

Распознавание образов

Одно из основных направлений искусственно интеллекта

Основные определения:

Любой физический объект или процесс характеризуется набором некоторых параметров (характеристик, свойств), которые, собственно, и позволяют отличать один объект от другого. Объекты, имеющие похожие параметры, можно объединить в группу или класс. Отнесение некоторого объекта к одной из известных групп называется **распознаванием** или классификацией.

Измеряемые или вычисляемые свойства объектов, позволяющие как классифицировать объекты, так и отличить классы друг от друга, называются **признаками**.

Совокупность конкретных значений признаков, относящихся к одному объекту, называется **образом** объекта. Тогда образ это информационная модель объекта, позволяющая отличить его от других объектов в процессе распознавания.

Понятие **класс** в распознавании образов можно определить как множество образов, обладающих близкими значениями признаков. Такие образы называются **элементами класса**.

Этапы создания системы распознавания

1. Выбор характеристик (свойств) распознаваемых объектов, перспективных для распознавания – полный набор признаков
2. Выбор объектов, для обучения системы распознавания и получение их признаков – получение обучающего множества образов
3. Выбор метрики пространства признаков и способов определения расстояния между образами, между образом и классом
3. Выбор метода и проведение кластеризации (если это необходимо) обучающего множества или разделение образов на классы
4. Определение информативности признаков и на основе этого минимизация пространства признаков
5. Выбор метода классификации – по расстоянию в пространстве признаков или разбиением пространства признаков на области
6. Проверка эффективности системы распознавания на множестве контрольных образов
7. При недостаточной эффективности возврат к одному из этапов и повторение процесса

Классификация признаков

Детерминированные признаки – характеристики относящихся к одному классу образов, которые имеют конкретные и постоянные числовые значения (число отверстий или углов, цвет).

Замечание. При распознавании с детерминированными признаками ошибки их измерения не играют роли, если точность измерения признака выше, чем различие этого признака у образов, отнесенных к разным классам. В системе с использованием **только** детерминированных признаков, распознавание производится путем сравнения полученных значений признаков распознаваемого образа со значениями признаков уже классифицированных образов. Решение принимается только при полном совпадении этих значений.

Вероятностные признаки – характеристики образа, носящие случайный характер, что обусловлено природой объекта или способом получения значения признака (площадь, длина контура, масса объекта).

Замечание. В силу случайности соответствующей величины признак у различных образов, относящихся к одному классу, может принимать различные значения. Возможны и такие случаи, когда значение признака образа из одного класса практически совпадает со значением признака у образа из другого класса, т.е. области изменения признака у разных классов пересекаются. В таком случае можно говорить только о системе распознавания с минимальной ошибкой.

Классификация признаков

Характеристика вероятностного признака

Любая случайная величина характеризуется законом распределения вероятностей – функцией распределения случайной величины $F_a^b(x)$, т.е. вероятностью нахождения случайной величины x в диапазоне $a - b$, или плотностью распределения вероятностей (ПРВ) $p(x)$, где **x непрерывно**.

$$F_a^b(x) = \int_a^b p(x) dx$$

Замечание. На практике непрерывный признак часто можно заменить на дискретный (многоградационный). В таком случае ПРВ превращается в дискретную функцию – вероятность появления каждого возможного значения признака.

ПРВ нормального или Гауссова закона распределения

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

где m – математическое ожидание или среднее значение случайной величины x ,
 σ – среднеквадратичное отклонение x .

Замечание. Нахождение ПРВ признака - достаточно сложная задача, особенно в части формирования репрезентативной выборки, гарантирующей требуемую достоверность.

Классификация признаков

Логические признаки – характеристики образа, представленные в бинарном виде (0 или 1), т.е. показывающие наличие или отсутствие свойства у данного образа (есть отверстие). К ним относятся и признаки, у которых важен только факт попадания или нет некоторой величины в заданный интервал (вес превышает норму).

Замечание. При наличии нескольких интервалов признак является дискретным, но при необходимости его можно преобразовать в совокупность логических признаков. Для этого каждый интервал дискретного признака нужно представить как отдельный логический признак, который получит значение 1 при попадании в него значения признака конкретного образа (недостаточный, нормальный, повышенный, избыточный вес). Недостаток - данный метод приводит к появлению **зависимости** между такими бинарными признаками, т.к. два и более из них не могут одновременно принимать значения 1, или все иметь значение 0.

Структурные признаки – примитивы (непроизводные или элементарные, т.е. не производимые из других элементарных признаков элементы) объекта распознавания.

Замечание. Для получения образа примитивы объединяются в цепочки (предложения). Отсюда структурные признаки носят еще название лингвистических или синтаксических признаков, а распознавание образов, описываемых структурными признаками, называют лингвистическим. Далее такие образы и методы распознавания не рассматриваются.

Пространство признаков

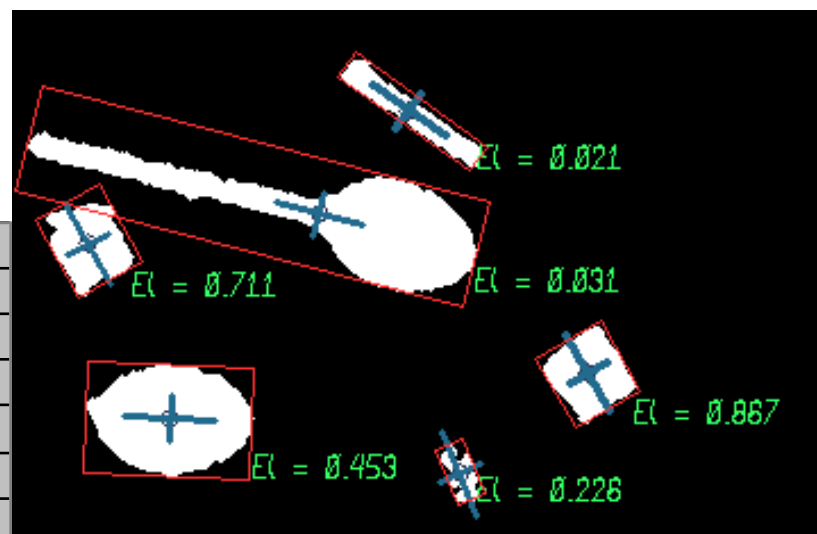
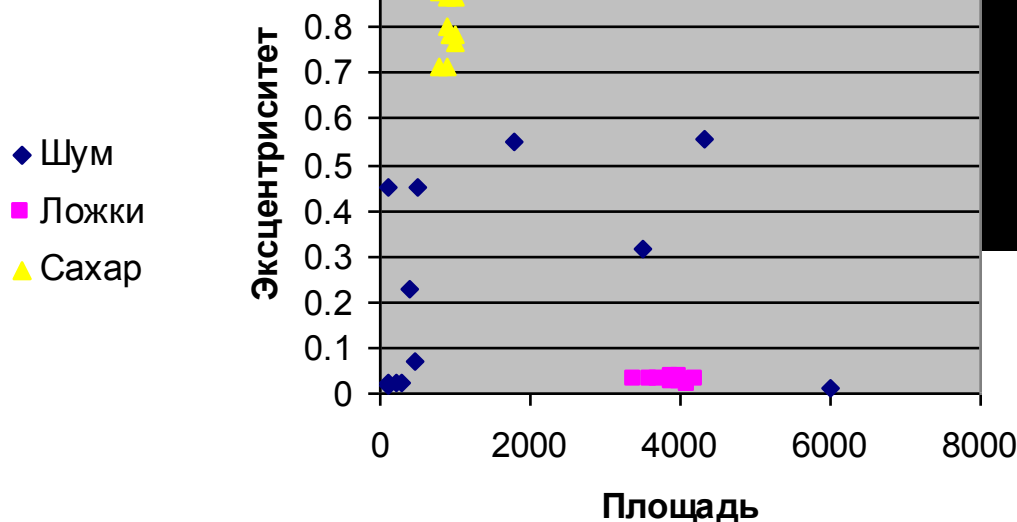
Числовые значения признаков можно интерпретировать как координаты точек - образов в многомерном пространстве признаков.

Образ можно представить в виде некоторого упорядоченного набора значений признаков или **вектора признаков** вида $\mathbf{x} = (x_1, \dots, x_n)$

где x_i - значение i -го признака данного образа, n - число признаков



Пример. Надо различить ложки и сахар



Замечание. Класс занимает область в пространстве признаков. Области разных классов могут пересекаться.

Меры близости точек

Мера близости **метрическая**, если вычисленное с ее помощью **расстояние** d_{lp} между точками l и p в пространстве признаков обеспечивает выполнение аксиом:

- симметричность ($d_{lp} = d_{pl}$)
- положительность ($d_{lp} \geq 0$, причем $d_{lp} = 0$ только если $l = p$)
- правило треугольника ($d_{lh} + d_{hp} \geq d_{lp}$)

Замечание. Если одна из аксиом не выполняется – мера близости неметрическая.

Евклидово расстояние

$$d_{lp} = \sqrt{\sum_{i=1}^n (x_{il} - x_{ip})^2}$$

где x_{il} , x_{ip} – i -ые координаты точек l и p соответственно.

В векторной форме (без извлечения квадратного корня)

$$d_{lp} = (\mathbf{x}_l - \mathbf{x}_p) \times (\mathbf{x}_l - \mathbf{x}_p)^T$$

Меры близости точек

Манхеттенское расстояние

$$d_{lp} = \sum_{i=1}^n |x_{il} - x_{ip}|$$

Расстояние доминирования

$$d_{lp} = \max_{i=1,n} (|x_{il} - x_{ip}|)$$

Расстояние Минковского

$$d_{lp} = \sum_{i=1}^n |x_{il} - x_{ip}|^\lambda$$

где λ – целое положительное число

Замечания. 1) Евклидово и Манхэттенское – частные случаи расстояния Минковского.

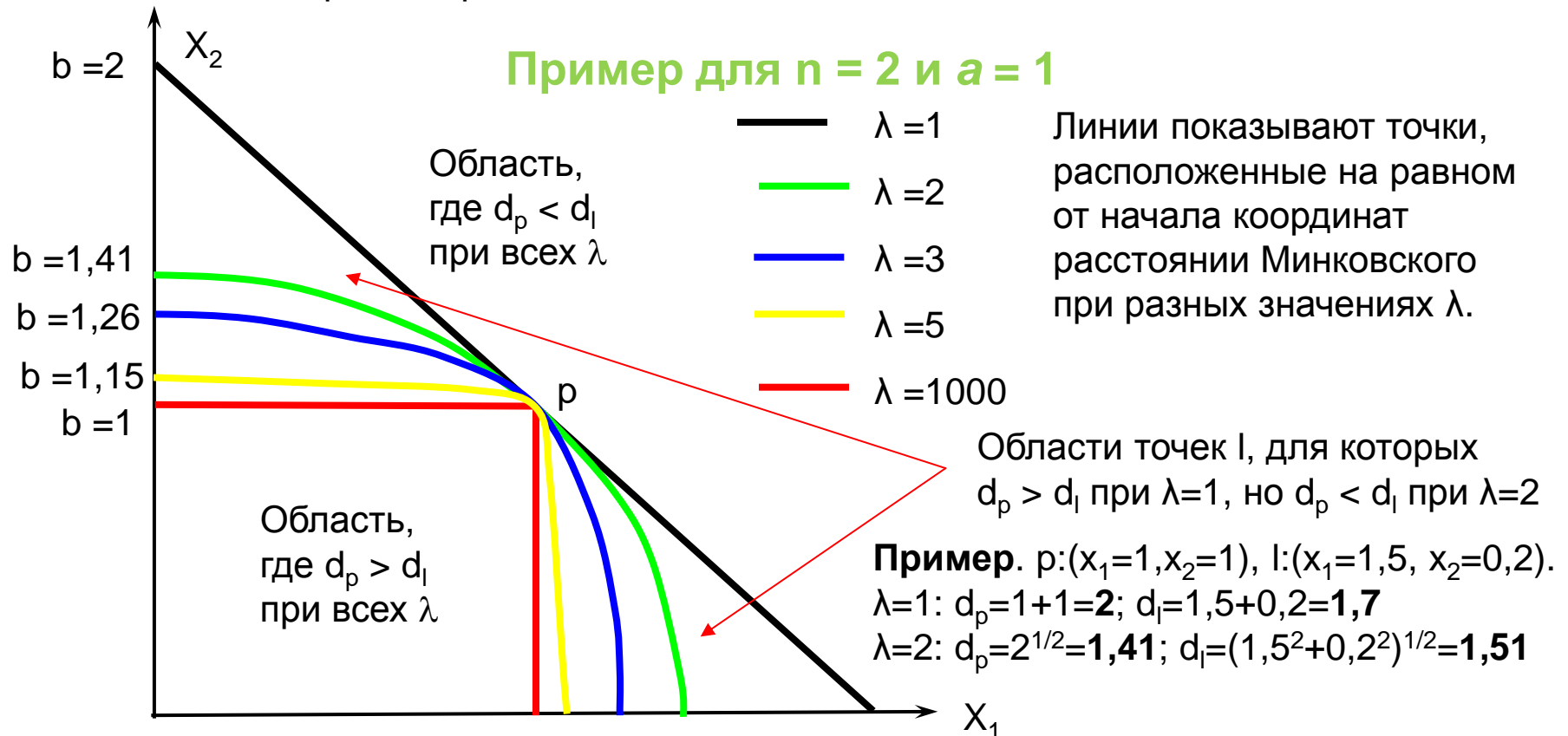
2) При $\lambda \rightarrow \infty$ расстояние Минковского – расстояние доминирования.

3) В n -мерном пространстве расстояние d_{hp} между точками h и p может быть как больше расстояния d_{hl} между точками h и l так и меньше, в зависимости от координат этих точек и выбранного для вычисления расстояния значения λ . Это обстоятельство иногда может повлиять на результаты распознавания

Расстояние Минковского

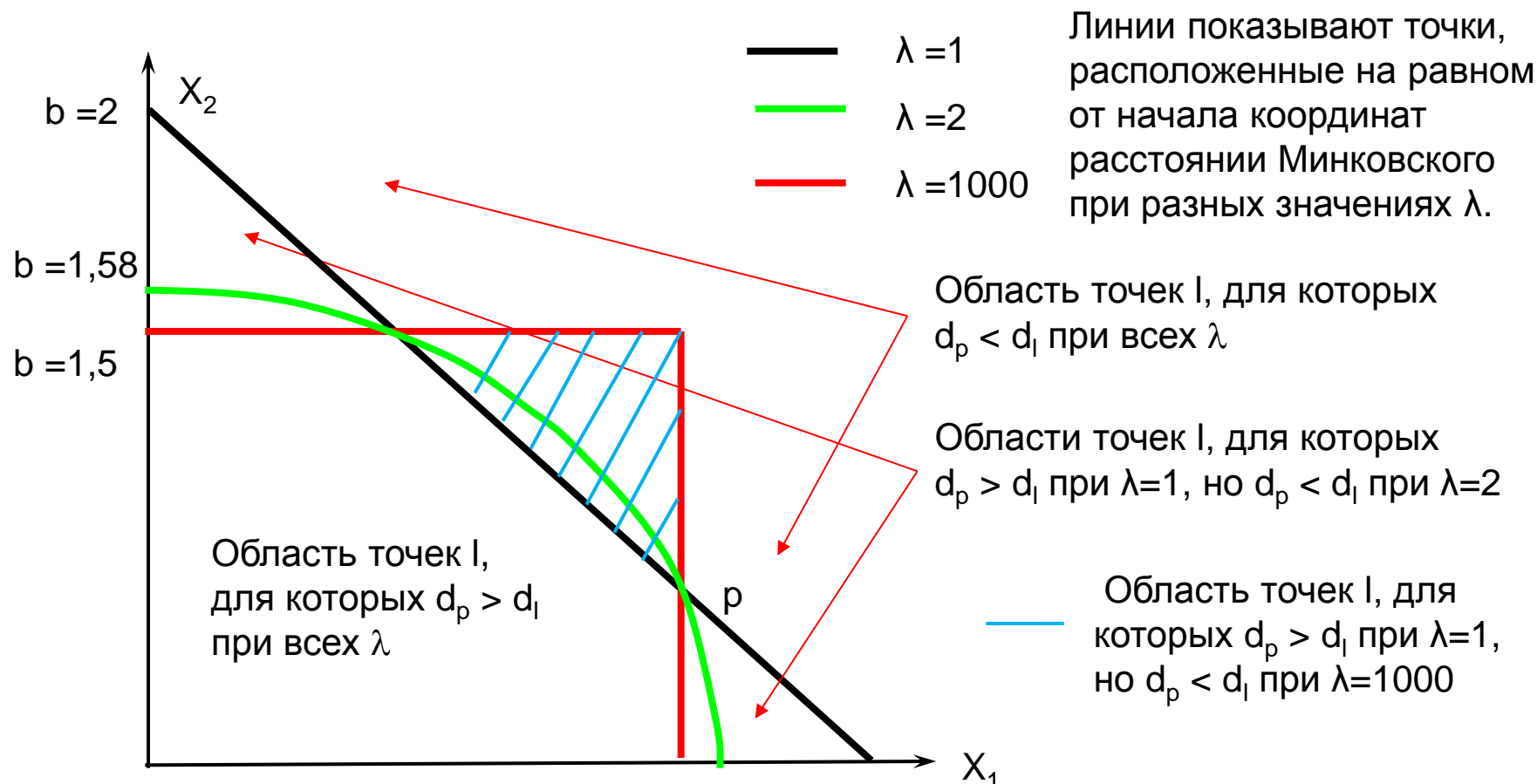
Пусть в n -мерном пространстве заданы: точка p с координатами $(x_1 = x_2 = \dots x_n = a)$ и точка l с координатами $(x_1 = b, x_2 = \dots x_n = 0)$. Расстояние Минковского до этих точек от начала координат: $d_p = a(n)^{1/\lambda}$, $d_l = b$. Точки p и l находятся на одинаковом от начала координат расстоянии Минковского если $a(n)^{1/\lambda} = b$. При фиксированном λ чем больше n , тем больше b отличается от a . При фиксированном n чем больше λ , тем меньше b отличается от a , причем при $\lambda \rightarrow \infty$ $b \rightarrow a$.

Пример для $n = 2$ и $a = 1$



Расстояние Минковского

Пример для $n = 2$ и точки p с координатами: $x_1 = 1,5$; $x_2 = 0,5$



$$\lambda = 1: d_p = 1,5 + 0,5 = 2; \lambda = 2: d_p = (1,5^2 + 0,5^2)^{1/2} = 1,58.$$

Для $n = 3$ поверхность с точками равноудаленными от начала координат при $\lambda = 1$ это основание трехгранной пирамиды с вершиной в начале координат и ребрами длиной $d = x_1 + x_2 + x_3$ – координаты точки, относительно которой определяется расстояние

Нормирование признаков

Если значения разных признаков по всему множеству образов существенно отличаются, то признаки следует нормировать. В противном случае одинаковые отклонения будут оказывать одинаковое влияние на результат.

Пример. три образа описываются двумя признаками x_1, x_2 :

первый образ – $x_1 = 9, x_2 = 1000$, второй – $x_1 = 4, x_2 = 1010$, третий – $x_1 = 10, x_2 = 986$.

Манхэттенское расстояние: $d_{12} = |9 - 4| + |1000 - 1010| = 5 + 10 = 15$

$$d_{13} = |9 - 10| + |1000 - 986| = 1 + 14 = 15$$

Второй и третий образы находятся на одинаковом расстоянии от первого образа, но из сравнения значений признаков видно, что отклонение в признаке x_2 лежит в пределах погрешности его определения.

Нормирование признака x

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

где x' – нормированное значение признака, x_{\max}, x_{\min} – возможные макс и мин значения

Тогда для первого образа – $x_1' = \frac{9 - 4}{10 - 4} = 0,833, x_2' = \frac{1000 - 986}{1010 - 986} = 0,583$

для второго – $x_1' = 0, x_2' = 1$; третий образ – $x_1' = 1, x_2' = 0$.

Манхэттенское расстояние $d_{12} = |0,833 - 0| + |0,583 - 1| = 1,25$

$$d_{13} = |0,833 - 1| + |0,583 - 0| = 0,75$$

Второй образ находится дальше от первого, чем третий, что более реалистично.

Нормирование признаков

Расстояние Канберра

$$d_{lp} = \sum_{i=1}^n \frac{|x_{il} - x_{ip}|}{|x_{il}| + |x_{ip}|}$$

Учитывает различие диапазонов изменений признаков, без нахождения возможных экстремальных значений. **Замечание.** Если значения x_{il} и x_{ip} разных знаков, то $d_i = 1$.

Для предыдущего примера

(первый образ – $x_1 = 9$, $x_2 = 1000$, второй – $x_1 = 4$, $x_2 = 1010$, третий – $x_1 = 10$, $x_2 = 986$)

$$d_{12} = \frac{|9-4|}{9+4} + \frac{|1000-1010|}{1000+1010} = 0,39 \qquad d_{13} = \frac{|9-10|}{9+10} + \frac{|1000-986|}{1000+986} = 0,06$$

Соответствует отношению между образами по нормированному Манхэттенскому

Веса признаков позволяют учесть их важность. Например, для Евклидова расстояния

$$d_{lp} = \sqrt{\sum_{i=1}^n \eta_i (x_{il} - x_{ip})^2}$$

η_i – вес i -го признака.

Замечание. Метод часто используется для ранжирования однотипных образов.

Например, выбор наилучшего места работы, проживания (многофакторный анализ).

Меры близости точек

Косинусное расстояние

$$\alpha_{lp} = \arccos \left(\frac{\mathbf{x}_l \times \mathbf{x}_p^T}{(\mathbf{x}_l \times \mathbf{x}_l^T)^{1/2} (\mathbf{x}_p \times \mathbf{x}_p^T)^{1/2}} \right)$$

α_{lp} – угол между векторами \mathbf{x}_l и \mathbf{x}_p . Можно использовать и величину обратную $\cos \alpha_{lp}$.

Замечание. Метрика дает хорошие результаты при распознавании классов, образы которых вытянуты вдоль радиус-вектора в пространстве признаков.

Если у рассматриваемых образов **все признаки логические**, то можно использовать **расстояние Хемминга** или **Меру Танимото**

$$d_{lp} = \sum_{i=1}^n x_{il} \oplus x_{ip}$$

где \oplus – сложение по модулю два

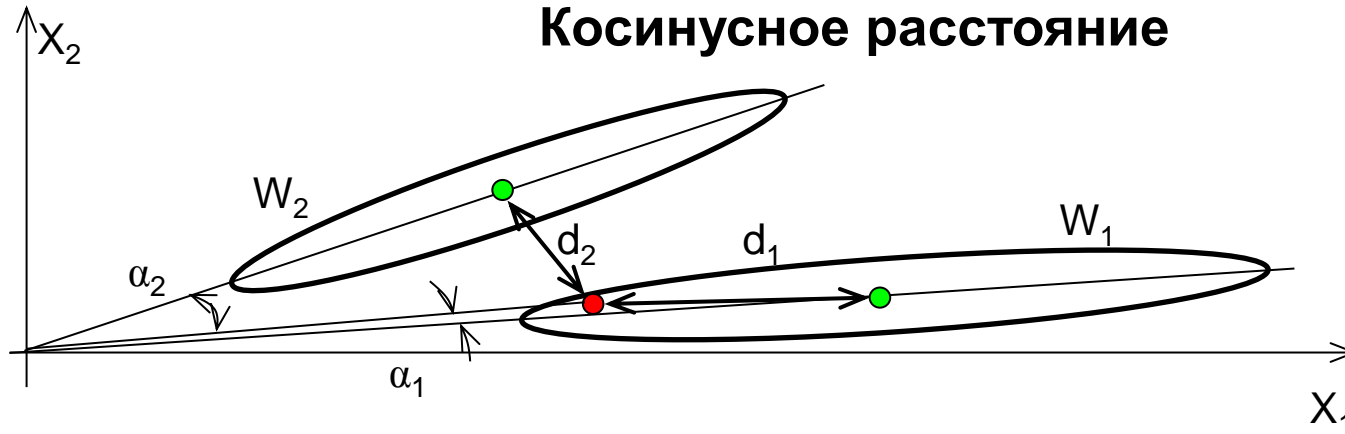
$$d_{lp} = \frac{\mathbf{x}_l \mathbf{x}_p^T}{\mathbf{x}_l \mathbf{x}_l^T + \mathbf{x}_p \mathbf{x}_p^T - \mathbf{x}_l \mathbf{x}_p^T}$$

$\mathbf{x}_l \mathbf{x}_p^T$ – векторное произведение

Расстояние Хемминга – число признаков (нулевых и единичных), значения которых у сравниваемых образов не совпадают.

Мера Танимото – отношение числа совпадающих и несовпадающих единичных признаков. Меняется от единицы (вектора \mathbf{x}_l и \mathbf{x}_p совпадают) до нуля (совпадающие **единичные** признаки в векторах \mathbf{x}_l и \mathbf{x}_p отсутствуют)

Меры близости точек. Пример



По Эвклидову
расстоянию

$$d_1 > d_2$$

По косинусному
расстоянию

$$\alpha_1 < \alpha_2$$

В логическом пространстве признаков заданы 7 точек. Расстояния от них до точки А.

Точки:	Хемминга:	мера Танимото	Соответствие
A <u>11111100</u>	0	$6 / (6 + 6 - 6) = 1$	совпадает
B 11110000	2	$4 / (6 + 4 - 4) = 2/3$	совпадает (точка
C 11000011	6	$2 / (6 + 4 - 2) = 1/4$	В ближе к А чем С)
D 11000000	4	$2 / (6 + 2 - 2) = 1/3$	<u>не совпадает</u> (точка
E 11110011	4	$4 / (6 + 6 - 4) = 1/2$	Е ближе к А чем D)
F 11100011	5	$3 / (6 + 5 - 3) = 3/8 = 0,375$	<u>не совпадает</u> (точка
G 00001101	5	$2 / (6 + 3 - 2) = 2/7 = 0,286$	Ф ближе чем G)

Расстояние Хемминга вычисляется быстрее. Мера Танимото полагает, что совпадение имеющихся признаков важнее, чем несовпадение отсутствующих.

Замечание. Точное число НЕсовпадающих единиц: $\mathbf{x}_i \mathbf{x}_i^T + \mathbf{x}_p \mathbf{x}_p^T - \underline{2\mathbf{x}_i \mathbf{x}_p^T}$ может давать 0.

Характеристики множества точек

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & \dots & x_{1m} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & \dots & x_{nm} \end{pmatrix}$$

Множество n -мерных векторов признаков, описывающих m образов

$$\mathbf{m} = (\mu_1, \mu_2, \dots, \mu_n) : \mu_i = \frac{1}{m} \sum_{k=1}^m x_{ik}$$

Вектор средних значений признаков (математических ожиданий)

$$\mathbf{Cov} = \begin{pmatrix} D_{11} & D_{12} & \dots & D_{1n} \\ D_{21} & D_{22} & \dots & D_{2n} \\ \dots & \dots & \dots & \dots \\ D_{n1} & D_{n2} & \dots & D_{nn} \end{pmatrix} : \begin{cases} D_{ii} = \frac{1}{m} \sum_{k=1}^m (x_{ik} - \mu_i)^2; \\ D_{ij} = \frac{1}{m} \sum_{k=1}^m (x_{ik} - \mu_i)(x_{jk} - \mu_j) \end{cases}$$

Ковариационная матрица

где x_{ik} – значение i -го признака k -го образа ($k=1, \dots, m$), μ_i – среднее i -ой компоненты вектора признаков, D_{ii} – дисперсия i -го признака, D_{ij} – ковариация i -го и j -го признаков.

Замечание. Ковариационная матрица симметрична относительно главной диагонали. Ковариация характеризует степень линейной зависимости случайных величин. Если ковариация равна нулю, то величины называются некоррелированными.

Расстояние между множеством точек и отдельной точкой

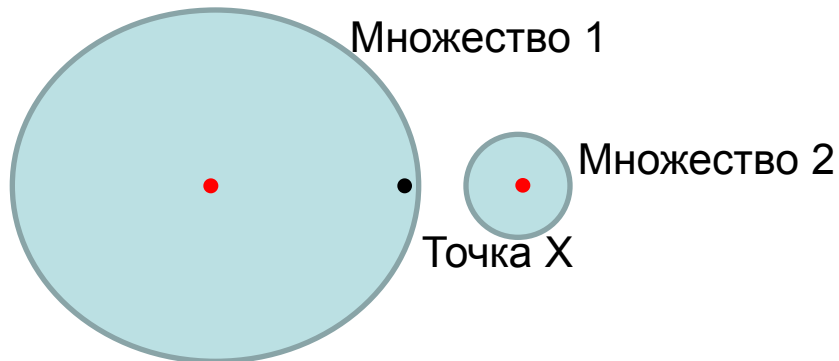
1. Найти расстояние между **центром** множества (вектор \mathbf{m}) и точкой

Наиболее просто, можно использовать рассмотренные ранее формулы. **Недостаток** - не учитывается степень компактности множества (определяется дисперсиями его признаков)

2. **Взвешенное** расстояние (Веса признаков – обратные значения их дисперсии) между центром множества и точкой

Пример. Взвешенное Евклидово расстояние
$$d = (\mathbf{x} - \mathbf{m}) \mathbf{D}^{-1} (\mathbf{x} - \mathbf{m})^T,$$

где \mathbf{x} – вектор признаков точки, \mathbf{m} – вектор средних значений признаков класса, \mathbf{D} – диагональная матрица (диагональные элементы – дисперсии признаков).



Если использовать расстояние между центрами, то точка X ближе ко второму множеству.

Если учесть дисперсии, то - к первому.

Расстояние между множеством точек и отдельной точкой

3. Расстояние **Махаланобиса**:

$$d = (\mathbf{x} - \mathbf{m}) \mathbf{Cov}^{-1} (\mathbf{x} - \mathbf{m})^T,$$

\mathbf{Cov}^{-1} – обратная ковариационная матрица

При $\mathbf{Cov}^{-1} = \mathbf{I}$ расстояние превращается в Евклидово

Замечание. Обратная ковариационная матрица существует только при выполнении условия $m > n$ (число образов больше числа признаков), т.к. $\mathbf{Cov} = \mathbf{x}_c \times \mathbf{x}_c^T$, где

$$\mathbf{x}_c = \begin{bmatrix} \mathbf{x}_1 - \mathbf{m} \\ \dots \\ \mathbf{x}_m - \mathbf{m} \end{bmatrix}$$

Расстояния 1-3 дают хорошие результаты для множества точек, имеющего эллипсоидную форму в пространстве признаков.

4. Расстояние от точки до **ближайшей точки множества**

Применяется для множества точек, имеющего более сложную форму.

Можно применить наиболее подходящий способ нахождения расстояния между точками.

Требует много времени при большом числе точек в множестве.

Расстояние между двумя множествами точек

1. Найти расстояние между **центрами** множеств W_1 и W_2 (вектора \mathbf{m}_1 и \mathbf{m}_2)

Дает хорошие результаты для компактных множеств, имеющих эллипсоидную форму

Для множеств точек, имеющих более сложную форму:

2. **Расстояние ближнего соседа** – расстояние между ближайшими точками, принадлежащими разным множествам

$$d(w_1, w_2) = \min(d_{lp}), \text{ где } (l = 1, m_1; p = 1, m_2)$$

3. **Расстояние дальнего соседа** $d(w_1, w_2) = \max(d_{lp})$

4. **К-расстояние или расстояние по Колмогорову**

$$d(w_1, w_2) = \left(\frac{1}{m_1 m_2} \sum_{l=1}^{m_1} \sum_{p=1}^{m_2} d_{lp}^\lambda \right)^{1/\lambda}$$

λ – целое число в диапазоне $\pm \infty$. При $\lambda \rightarrow \infty$ К-расстояние вырождается в расстояние дальнего соседа, а при $\lambda \rightarrow -\infty$ – в расстояние ближнего соседа

Кластеризация

Задача разделения множества образов на группы (подмножества) с близкими значениями признаков в образах каждой группы называется **кластеризацией**. Полученные группы называются **кластерами**.

Замечание. Различие между классом и кластером: кластер – результат разбиения на подмножества множества образов с заранее неизвестной классификацией, в классе собраны образы, принадлежность которых к данному классу задана изначально.

Цели кластеризации

- Понимание данных при помощи выявления кластерной структуры. Разбиение выборки на группы похожих объектов дает возможность упростить обработку данных в дальнейшем и принятие решений, к каждому кластеру применяя собственный метод анализа (стратегия «разделяй и властвуй»).
- Сжатие данных. Когда исходное множество данных о каких-то объектах большая, то можно её сократить, оставив от каждого кластера по одному самому типичному представителю.
- Обнаружение новизны. Выделяют нетипичные объекты, которые не получается присоединить ни к одному из кластеров

Источник: <https://biznes-prost.ru/analiz-klasternyj.html>

Кластеризация

Области применения кластеризации

- Биология. Поиск новых видов, классификации, теорий происхождения
- Социология. Поиск эффективных рабочих групп
- Информатика. Поиск групп похожих данных (документов, изображений)

А также - психология, медицина, химия, маркетинг, логистика, управление процессами.

Пороговый алгоритм кластеризации

В пространстве признаков задано множество образов $M = \{x_1, \dots, x_k, \dots, x_m\}$, где x_k – вектор признаков k -го образа, m – мощность множества. За центр первого кластера w_1 принимаем любой из образов, например x_1 , т.е. $w_1 = \{x_1\}$. Далее вычисляется расстояние d_{21} между образом x_2 и центром кластера w_1 . Если значение расстояния больше заранее заданной **пороговой величины** t , то образ x_2 принимается за центр нового кластера w_2 . В противном случае образ x_2 включается в кластер w_1 . Далее для каждого следующего образа из M оцениваются расстояния от него до уже имеющихся кластеров. Если **все** расстояния **больше** порога, то образ относится к новому кластеру. Если часть расстояний **меньше** порога, то образ относится к **ближайшему** кластеру. Процедура продолжается пока не будут исчерпаны все образы из множества M .

Замечание. Способы определения расстояния между отдельными образами и кластером и предъявляемым образом выбираются разработчиком. Образы предъявляются один раз. Результат зависит от последовательности предъявления

Кластеризация

Пример пороговой кластеризации



Расстояние между образами – Евклидово (можно и другое)

Расстояние между множеством и образом – до ближайшей точки множества (или другое)

Порог отнесения к новому кластеру – t_1

Порядок предъявления образов 1:

1	2	3	4		5	6	7	8
---	---	---	---	--	---	---	---	---

Результат пороговой кластеризации 1:

I	I	I	I		II	II	II	II
---	---	---	---	--	----	----	----	----

Порядок предъявления образов 2:

1	4	3	2		7	6	5	8
---	---	---	---	--	---	---	---	---

Результат пороговой кластеризации 2:

I	II	II	II		III	III	III	III
---	----	----	----	--	-----	-----	-----	-----

Порядок предъявления образов 3:

1	2	4	3		5	8	6	7
---	---	---	---	--	---	---	---	---

Результат пороговой кластеризации 3:

I	I	I	II		III	?	IV	IV
---	---	---	----	--	-----	---	----	----

Замечание. Изменив порог на t_2 можно получить «правильный» результат.

Кластеризация

Алгоритм цепной кластеризации

Расстояние между множеством и образом – до ближайшей точки множества.

Расстояние между образами – на усмотрение разработчика.

В начале кластеризации некоторый образ считается принадлежащим к первому кластеру. К данному кластеру присоединяются **все** образы, принадлежность которых к какому-либо кластеру еще не установлена, и расстояние от которых до исходного образа меньше заранее заданного порога t . Затем для каждого из присоединенных образов данная процедура повторяется. После того как к первому кластеру больше нельзя отнести ни одного образа, в качестве исходного образа для второго кластера из оставшихся в **M** образов берется произвольный образ. Процедура повторяется до тех пор, пока не будут исчерпаны все образы из множества **M**.

Замечание. Образы предъявляются многократно. Результат не зависит от предъявления.

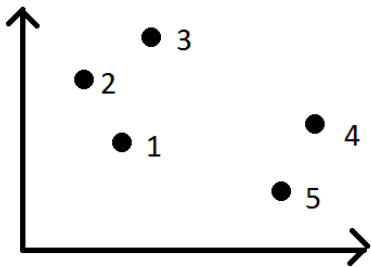
Алгоритм слияния

В начале каждый образ считается отдельным кластером, далее вычисляются расстояния между **всеми** кластерами. На каждом шаге кластеризации сливаются два самых близких кластера, после чего вычисляются **новые** расстояния между изменившимися по составу кластерами. Процесс прекращается если достигнуто заранее заданное число кластеров или расстояние между ближайшими кластерами больше заранее заданного порога.

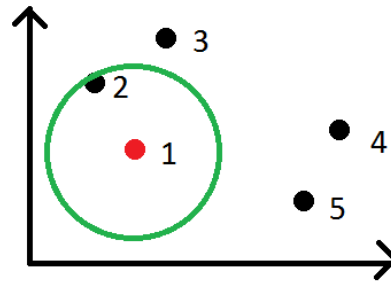
Замечание. Требуется задание способа определения расстояния между множествами. На каждом шаге требуется пересчитывать характеристики изменившегося кластера.

Пример кластеризации

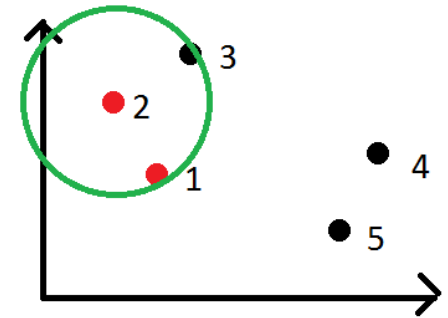
Цепная кластеризация



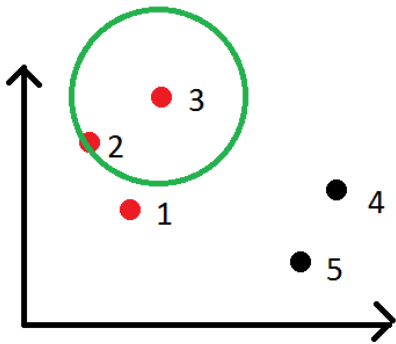
Исходные данные.
Пять образов



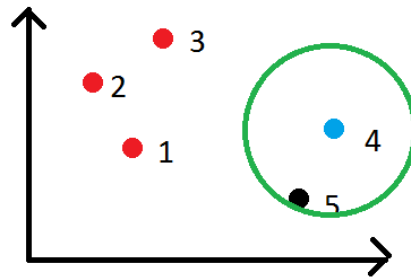
Первый кластер - Образ 1.
Круг - пороговое расстояние.
Образ 2 попадает в кластер 1



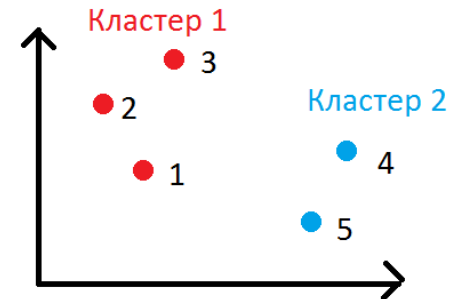
Ищем образы около образа 2.
Образ 3 близок к образу 2.
Присоединяется к кластеру 1



Образов, близких к № 3
нет. Кластер №1 закончен



Второй кластер - Образ 4.
Образ 5 попадает в кластер 2



Результат цепной
кластеризации

Кластеризация

Алгоритм кластеризации по k средним

Задано число кластеров – k . В начале произвольно выбирается положение k центров кластеров, не обязательно совпадающих с какими-либо образами. Далее на каждом шаге каждый образ относится к тому кластеру, расстояние до центра которого для него минимально, а после распределения **всех** образов по кластерам – перерасчет положения центров кластеров. Процесс продолжается пока не стабилизируется состав кластеров.

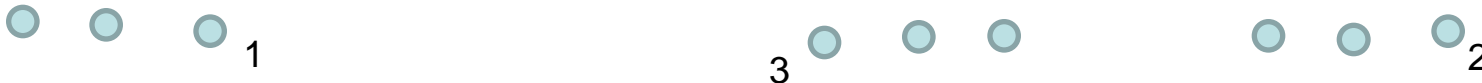
Замечание. Цель – минимизировать суммарное расстояние от центров кластеров до отнесенных к ним образов по всем кластерам. Возможно схождение процесса к локальному минимуму, а также отсутствие образов в некоторых кластерах. Изменяя k и сравнивая результаты, можно найти подходящее число кластеров.

Результат кластеризации зависит от числа кластеров и от выбора **первоначального** расположения центров кластеров.

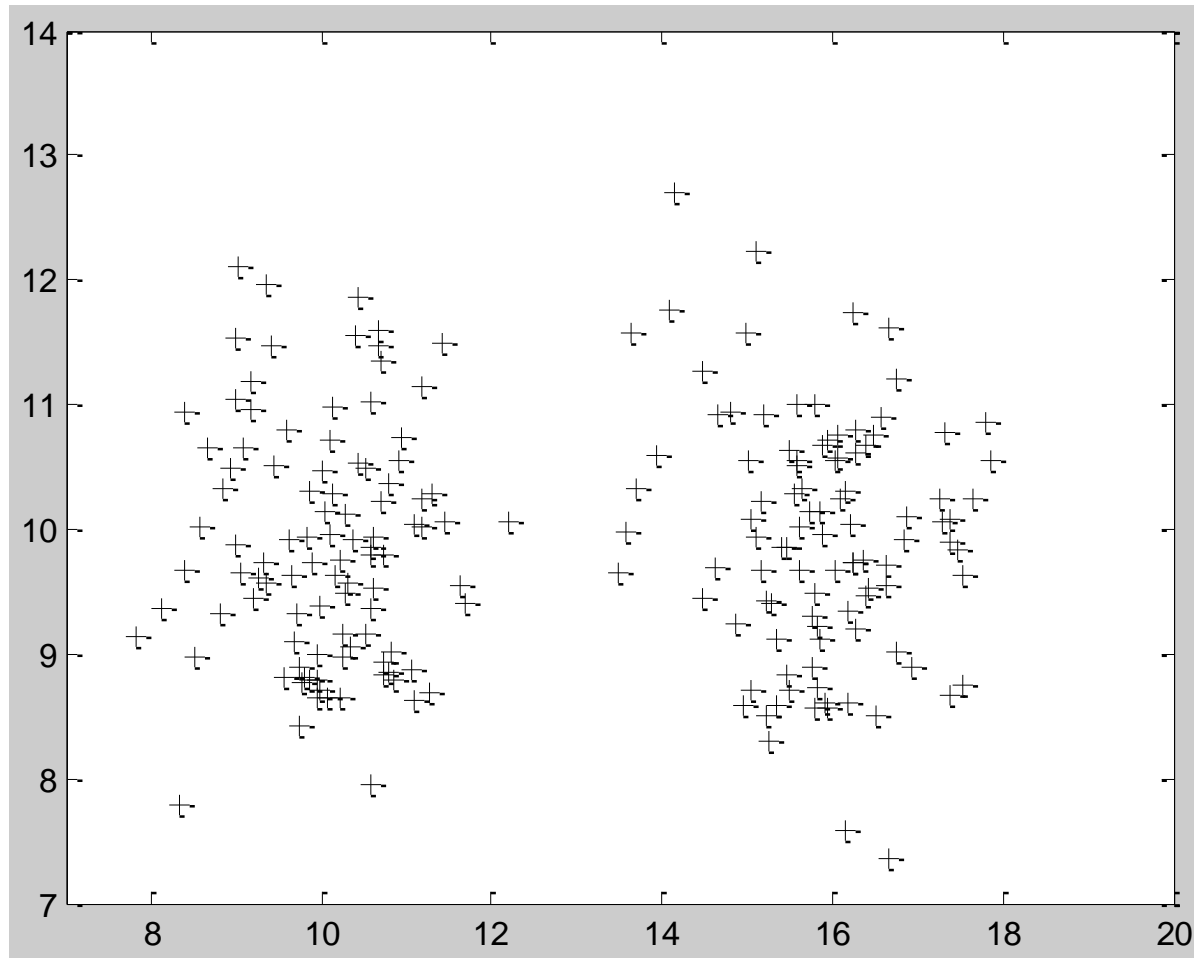
Способ выбора начальных центров кластеров

В качестве первого центра c_1 выбирается произвольный образ из **M**. В качестве второго центра c_2 выбирается образ, который находится на **наибольшем** расстоянии от c_1 . В качестве следующего центра c_j выбирается образ который находится на **наибольшем**

расстоянии от **ближайшего** к нему **уже выбранного центра**, т.е.
$$d_{pj} = \max_{l=j-1, m}^{c=1, j-1} (\min(d_{lc}))$$

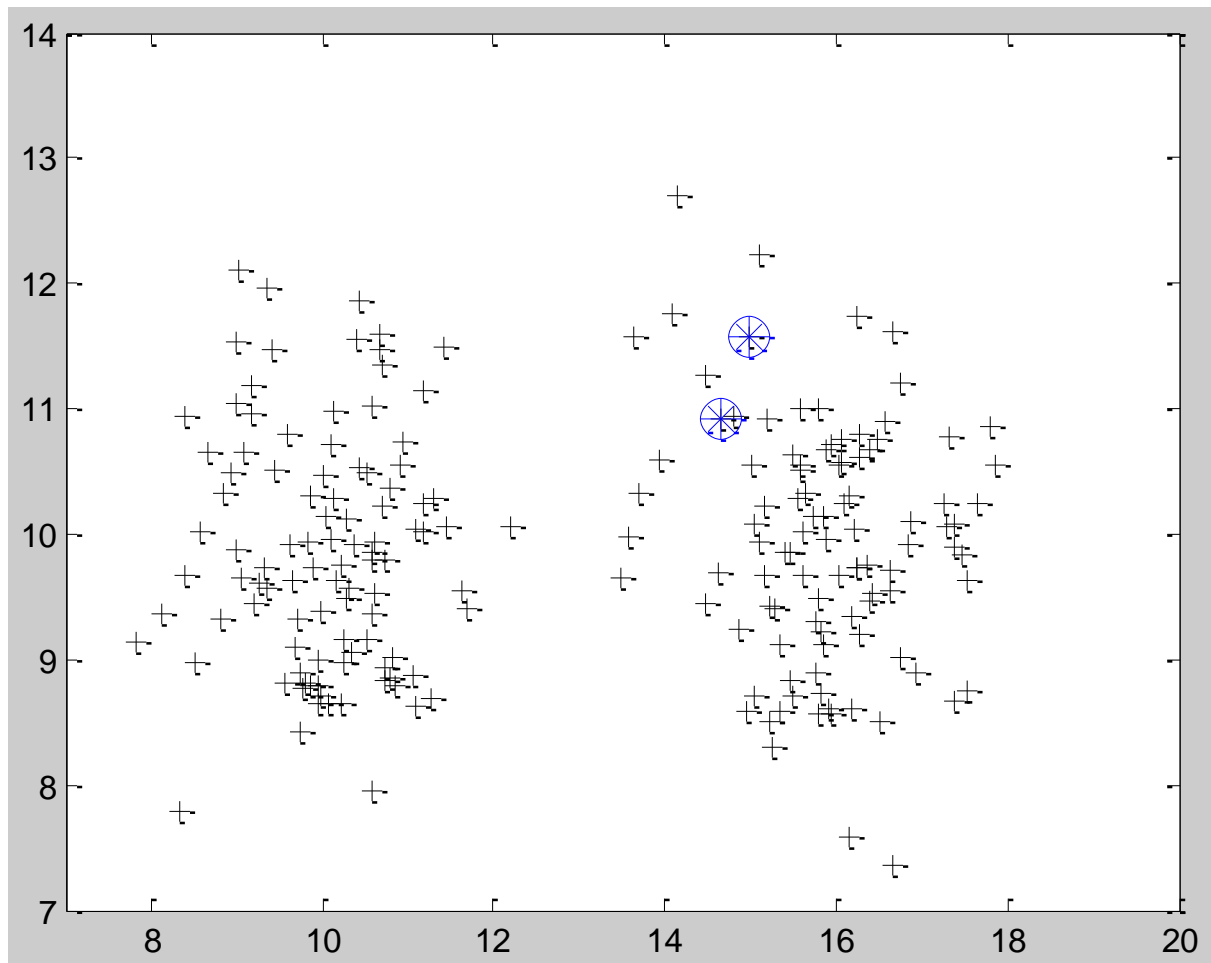


Пример кластеризации



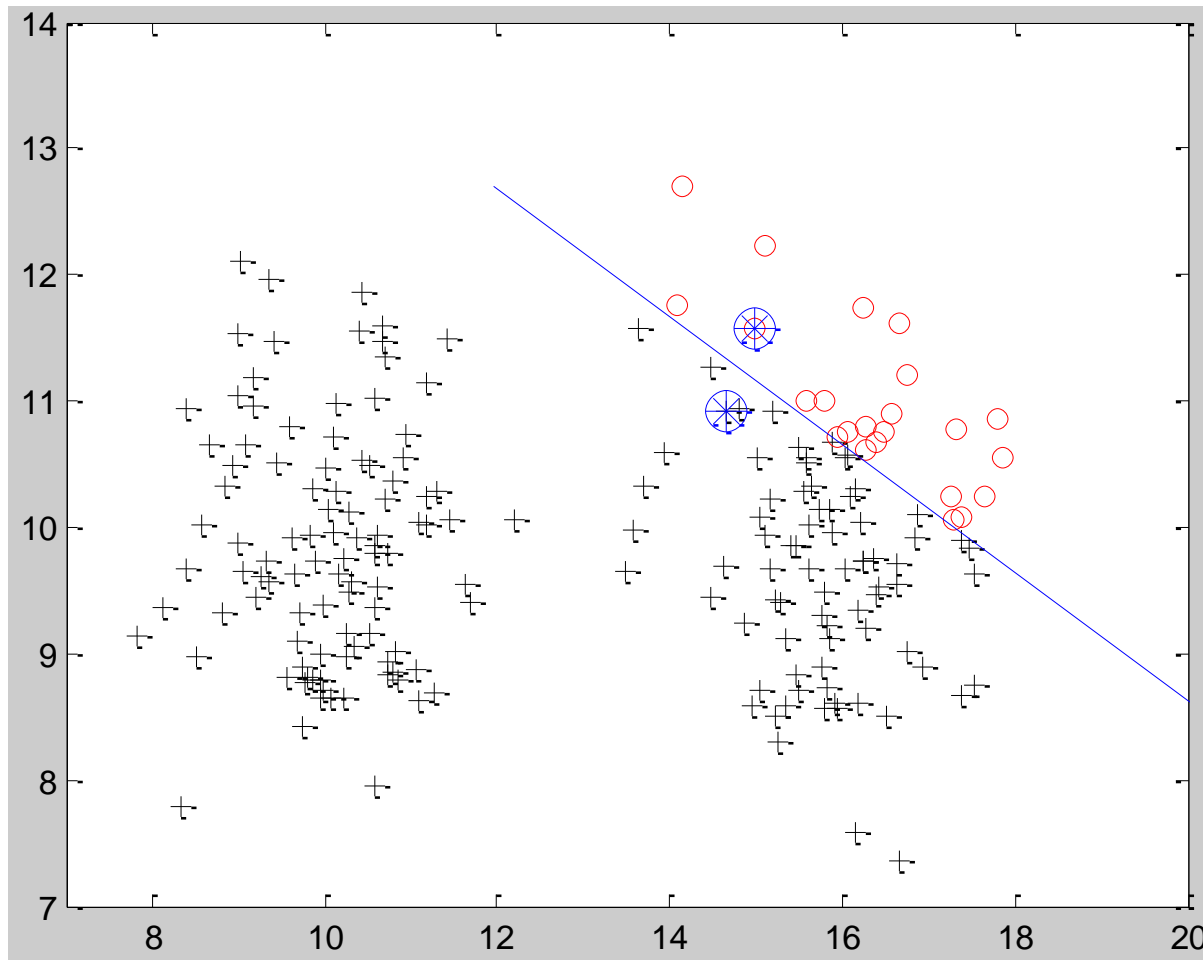
Исходные данные

Пример кластеризации



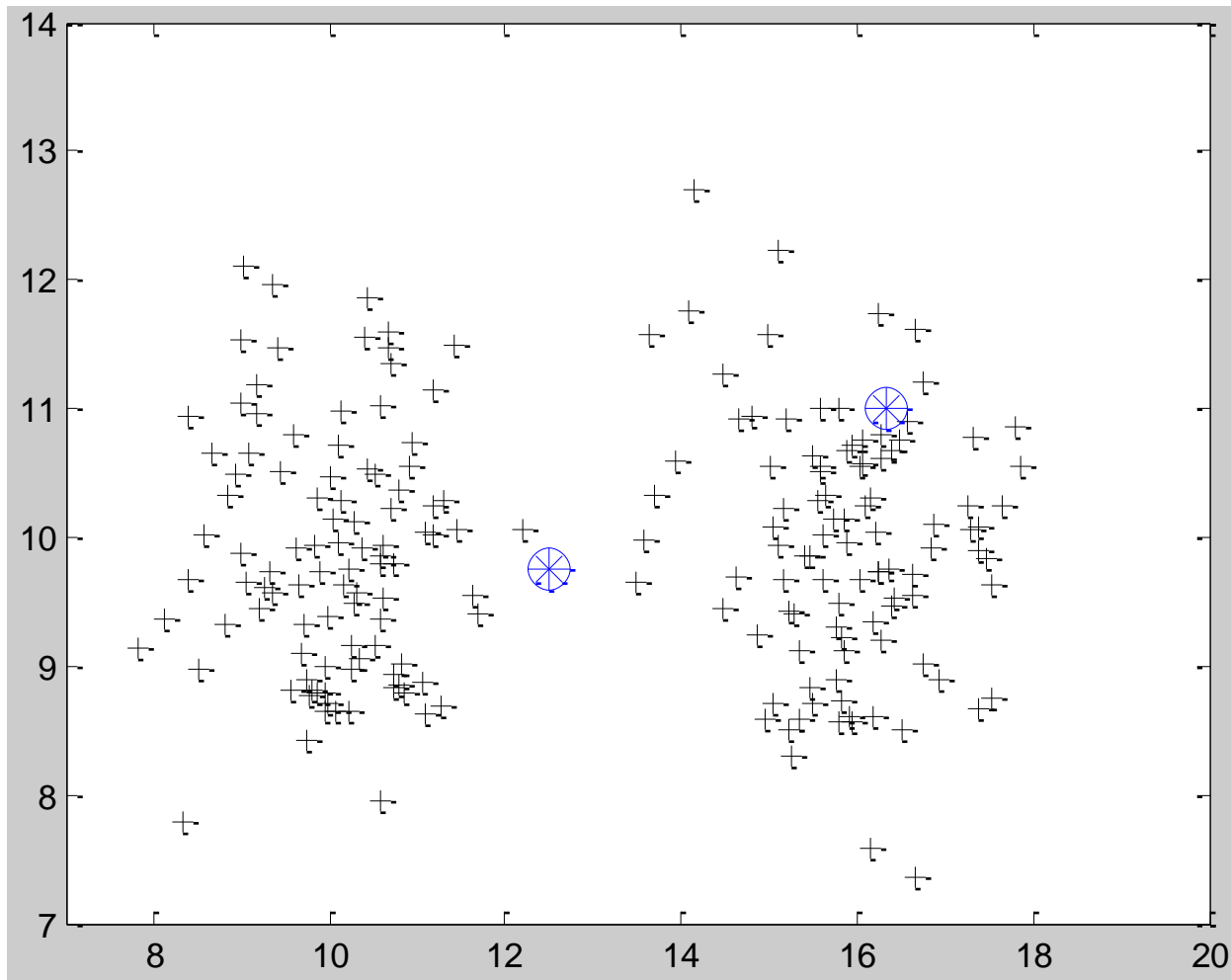
Случайная инициализация центров кластеров (шаг 1)

Пример кластеризации



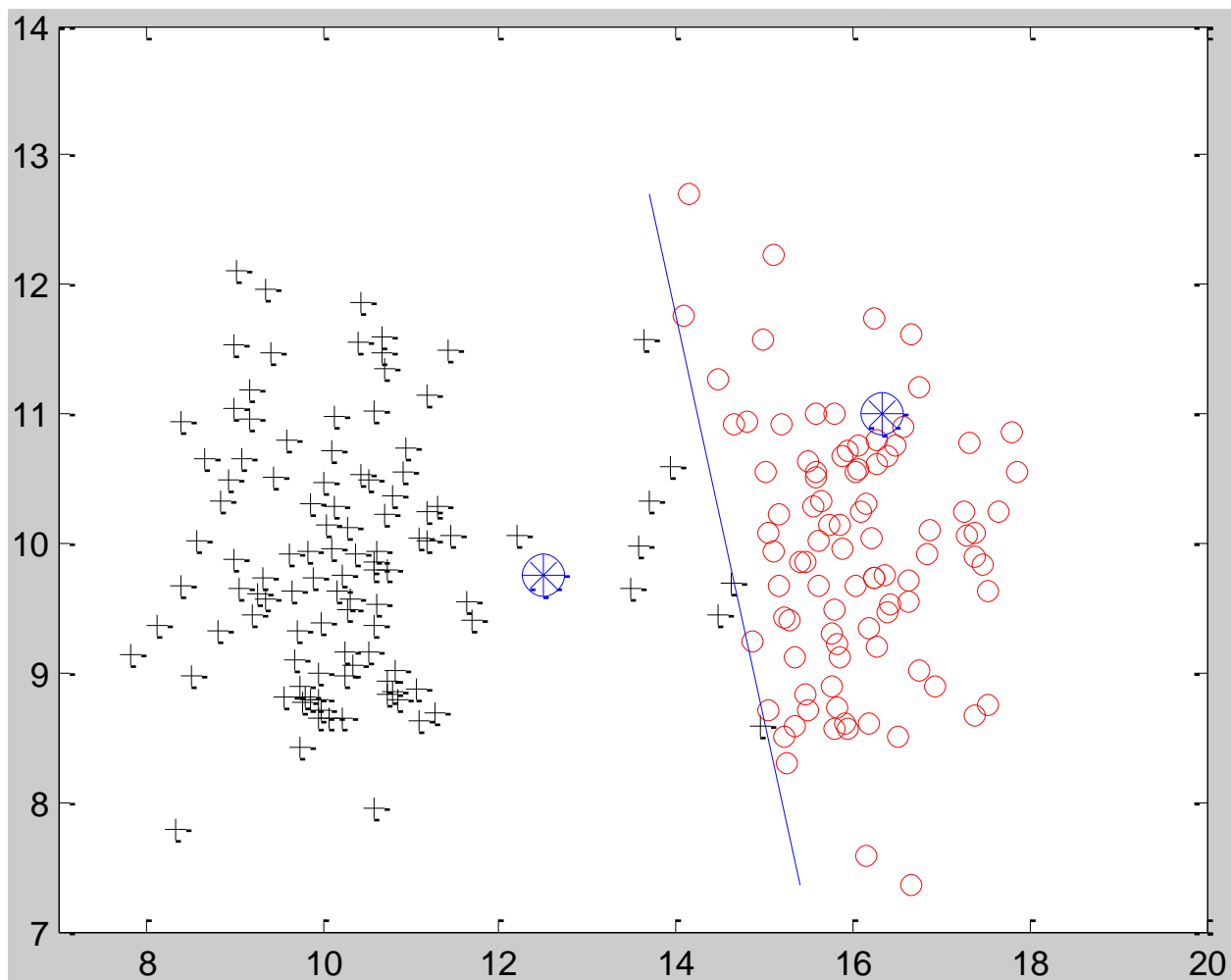
Кластеры после первой итерации (шаг 2)

Пример кластеризации



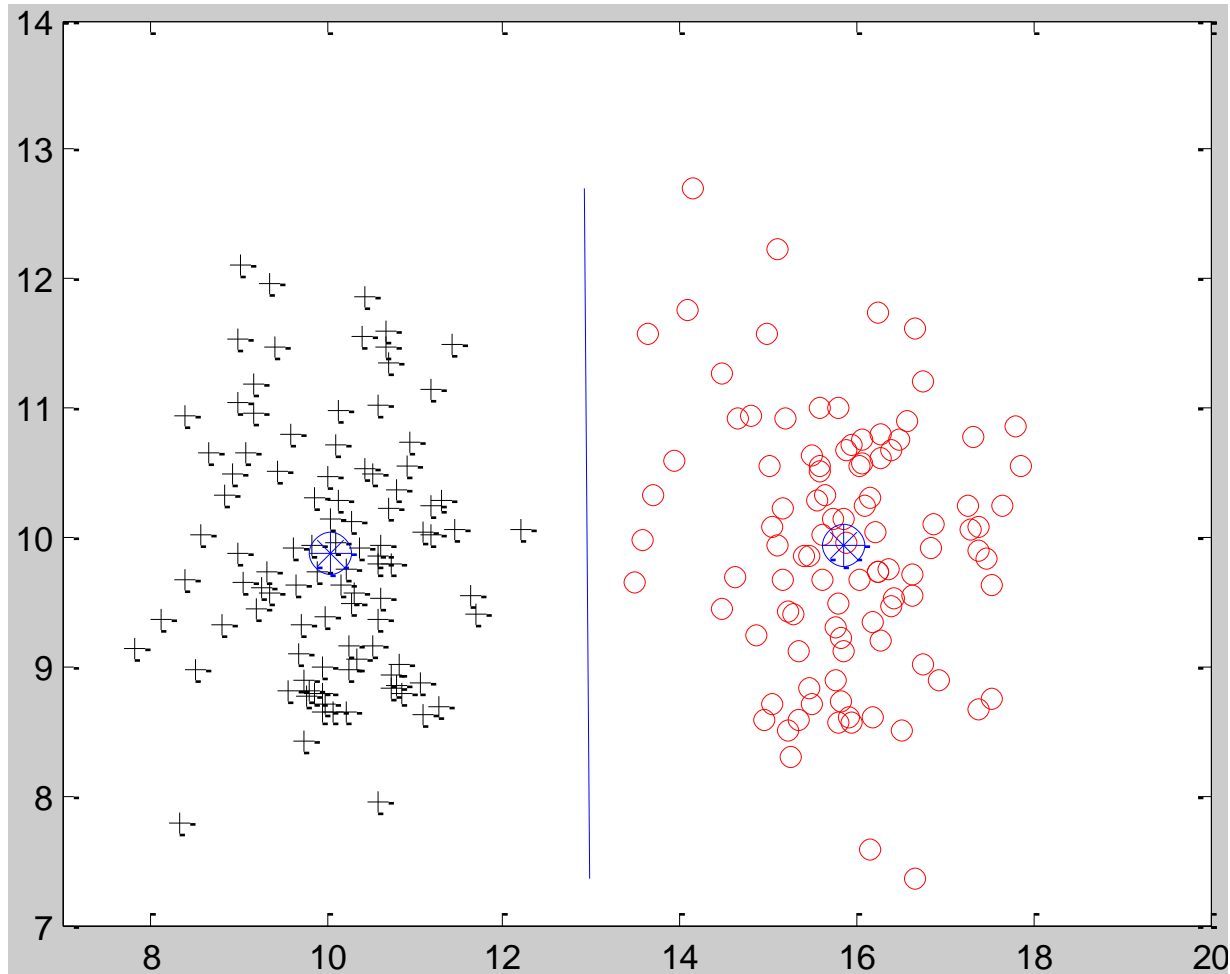
Пересчет центров кластеров после первой итерации (шаг 3)

Пример кластеризации



Кластеры после второй итерации (шаг 2)

Пример кластеризации



Стабильная конфигурация после четвертой итерации

Кластеризация

Алгоритм ISODATA

Iterative Self-Organizing Data Analysis Techniques – Интерактивный самоорганизующийся метод анализа данных (Ball G., Hall D. 1965). Разновидность кластеризации по k средним. Дополнен возможностью изменения числа кластеров (удаление, слияние, разделение) на каждом шаге. Требуется задания для каждой операции пороговых значений. Возможно задание дополнительных параметров, например, число итераций.

Удаление кластера.

Если число образов в кластере меньше порогового, то этот кластер удаляется (k уменьшается на единицу), а его образы распределяются между другими кластерами.

Слияние кластеров.

Если расстояние между центрами двух кластеров меньше порогового, то эти кластеры сливаются в один (k уменьшается на единицу), вычисляется новый центр кластера.

Разделение кластеров.

Если дисперсия образов в кластере больше порога, то этот кластер разделяется на два (k увеличивается на единицу). Для этого в пространстве признаков находится направление, в котором дисперсия образов кластера максимальная и образы кластера разделяются перпендикулярной к этому направлению гиперплоскостью, проходящей через центр кластера. Далее вычисляются новые центры кластеров.

Основные недостатки : – требует «эвристического» задания большего числа параметров
– более медленный, по сравнению с базовым алгоритмом