transported (million ton-km) 2. Roads, goods transported (million ton-km)	2007
3	1. Air transport, freight (million ton-km) 2. Roads, goods transported (million ton-km)	2006
4	1. Railways, goods transported (million ton-km) 2. Railways, passengers carried (million passenger-km)	2009
5	1. Air transport, freight (million ton-km) 2. Air transport, passengers carried	2005
6	1. Roads, goods transported (million ton-km) 2. Roads, passengers carried (million passenger-km)	2002
7	1. Roads, total network (km) 2. Rail lines (total route-km)	2002
8	1. Internet users (per 100 people) 2. Mobile cellular subscriptions (per 100 people)	2011
9	1. Internet users (per 100 people) 2. Passenger cars (per 1,000 people)	2010
10	1. Mobile cellular subscriptions (per 100 people) 2. Passenger cars (per 1,000 people)	2009
11	1. GDP per capita (current US$) 2. Passenger cars (per 1,000 people)	2008
12	1. GDP per capita (current US$) 2. Internet users (per 100 people)	2007
13	1. GDP per capita (current US$) 2. Life expectancy at birth, total (years)	2011
14	1. GDP per capita (current US$) 2. Physicians (per 1,000 people)	2010

Окончание табл. 27

Вариант	Показатели	Год
15	1. Access to electricity (% of population) 2. Life expectancy at birth, total (years)	2009
16	1. Average precipitation in depth (mm per year) 2. Cereal yield (kg per hectare)	2011
17	1. GDP per capita (current US$) 2. Cereal yield (kg per hectare)	2011
18	1. GDP per capita (current US$) 2. GDP per unit of energy use (constant 2005 PPP $ per kg of oil equivalent)	2010
19	1. Researchers in R&D (per million people) 2. GDP per unit of energy use (constant 2005 PPP $ per kg of oil equivalent)	2008
20	1. Agricultural land (sq. km) 2. Land area (sq. km)	2011
21	1. Forest area (sq. km) 2. Land area (sq. km)	2010
22	1. Population, total 2. Land area (sq. km)	2009
23	1. Health expenditure, total (% of GDP) 2. Military expenditure (% of GDP)	2010
24	1. High-technology exports (% of manufactured exports) 2. GDP per capita (current US$)	2010
25	1. Hospital beds (per 1,000 people) 2. Life expectancy at birth, total (years)	2005

Порядок выполнения

1 Подготовка:

1.1 Выберите задание (табл. 27).

1.2 Загрузите программу *kmeans* из курса «Анализ данных» СДО университета [2].

1.3 Подготовьте исходные данные для кластеризации:

1.3.1 На сайте Всемирного банка [16] найдите данные по странам мира согласно заданию.

1.3.2 Загрузите соответствующие файлы на компьютер (Download data – Excel file).

1.3.3 Соберите данные из двух загруженных файлов в один файл в формате CSV. Файл должен содержать три столбца: название страны, показатель №1, показатель №2. Схема подготовки файла с исходными данными приведена ниже (рис. 50).

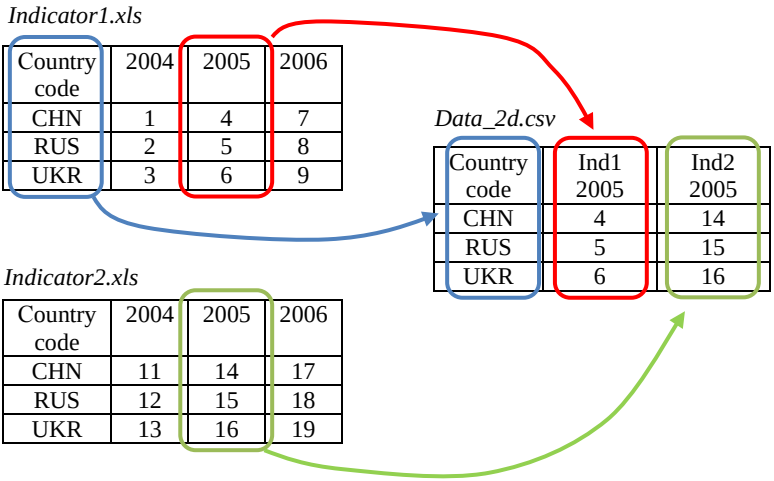


Рис. 50. Схема подготовки исходных данных

2 Эксперименты:

2.1 Запустите программу *kmeans*.

2.2 Установите количество кластеров, равное единице.

2.3 Выберите файл с исходными данными (кнопка «Загрузить данные»).

2.4 Нажмите кнопку «Старт».

2.5 Нажимая соответствующие кнопки, сохраните диаграмму с кривой обучения, диаграмму кластеров и файл с кластеризованными объектами.

2.6 Запишите номер эксперимента и значение функции штрафа в таблицу экспериментальных данных (табл. 28).

Таблица 28. Форма журнала экспериментальных данных

Номер эксперимента	Число кластеров	Функция штрафа

2.7 Повторите эксперимент (шаги 2.2-2.6) пять раз. В результате журнал экспериментальных данных будет содержать пять записей.

2.8 Последовательно увеличивая число кластеров до восьми, проведите серии экспериментов (шаги 2.2 – 2.7). В результате журнал экспериментальных будет содержать 40 записей (таблица 29).

3 Обработка экспериментальных данных.

3.1 Выберите эксперименты, в которых достигнуто минимальное значение функции штрафа для каждого числа кластеров, запишите эти данные в таблицу обработанных экспериментальных данных (табл. 29).

Таблица 29. Обработанные экспериментальные данные

Номер эксперимента	Число кластеров	Функция штрафа
	1	
	2	
	…	
	8	

3.2 На основе полученной таблицы обработанных экспериментальных данных постройте график зависимости минимального значения функции штрафа от числа кластеров. Пример такого графика приведен ниже (рис. 51).

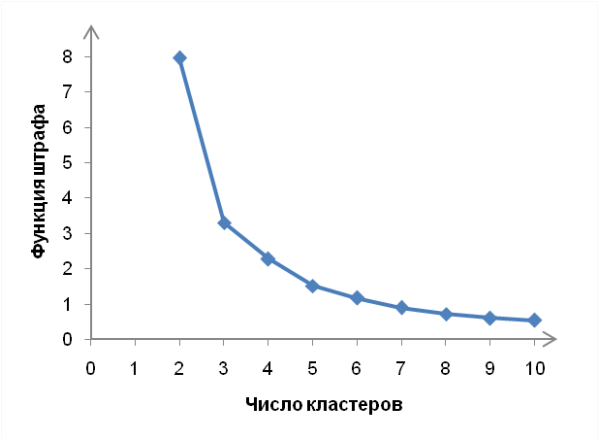


Рис. 51. Зависимость функции штрафа от числа кластеров

3.3 По построенному графику, пользуясь методом локтя, определите оптимальное число кластеров. Для приведенного выше графика характерный излом происходит при числе кластеров, равном трем, соответственно в данном случае оптимальное число кластеров равно трем.

3.4 Сделайте выводы по работе:

4 Отчет по работе.

4.1 Составьте отчет о работе.

4.2 Преобразуйте отчет в формат PDF.

4.3 Запакуйте отчет (PDF), два исходных файла с индикаторами (XLS) и объединенный файл данных (CSV) двумерного анализа в один архив формата ZIP.

4.4 Прикрепите созданный архив в раздел «Отчет по лабораторной работе №5 (кластерный анализ)» курса «Анализ данных» СДО университета [2].

Требования к отчету

Отчет должен содержать:

1 Титульный лист: наименование работы, вариант задания, ФИО студента, номер учебной группы, дата выполнения работы.

2 Реферат.

3 Оглавление.

4 Задание.

5 Журнал экспериментальных данных.

6 Обработанные экспериментальные данные.

7 Диаграмму функции штрафа.

8 Определение числа кластеров.

9 Диаграмма кластеров.

10 Выводы.

11 Список использованных источников:

11.1 Источники данных.

11.2 Нормативные документы.

Отчет должен быть оформлен в соответствии с действующими стандартами университета [18, 19].

БЫСТРОДЕЙСТВИЕ СИСТЕМ АНАЛИЗА ДАННЫХ

Общие сведения

При проектировании систем анализа данных важно представлять, какое время будет затрачено на решение вычислительных задач. В общем случае быстродействие систем анализа данных зависит от многих факторов:

- вычислительной сложности использованных алгоритмов;
- способа программной реализации алгоритмов;
- аппаратного обеспечения.

Вычислительная сложность

Вычислительная сложность – это зависимость трудоемкости вычислений от объема обрабатываемых данных. Например, вычислительная сложность алгоритма линейного поиска $O(n) = n$, а алгоритма бинарного поиска – $O(n) = \log n$ (рис. 52).

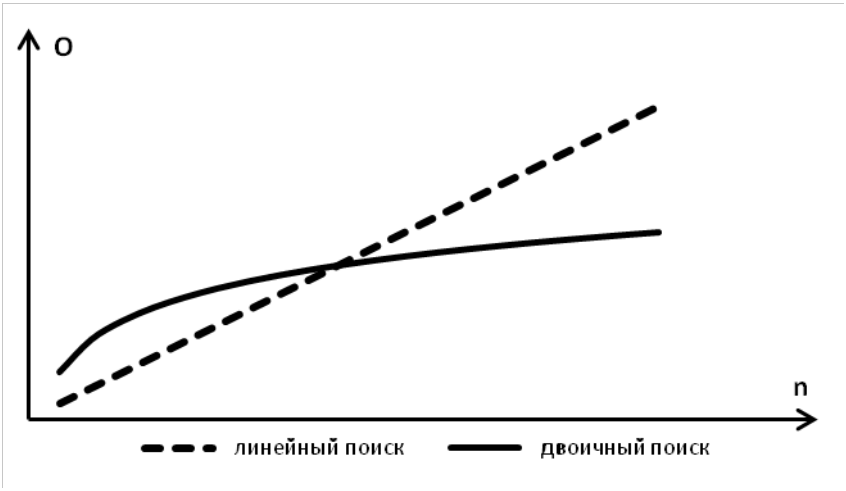


Рис. 52. Вычислительная сложность алгоритмов поиска

Очевидно, что один и тот же объем данных будет обработан быстрее алгоритмом с меньшей вычислительной сложностью.

Например, вычислительная сложность алгоритма вычисления произведения матриц «по определению» $O(n) = n^3$, а алгоритма

Штрассена – $O(n) = n^{2,81}$. Очевидно, что один и тот же объем данных будет обработан быстрее алгоритмом с меньшей вычислительной сложностью

Предположим, что имеются два алгоритма, имеющие соответственно вычислительную сложность $O_1(n)$ и $O_2(n)$. Для сравнения их вычислительной сложности необходимо найти значения предела

$$\lim_{n\to\infty} \frac{O_1(n)}{O_2(n)}. \qquad (24)$$

При этом если значение выражения (24):

- равно ∞, то вычислительная сложность функции $O_1(n)$ больше, чем вычислительная сложность функции $O_2(n)$;

- равно 0, то вычислительная сложность функции $O_2(n)$ больше, чем вычислительная сложность функции $O_1(n)$;

- равно некоторому числу, то функции $O_1(n)$ и $O_2(n)$ имеют одинаковую вычислительную сложность.

Пример

Пусть имеются два алгоритма с вычислительной сложностью соответственно $O_1(n) = e^n$, $O_2(n) = n^2$, тогда

$$\lim_{n\to\infty} \frac{O_1(n)}{O_2(n)} = \lim_{n\to\infty} \frac{e^n}{n^2} = \lim_{n\to\infty} \frac{\left(e^n\right)'}{\left(n^2\right)'} = \lim_{n\to\infty} \frac{e^n}{2n} = \infty.$$

Следовательно, вычислительная сложность первого алгоритма больше вычислительной сложности второго алгоритма.

Экспериментальное определение вычислительной сложности

Допустим, что реализован некоторый модуль системы анализа данных, предназначенный для обработки информации. Для использования этого модуля в реальных задачах необходимо четко представлять его вычислительную сложность.

Для определения вычислительной сложности необходимо провести эксперимент, в ходе которого нужно измерить время обработки этим модулем некоторого объема исходных данных.

В результате эксперимента будут получены точки вида «объем данных - время обработки» (рис. 53, а). Аппроксимировав экспериментальные данные какой-либо функцией, можно построить прогноз трудоемкости вычислений для любого, в том числе большего объема данных (рис. 53, б).

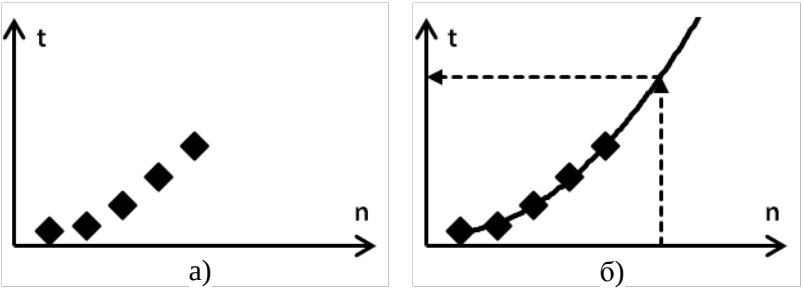


Рис. 53. Прогнозирование вычислительной сложности

Программное обеспечение square_matrix_multiply

Программное обеспечение *square_matrix_multiply* предназначено для исследования факторов быстродействия систем анализа данных на примере операции умножения квадратных матриц. Пользователь может варьировать размеры матриц, количество вычислительных потоков (рис. 54). Дополнительно может быть задан размер серии экспериментов. Программное обеспечение выполняет ряд экспериментов, в ходе которых проводится измерение времени, затраченного на умножение матриц.

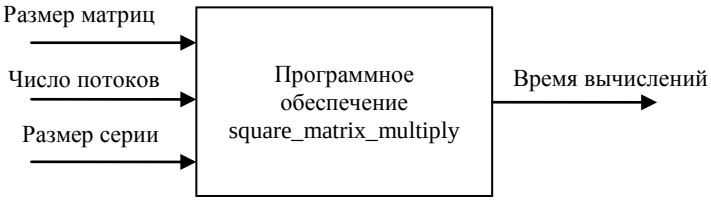


Рис. 54. Схема эксперимента

Программное обеспечение реализовано в виде приложения для операционной системы *Windows* (рис. 55).

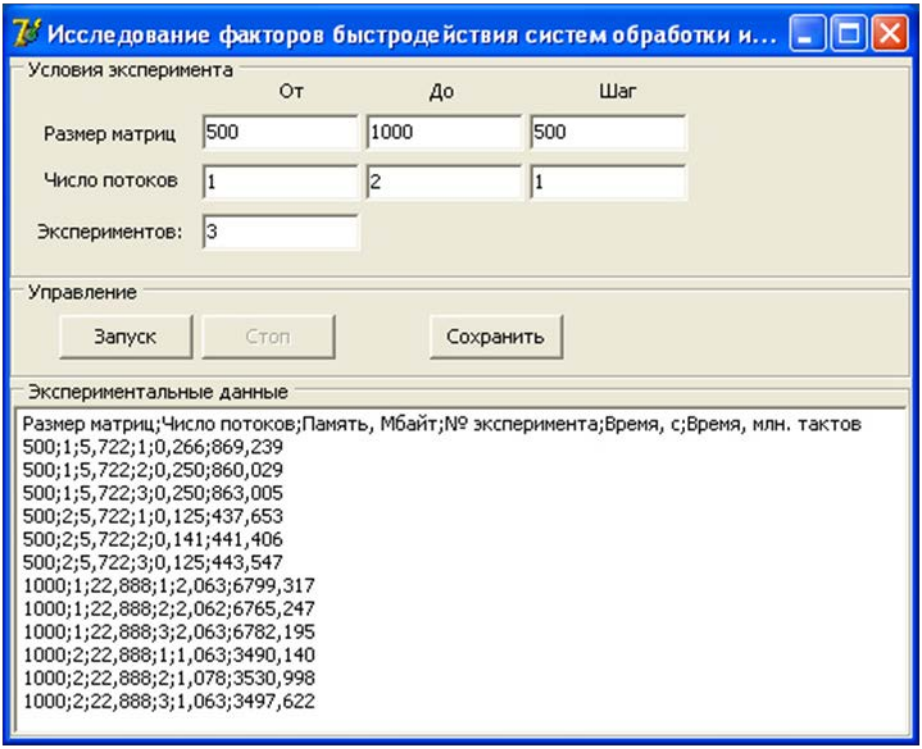


Рис. 55. Интерфейс программного обеспечения *square_matrix_multiply*

Запуск эксперимента производится нажатием кнопки «Запуск». Результатом эксперимента является набор данных, включающий как условия эксперимента, так и экспериментальные данные. Результаты эксперимента можно сохранить в текстовый файл формата *csv* с помощью нажатия кнопки «Сохранить». Прервать выполнение экспериментов можно с помощью кнопки «Стоп».

Вопросы для самоконтроля

1 Перечислите факторы быстродействия систем анализа данных.

2 Дайте определение понятия «вычислительная сложность».

3 Приведите принцип сравнения вычислительной сложности алгоритмов.

4 Приведите пример сравнения вычислительной сложности алгоритмов.

5 Приведите алгоритм экспериментального определения вычислительной сложности.

6 Приведите алгоритм прогнозирования быстродействия.

Задачи

1 Сравните вычислительную сложность алгоритмов с $O(n) = n\log n$ и $O(n) = n^2$.

2 Решите задачу.

Проведен ряд экспериментов по измерению времени, затрачиваемого на обработку массива данных (табл. 30).

Таблица 30. Экспериментальные данные

Размер массива данных, кбайт	Время обработки, с
10	6
20	10
30	19
40	26
50	36
60	48
70	62
80	88
90	96
100	115

Найдите функциональную зависимость времени обработки данных от размера массива данных.

3 Решите задачу.

Экспериментальным путем определено, что зависимость времени t (секунды), затрачиваемого на обработку файла данных объемом S Мбайт, подчиняется закону

$$t = 5{,}2 \cdot S^2 . \qquad (25)$$

Определите:

1) время, которое будет затрачено на обработку файла объемом 5 Мбайт;

2) объем файла, который будет обработан за 20 с.

Лабораторная работа «Быстродействие систем анализа данных»

Общие сведения

Целью работы является приобретение навыка анализа быстродействия систем обработки данных.

Задачи:

1 Определение вычислительной сложности алгоритма.

2 Прогнозирование затрат времени на обработку данных.

В качестве инструментального средства используется программное обеспечение *square_matrix_multiply*, описанное ранее на с. 77.

Исходные данные

Таблица 31. Варианты задания

Вариант	Размер матриц			Количество вычислительных потоков		
	Мин.	Макс.	Шаг	Мин.	Макс.	Шаг
1	100	1000	100	1	3	1
2	200	1100	100	1	4	1
3	400	1200	100	1	5	1
4	600	1300	100	1	3	1
5	800	1400	100	1	4	1
6	500	1800	200	1	5	1
7	300	1900	200	1	3	1
8	400	1900	200	1	4	1
9	500	2000	200	1	5	1
10	300	2000	200	1	3	1

Порядок выполнения

1 Подготовка:

1.1 Выберите задание (табл. 31).

1.2 Загрузите программное обеспечение *square_matrix_multiply* из курса «Анализ данных» СДО университета [2].

1.3 Опишите вычислительную систему: процессор, оперативная память, операционная система.

2 Проведение эксперимента:

2.1 Запустите программу *square_matrix_multiply*.

2.2 Задайте условия эксперимента (размер матриц, количество вычислительных потоков) в соответствии с заданием. Установите число экспериментов, равное трём.

2.3 Проведите эксперименты и сохраните их результаты. В ходе экспериментов сделайте копию экрана с изображением вкладок «Процессы» и «Быстродействие» диспетчера задач Windows.

2.4 Выполните предварительную обработку экспериментальных данных – усреднение результатов по серии экспериментов. Предварительную обработку удобно проводить с помощью инструмента «Сводные таблицы» программного обеспечения Microsoft Excel.

2.5 Постройте графики зависимости (для разного числа потоков) времени выполнения вычислений от размера матриц. Пример приведен ниже (рис. 56).

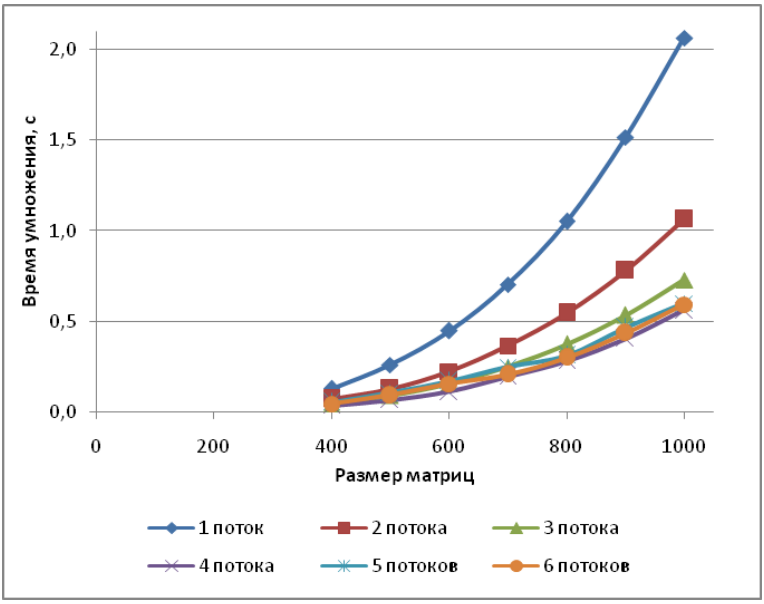


Рис. 56. Зависимость времени вычислений от размера матриц

2.6 Определите функцию, описывающую вычислительную сложность использованного в программе *square_matrix_multiply* алгоритма. Для этого можно воспользоваться инструментом «Линия

тренда» при построении диаграмм в *Microsoft Excel*. Построить тренд для случая одного вычислительного потока. Пример приведен ниже (рис. 57).

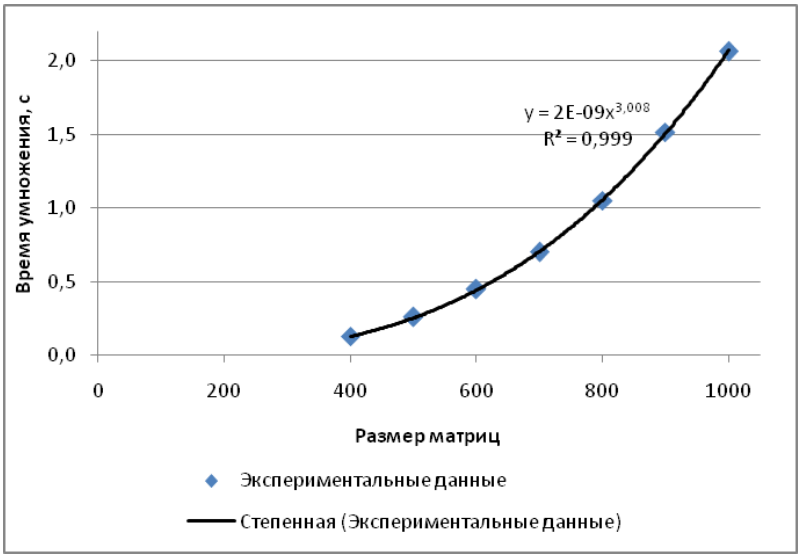


Рис. 57. Аппроксимация экспериментальных данных

3 Анализ результатов:

3.1 Сделайте выводы о влиянии объема исходных данных и фактора распараллеливания на время решения вычислительной задачи.

3.2 Сделайте выводы о влиянии объема исходных данных и фактора распараллеливания на время решения вычислительной задачи.

3.3 Сравните время, необходимое для решения задачи умножения матриц размером 10000x10000 при одном вычислительном потоке с помощью алгоритма, использованного в программе *square_matrix_multiply* и с помощью алгоритма Штрассена ($O(n) = n^{2,81}$).

3.2 Продемонстрируйте преподавателю полученные результаты. При наличии замечаний провести повторные эксперименты.

4 Отчет о работе:

4.1 Составьте отчет.

4.2 Преобразуйте отчет в формат PDF.

4.3 Создайте архив в формате ZIP, содержащий отчет и таблицу с расчетами и графиками (файл Excel).

4.4 Прикрепите архив в раздел «Отчет по лабораторной работе №6» (быстродействие систем анализа данных) курса «Анализ данных» СДО университета [2].

4.5 При наличии замечаний от преподавателя скорректируйте отчет.

Содержание отчета

Отчет должен содержать:

1 Титульный лист: наименование работы, вариант задания, ФИО студента, номер учебной группы, дата выполнения работы.

2 Реферат.

3 Оглавление.

4 Задание.

5 Описание выполненной работы.

6 Условия эксперимента:

6.1 Исходные данные.

6.2 Описание вычислительной системы.

7 Эксперименты:

7.1 Копия экрана (диспетчер задач, вкладка «Процессы»).

7.2 Копия экрана (диспетчер задач, вкладка «Быстродействие»).

7.3 Экспериментальные данные.

8 Обработка экспериментальных данных:

8.1 Результаты усреднения экспериментальных данных.

8.2 Графики зависимости времени выполнения вычислений от размера матриц, зависимости времени выполнения вычислений от количества вычислительных потоков.

8.3 Функция, описывающая вычислительную сложность использованного алгоритма.

9 Выводы.

10 Список использованных источников (нормативные документы).

11 Приложения.

Отчет должен быть оформлен в соответствии с действующими стандартами университета [18, 19].

ЗАКЛЮЧЕНИЕ

В данном учебном пособии рассмотрены основы анализа данных.

Об интересе к сфере анализа данных свидетельствует не только рост числа научных исследований, но и еще два фактора:

1 Популярность соответствующих массовых онлайн-курсов. По состоянию на декабрь 2015 года на образовательной платформе Coursera [24] было представлено 1764 курса, более 100 из которых посвящено сфере анализа данных.

2 Появление большого количества конкурсов по анализу открытых данных. Наиболее полная информация о подобных конкурсах содержится в разделе «Конкурсы» портала «Открытые данные России» [25].

Читателям, заинтересованным в систематизации и углублении знаний в сфере анализа данных, можно рекомендовать прохождение курсов «Machine Learning» Стэнфордского университета (автор Andrew Ng) [25] и «Введение в машинное обучение» Высшей школы экономики (авторы К.В. Воронцов и Е. Соколов) [27].

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Российская Федерация. Министерство образования и науки. Приказ от 14 января 2010 г. №27 «Об утверждении и введении в действие федерального государственного образовательного стандарта высшего профессионального образования по направлению подготовки 080500 бизнес-информатика (квалификация (степень) "бакалавр")» [Текст] : офиц. текст. — М. : Минюст РФ, 2010. — 15 с.

2. Курс: Анализ данных [Электронный ресурс] : [б. и.], 2013. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://do.ssau.ru/moodle/course/view.php?id=442 (Дата обращения 23.12.2015).

3. Hilbert, M. The world's technological capacity to store, communicate, and compute information [Text] / M. Hilber, P. López. // Science, Volume 332, Issue 6025, April 2011, Pages 60-65.

4. CMS releases new batch of research data from LHC | CMS Experiment [Электронный ресурс] : [б. и.], 2016. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://cms.web.cern.ch/news/cms-releases-new-batch-research-data-lhc (Дата обращения 25.04.2016).

5. Boeing 787s to create half a terabyte of data per flight, says Virgin Atlantic [Электронный ресурс] : [б. и.], 2016. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://www.computerworlduk.com/news/data/boeing-787s-create-half-terabyte-of-data-per-flight-says-virgin-atlantic-3433595/ (Дата обращения 25.04.2016).

6. Machine learning | artificial intelligence | Britannica.com [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://global.britannica.com/technology/machine-learning (Дата обращения 23.12.2015).

7. Анализ данных – обучение в Яндексе [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : https://academy.yandex.ru/events/data_analysis (Дата обращения 23.12.2015).

8. Deep Blue [Text] / M. Campbell, J. Hoane Jr., F.-h. Hsu [et al.] // Artificial Intelligence, Volume 134, Issue 1-2, January 2002, Pages 57-83.

9. Building watson: An overview of the deepQA project [Text] / D. Ferrucci, E. Brown, J. Chu-Carroll [et al.] // AI Magazine. Volume 31, Issue 3, September 2011, Pages 59-79.

10. Google Self-Driving Car Project [Электронный ресурс] : [б. и.], 2016. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. – URL : https://www.google.com/selfdrivingcar (Дата обращения 23.12.2015).

11. Fast Artificial Neural Network Library (FANN) [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://leenissen.dk/fann/wp (Дата обращения 23.12.2015).

12. OpenCV | OpenCV [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://opencv.org (Дата обращения 23.12.2015).

13. UCI Machine Learning Repository: Bank Marketing Data Set [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://archive.ics.uci.edu/ml/datasets/Bank+Marketing (Дата обращения 23.12.2015).

14. Data | The World Bank [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://data.worldbank.org (Дата обращения 23.12.2015).

15. Data.gov.ru: открытые данные России [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://data.gov.ru (Дата обращения 23.12.2015).

16. Федеральный закон от 27 июля 2006 г. № 149-ФЗ «Об информации, информационных технологиях и о защите информации» (http://base.garant.ru/12148555).

17. Главная Федеральная служба государственной статистики [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : https://www.gks.ru (Дата обращения 23.12.2015).

18. СТО СГАУ 02068410-004-2007. Общие требования к учебным текстовым документам [Текст]. — Введ. 2007—10—09. — Самара. : СГАУ, 2007. —30 с.

19. Приказ «Об актуализации стандартов СГАУ (СТО СГАУ) от 14.11.2014 № 381-О. - URL :

http://www.ssau.ru/files/science/org/no/osm/changes_sto_ssau.pdf (Дата обращения 21.01.2016).

20. Fern´andez-Delgado, M. Do we need hundreds of classifiers to solve real world classification problems? / M. Fern´andez-Delgado, E. Cernadas, S. Barro. [Text] // Journal of Machine Learning Research 15 (2014) 3133-3181 URL: http://jmlr.csail.mit.edu/papers/volume15/delgado14a/delgado14a.pdf.

21. UCI Machine Learning Repository [Электронный ресурс]: [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://archive.ics.uci.edu/ml/ (Дата обращения 23.12.2015).

22. Data clustering: A review [Text] / Jain, A.K.a, M.N.b Murty, P.J.c Flynn [et al.] // ACM Computing Surveys, Volume 31, Issue 3, 1999, Pages 264-323.

23. Xu, R. Survey of clustering algorithms (Review) [Text] / R. Xu, D. Wunsch II, // IEEE Transactions on Neural Networks, Volume 16, Issue 3, May 2005, Pages 645-678.

24. Coursera – бесплатные онлайн-курсы от ведущих университетов мира | Coursera [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : https://www.coursera.org (Дата обращения 23.12.2015).

25. Конкурсы | Data.gov.ru [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : http://data.gov.ru/competitions (Дата обращения 23.12.2015).

26. Машинное обучение – Стэнфордский университет | Cousera [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : https://www.coursera.org/learn/machine-learning (Дата обращения 23.12.2015).

27. Введение в машинное обучение – Высшая школа экономики | Cousera [Электронный ресурс] : [б. и.], 2015. - Электрон. текстовые дан. on-line. - Загл. с титул. экрана. - URL : https://www.coursera.org/learn/introduction-machine-learning (Дата обращения 23.12.2015).

Учебное издание

Михаил Алексеевич Поручиков

АНАЛИЗ ДАННЫХ

Учебное пособие

Редактор Т.К. Кретинина
Доверстка Т.С. Зинкина

Подписано в печать 02.06.2016. Формат 60x84 1/16.
Бумага офсетная. Печать офсетная. Печ. л. 5,5.
Тираж 100 экз. Заказ . Арт. 31/2016.

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.