УНИВЕРСИТЕТ ИТМО

Количественные методы анализа информации

Кластерный анализ

Основы анализа данных.

Лекция 12.

Секции

Основная цель

Функции расстояния

Методы кластеризации

К-средних

Пример применения

Определение

Кластерный анализ представляет собой класс методов, используемых для классификации объектов или событий в относительно однородные группы.

Группы называют кластерами (clusters).
Объекты в каждом кластере должны быть похожи между собой и отличаться от объектов в других кластерах.

Типы входных данных

1. Признаковое описание объектов.
Каждый объект описывается набором своих характеристик, называемых признаками. Признаки могут быть числовыми или нечисловыми.

2. Матрица расстояний между объектами.
Каждый объект описывается расстояниями до всех остальных объектов обучающей выборки.

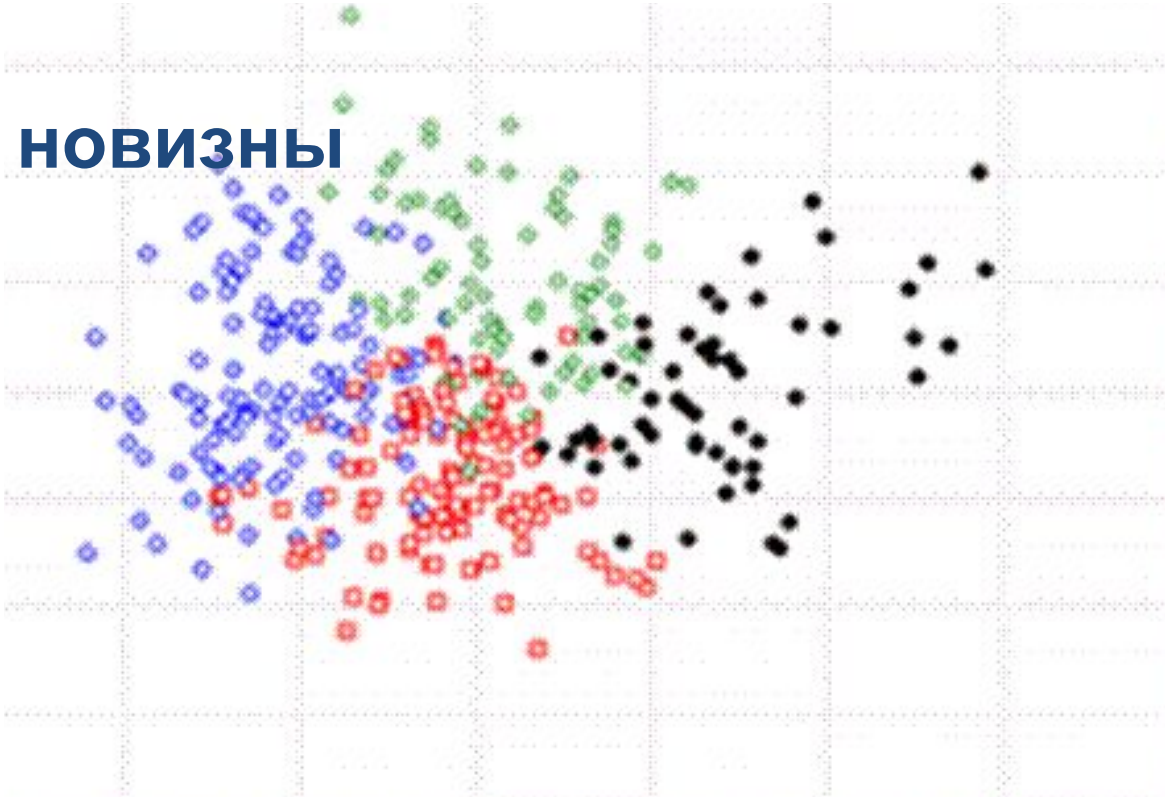
*постановка задачи кластеризации по матрице расстояний является более общей

Цели кластеризации

1. **Понимание данных путём выявления кластерной структуры.**

2. **Сжатие данных.**

3. **Обнаружение новизны**



Формальная постановка задачи кластеризации

$\boldsymbol{X}$ - множество объектов;

$\boldsymbol{Y}$ - множество номеров (имён, меток) кластеров;

$\boldsymbol{\rho(x_i, x_j)}$ - функция расстояния между объектами;

$\boldsymbol{X^m = (x_1, \ldots, x_m) \subset X}$ – обучающая выборка

Алгоритм кластеризации:

$a: X \to Y$ кластеризация

$a: X| < \boldsymbol{X^m, Y} > \to Y$ классификация

Статистики, связанные с кластерным анализом

1. **План агломерации,** объединения (agglomeration schedule).
Дает информацию об объектах (событиях, случаях), которые должны быть объединены на каждой стадии процесса иерархической кластеризации.
2. **Кластерный центроид** (cluster centroid).
Среднее значение переменных для всех случаев или объектов в конкретном кластере,
3. **Кластерные центры** (cluster centers).
Исходные начальные точки в неиерархической кластеризации. Кластеры строят вокруг этих центров, или зерен кластеризации.
4. **Принадлежность кластеру** (cluster membership).
Указывает кластер, которому принадлежит каждый случай или объект.
5. **Древовидная диаграмма** (дендрограмма) (dendrogram).
Ее также называют древовидный граф— графическое средство для показа результатов кластеризации.
6. **Расстояния между кластерными центрами** (distances between cluster centres). Указывают, насколько разнесены отдельные пары кластеров, Кластеры, которые разнесены широко, ясно выражены и поэтому желательны.
7. **Сосульчатая диаграмма** (icicle diagram),
8. **Матрица сходства/матрица расстояний** между объединяемыми объектами (similarity/distance coefficient matrix). Матрица сходства (расстояний) — это нижняя треугольная матрица, содержащая значения расстояния между парами объектов или случаев.

Неоднозначность решения задачи кластеризации

Причины неоднозначности:

1. Не существует однозначно наилучшего критерия качества кластеризации.

2. Число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторым субъективным критерием.

3. Результат кластеризации существенно зависит от метрики, выбор которой, как правило, также субъективен и определяется экспертом.

Этапы выполнения кластерного анализа

1. формулировка проблемы
2. выбор меры расстояния
3. выбор метода кластеризации
4. принятие решения о количестве кластеров
5. интерпретация и профилирование кластеров
6. оценка достоверности кластеризации

Секции

Основная цель

Функции расстояния

Методы кластеризации

К-средних

Пример применения

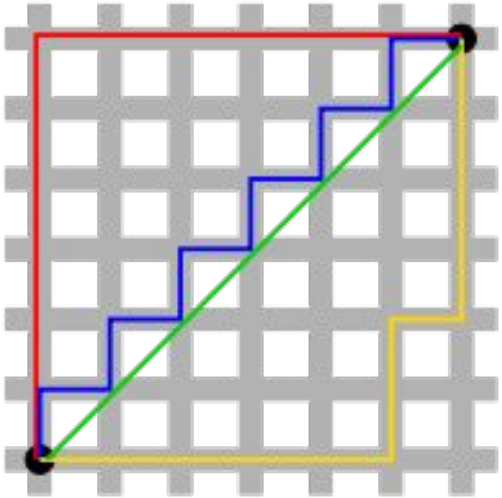
Выбор способа измерения расстояния

1. Евклидово расстояние:

$$d(a,b) = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2}$$

2. Расстояние городских кварталов

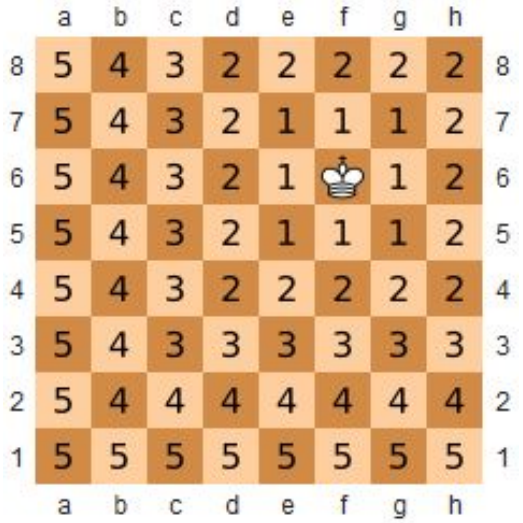
$$d(a,b) = \sum_{i=1}^{n}|a_i - b_i|$$



Выбор способа измерения расстояния

3. Расстояние Чебышева:

$$d(a,b) = \max_{i=1\ldots n} |a_i - b_i|$$



4. Метрика Минковского

$$d(a,b) = \left(\sum_{i=1}^{n} |a_i - b_i|^p \right)^{1/p}$$

5. Взвешенная евклидова метрика:

$$d(a,b) = \sqrt{\sum_{i=1}^{n} w \cdot (a_i - b_i)^2}$$

6. Расстояние Махланобиса

$$d(\vec{a},\vec{b}) = \sqrt{(\vec{a}-\vec{b})^T S^{-1} (\vec{a}-\vec{b})}$$

Секции

Основная цель

Функции расстояния

Методы кластеризации

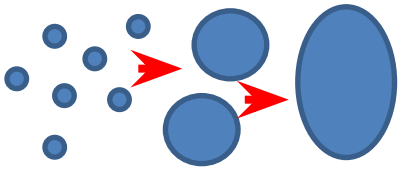
К-средних

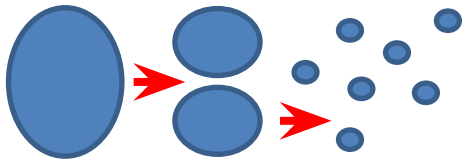
Пример применения

Методы кластеризации

Иерархические методы

- Агломеративные
  - Методы связи
  - Дисперсионные методы
  - Центроидные методе
- Дивизивные

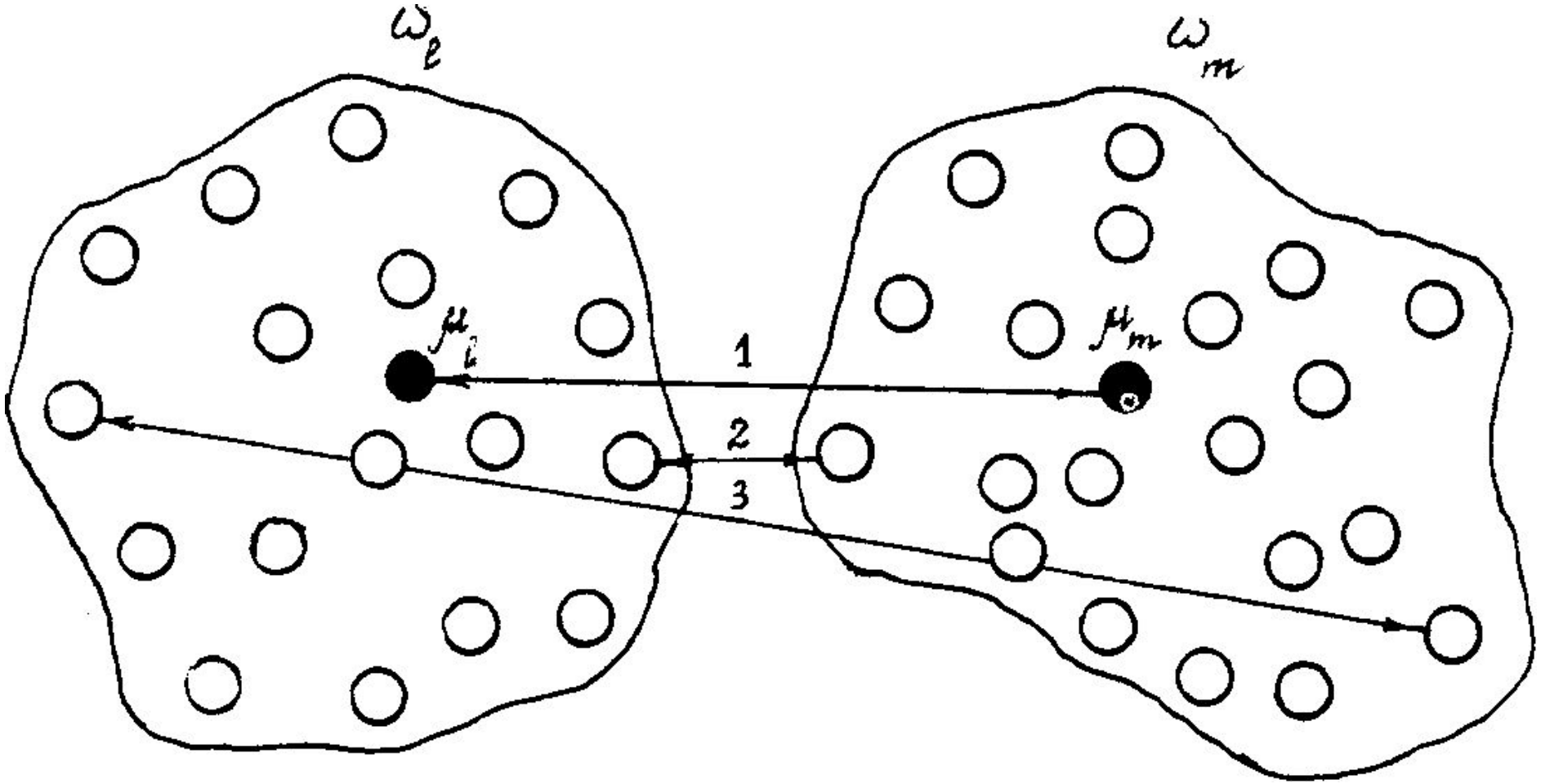




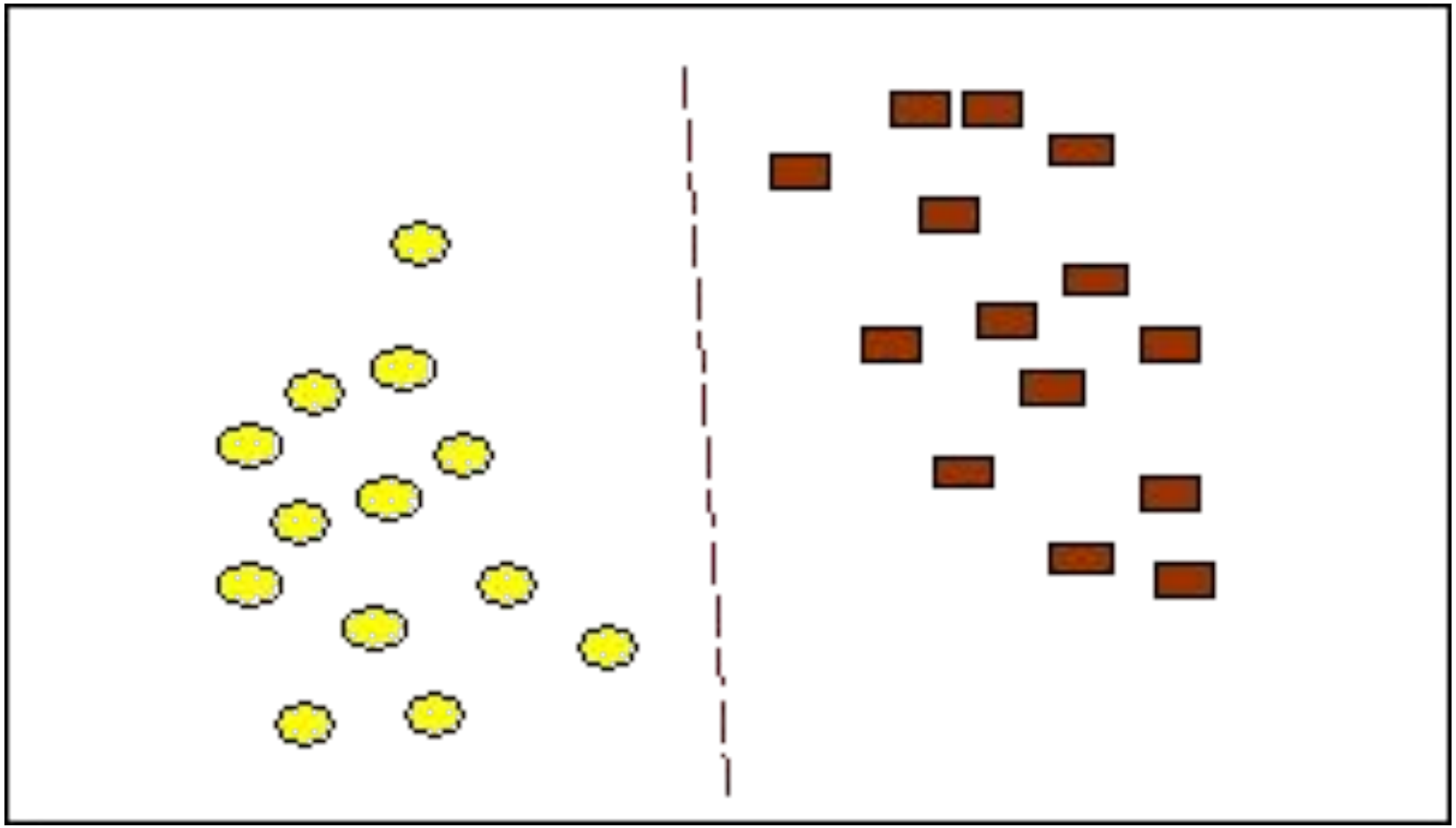
Неиерархические методы

- К-средних
- Последовательный пороговый метод
- Параллельный пороговый метод
- Метод оптимизирующего распределения

1. Метод ближайшего соседа и центроидный метод



2. Дисперсионный метод Варда.



Принятие решения о количестве кластеров

Вопрос о количестве кластеров – главный вопрос кластерного анализа.

Рекомендации:

1. При определении количества кластеров руководствуются теоретическими и практическими соображениями.
2. В иерархической кластеризации в качестве критерия можно использовать расстояния, при которых объединяют кластеры.
3. В неиерархической кластеризации чертят график зависимости отношения суммарной внутригрупповой дисперсии к межгрупповой дисперсии от числа кластеров.
4. Относительные размеры кластеров должны быть достаточно выразительными.

Интерпретация и профилирование кластеров

Процедуры проверки качества кластерного анализа:

1. Выполняйте кластерный анализ на основании одних и тех же данных, но с использованием различных способов измерения расстояния..
2. Используйте разные методы кластерного анализа и сравните полученные результаты.
3. Разбейте данные на две равные части случайным образом. Выполните кластерный анализ отдельно для каждой половины.
4. Случайным образом удалите некоторые переменные. Выполните кластерный анализ по сокращенному набору переменных.
5. В неиерархической кластеризации решение может зависеть от порядка случаев в наборе данных. Выполните анализ несколько раз, меняя порядок случаев, до получения стабильного решения.

Кластеризация переменных

Метрика расстояния: коэффициент корреляции

Цель:

1. идентификация характерных переменных или переменных, которые вносят уникальный вклад в данные;

2. уменьшение числа переменны (замена переменных на кластерные компоненты).

Задание на самостоятельную работу

1. Разбиться на группы по 1-3 человека.
2. Подготовить доклад на одну из тем:
   - Метод кластеризации ближайшего соседа
   - Кластеризация методом полной связи
   - Кластеризация методом средней связи
     - Невзвешенный
     - Взвешенный
   - Центроидный метод кластеризации
     - Невзвешенный
     - Взвешенный
   - Кластеризация методов Варда
   - К-средних
3. Подготовить пример использования и реализации метода
4. Подготовить презентацию.

* Можно рассматривать один метод с разными метриками расстояний

Секции

Основная цель

Функции расстояния

Методы кластеризации

К-средних

Пример применения

K-средних

k-means (метод k-средних) — метод кластеризации, предполагающий минимизацию суммарное квадратичное отклонение точек кластеров от центров этих кластеров.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

k – количество кластеров;

μ_i - центры (масс) кластеров (также – «главные точки»);

S_i - кластеры

К-средних

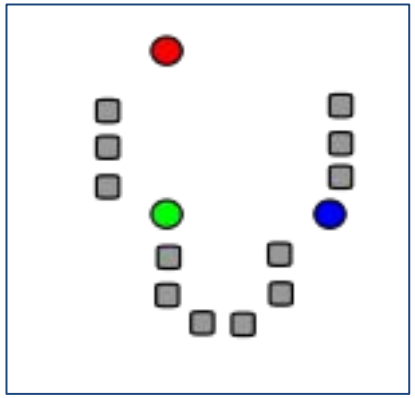
Алгоритм:

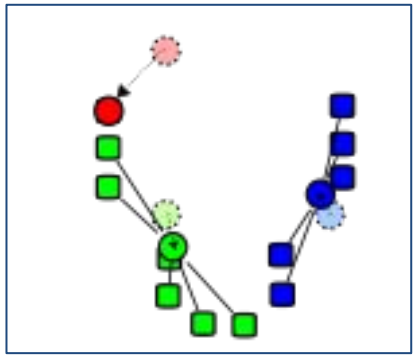
1. Выбор центров масс кластеров (на первой итерации случайный).

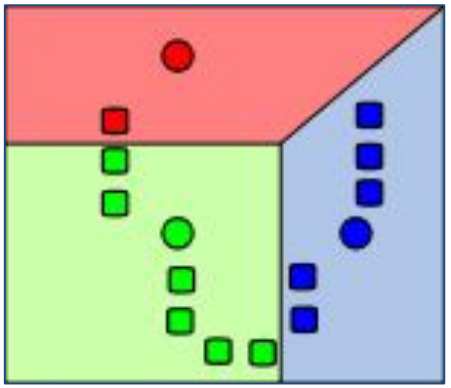
2. Прикрепление точек к кластерам, центр которого ближе других.

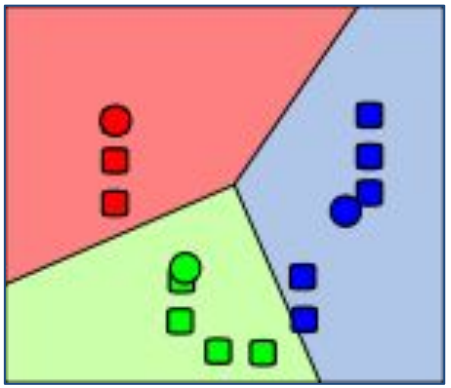
3. Вычисление новых центров масс кластеров

4. Возврат на шаг 1 или конец, если центр масс более не меняется.









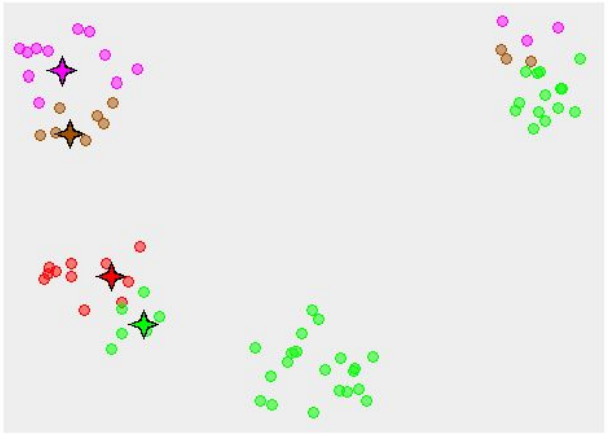
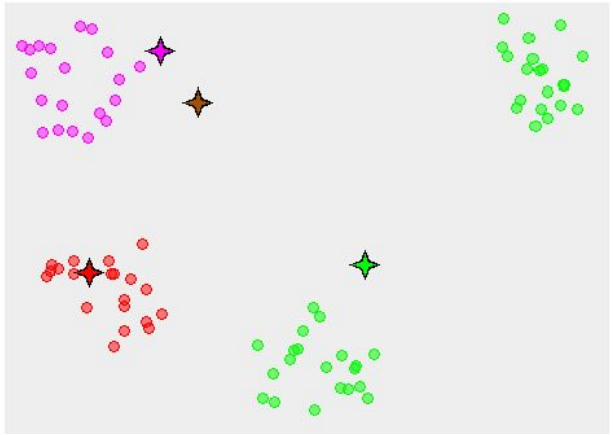
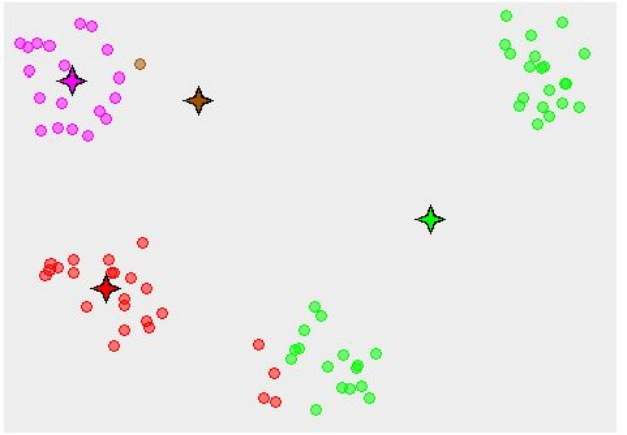
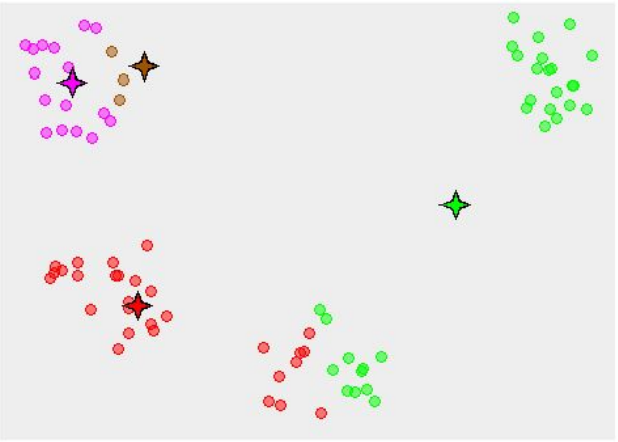
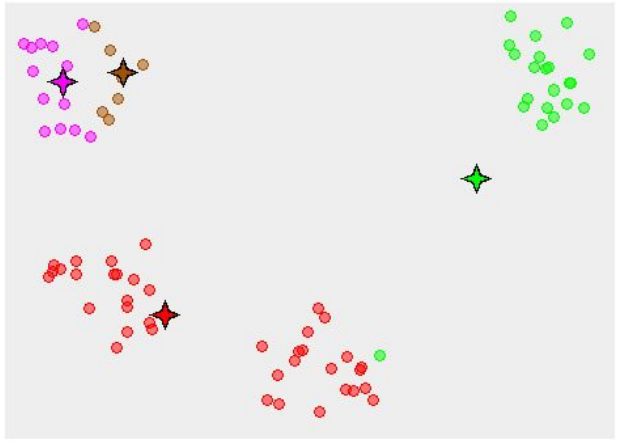
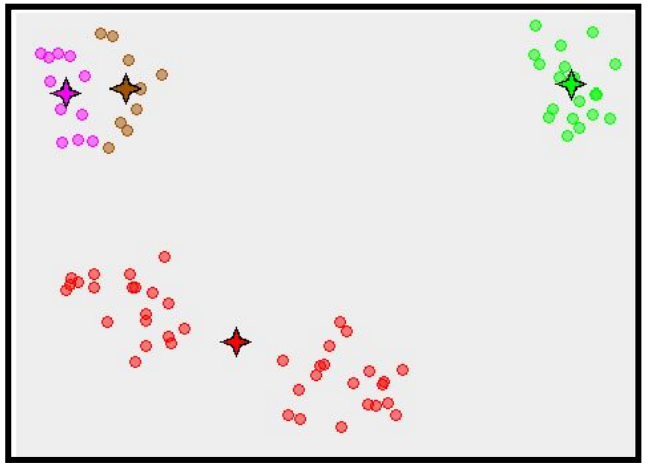
Проблемы К-средних

1. Не гарантируется достижение глобального минимума суммарного квадратичного отклонения V, а только одного из локальных минимумов.

2. Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.

3. Число кластеров надо знать заранее.

Пример неправильного применения К-средних

Секции

Основная цель

Функции расстояния

Методы кластеризации

К-средних

Пример применения

Пример применения K-средних

Исходный файл данных содержит следующую информацию об автомобилях и их владельцах:

1. марка автомобиля – первая переменная;
2. стоимость автомобиля – вторая переменная;
3. возраст водителя – третья переменная;
4. стаж водителя – четвертая переменная;
5. возраст автомобиля – пятая переменная;

Целью данного анализа является разбиение автомобилей и их владельцев на классы, каждый из которых соответствует определенной рисковой группе.

Наблюдения, попавшие в одну группу, характеризуются одинаковой вероятностью наступления страхового случая, которая впоследствии оценивается страховщиком.

Пример применения K-средних

Данные: Cars_my.sta* (5v * 22c)

	1 car	2 PRICE	3 age	4 experience	5 age_car
1	Acura	0,521	25	3	10
2	Audi	0,866	24	3	1
3	BMW	0,496	29	3	4
4	Buick	0,614	50	25	9
5	Corvette	1,235	62	38	15
6	Chrysler	0,614	43	21	9
7	Dodge	0,706	26	1	5
8	Eagle	0,614	20	1	1
9	Ford	0,706	54	10	11
10	Honda	0,429	38	8	7
11	Isuzu	0,798	27	5	3
12	Mazda	0,126	51	20	10
13	Mercedes	1,051	46	25	4
14	Mitsub.	0,614	28	2	7
15	Nissan	0,429	31	6	6
16	Olds	0,614	45	16	4
17	Pontiac	0,614	40	16	2
18	Porsche	3,454	41	8	8
19	Saab	0,588	29	5	2
20	Toyota	0,059	36	13	1
21	VW	0,706	38	15	6
22	Volvo	0,219	42	19	4

Фрагмент исходных данных

Пример применения K-средних

Шаг 1. Масштаб измерений.

Поскольку различные измерения используют абсолютно различные типы шкал, данные необходимо стандартизовать - каждая переменная должна иметь среднее 0 и стандартное отклонение 1.

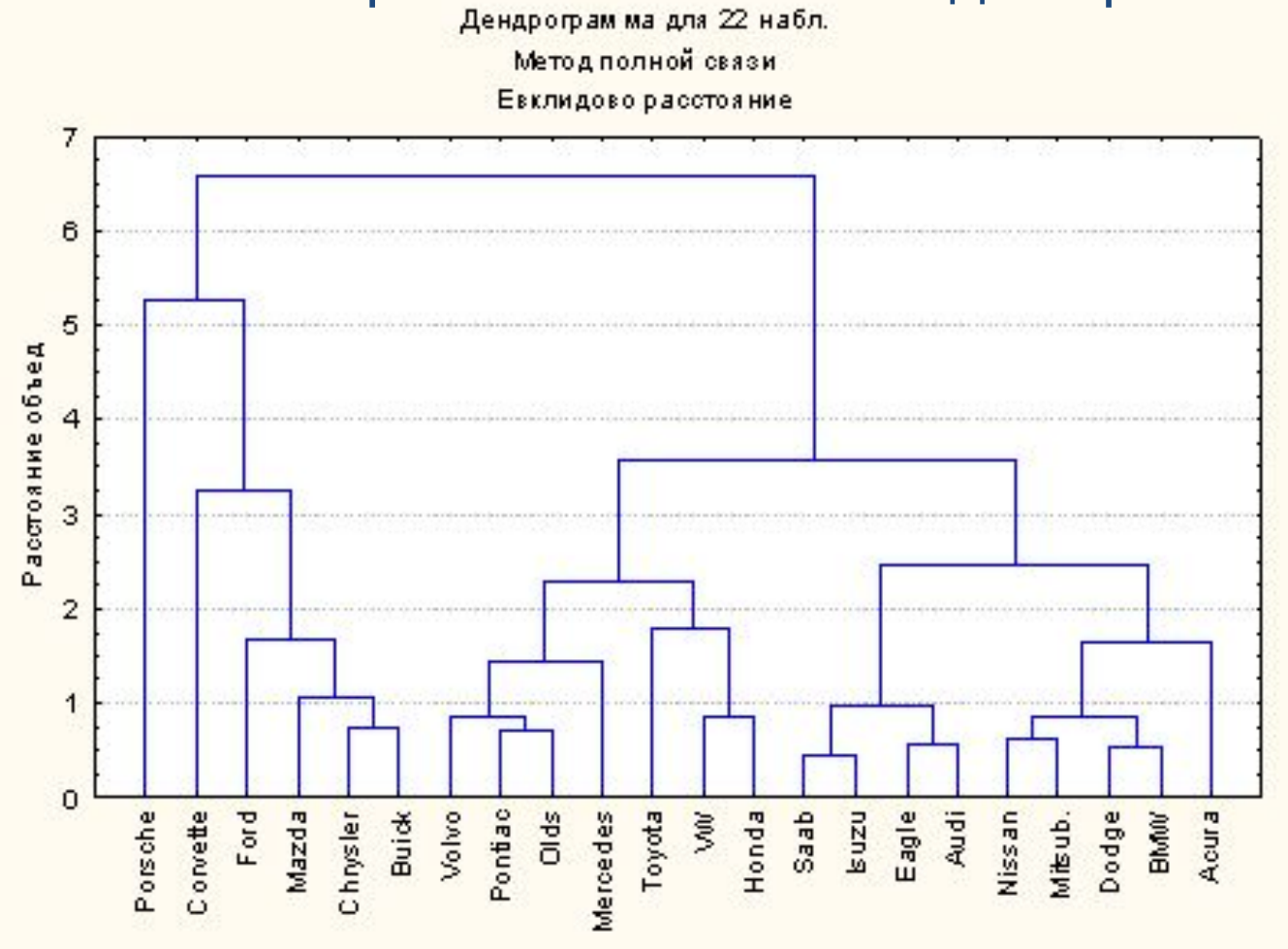
Данные: Cars_my_standartiz.sta* (4v * 22c)

	1 PRICE	2 age	3 experience	4 age_car
Acura	-0,315	-1,128727	-1	1
Audi	0,204	-1,219025	-1	-1
BMW	-0,353	-0,767534	-1	-0
Buick	-0,176	1,1287269	1	1
Corvette	0,760	2,2123048	3	2
Chrysler	-0,176	0,4966398	1	1
Dodge	-0,037	-1,038429	-1	-0
Eagle	-0,176	-1,580218	-1	-1
Ford	-0,037	1,4899195	-0	1
Honda	-0,454	0,0451491	-0	0
Isuzu	0,102	-0,948131	-1	-1
Mazda	-0,910	1,2190251	1	1
Mercedes	0,482	0,7675343	1	-0
Mitsub	-0,176	-0,857832	-1	0
Nissan	-0,454	-0,586938	-1	0
Olds	-0,176	0,6772362	0	-0
Pontiac	-0,176	0,2257454	0	-1
Porsche	4,101	0,3160435	-0	1
Saab	-0,214	-0,767534	-1	-1
Toyota	-1,011	-0,135447	0	-1
VW	-0,037	0,0451491	0	0
Volvo	-0,771	0,4063417	1	-0

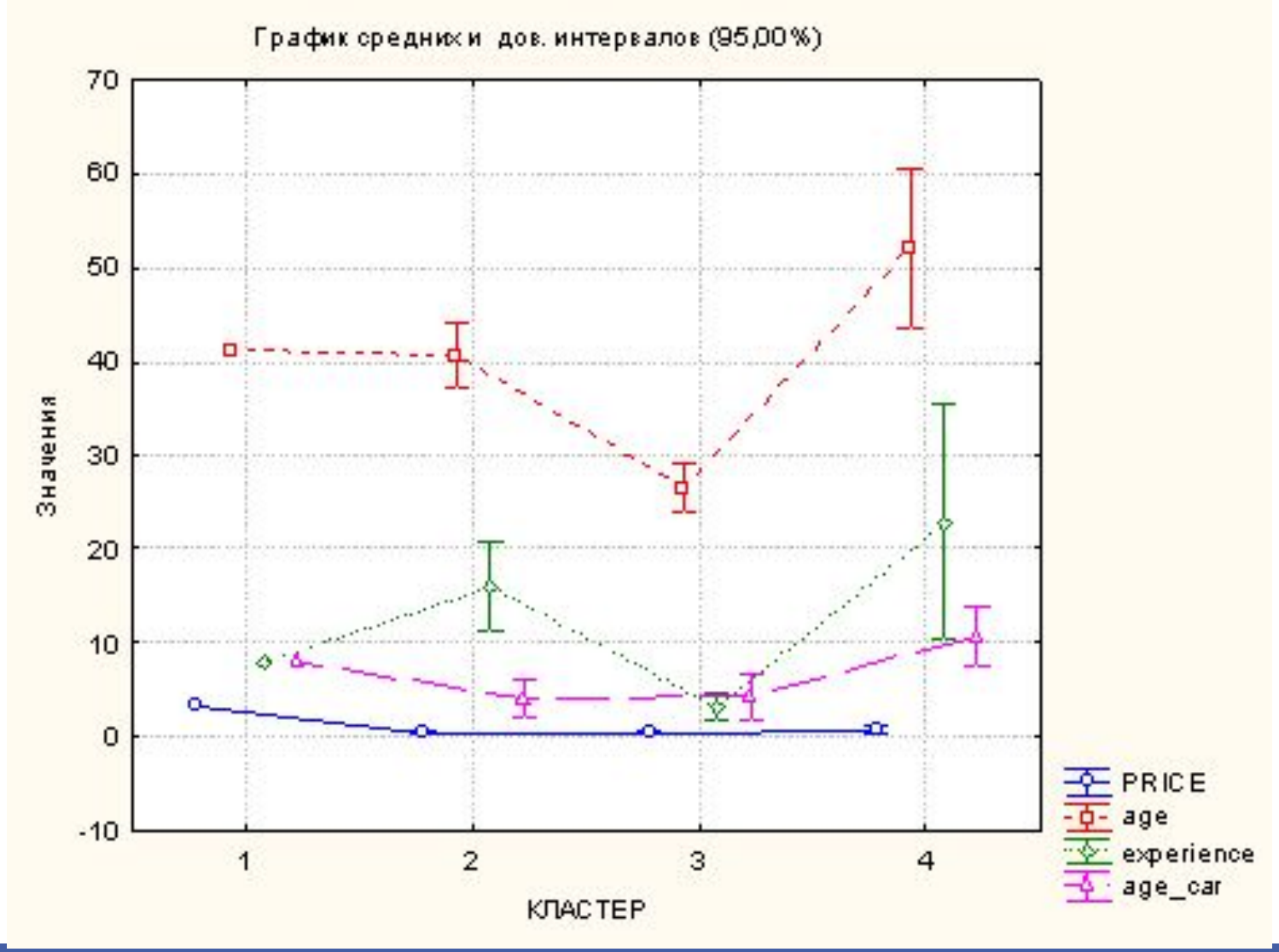
Пример применения Иерархического алгоритма

Шаг 2. Иерархическая классификация .

В качестве правила объединения отметим Метод полной связи, в качестве меры близости – Евклидово расстояние.



Шаг 3. Кластеризация методом К средних.



Шаг 3. Кластеризация методом К средних.

Первый кластер:

	Элементы кластера номер 1 (Cars_my_standartiz.sta) и расстояния до центра кластера. Кластер содержит 1 набл.
	Porsche
Расст.	0,00

Второй кластер:

	Элементы кластера номер 2 (Cars_my_standartiz.sta) и расстояния до центра кластера. Кластер содержит 7 набл.						
	Honda	Mercedes	Olds	Pontiac	Toyota	VW	Volvo
Расст.	0,590329	0,651432	0,204147	0,276696	0,594585	0,327201	0,284429

Шаг 3. Кластеризация методом К средних.

Третий кластер:

	Элементы кластера номер 3 (Cars_my_standartiz.sta) и расстояния до центра кластера. Кластер содержит 9 набл.								
	Acura	Audi	BMW	Dodge	Eagle	Isuzu	Mitsub.	Nissan	Saab
Расст.	0,763888	0,494066	0,154399	0,158544	0,546472	0,239325	0,367393	0,363288	0,343910

Четвертый кластер:

	Элементы кластера номер 4 (Cars_my_standartiz.sta) и расстояния до центра кластера. Кластер содержит 5 набл.				
	Buick	Corvette	Chrysler	Ford	Mazda
Расст.	0,282375	1,145248	0,482189	0,662676	0,441467

Шаг 4. Описательный статистики кластеров.

Ниже приведены таблицы описательных статистик для каждого из показателей:

Цена:

Итоговая таблица средних (Cars_my.sta)
N=22 (Нет пропусков в завис. перем.)

КЛАСТЕР	PRICE Среднее	PRICE N	PRICE Ст.откл.	PRICE Минимум	PRICE Максимум
1	3,454206	1	0,000000	3,454206	3,454206
2	0,527076	7	0,327905	0,058831	1,050549
3	0,625660	9	0,142386	0,428624	0,865652
4	0,658904	5	0,394541	0,126066	1,235446
Всего	0,730418	22	0,664142	0,058831	3,454206

Шаг 4. Описательный статистики кластеров.

Ниже приведены таблицы описательных статистик для каждого из показателей:

Возраст:

Итоговая таблица средних (Cars_my.sta)
N=22 (Нет пропусков в завис. перем.)

КЛАСТЕР	age Среднее	age N	age Ст.откл.	age Минимум	age Максимум
1	41,00000	1	0,00000	41,00000	41,00000
2	40,71429	7	3,77334	36,00000	46,00000
3	26,55556	9	3,28295	20,00000	31,00000
4	52,00000	5	6,89202	43,00000	62,00000
Всего	37,50000	22	11,07442	20,00000	62,00000

Шаг 4. Описательный статистики кластеров.

Ниже приведены таблицы описательных статистик для каждого из показателей:

Опыт:

Итоговая таблица средних (Cars_my.sta)
N=22 (Нет пропусков в завис. перем.)

КЛАСТЕР	experience Среднее	experience N	experience Ст.откл.	experience Минимум	experience Максимум
1	8,00000	1	0,00000	8,00000	8,00000
2	16,00000	7	5,22813	8,00000	25,00000
3	3,22222	9	1,78730	1,00000	6,00000
4	22,80000	5	10,13410	10,00000	38,00000
Всего	11,95455	22	9,77108	1,00000	38,00000

Шаг 4. Описательный статистики кластеров.

Ниже приведены таблицы описательных статистик для каждого из показателей:

Возраст автомобиля:

Итоговая таблица средних (Cars_my.sta)
N=22 (Нет пропусков в завис. перем.)

КЛАСТЕР	age_car Среднее	age_car N	age_car Ст.откл.	age_car Минимум	age_car Максимум
1	8,00000	1	0,000000	8,000000	8,00000
2	4,00000	7	2,081666	1,000000	7,00000
3	4,33333	9	3,000000	1,000000	10,00000
4	10,80000	5	2,489980	9,000000	15,00000
Всего	5,86364	22	3,745416	1,000000	15,00000

Шаг 5*. Дисперсионный анализ.

для определения значимости различия между полученными кластерами.

перемен.	Дисперсионный анализ (Cars_my_standartiz.sta					
	Между SS	cc	Внутри SS	cc	F	значим. p
PRICE	17,75805	3	3,241945	18	32,86555	0,000000
age	18,05119	3	2,948812	18	36,72906	0,000000
experience	14,71184	3	6,288156	18	14,03767	0,000058
age_car	12,24617	3	8,753834	18	8,39369	0,001058

Итак, значение $p<0.05$, что говорит о значимом различии.

Вопросы

1. Определение кластерного анализа. Цели кластеризации.
2. Типы входных данных. Подготовка исходных данных для кластеризации.
3. Причины неоднозначности решения задачи кластеризации.
4. Этапы кластерного анализа.
5. Функции расстояния в кластерном анализе: евклидово, взвешенное евклидово, расстояние Минковского.
6. Функции расстояния в кластерном анализе: расстояние городских кварталов, расстояние Чебышева, расстояние Махланобиса.
7. Методы кластеризации.
8. Метод кластеризации K-средних.