

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ  
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ  
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ  
УНИВЕРСИТЕТ имени академика С.П.КОРОЛЕВА  
(НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ)»

**Автоматическая классификация**

*Электронные методические указания  
к лабораторной работе № 4*

**САМАРА**

**2010**

Составители:       БАВРИНА Алина Юрьевна,  
                              МЯСНИКОВ Владислав Валерьевич

В лабораторной работе № 4 по дисциплине «Математические методы распознавания образов и понимания изображений» изучаются методы автоматической классификации. Рассматривается постановка алгоритма автоматической классификации, меры сходства, критерии кластеризации и два известных алгоритма: минимаксный и  $K$  внутригрупповых средних.

Методические указания предназначены для магистров направления 010400.68 «Прикладная математика и информатика», обучающихся по программе «Математические и компьютерные методы обработки изображений и геоинформатики».

**Цель работы** - изучение теоретических основ и экспериментальное исследование методов автоматической классификации для распознавания образов.

## 1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЛАБОРАТОРНОЙ РАБОТЫ

### 1.1. Постановка задачи автоматической классификации

Пусть классификации подлежат  $N$  объектов, каждый из которых характеризуется  $n$ -мерным вектором признаков  $\bar{x}$ , то есть дано множество векторов  $\{\bar{x}_i\}_{i=0}^{N-1}$ . Эти вектора рассматриваются как фиксированные. Каждый объект должен быть отнесен к одному из  $L$  классов:  $\Omega_0, \Omega_1, \dots, \Omega_{L-1}$ , где число классов  $L$  может быть известно заранее или может быть не известно. Таким образом основной вопрос задачи автоматической классификации (АК), так же как и в других задачах классификации, это вопрос об определении класса<sup>1</sup>.

Один из возможных подходов к определению класса состоит в его понимании как *кластера (таксона)*, то есть компактной в некотором смысле области в признаковом пространстве. В этом случае задача АК является задачей *кластер-анализа* и представляет собой задачу идентификации групп схожих образов в анализируемом множестве данных (или задачу выделения кластеров).

Процесс выделения кластеров является искусством весьма "эмпирическим", так как работа конкретного алгоритма зависит не только от характера анализируемых данных, но в значительной степени и от выбранной меры подобия образов и метода, используемого для идентификации кластеров, и даже от последовательности просмотра образов.

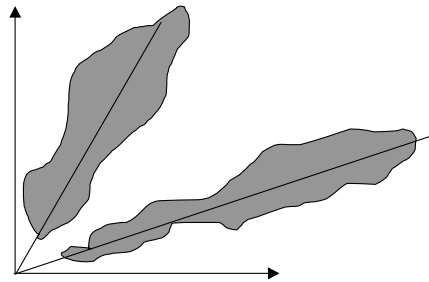
### 1.2. Меры сходства

Для того чтобы определить на множестве данных кластеры, необходимо в первую очередь ввести меру сходства (подобия) образов (векторов признаков), которая может быть положена в основу правила отнесения образов характеризуемой некоторым центром кластера. В качестве мер сходства широко используются следующие:

---

<sup>1</sup> Речь идет не о задании самого класса, а о задании областей признакового пространства которые "соответствуют" классам.

- *евклидово расстояния*  $d(\bar{x}, \bar{z}) = \|\bar{x} - \bar{z}\|$  (меньше  $d$  - больше сходство);
- *расстояние Махаланобиса*  $d(\bar{x}, \bar{z}) = (\bar{x} - \bar{z})^T B^{-1}(\bar{x} - \bar{z})$  (меньше  $d$  - больше сходство). Используется, когда известны статистические характеристики кластеров (матрицы разброса);
- *косинус угла между векторами*  $d(\bar{x}, \bar{z}) = \frac{\bar{x}^T \bar{z}}{\|\bar{x}\| \cdot \|\bar{z}\|}$ . Данную меру удобно использовать, когда кластеры имеют тенденцию располагаться вдоль главных осей, как например изображено на рисунке:



- *мера Такимото*  $d(\bar{x}, \bar{z}) = \frac{\bar{x}^T \bar{z}}{\bar{x}^T \bar{x} + \bar{z}^T \bar{z} - \bar{x}^T \bar{z}}$ .

В дальнейшем ограничимся евклидовой мерой подобия.

### 1.3. Критерии кластеризации

После выбора меры сходства необходимо определить *критерий кластеризации*. Критерий кластеризации может либо воспроизводить некие эвристические соображения, либо основываться на минимизации (или максимизации) какого-нибудь показателя качества.

При *эвристическом подходе* решающую роль играют интуиция и опыт. Он предусматривает задание набора правил, которые обеспечивают использование выбранной меры сходства для отнесения образов к одному из кластеров.

Подход к кластеризации, предусматривающий *использование показателя качества*, связан с разработкой процедур, которые обеспечат минимизацию или максимизацию выбранного показателя качества. Одним из наиболее популярных показателей является сумма квадратов ошибки:

$$\mathfrak{S} = \sum_{l=0}^{L-1} \sum_{\bar{x} \in S_l} \|\bar{x} - \bar{M}_j\|^2 \quad (1)$$

где  $L$  — число кластеров,  $S_l$  — множество образов (векторов признаков), относящихся к  $l$ -му кластеру, а  $\bar{M}_l = \frac{1}{N_l} \sum_{\bar{x} \in S_l} \bar{x}$  — вектор выборочных средних значений для множества образов  $S_l$ , величина  $N_l$  характеризует количество образов, входящих во множество  $S_l$ . Как видно, показатель качества (1) определяет общую сумму квадратов отклонений характеристик всех образов, входящих в некоторый кластер, от соответствующих средних значений по этому кластеру. Алгоритм, основанный на этом показателе качества, рассматривается в ниже. Естественно, существует масса показателей качества помимо рассмотренного. Нередко применяются алгоритмы отыскания кластеров, основанные на совместном использовании эвристического подхода и показателя качества. Подобной комбинацией является алгоритм ИСОМАД (ISODATA) [2].

Ниже рассмотрены три наиболее известные алгоритма [2], которые являются примерами как эвристического подхода, так и подхода, использующего показатель качества.

#### 1.4. Простой алгоритм выделения кластеров

Пусть задано множество  $N$  образов  $\{\bar{x}_0, \bar{x}_1, \dots, \bar{x}_{N-1}\}$ . Пусть также центр первого кластера  $\bar{M}_0$  совпадает с любым из заданных образов и определена произвольная неотрицательная пороговая величина  $T$ ; для удобства можно считать, что  $\bar{M}_0 = \bar{x}_0$ . После этого вычисляется расстояние  $d(\bar{M}_0, \bar{x}_1)$  между существующим центром кластера  $\bar{M}_0$  и образом  $\bar{x}_1$ . Если это расстояние больше значения пороговой величины  $T$  ( $d(\bar{M}_0, \bar{x}_1) > T$ ), то учреждается новый кластер с центром  $\bar{M}_1 = \bar{x}_1$ . В противном случае образ  $\bar{x}_1$  включается в кластер, центром которого является  $\bar{M}_0$ . Пусть условие  $d(\bar{M}_0, \bar{x}_1) > T$  выполнено, и  $\bar{M}_1$  — центр нового кластера.

На следующем шаге вычисляются расстояния  $d(\bar{M}_0, \bar{x}_2)$  и  $d(\bar{M}_1, \bar{x}_2)$  до образа  $\bar{x}_2$  от центров кластеров  $\bar{M}_0$  и  $\bar{M}_1$ . Если оба расстояния оказываются больше порога  $T$  ( $d(\bar{M}_0, \bar{x}_2) > T$  и  $d(\bar{M}_1, \bar{x}_2) > T$ ), то учреждается новый кластер с центром  $\bar{M}_2 = \bar{x}_2$ . В противном случае образ  $\bar{x}_2$  зачисляется в тот кластер  $l$  ( $l = \overline{0,1}$ ), чей центр  $\bar{M}_l$  к нему ближе.

Подобным же образом расстояния от каждого нового образа  $\bar{x}_j$  ( $j = \overline{1, N-1}$ ) до каждого известного центра кластера  $\bar{M}_l$  ( $l = \overline{0, L-1}$ ) вычисляются и сравниваются с пороговой величиной. Если все эти расстояния превосходят значение порога  $T$  ( $\forall l = \overline{0, L-1} \quad d(\bar{M}_l, \bar{x}_j) > T$ ), то учреждается новый кластер с центром  $\bar{M}_L = \bar{x}_j$  (и число кластеров увеличивается на единицу). В противном случае образ зачисляется в кластер с самым близким к нему центром.

Результаты описанной процедуры определяются выбором первого центра кластера, порядком просмотра образов, значением пороговой величины  $T$  и, конечно, геометрическими характеристиками данных.

### 1.5. Алгоритм максиминного расстояния

Алгоритм, основанный на принципе *максиминного* (максимально-минимального) *расстояния*, представляет собой еще одну простую эвристическую процедуру, использующую евклидово расстояние. Этот алгоритм в принципе аналогичен схеме из п.1.4, за исключением того обстоятельства, что в первую очередь он выявляет наиболее удаленные кластеры.

Алгоритм состоит из нескольких шагов.

Шаг 1. Произвольным образом выбирается центр первого кластера  $\bar{M}_0$ . Удобно выбирать в качестве центра кластера  $\bar{M}_0$  тот вектор признаков  $\bar{x} \in \{\bar{x}_j\}_{j=0}^{N-1}$ , который обладает некоторыми «экстремальными» свойствами. Например, удобным является тот вектор, удаление которого от среднего всех векторов из выборки максимально:

$$\bar{M}_0 = \arg \max_{\bar{x} \in \{\bar{x}_j\}_{j=0}^{N-1}} d\left(\bar{x}, \frac{1}{N} \sum_{i=0}^{N-1} \bar{x}_i\right).$$

Шаг 2. Выбирается центр второго кластера  $\bar{M}_1$ . В качестве центра используется тот вектор  $\bar{x} \in \{\bar{x}_j\}_{j=0}^{N-1}$ , который наиболее удален от первого центра кластера:

$$\bar{M}_1 = \arg \max_{\bar{x} \in \{\bar{x}_j\}_{j=0}^{N-1}} d(\bar{M}_0, \bar{x}).$$

Шаг  $L$  ( $L > 2$ ). Выбирается цент кластера  $\bar{M}_{L-1}$ . Для этого вычисляются все расстояния между оставшимися образами (векторами признаков)  $\bar{x} \in \{\bar{x}_j\}_{j=0}^{N-1} \setminus \{\bar{M}_l\}_{l=0}^{L-2}$  и уже существующими центрами кластеров  $\bar{M}_0, \bar{M}_1, \dots, \bar{M}_{L-2}$ . Для каждого оставшегося образа  $\bar{x}$  находится тот центр кластера  $l$ , расстояние до которого минимально:  $l = \arg \min_{l=0, L-2} d(\bar{M}_l, \bar{x})$  (вектора распределяются по кластерам по критерию близости к их центру). В качестве претендента на новый центр кластера  $\bar{M}_{L-1}$  берется тот вектор признаков  $\bar{x}$ , у которого это минимальное расстояние (расстояние до центра «своего» кластера) максимально:

$$\tilde{M}_{L-1} = \arg \max_{\bar{x} \in \{\bar{x}_j\}_{j=0}^{N-1}} \min_{l=0, L-2} d(\bar{M}_l, \bar{x}).$$

Полученное для выбранного вектора расстояние  $d_{\min} = \min_{l=0, L-2} d(\bar{M}_l, \tilde{M}_{L-1})$  от него до ближайшего кластера сравнивается с некоторым «типичным» расстоянием  $d_{\text{typical}}$  между кластерами. Если полученное расстояние больше «типичного», то этот вектор становится новым центром кластера:

$$d_{\min} > d_{\text{typical}} \Rightarrow \bar{M}_{L-1} = \tilde{M}_{L-1}.$$

В противном случае – процесс выделения кластеров останавливается.

Выбор типичного расстояния может быть осуществлен различными способами. Один из наиболее типичных способов, это задание типичного расстояния равного некоторой части  $\eta$  от среднего расстояния между уже существующими кластерами:

$$d_{\text{typical}} = \eta \left( \frac{2}{(L-1)(L-2)} \right) \sum_{l=0}^{L-2} \sum_{j=l+1}^{L-2} d(\bar{M}_l, \bar{M}_j).$$

Величина  $\eta$  выбирается из условия  $0 < \eta < 1$ . Типичное значение:  $\eta = \frac{1}{2}$ .

## 1.6. Алгоритм $K$ внутригрупповых средних

Алгоритмы, рассмотренные в п.1.4 и п.1.5 являются, в сущности, эвристическими процедурами. Алгоритм, представленный ниже, минимизирует показатель качества, заданный как сумма квадратов расстояний всех точек, входящих

в кластерную область, до центра кластера. Эта процедура, которую часто называют алгоритмом, основанным на вычислении  $K$  внутригрупповых средних, состоит из следующих шагов.

*Шаг 1.* Выбираются  $K$  исходных центров кластеров на первой итерации ( $r = 1$ ):  $\bar{M}_0(r), \bar{M}_1(r), \dots, \bar{M}_{K-1}(r)$ . Этот выбор производится произвольно, и обычно в качестве исходных центров используются первые  $K$  образов из заданного множества  $\{\bar{x}_j\}_{j=0}^{N-1}$  ( $N > K$ ):

$$\bar{M}_0(1) = \bar{x}_0, \quad \bar{M}_1(1) = \bar{x}_1, \quad \dots, \quad \bar{M}_{K-1}(1) = \bar{x}_{K-1}.$$

Номер итерации увеличивается:  $r := r + 1$ .

*Шаг 2.* На  $r$ -ой итерации ( $r = 2, 3, \dots$ ) исходное множество образов  $\{\bar{x}_j\}_{j=0}^{N-1}$  распределяется по  $K$  кластерам по правилу близости. То есть некоторый образ  $\bar{x} \in \{\bar{x}_j\}_{j=0}^{N-1}$  относят в кластер  $S_k(r)$  ( $k = \overline{0, K-1}$ ) с центром  $\bar{M}_k(r-1)$ , рассчитанным на предыдущем шаге, если этот центр - ближайший:

$$\bar{x} \in S_k(r), \quad k = \arg \min_{j=0, K-1} d(\bar{x}, \bar{M}_j(r-1)).$$

Таким образом  $S_k(r)$  - множество образов, входящих в кластер с номером  $k$  на  $r$ -ой итерации алгоритма. В случае равенства расстояний от некоторого образа до нескольких центров кластеров решение об отнесении этого образа к одному из них принимается произвольным образом.

*Шаг 3.* На основе результатов шага 2 определяются новые центры кластеров  $\bar{M}_0(r), \bar{M}_1(r), \dots, \bar{M}_{K-1}(r)$  на  $r$ -ой итерации алгоритма. Они определяются из условия, чтобы сумма квадратов расстояний между всеми образами, принадлежащими кластеру  $S_k(r)$ , и новым центром кластера  $\bar{M}_k(r)$  должна быть минимальной. Другими словами, новые центры кластеров выбираются таким образом, чтобы минимизировать частные показатели

$$J_k(r) = \sum_{\bar{x} \in S_k(r)} d(\bar{x}, \bar{M}_k(r))^2, \quad k = 0, 1, \dots, K-1$$

и, следовательно, интегральный показатель качества кластеризации на  $r$ -ом шаге



$$J(r) = \sum_{k=0}^{K-1} J_k(r) = \sum_{k=0}^{K-1} \sum_{\bar{x} \in S_k(r)} d(\bar{x}, \bar{M}_k(r))^2.$$

Для случая евклидова расстояния новый центр кластера  $\bar{M}_k(r)$ , обеспечивающий минимизацию соответствующего частного показателя  $J_k(r)$ , является, в сущности, выборочным средним, определенным по множеству образов в кластере  $S_k(r)$ :

$$\bar{M}_k(r) = \frac{1}{N_k} \sum_{x \in S_k(r)} \bar{x}, \quad k = 0, 1, \dots, K-1,$$

где  $N_k = |S_k(r)|$  - число образов, входящих в кластер  $S_k(r)$  на  $r$ -ом шаге. Как видно, что название алгоритма « $K$  внутригрупповых средних» определяется способом, принятым для последовательной коррекции назначения центров кластеров.

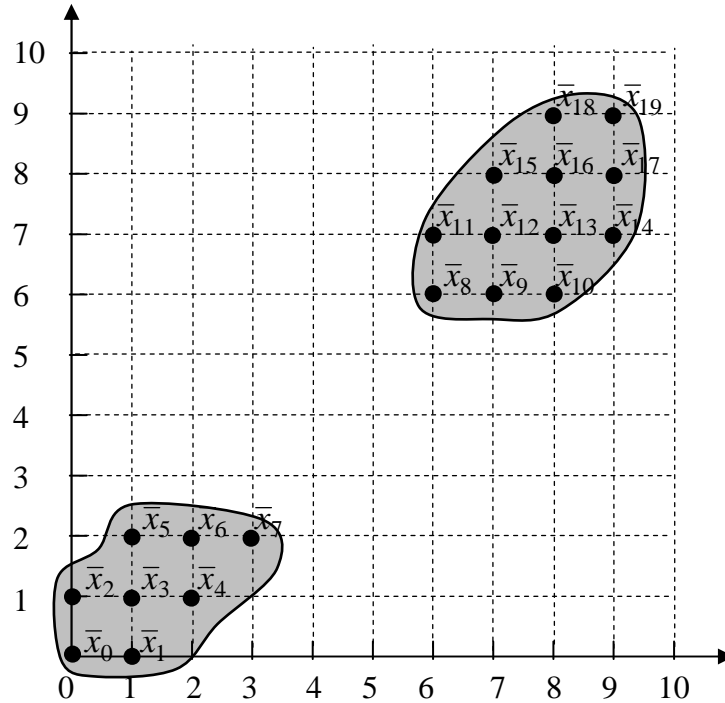
*Шаг 4.* Равенство центров кластеров на соседних шагах  $\bar{M}_k(r) = \bar{M}_k(r-1)$  ( $k = \overline{0, K-1}$ ) является условием сходимости алгоритма, и при его достижении выполнение алгоритма заканчивается. В противном случае алгоритм повторяется с шага 2 с новым номером итерации  $r := r + 1$ .

Качество работы алгоритма  $K$  внутригрупповых средних зависит от числа выбираемых центров кластеров  $K$ , от выбора центров кластеров на первой итерации  $\bar{M}_0(1), \bar{M}_1(1), \dots, \bar{M}_{K-1}(1)$  и, естественно, от геометрических особенностей данных. От последовательности просмотра данных результаты не зависят.

Хотя для этого алгоритма общее доказательство сходимости не известно, получения приемлемых результатов можно ожидать в тех случаях, когда данные образуют характерные гроздьи, отстоящие друг от друга достаточно далеко. В большинстве случаев практическое применение этого алгоритма потребует проведения экспериментов, связанных с выбором различных значений параметра  $K$  и расположения первоначальных центров кластеров.

**Пример.** В качестве простой числительной иллюстрации алгоритма  $K$  внутригрупповых средних рассмотрим образы, представленные на рисунке 1. Процедура кластеризации протекает следующим образом:

*Шаг 1.* Задается  $K = 2, r = 1$  и выбирается  $\bar{M}_0(1) = \bar{x}_0 = (0, 0)'$ ,  
 $\bar{M}_1(1) = \bar{x}_1 = (1, 0)'$ . Увеличивается номер итерации:  $r = 2$ .



*Рисунок 1.* Выборка образов, иллюстрирующая работу алгоритма  $K$  внутригрупповых средних

*Шаг 2.* Так как  $\|\bar{x}_0 - \bar{M}_0(1)\| < \|\bar{x}_0 - \bar{M}_1(1)\|$  и  $\|\bar{x}_2 - \bar{M}_0(1)\| < \|\bar{x}_2 - \bar{M}_1(1)\|$ , то  $S_0(2) = \{\bar{x}_0, \bar{x}_2\}$ . Аналогично устанавливается, что остальные образы расположены ближе к центру кластера  $\bar{M}_1(1)$ , и поэтому  $S_1(1) = \{x_1, x_3, x_4, \dots, x_{19}\}$ .

*Шаг 3.* Коррекция центров кластеров:

$$\bar{M}_0(2) = \frac{1}{N_0} \sum_{\bar{x} \in S_0(2)} \bar{x} = \frac{1}{2} (\bar{x}_0 + \bar{x}_2) = \begin{pmatrix} 0.0 \\ 0.5 \end{pmatrix},$$

$$\bar{M}_1(2) = \frac{1}{N_1} \sum_{\bar{x} \in S_1(2)} \bar{x} = \frac{1}{18} (\bar{x}_1 + \bar{x}_3 + \dots + \bar{x}_{19}) = \begin{pmatrix} 5.67 \\ 5.33 \end{pmatrix}.$$

*Шаг 4.* Так как  $\bar{M}_k(2) \neq \bar{M}_k(1)$  ( $k = 1, 2$ ), то

- увеличивается номер итерации  $r = 3$  и
- производится возврат к шагу 2.

*Шаг 2.* Выбор новых центров кластеров приводит к неравенствам:

- $\|\bar{x}_j - \bar{M}_0(2)\| < \|\bar{x}_j - \bar{M}_1(2)\|$  для  $j = 0, 1, \dots, 7$  и
- $\|\bar{x}_j - \bar{M}_0(2)\| > \|\bar{x}_j - \bar{M}_1(2)\|$  для  $j = 8, 9, \dots, 19$ .

Следовательно,  $S_0(3) = \{\bar{x}_0, \bar{x}_1, \dots, \bar{x}_7\}$  и  $S_1(3) = \{\bar{x}_8, \bar{x}_9, \dots, \bar{x}_{19}\}$ .

*Шаг 3.* Коррекция центров кластеров:

$$\bar{M}_0(3) = \frac{1}{N_{0 \bar{x} \in S_0(3)}} \sum_{\bar{x} \in S_0(3)} \bar{x} = \frac{1}{8} (\bar{x}_0 + \bar{x}_1 + \dots + \bar{x}_7) = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix},$$

$$\bar{M}_1(3) = \frac{1}{N_{1 \bar{x} \in S_1(3)}} \sum_{\bar{x} \in S_1(3)} \bar{x} = \frac{1}{12} (\bar{x}_9 + \bar{x}_{10} + \dots + \bar{x}_{19}) = \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}.$$

*Шаг 4.* Так как  $\bar{M}_k(3) \neq \bar{M}_k(2)$  ( $k = 1, 2$ ), то , то

- увеличивается номер итерации  $r = 4$  и
- производится возврат к шагу 2.

*Шаг 2.* Получаем те же результаты, что и на предыдущей итерации, то есть

$$S_0(4) = S_0(3) = \{\bar{x}_0, \bar{x}_1, \dots, \bar{x}_7\} \text{ и } S_1(4) = S_1(3) = \{\bar{x}_8, \bar{x}_9, \dots, \bar{x}_{19}\}.$$

*Шаг 3.* Также получаем идентичные результаты.

*Шаг 4.* Так как  $\bar{M}_k(4) = \bar{M}_k(3)$  ( $k = 1, 2$ ), то алгоритм сошелся и в результате

получены следующие центры кластеров

$$\bar{M}_0(3) = \begin{pmatrix} 1.25 \\ 1.13 \end{pmatrix}, \quad \bar{M}_1(3) = \begin{pmatrix} 7.67 \\ 7.33 \end{pmatrix}.$$

Эти результаты согласуются с человеческим интуитивным представлением о группировке искомым образов в пространстве признаков.

## 2. РЕАЛИЗАЦИЯ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ В СРЕДЕ MATHCAD

### 2.1. Генерация исходных данных

Для генерации данных, которые подлежат кластеризации, используйте алгоритм генерации случайного двумерного вектора, описанный в методических указаниях для первой лабораторной работы [4]. Ниже представлен код MathCAD программы для генерации 200 значений-реализаций случайного вектора признаков.

Текст программы в MathCad	Комментарии
$M := \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad B := \begin{bmatrix} 5 & 2 \\ 2 & 1 \end{bmatrix}$	Задание параметров нормального закона распределения
$A_{0,0} := \sqrt{R_{0,0}} \quad A_{0,1} := 0$	Определение параметров линейного преобразования
$A_{1,0} := \frac{R_{0,1}}{\sqrt{R_{0,0}}} \quad A_{1,1} := \sqrt{R_{1,1} - \frac{(R_{0,1})^2}{R_{0,0}}}$	Отображение полученного результата для матрицы линейного преобразования
$A = \begin{bmatrix} 2.236 & 0 \\ 0.894 & 0.447 \end{bmatrix}$	Отображение полученного результата для матрицы линейного преобразования
$n := 2 \quad l := 0..n-1 \quad k := 0..n-1$	Вспомогательные переменные, отвечающие за
$N := 200 \quad i := 0..N-1$	двухкомпонентность вектора $(n, l, k)$ , число
$j := 0..11$	выборочных значений $(N, i)$ и за процесс
$y_{1,i} := \sum_j (\text{rnd}(1) - 0.5)$	генерации стандартной нормально распределенной
$X_{k,i} := \sum_1 A_{k,l} \cdot y_{1,i} + M_k$	случайной величины $(j)$ .
	Генерация $N$ реализаций случайного вектора, компоненты которого – суть независимые и нормально распределенные $N(0,1)$ случайные величины.
	Генерация $N$ реализаций случайного вектора с требуемым нормальным законом распределения $N(\bar{M}, B)$ .

## 2.2. Основные функции кластеризации

Большинство алгоритмов кластеризации состоит из трех блоков: блока задания начальных центров кластеров, блока отнесения вектора к некоторому кластеру/классу и блока пересчета центров кластеров по данным векторам признаков и их номерам классов. Для этой цели удобно использовать следующий код в системе MathCad.

$L := 2$	$l := 0..L - 1$	$i := 0..N - 1$	Инициирование счетчиков и первоначальных центров кластеров
$M_{0,l} := x_{0,l}$	$M_{1,l} := x_{1,l}$		
$d_{i,l} := \sqrt{(M_{0,l} - x_{0,i})^2 + (M_{1,l} - x_{1,i})^2}$			Получение массива, содержащего номера классов для каждого вектора признаков
$ClassNum_i := \min\_ind(GetValue(d, i, L), L)$			
$k_l := 0$	$k(ClassNum_i) := k(ClassNum_i) + 1$		Пересчет центров кластеров
$M_{0,l} := 0$	$M_{0,(ClassNum_i)} := M_{0,(ClassNum_i)} + x_{0,i}$		
$M_{1,l} := 0$	$M_{1,(ClassNum_i)} := M_{1,(ClassNum_i)} + x_{1,i}$		
$M_{0,l} := \frac{M_{0,l}}{k_l}$	$M_{1,l} := \frac{M_{1,l}}{k_l}$		

Описание дополнительных функций и способ визуализации результатов кластеризации приведены в п.2.4.

## 2.3. Вспомогательные функции, используемые для кластеризации

Среда математического программирования MathCAD предназначена для математического программирования. Поэтому реализацию подпрограммы в ней удобно выполнить в виде функции, объявление которой должно предшествовать ее первому вызову. Для кластеризации удобно использовать следующие функции.

### 2.3.1. Функция поиска и отбора

$$\text{AnyIndexOfType}(\text{array}, \text{value}, \text{Len}) := \left| \begin{array}{l} \text{ind} \leftarrow 0 \\ i \leftarrow 1 \\ \text{while } i < \text{Len} \\ \quad \left| \begin{array}{l} \text{ind} \leftarrow i \text{ if } \text{array}_i = \text{value} \\ i \leftarrow i + 1 \end{array} \right. \\ \text{return ind} \end{array} \right.$$

Функция нахождения некоторого (последнего) индекса элемента  $\text{value}$  в массиве  $\text{array}$  длины  $\text{Len}$ .

$$\text{min\_ind}(x, \text{Len}) := \left| \begin{array}{l} \text{ind} \leftarrow 0 \\ \text{min} \leftarrow x_0 \\ i \leftarrow 1 \\ \text{while } i < \text{Len} \\ \quad \left| \begin{array}{l} \text{if } x_i < \text{min} \\ \quad \left| \begin{array}{l} \text{min} \leftarrow x_i \\ \text{ind} \leftarrow i \end{array} \right. \\ i \leftarrow i + 1 \end{array} \right. \\ \text{return ind} \end{array} \right.$$

Функция выдачи индекса минимального элемента в массиве  $x$  размером  $\text{Len}$ .

$$\text{Max\_Ind}(x, \text{Len}) := \left| \begin{array}{l} \text{ind} \leftarrow 0 \\ \text{max} \leftarrow x_0 \\ i \leftarrow 1 \\ \text{while } i < \text{Len} \\ \quad \left| \begin{array}{l} \text{if } x_i > \text{max} \\ \quad \left| \begin{array}{l} \text{max} \leftarrow x_i \\ \text{ind} \leftarrow i \end{array} \right. \\ i \leftarrow i + 1 \end{array} \right. \\ \text{return ind} \end{array} \right.$$

Функция выдачи индекса максимального элемента в массиве  $x$  размером  $\text{Len}$ .

$$\text{GetVector}(x, i, \text{Len}) := \left| \begin{array}{l} j \leftarrow 0 \\ \text{while } j < \text{Len} \\ \quad \left| \begin{array}{l} y_j \leftarrow x_{i,j} \\ j \leftarrow j + 1 \end{array} \right. \\ \text{return } y \end{array} \right.$$

Функция выдачи вектора, сформированного из элементов  $i$ -ой строки матрицы  $x$  размером  $\text{Len}$  элементов по горизонтали

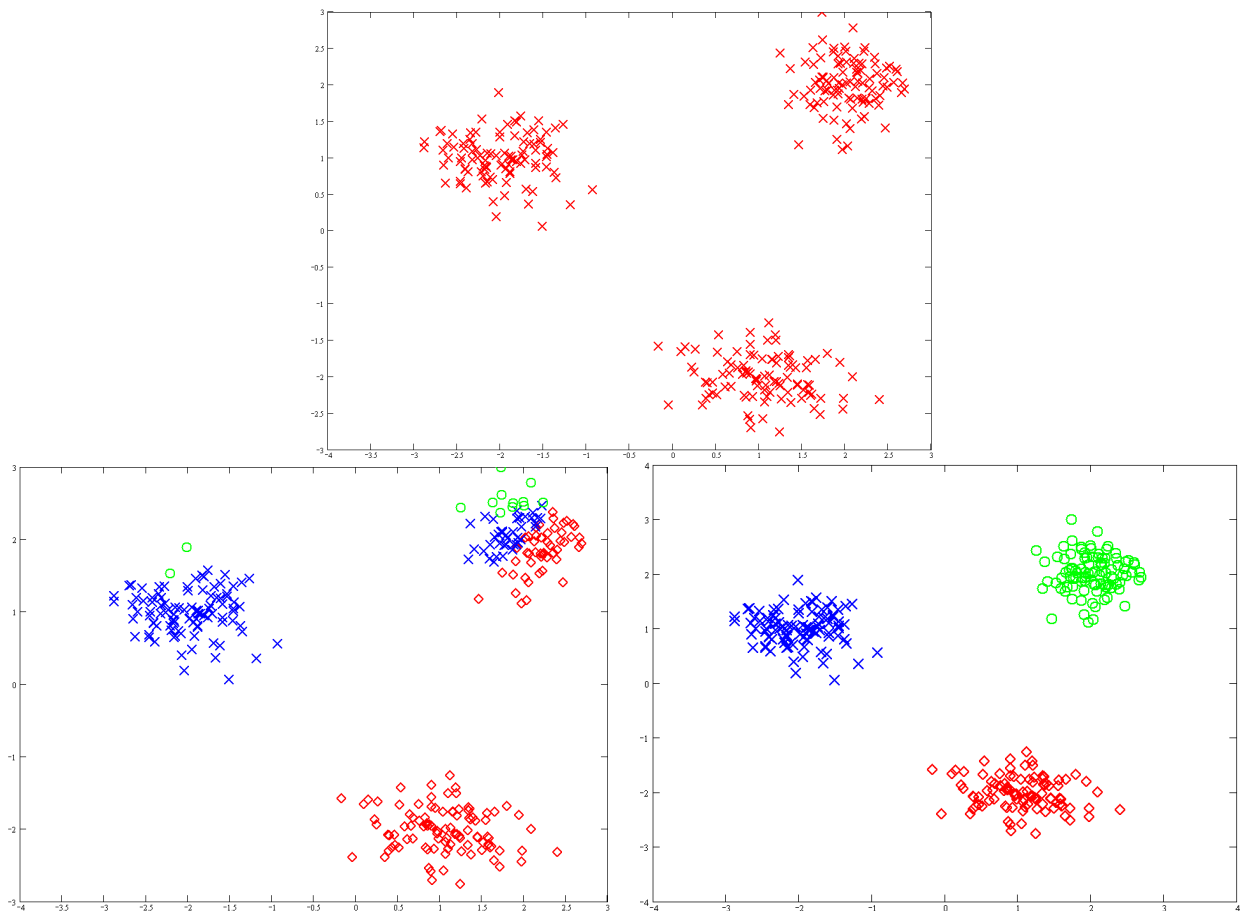
### 2.3.2. Метод отображения результатов кластеризации

Результатом кластеризации является массив  $ClassNum$ , элементы которого содержат номер класса/кластера, к которому отнесен соответствующий вектор из набора векторов  $x$ . Для отображения векторов, отнесенных к различным классам, следует распределить весь набор векторов на два набора с векторами, принадлежащими только какому-либо одному классу. Для формирования списка векторов, отнесенных к некоторому классу, например к классу «0» следует использовать следующий код MathCad:

$$ind1 := AnyIndexOfValue(ClassNum, 1, N)$$

$$x1_{s,i} := if(ClassNum_i = 1, x_{s,i}, x_{s,ind1})$$

При таком формировании списка вектора, отнесенные к необходимому классу, переносятся в новый вектор, а на место остальных векторов записывается некоторый наперед заданный вектор нужного класса.



**Рисунок 2.** Пример кластеризации с использованием алгоритма  $K$  внутригрупповых средних (2 шага алгоритма)

### 3. ЛИТЕРАТУРА

1. Дуда Р., Харт П. Распознавание образов и анализ сцен: Пер. с англ. - М.: Мир, 1976. - 512 с.
2. Ту Дж., Гонсалес Р. Принципы распознавания образов: Пер. с англ. - М.: Мир, 1978. - 412с.
3. Фукунага К. Введение в статистическую теорию распознавания образов: Пер. с англ. - М.: Наука, 1979. - 368с.
4. Коломиец Э.И., Мясников В.В. Моделирование экспериментальных данных для решения задач распознавания образов: Методические указания к лабораторной работе № 1 по курсу «Методы распознавания образов».



## 4. ПОРЯДОК ВЫПОЛНЕНИЯ ЛАБОРАТОРНОЙ РАБОТЫ

### 4.1. Исходные данные

- математические ожидания для пяти случайных векторов признаков, задаются по номеру варианта учащегося в п.5 настоящего руководства;
- исполняемые в системе MathCad файлы, необходимые при выполнении лабораторной работы (предоставляются преподавателем).

### 4.2. Общий план выполнения работы

1. Смоделировать и изобразить графически обучающие выборки объема  $N=50$  для пяти нормально распределенных двумерных случайных векторов с заданными математическими ожиданиями и самостоятельно подобранными корреляционными матрицами, которые обеспечивают линейную делимость классов.
2. Объединить пять выборок в одну. Общее количество векторов в объединенной выборке должно быть 250. Полученная объединенная выборка используется для выполнения пунктов 3 и 4 настоящего плана.
3. Разработать программу кластеризации данных с использованием минимаксного алгоритма. В качестве типичного расстояния взять половину среднего расстояния между существующими кластерами. Построить отображение результатов кластеризации для числа кластеров, начиная с двух. Построить график зависимости максимального (из минимальных) и типичного расстояний от числа кластеров.
4. Разработать программу кластеризации данных с использованием алгоритма  $K$  внутригрупповых средних для числа кластеров равного 3 и 5. Для ситуации 5 кластеров подобрать начальные условия так, чтобы получить два результата: а) чтобы кластеризация максимально **соответствовала** первоначальному разбиению на классы («правильная» кластеризация); б) чтобы кластеризация максимально **не соответствовала** первоначальному разбиению на классы («неправильная» кластеризация). Для всех случаев построить графики зависимости числа векторов признаков, сменивших номер кластера, от номера итерации алгоритма.

### 4.3. Содержание отчета

Отчет по работе должен содержать:

1. Исходные данные генерируемых векторов признаков – средние и ковариационные матрицы.
2. Для минимаксного алгоритма кластеризации включить в отчет: число классов, полученное в результате работы алгоритма; график зависимости максимального (из минимальных) и типичного расстояний от числа кластеров.
3. Для алгоритма  $K$  внутригрупповых средних: нарисовать график зависимости числа векторов признаков, сменивших номер кластера, от номера итерации алгоритма (для 3 и 5 кластеров, в последнем случае - для ситуаций «правильной» и «неправильной» кластеризации).

## 5. КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Меры сходства, используемые при автоматической классификации.
2. Критерии, используемые при автоматической классификации.
3. Простой алгоритм выделения кластеров. Достоинства и недостатки.
4. Минимаксный алгоритм. Достоинства и недостатки.
5. Алгоритм  $K$  внутригрупповых средних. Достоинства и недостатки.
6. Алгоритм ИЗОМАД.

## 6. ВАРИАНТЫ ЗАДАНИЙ

Вариант

Математические ожидания пяти наборов  
нормально распределенных случайных векторов

1.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$
2.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 3 \\ 1 \end{pmatrix}.$
3.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$
4.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}.$
5.  $\bar{M}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}.$
6.  $\bar{M}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} -2 \\ -1 \end{pmatrix}.$
7.  $\bar{M}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$
8.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}.$
9.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$
10.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$
11.  $\bar{M}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$
12.  $\bar{M}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$
13.  $\bar{M}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 1 \\ -2 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$
14.  $\bar{M}_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$
15.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} -1 \\ -2 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$
16.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$
17.  $\bar{M}_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$

18.  $\bar{M}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$
19.  $\bar{M}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 2 \\ -2 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}.$
20.  $\bar{M}_1 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$
21.  $\bar{M}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -2 \\ 2 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 0 \\ 3 \end{pmatrix}.$
22.  $\bar{M}_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 3 \\ 0 \end{pmatrix}.$
23.  $\bar{M}_1 = \begin{pmatrix} 0 \\ 2 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 2 \\ -1 \end{pmatrix}.$
24.  $\bar{M}_1 = \begin{pmatrix} -2 \\ 0 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} 0 \\ -3 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$
25.  $\bar{M}_1 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}, \bar{M}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \bar{M}_3 = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \bar{M}_4 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \bar{M}_5 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}.$

## СОДЕРЖАНИЕ

<b>1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЛАБОРАТОРНОЙ РАБОТЫ.....</b>	<b>3</b>
1.1. ПОСТАНОВКА ЗАДАЧИ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ.....	3
1.2. МЕРЫ СХОДСТВА .....	3
1.3. КРИТЕРИИ КЛАСТЕРИЗАЦИИ .....	4
1.4. ПРОСТОЙ АЛГОРИТМ ВЫДЕЛЕНИЯ КЛАСТЕРОВ.....	5
1.5. АЛГОРИТМ МАКСИМИННОГО РАССТОЯНИЯ.....	6
1.6. АЛГОРИТМ $K$ ВНУТРИГРУППОВЫХ СРЕДНИХ.....	7
<b>2. РЕАЛИЗАЦИЯ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ В СРЕДЕ МАТНСАД.....</b>	<b>12</b>
2.1. ГЕНЕРАЦИЯ ИСХОДНЫХ ДАННЫХ .....	12
2.2. ОСНОВНЫЕ ФУНКЦИИ КЛАСТЕРИЗАЦИИ.....	13
2.3. ВСПОМОГАТЕЛЬНЫЕ ФУНКЦИИ, ИСПОЛЬЗУЕМЫЕ ДЛЯ КЛАСТЕРИЗАЦИИ.....	13
2.3.1. <i>Функция поиска и отбора</i> .....	14
2.3.2. <i>Метод отображения результатов кластеризации</i> .....	15
<b>3. ЛИТЕРАТУРА .....</b>	<b>16</b>
<b>4. ПОРЯДОК ВЫПОЛНЕНИЯ ЛАБОРАТОРНОЙ РАБОТЫ.....</b>	<b>17</b>
4.1. ИСХОДНЫЕ ДАННЫЕ .....	17
4.2. ОБЩИЙ ПЛАН ВЫПОЛНЕНИЯ РАБОТЫ .....	17
4.3. СОДЕРЖАНИЕ ОТЧЕТА.....	18
<b>5. КОНТРОЛЬНЫЕ ВОПРОСЫ.....</b>	<b>18</b>
<b>6. ВАРИАНТЫ ЗАДАНИЙ .....</b>	<b>19</b>