Основы математических методов распознавания образов

Введение

Введение. Общая схема взаимосвязей различных областей знаний

Распознавание образов относится к одной из важнейших задач искусственного интеллекта. Она связана со множеством различных областей исследований.

Введение. Основные термины

Распознавание образа — это отнесение объекта или события к одному или нескольким предопределенным категориям.

Образ — это объект, процесс или событие, которому можно присвоить имя.

Класс образов (или **категория**) — это набор образов, имеющих общие свойства, обычно произошедших от одного источника.

Способность восприятия внешнего мира в форме образов позволяет с определенной достоверностью узнавать бесконечное число объектов на основании ознакомления с конечным их числом, а объективный характер основного свойства образов позволяет моделировать процесс их распознавания.

Пример

- Собака узнает хозяина или другую собаку

Пример

- По рзезульаттам илссеовадний одонго анлигйсокго унвиертисета, не иеемт занчнеия, в кокам пряокде рсапожолены бкувы в солве. Галвоне, чотбы преавя и пслоендяя бквуы блыи на мсете. Осатьлыне бкувы мгоут селдовтаь в плоонм бсепордяке, все-рвано ткест чтаитсея без побрелм. Пичрионй эгото ялвятеся то, что мы не чиатем кдаужю бкуву по отдльенотси, а все солво цликеом.

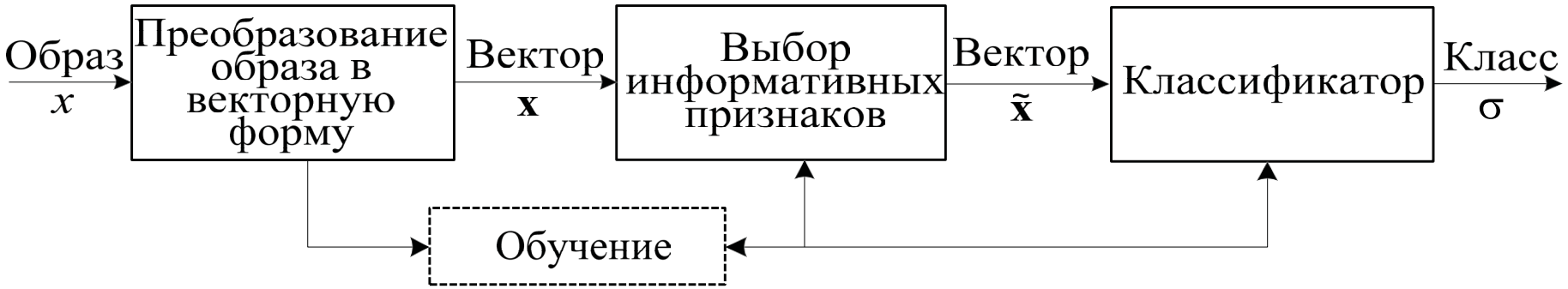
История

- Нейрофизиология и психология конец 19 века, начало 20-го века (Павлов - собака)
- Р.Фишер – дискриминантный анализ – 1936 г. (направление наибольшей различимости)
- Колмогоров А.Н. – Разделение смеси двух распределений 1936-1940
- Кибернетика – Н.Виннер - 1948г.
- Кластерный анализ –начало 20-го века
- Нейронные сети 50-е
- Многомерное шкалирование 70-е

Основные цели разработки систем распознавания

- Освобождение человека от однообразных рутинных операций для решения других более важных задач.
- Повышение качества выполняемых работ.
- Повышение скорости решения задач.

Общая схема системы распознавания образов



Классификация систем распознавания

- Однородность:
  - -простые;
  - -сложные
- Способ получения апостериорной информации
  - -одноуровневые;
  - -многоуровневые.
- Количество первоначальной априорной информации
  - Без обучения
  - С обучением
  - Самообучаемые
- Характер информации о признаках распознавания
  - детерминированные;
  - вероятностные;
  - Логические;
  - структурные (лингвистические);
  - комбинированные.

Образ (описание) не объект!

- Описание не полностью представляет объект
- Описание зависит от задач
- Описание содержит погрешности представления

- Измерения, используемые для классификации образов, называются **признаками.**

- Любой образ представляется некоторым **набором признаков**

- Основное назначение описаний (образов) - это их использование в процессе установления соответствия объектов

Образ (описание) не объект!

- Совокупность признаков, относящихся к одному образу, называется **вектором признаков.**
- Вектора признаков принимают значения в **пространстве признаков**

Класс

- **классы -** это объединения объектов (явлений), отличающиеся общими свойствами, интересующими человека.

- **цель распознавания –** принятие решения об отнесении объекта к тому или иному классу.

- **классификатором или решающим правилом** называется правило отнесения образа к одному из классов на основании его вектора признаков.

Основная идея

- Разделение образов основывается на прецедентах.
- Прецедент – это образ, правильное отнесение к категории которого известно.
- Прецедент – объект, принимаемый как образец при решении задач образов разделения по категориям.

- Идея принятия решений на основе прецедентности – основополагающая в естественно-научном мировоззрении.

Классификация методов распознавания

Д.А.Поспелов (1990) выделяет два основных способа представления знаний :

1. Интенсиональное представление - в виде схемы связей между атрибутами (признаками).
2. Экстенсиональное представление - с помощью конкретных фактов (объекты, примеры).

Классификация методов распознавания

Описанные выше два фундаментальных способа представления знаний позволяют предложить следующую классификацию методов распознавания образов:

Интенсиональные методы распознавания образов - методы, основанные на операциях с признаками.

Экстенсиональные методы распознавания образов - методы, основанные на операциях с объектами.

Классификация методов распознавания (2)

Классификация методов распознавания (3)

Классификация методов распознавания			Область применения	Ограничения (недостатки)
Методы распознавания	Интенсиальные методы	Методы, основанные на оценках плотностей распределения значений признаков (или сходства и различия объектов)	Задачи с известным распределением, как правило нормальным, необходимость набора большой статистики.	Отсутствие обобщения. Необходимость перебора всей обучающей выборки при распознавании, высокая чувствительность к непредставительности обучающей выборки и артефактам.
		Методы, основанные на предположениях о классе решающих функций	Классы должны быть хорошо разделяемыми, система признаков - ортонормированной	Отсутствие обобщения. Должен быть заранее известен вид решающей функции. Невозможность учета новых знаний о корреляциях между признаками.
		Логические методы	Задачи небольшой размерности пространства признаков.	Отсутствие обобщения. При отборе логических решающих правил (коньюнкций) необходим полный перебор. Высокая вычислительная трудоемкость.
		Лингвистические (структурные) методы	Задачи небольшой размерности пространства признаков.	Отсутствие обобщения. Задача восстановления (определения) грамматики по некоторому множеству высказываний (описаний объектов), является трудно формализуемой. Нерешенность теореических проблем.

Классификация методов распознавания (4)

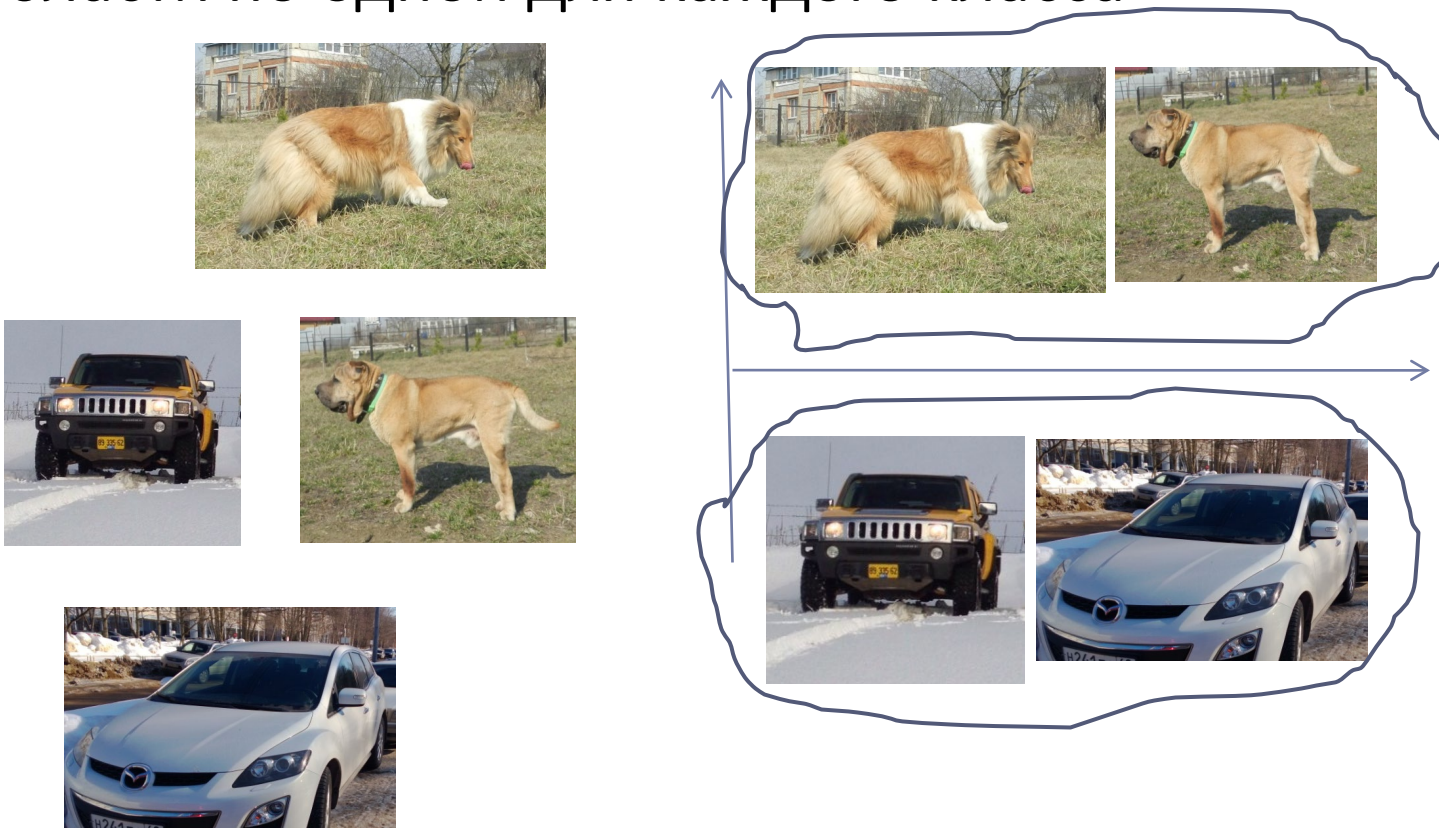
Классификация методов распознавания			Область применения	Ограничения (недостатки)
Методы распознавания	Экстенсиальные методы	Метод сравнения с прототипом	Задачи небольшой размерности пространства признаков.	Отсутствие обобщения. Высокая зависимость результатов классификации от меры расстояния (метрики).
		Метод k-ближайших соседей	Задачи небольшой размерности по количеству классов и признаков.	Отсутствие обобщения. Высокая зависимость результатов классификации от меры расстояния (метрики). Необходимость полного перебора обучающей выборки при распознавании. Вычислительная трудоемкость.
		Алгоритмы вычисления оценок (голосования) АВО	Задачи небольшой размерности по количеству классов и признаков.	Отсутствие обобщения. Зависимость результатов классификации от меры расстояния (метрики). Необходимость полного перебора обучающей выборки при распознавании. Высокая техническая сложность метода.
		Коллективы решающих правил	Задачи небольшой размерности по количеству классов и признаков.	Отсутствие обобщения. Очень высокая техническая сложность метода, нерешенность ряда теоретических проблем, как при определении областей компетенции частных методов, так и в самих частных методах.

Задача классификации (что делает?)

- Разделить объекты на 2 группы и сказать к какой из них относиться новый объект:

Задача классификации (по существу)

- Разбиение пространства признаков на области по одной для каждого класса

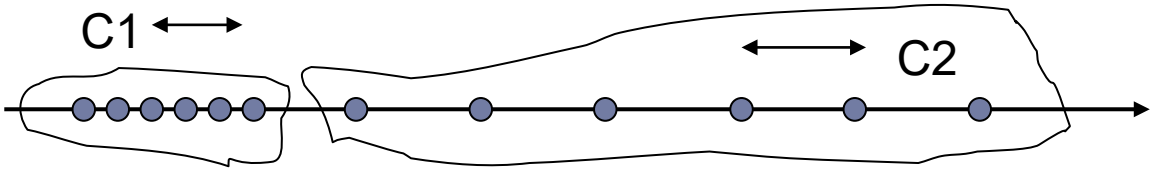


- Дуда Р., Харт П. Распознавание образов и анализ сцен. 1976

Гипотеза компактности

- *Классическая. Реализация одного и того же образа, обычно, отображается признаком пространства геометрически близкими точками.*
- Гипотеза λ-компактности

Расстояние мало, но есть неоднородность.



появление гипотезы компактности, которая гласит: образам соответствуют компактные множества в пространстве признаков. Под компактным множеством пока будем понимать некие "сгустки" точек в пространстве изображений, предполагая, что между этими сгустками существуют разделяющие их разряжения.

Описание классов по примерам и по признакам (эталоны)

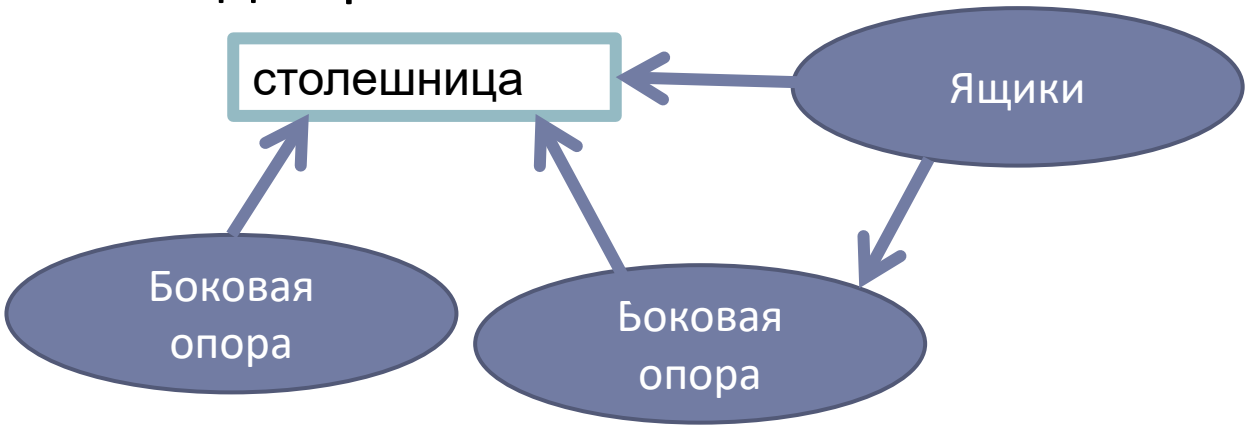
- Столы для работы

признак	Длина, м	Ширина, м	Число ящиков
Стол 1	1	0.6	3
Стол 2	1.5	0.7	5
Стол 3	3	0.7	4

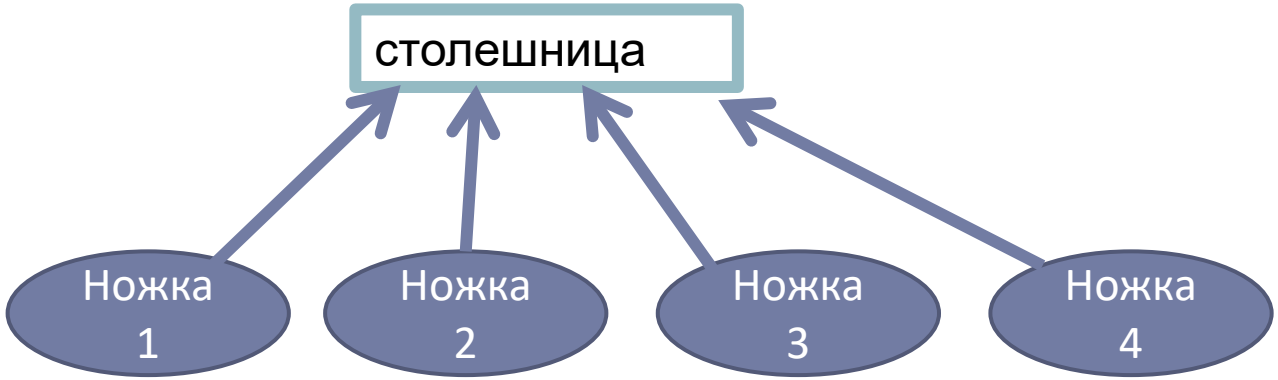
признак	Длина, м	Ширина, м	Число ящиков
Стол 1	1. 6	1.2	1
Стол 2	1.5	0.8	0
Стол 3	3	1.25	0

Описание классов структурами

- Столы для работы



- Столы для обеда

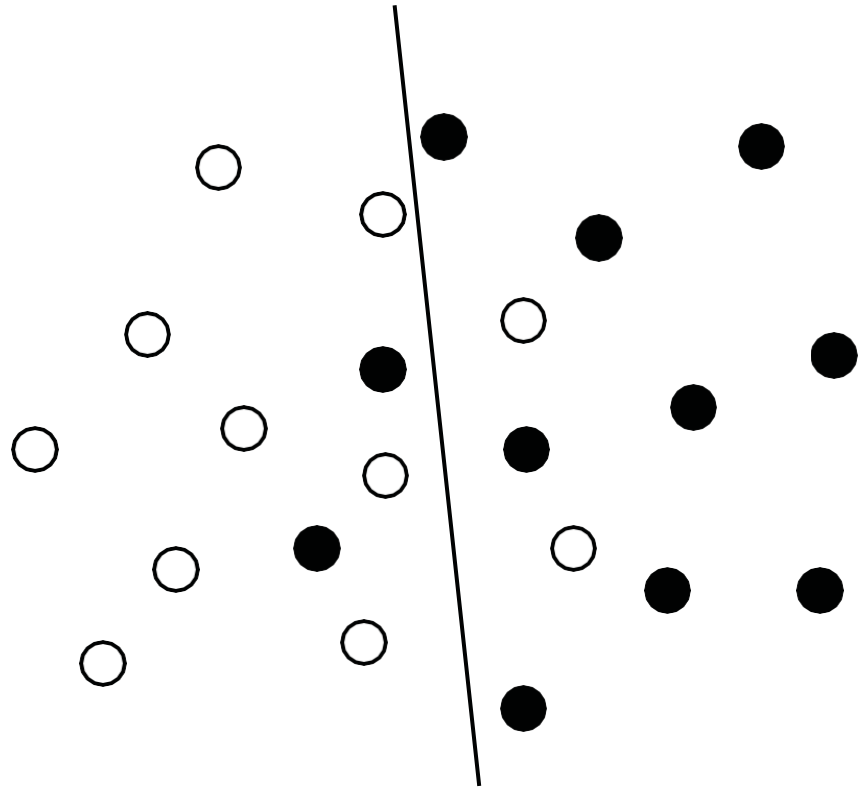


Логическое описание образа

- Обеденный стол содержит несколько
- (не менее 1) ножки и немного ящиков
- (не более 2), его столешница имеет отношение ширины к длине не более 1/2

Обобщенная постановка задачи линейного разделения классов

Найти такую ЛРФ, чтобы **ошибка** неправильной классификации **была минимальной**.



Критерии ошибки неправильной классификации:

1) число неправильно классифицируемых векторов

$$F(\mathbf{w}) = E(\mathbf{w}) = \sum_{\mathbf{x}' \in \Xi'} \eta(-(\mathbf{w}, \mathbf{x}')),$$

где Ξ' – множество унифицированных векторов обучающей выборки;

2) квадрат среднеквадратичной ошибки

$$F(\mathbf{w}) = \sum_{\mathbf{x}' \in E(\mathbf{w})} (\mathbf{w}, \mathbf{x}')^2 = \sum_{\mathbf{x}' \in \Xi'} \eta(-(\mathbf{w}, \mathbf{x}'))(\mathbf{w}, \mathbf{x}')^2,$$

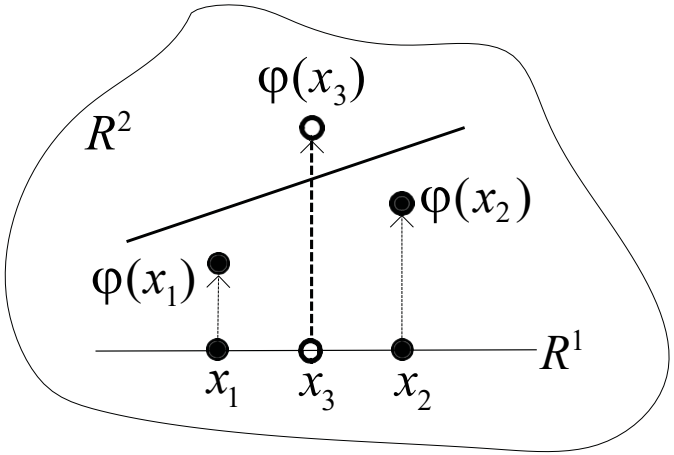
если $\mathbf{w} = (w_1, ..., w_n, w_{n+1})^T$ такой, что $w_1^2 + ... + w_n^2 = 1.$

Обобщенные решающие функции (ОРФ)

Пусть классы $\Xi_1 = \{\mathbf{x}_i : i \in I_1\}$, $\Xi_2 = \{\mathbf{x}_i : i \in I_2\}$ линейно **неразделимы** в R^n.

Задача: найти такое отображение $\varphi : R^n \to R^l$, $l > n$, чтобы классы $\tilde{\Xi}_1 = \{\varphi(\mathbf{x}_i) : i \in I_1\}$ и $\tilde{\Xi}_2 = \{\varphi(\mathbf{x}_i) : i \in I_2\}$ были линейно **разделимы** в R^l.

Пример. $\Xi_1 = \{x_1, x_2\}$, $\Xi_2 = \{x_3\}$ – линейно неразделимы в R^1, а $\tilde{\Xi}_1 = \{\varphi(x_1), \varphi(x_2)\}$, $\tilde{\Xi}_2 = \{\varphi(x_3)\}$ – линейно разделимы в R^2.



Обобщенные решающие функции (ОРФ) (2)

Важный класс ОРФ – *мономиальные* функции вида

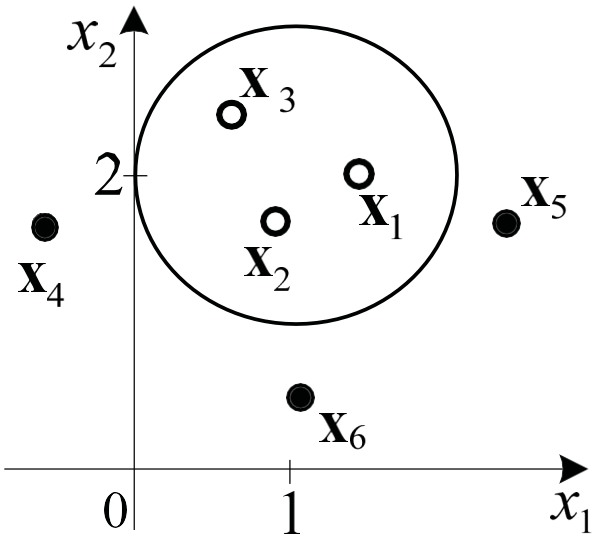
$$f_{s_1 \ldots s_n}(x_1, \ldots, x_n) = x_1^{s_1} \cdot \ldots \cdot x_n^{s_n}.$$

Пример. Классы $\Xi_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$, $\Xi_2 = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$ линейно неразделимы, но разделимы с помощью ОРФ

$$d(\mathbf{x}) = (x_1 - 1)^2 + (x_2 - 2)^2 - 1 = x_1^2 + x_2^2 - 2x_1 - 4x_2 + 4$$

или

$$d(\mathbf{x}^*) = 1 \cdot x_1^* + 1 \cdot x_2^* + 0 \cdot x_3^* - 2 \cdot x_4^* - 4 \cdot x_5^* + 4 \cdot 1.$$

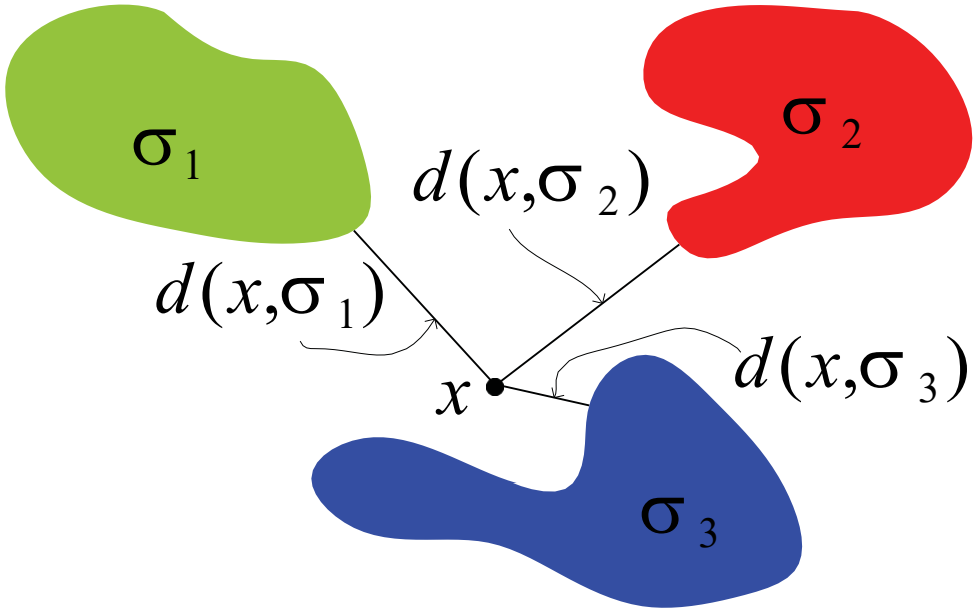


Классификация с помощью функций расстояния

- постановка задачи классификации с помощью функции расстояния;
- стандартизация признаков;
- способы измерения расстояний между векторами признаков;
- способы определения расстояния между вектором-образом и классом;
- способы определения расстояний между классами.

Постановка задачи

Найти такую функцию $d(x,\sigma_i)$, что $x \in \sigma_i$, если $d(x,\sigma_i) \le d(x,\sigma_j)$ для всех $j \ne i$



$d(x,\sigma_i)$ – функция расстояния.

Меры близости и аксиомы метрики

Мера близости между:

- двумя образами $d(x, y)$,
- образом и классом $d(x, \sigma)$,
- двумя классами $d(\sigma_i, \sigma_j)$.

Аксиомы метрики (метрического пространства):

1) $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (симметричность);
2) $d(\mathbf{x}, \mathbf{x}) = 0$;
3) $d(\mathbf{x}, \mathbf{y}) = 0 \Leftrightarrow \mathbf{x} = \mathbf{y}$ (определенность);
4) $d(\mathbf{x}, \mathbf{y}) \le d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$ (неравенство треугольника).

Способы стандартизации признаков

Дана выборка $\Xi = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ векторов $\mathbf{x}_i = (x_{i1}, ..., x_{in})^T$, x_{ik} – признаки (гены).

Задача: привести все признаки к единому масштабу.

Способы стандартизации признаков:

1) $x_{ik} \to \dfrac{x_{ik} - \tilde{m}_i}{\tilde{\sigma}_i}$, где $\tilde{m}_i = \frac{1}{N}\sum_{k=1}^{N} x_{ik}$ – среднее выборочное значение i-й координаты, $\tilde{\sigma}_i = \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_{ik} - \tilde{m}_i)^2}$ – выборочное среднеквадратичное отклонение;

2) $x_{ik} \to \dfrac{x_{ik} - \min\limits_k x_{ik}}{\max\limits_k x_{ik} - \min\limits_k x_{ik}}$.

Способы измерения расстояний между векторами признаков

а) метрика Евклида

$$d_2(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_2 = \sqrt{(x_1 - y_1)^2 + \ldots + (x_n - y_n)^2};$$

б) манхаттановскую метрика

$$d_1(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_1 = |x_1 - y_1| + \ldots + |x_n - y_n|$$

(если $x_i \in \{-1,1\}$, то d_1 – метрика Хэмминга);

в) равномерная метрика

$$d_\infty(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_\infty = \max_{1 \le i \le n} |x_i - y_i|;$$

г) метрика Минковского

$$d_p(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_p = \sqrt[p]{|x_1 - y_1|^p + \ldots + |x_n - y_n|^p} \quad (p \ge 1);$$

Способы измерения расстояний между векторами признаков (2)

д) метрика Махаланобиса

$$d_{S^{-1}}(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|_{S^{-1}} = \sqrt{(\mathbf{x}-\mathbf{y})^T S^{-1}(\mathbf{x}-\mathbf{y})} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i - y_i)s_{ij}^{-1}(x_j - y_j)},$$

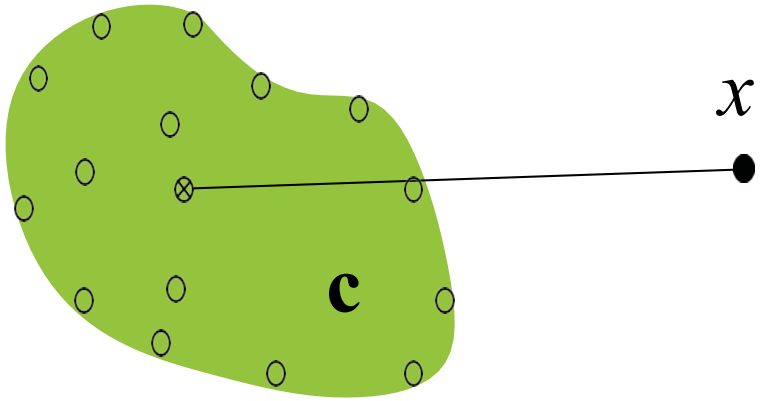
где $S = (s_{ij})$ – ковариационная матрица векторов обучающей выборки;

г′) $d_p^{\boldsymbol{\eta}}(\mathbf{x},\mathbf{y}) = \sqrt[p]{\eta_1 |x_1 - y_1|^p + \ldots + \eta_n |x_n - y_n|^p}$, $\boldsymbol{\eta} = (\eta_i)$, $\eta_i > 0$;

е) метрика Канберра $d_k(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$, если $\mathbf{x} \neq \mathbf{0}$ или $\mathbf{y} \neq \mathbf{0}$.

Способы определения расстояния между вектором-образом и классом

1. Определение расстояния до центра класса



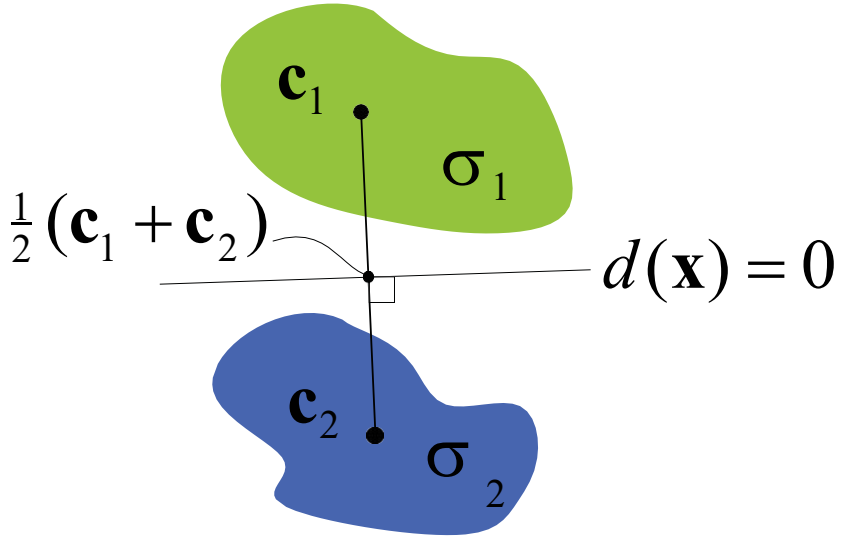
$$d(x,\sigma) = d(\mathbf{x},\mathbf{c})$$

Правило классификации:

$$x \in \sigma_i, \text{ если } d(\mathbf{x},\mathbf{c}_i) < d(\mathbf{x},\mathbf{c}_j)\ \forall j \neq i,\ \mathbf{c}_i = \frac{1}{|\sigma_i|}\sum_{\mathbf{x}\in\sigma_i}\mathbf{x}$$

Способы определения расстояния между вектором-образом и классом (2)

а) классификация по двум классам (евклидова метрика):



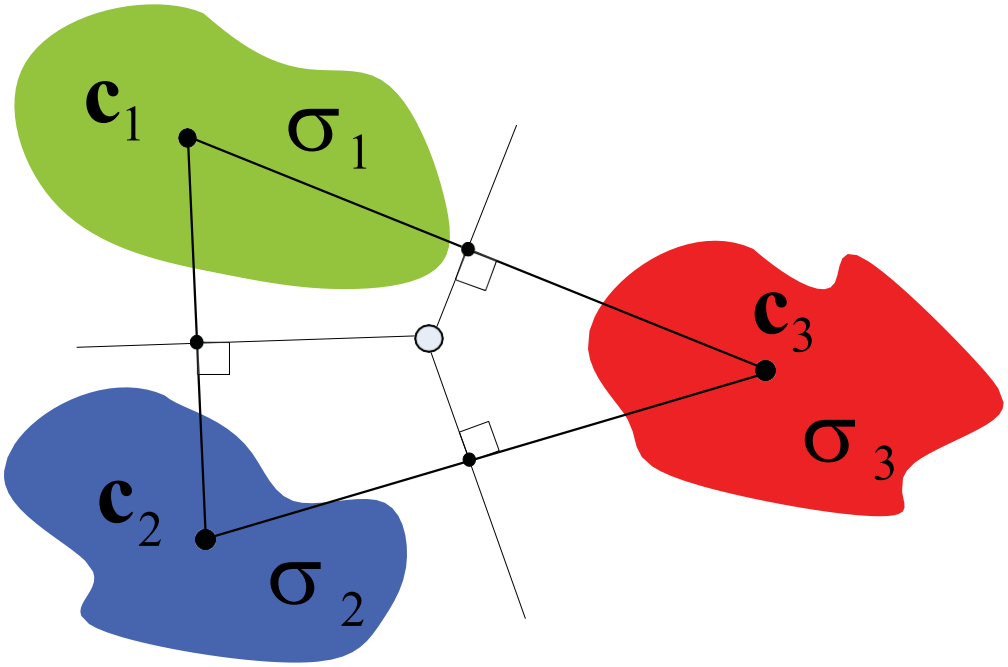
Решающая функция $d(\mathbf{x}) = \|\mathbf{x}-\mathbf{c}_1\|_2^2 - \|\mathbf{x}-\mathbf{c}_2\|_2^2 =$

$$(\mathbf{x}-\mathbf{c}_1,\mathbf{x}-\mathbf{c}_1)-(\mathbf{x}-\mathbf{c}_2,\mathbf{x}-\mathbf{c}_2)=\mathbf{x}^2-2\mathbf{c}_1\cdot\mathbf{x}+\mathbf{c}_1^2-$$

$$\mathbf{x}^2+2\mathbf{c}_2\cdot\mathbf{x}-\mathbf{c}_2^2=2(\mathbf{c}_2-\mathbf{c}_1)\cdot\left(\mathbf{x}-\tfrac{1}{2}(\mathbf{c}_1+\mathbf{c}_2)\right).$$

Способы определения расстояния между вектором-образом и классом (3)

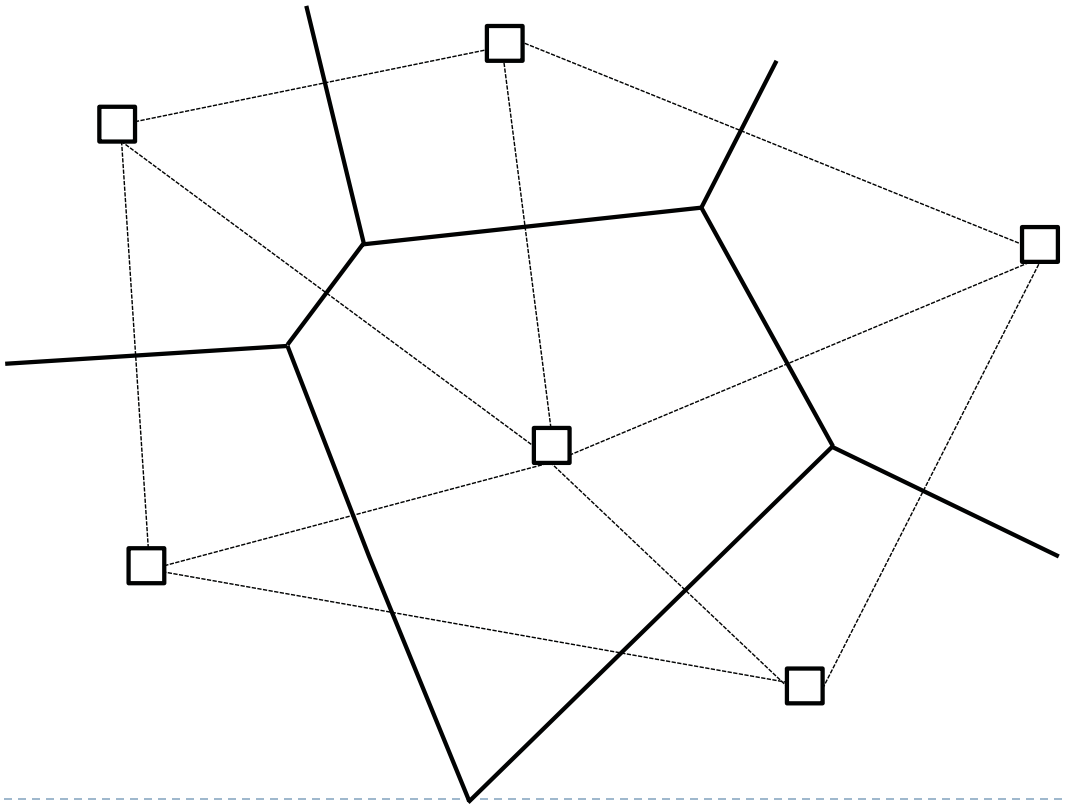
б) классификация по трем классам (евклидова метрика):



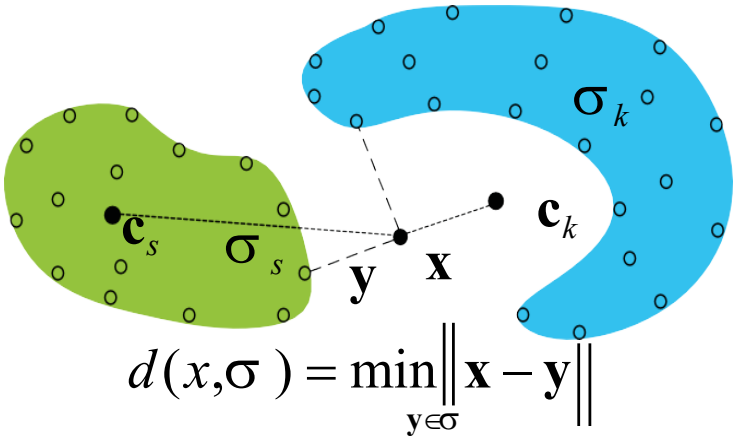
Способы определения расстояния между вектором-образом и классом (2)

в) классификация по m классам (евклидова метрика)

триангуляция Делоне (в R^2)



Метод ближайшего соседа



Правило классификации: $x \in \sigma_s$, если

$$\|\mathbf{x} - \mathbf{x}_s\| = \min\left\{\|\mathbf{x} - \mathbf{x}_i\| : i = 1,\ldots,N\right\} \text{ и } x_s \in \sigma_s.$$

Способы определения расстояний между классами

1) $d_1(\sigma_i,\sigma_j) = \min\limits_{x\in\sigma_i, y\in\sigma_j} d(\mathbf{x},\mathbf{y})$;

2) $d_2(\sigma_i,\sigma_j) = \frac{1}{|S_{ij}|}\sum\limits_{x\in\sigma_i, y\in\sigma_j} d(\mathbf{x},\mathbf{y})$, S_{ij} – множество всех пар элементов между классами σ_i и σ_j;

3) $d_3(\sigma_i,\sigma_j) = \max\limits_{x\in\sigma_i, y\in\sigma_j} d(\mathbf{x},\mathbf{y})$;

4) $d_4(\sigma_i,\sigma_j) = d(\mathbf{c}_i,\mathbf{c}_j)$, $\mathbf{c}_i, \mathbf{c}_j$ – центры классов σ_i и σ_j соответственно.

Постановка задачи кластеризации

Кластером называет группу образов {xi}, Удовлетворяющих условию:

$||x_i - x_k|| < d$, где $||.||$ - мера сходства между образами,

d - заданное пороговое ограничение по этой мере.

Задача. Найти такое разбиение обучающей выборки $\Xi = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ на непересекающиеся подмножества (кластеры) $X_1, ..., X_m$: $X_1 \cup ... \cup X_m = \Xi$,

$X_i \cap X_j = \varnothing \ \forall i \neq j$, чтобы все точки одного кластера состояли из «похожих» элементов, а точки разных кластеров существенно отличались.

Основные параметры кластеризации:

1) критерий «похожести» элементов Q;
2) используемая метрика d;
3) число кластеров.

Основные цели кластеризации

1) нахождение групп схожих элементов с целью дальнейшей независимой их обработки. Параметры кластеризации должны обеспечивать минимальность числа кластеров;

2) получение выборки эталонных элементов – типичных представителей кластеров. Параметры кластеризации должны обеспечивать формирование однородных кластеров;

3) нахождение элементов, не попадающих ни в один из кластеров, при этом сами кластеры должны быть небольшими;

4) формирование иерархической структуры выборки (*задача таксономии*).

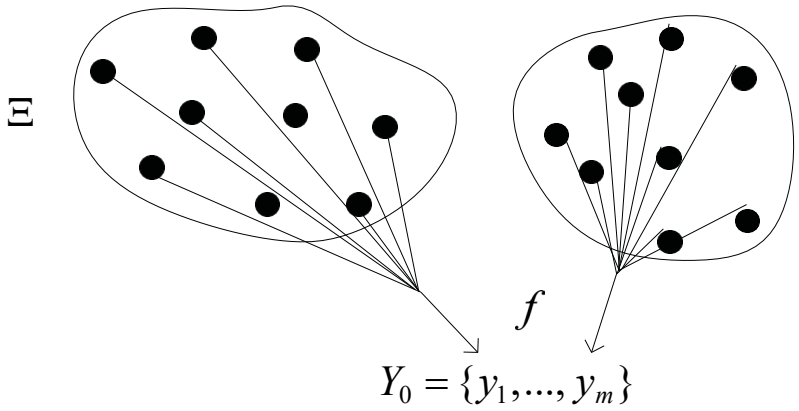
- Иногда кластеры называют **таксонами**, а кластерный анализ - **таксономией**.

Математическая постановка задачи кластеризации

Пусть $\Xi = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ – обучающая выборка, Ψ – множества допустимых меток (определяется целями кластеризации). Найти множества меток $Y_0 \in \Psi$ и функцию $f : \Xi \to Y_0$, чтобы

$$Y_0 = \arg \min_{Y \subseteq \Psi,\, f} Q(Y, f),$$

где $Q(Y, f)$ – выбранный критерий качества кластеризации.



Основные критерии качества кластеризации

1) среднее внутрикластерное расстояние

$$Q^{(1)} = \sum_i \sum_{\mathbf{x},\mathbf{y} \in X_i} d(\mathbf{x},\mathbf{y}) \to \min;$$

2) среднее межкластерное расстояние

$$Q^{(2)} = \sum_{i<j} \sum_{\substack{\mathbf{x} \in X_i, \\ \mathbf{y} \in X_j}} d(\mathbf{x},\mathbf{y}) \to \max;$$

3) суммарная выборочная дисперсия разброса элементов относительно центров кластеров

$$Q^{(3)} = \sum_i \frac{1}{|X_i|} \sum_{\mathbf{x} \in X_i} d^2(\mathbf{x},\mathbf{c}_i) \to \min,$$

где $\mathbf{c}_i = \frac{1}{|X_i|} \sum_{\mathbf{x} \in X_i} \mathbf{x}$ – центр кластера X_i;

4) комбинированные критерии, например,

$$Q^{(1)} \Big/ Q^{(2)} \to \min.$$

Алгоритм расстановки центров кластеров

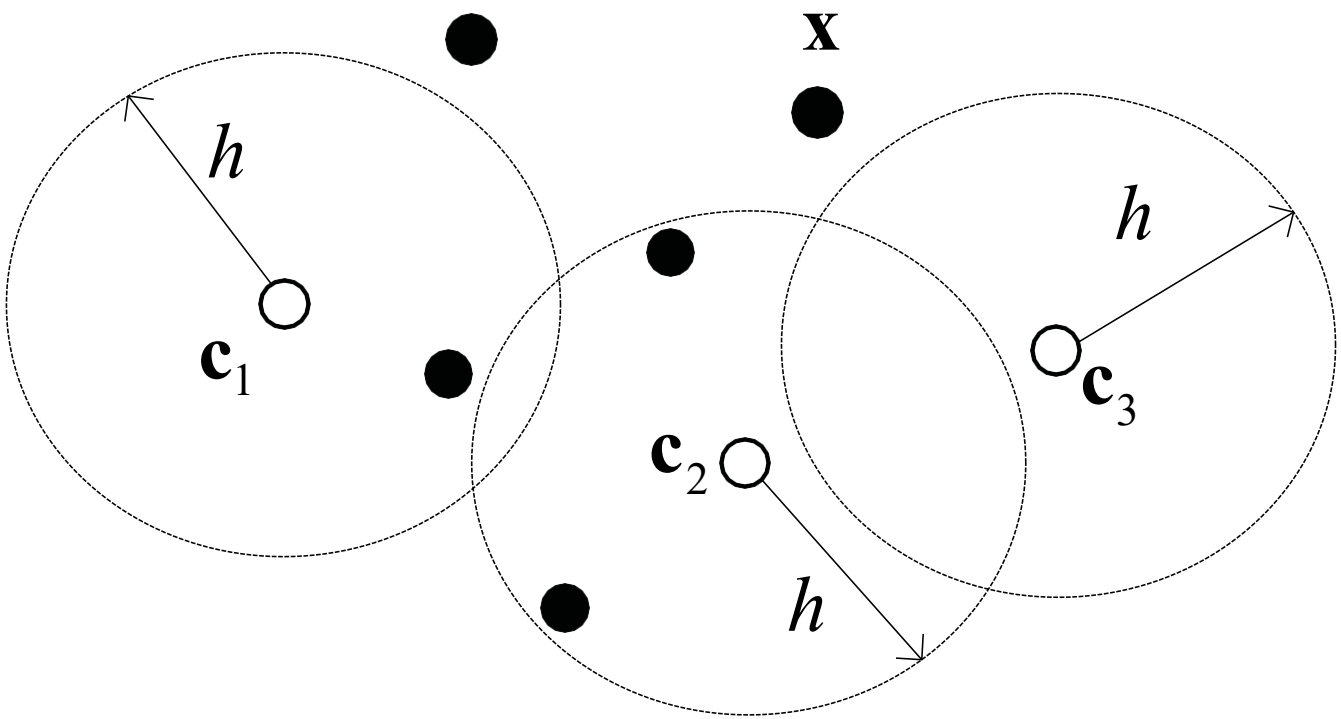
Алгоритм простейшей расстановки центров кластеров

Вводится некоторый порог $h > 0$.

1)в качестве первого центра кластера назначается первый элемент выборки $\mathbf{c}_1 = \mathbf{x}_1$.

2)если уже выбраны k центров кластеров, то в ка- честве $k+1$-го центра выбирается такой элемент выборки $\mathbf{x}_j$, что минимальное расстояние от $\mathbf{x}_j$ до центров $\mathbf{c}_i$, $i = 1,...,k$, будет больше h.

Алгоритм расстановки центров кластеров



Основные процедуры изменения числа кластеров.

1. Удаление кластеров. Если кластер содержит мало элементов $X_i < q_1$, то он удаляется, его элементы распределяются по другим кластерам, а центр кластера $\mathbf{c}_i$ удаляется из списка центров кластеров.

Алгоритм расстановки центров кластеров (2)

2. Разделение кластеров. Если дисперсия i -го кластера $D_i > q_2$, то i -й кластер разделяют на два. Для этого вычисляется «направление» в R^n , вдоль которого дисперсия кластера максимальна. Кластер разделяется на два гиперплоскостью, проходящей через центр кластера и перпендикулярной вычисленному направлению.

3. Слияние кластеров. Пусть $l_{ij} = \|\mathbf{c}_i - \mathbf{c}_j\|$ – расстояние между центрами кластеров. Если $l_{ij} < q_3$, то кластеры X_i и X_j объединяются. Новый центр кластера вычисляется по формуле $\mathbf{c} = \frac{1}{|X_i|+|X_j|}\left(\mathbf{c}_i |X_i| + \mathbf{c}_j |X_j|\right)$.