

Кластеризация данных

Использованы материалы:

Методы кластеризации. Воронцов К.В.

Кластерный анализ. Родионова Л.А.

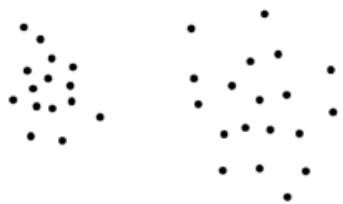
-
- “Всякий раз, когда необходимо классифицировать “горы” информации на пригодные для дальнейшей обработки группы, **кластерный анализ** оказывается весьма полезным и эффективным”

Кластерный анализ

ЗАДАЧА - разбить изучаемую совокупность объектов на группы схожих, близких в некотором смысле объектов, называемых *кластерами* (классами, таксонами).

Заранее не известно, к какому классу принадлежит каждое из наблюдений.

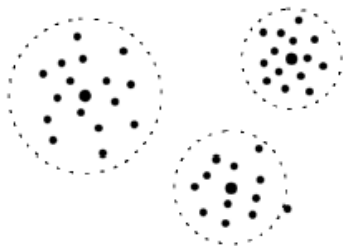
Типы кластерных структур



внутрикластерные расстояния, как правило,
меньше межкластерных



ленточные кластеры



кластеры с центром

Типы кластерных структур



кластеры могут образовываться не по сходству, а по иным типам регулярностей



кластеры могут вообще отсутствовать

- Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов.
- Понятие «тип кластерной структуры» зависит от метода и также не имеет формального определения.

Типы кластерных структур



кластеры могут соединяться перемычками



кластеры могут накладываться на разреженный фон из редко расположенных объектов



кластеры могут перекрываться

Цели кластеризации

- Упростить дальнейшую обработку данных, разбить множество X^{ℓ} на группы схожих объектов чтобы работать с каждой группой в отдельности (задачи классификации, регрессии, прогнозирования).
- Сократить объём хранимых данных, оставив по одному представителю от каждого кластера (задачи сжатия данных).
- ▶ ● Выделить нетипичные объекты, которые не подходят ни к одному из кластеров (задачи одноклассовой классификации).
- Построить иерархию множества объектов (задачи таксономии).

ПРИМЕНЕНИЕ

- В маркетинговых исследованиях
 - Сегментации рынка
 - Анализ поведения покупателей. Идентификация однородных групп покупателей.
 - Кластеризация торговых марок и товаров.
 - Выбор тестовых рынков.
- В управлении персоналом
 - Методика оценки качества работы интервьюеров (кластерный анализ).
- В финансовом анализе

Постановка задачи кластеризации

Дано:

X — пространство объектов;

$X^\ell = \{x_i\}_{i=1}^\ell$ — обучающая выборка;

$\rho: X \times X \rightarrow [0, \infty)$ — функция расстояния между объектами.

Найти:

Y — множество кластеров и

$a: X \rightarrow Y$ — алгоритм кластеризации, такие, что:

- каждый кластер состоит из близких объектов;
- объекты разных кластеров существенно различны.

Кластеризация — это *обучение без учителя*.

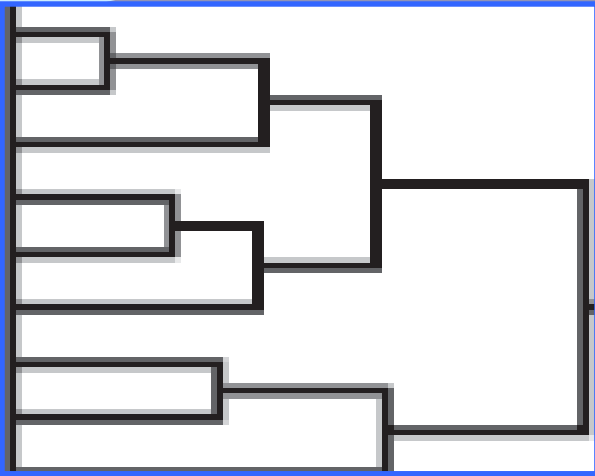
Исследуются n объектов,
каждый объект имеет m признаков.
Каждый объект можно представить в виде точки
в m -мерном пространстве

$$X_i = (X_{i1}, X_{i2}, \dots, X_{im}) \quad i = 1, \dots, n$$

ОБЪЕКТ

ПРИЗНАК (фактор)

Совокупность этих точек можно трактовать как выборку объема n
из многомерной генеральной совокупности.



КЛАСТЕРНЫЙ Анализ

1. Какую метрику выбрать для расчета расстояний?
 - Расстояние между объектами
 - Расстояние между кластерами
2. Какой метод кластеризации следует использовать?
3. Сколько кластеров необходимо сформировать?

Некорректность задачи кластеризации

Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет;
- существует много критериев качества кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров $|Y|$, как правило, неизвестно заранее;
- результат кластеризации существенно зависит от метрики ρ , которую эксперт задаёт субъективно.

ЭТАПЫ КЛАСТЕР-АНАЛИЗА

- **1 этап.** Проводятся кластеризация множества исследуемых объектов по показателям. На этом этапе предполагается использование методов иерархического кластерного анализа.
- **2 этап.** Проверка качества полученных кластеров.
- **3 этап.** Анализ полученных кластеров. Выявление наличия общих закономерностей распределения отдельных объектов в рамках полученных классификаций.

Интерпретация результатов

- Насколько полученное разбиение отличается от случайного?
- Является ли оно надежным и стабильным на подвыборках?
- Какова взаимосвязь между результатами кластеризации и переменными, не участвовавшими в процессе кластеризации?
- Можно ли проинтерпретировать полученные результаты?

ПОСТАНОВКА ЗАДАЧИ

ТРЕБУЕТСЯ

Провести объединение объектов в кластеры на основе вычисляемой меры сходства.

В качестве меры сходства используется **расстояние**.



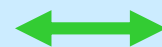
Расстояние между объектами – выбор формулы расчета расстояния.

Расстояние между кластерами - правило объединения в кластеры.

Расстояния

- Расстояние между **объектами** в кластере
- Расстояние между **кластерами**.

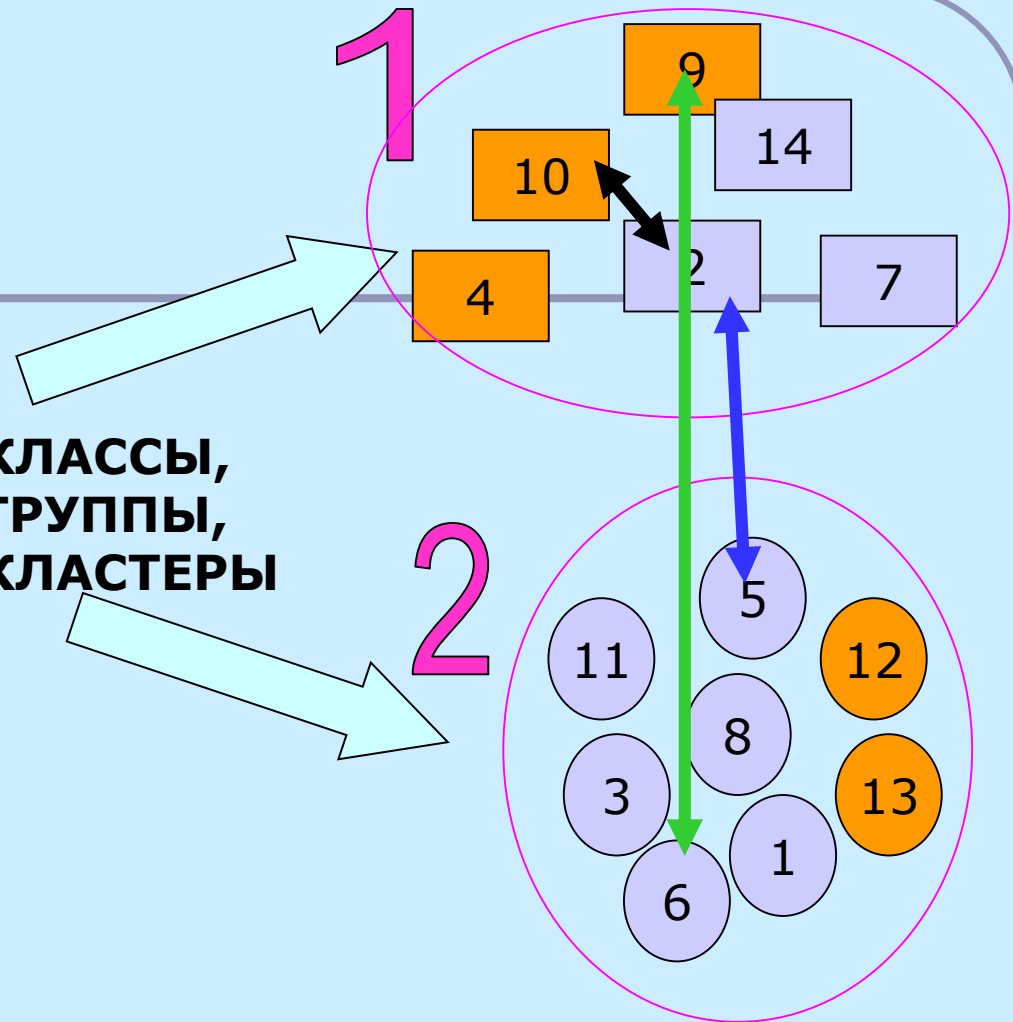
КЛАССЫ,
ГРУППЫ,
КЛАСТЕРЫ



Метод дальнего соседа

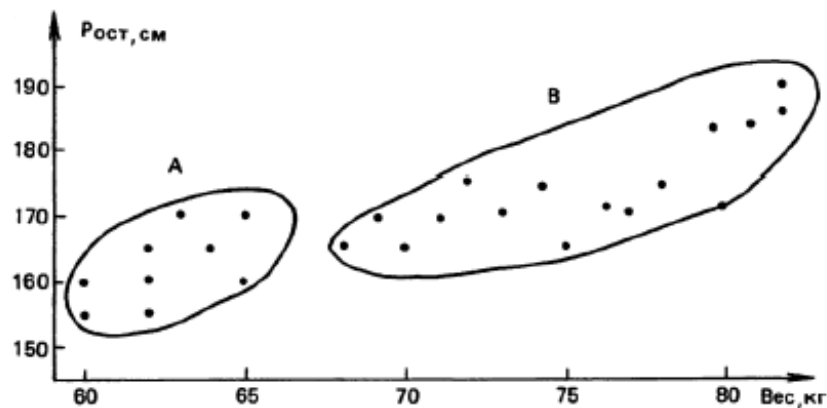


Метод ближайшего соседа

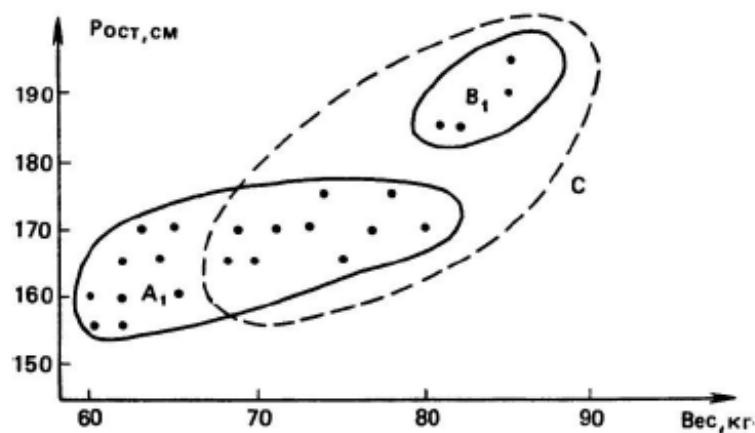


Проблема чувствительности к выбору метрики

Результат зависит от нормировки признаков:



A — студентки,
B — студенты



после перенормировки
(сжали ось «вес» вдвое)

ПРИМЕР

ПРИЗНАКИ
(факторы)

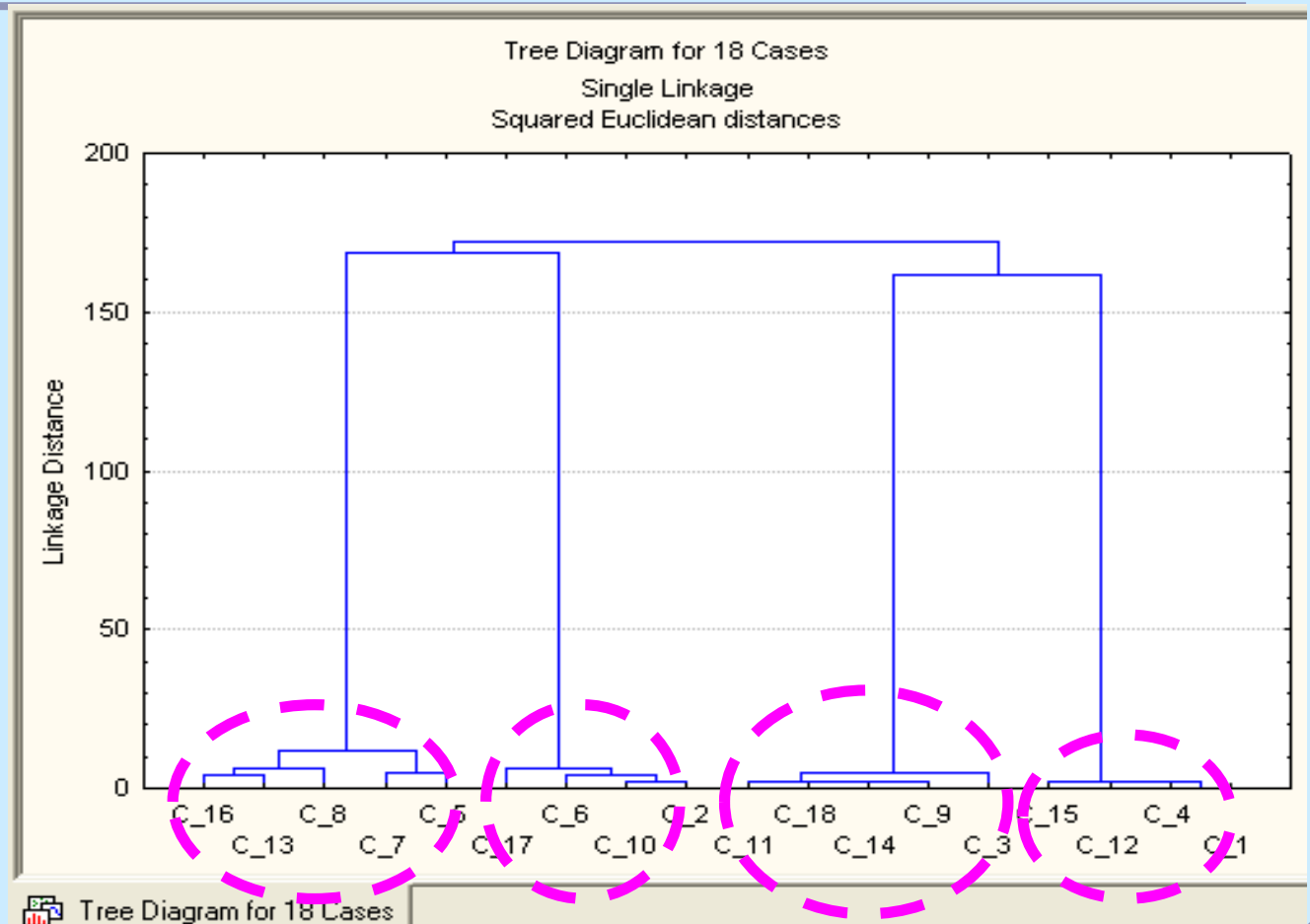
ПЕРЕМЕННЫЕ
Variables

ОБЪЕКТЫ

НАБЛЮДЕНИЯ
Cases

Data: klaster1 (117 by 18c)											
	1 NewVar	2 Test1	3 Test2	4 Test3	5 Test4	6 Test5	7 Test6	8 Test7	9 Test8	10 Test9	11 Test10
1	Volker R	10	10	9	10	10	10	9	10	10	9
2	Sigrid K	10	10	4	10	5	5	4	5	4	3
3	Elmar M	5	4	10	5	10	4	10	5	3	10
4	Peter B	10	10	9	10	10	10	9	10	10	9
5	Otto R	4	3	5	4	3	10	4	10	10	5
6	Elke M	10	10	4	10	5	4	3	4	5	5
7	Sarah K	4	4	5	5	4	10	5	10	10	6
8	Peter T	4	5	3	4	5	10	4	10	10	4
9	Gudrun M	4	5	10	4	10	5	10	4	3	10
10	Siglinde P	10	10	4	10	5	4	4	5	4	4
11	Werner W	4	5	10	5	10	4	10	4	5	10
12	Achim Z	10	10	9	10	10	9	9	10	10	10
13	Dieter K	6	5	4	3	5	10	5	10	10	5
14	Boris P	4	5	10	4	10	5	10	3	4	10
15	Silke W	10	10	9	10	10	9	10	9	10	10
16	Clara T	6	5	3	4	4	10	4	10	10	5
17	Manfred K	10	10	5	10	4	5	4	3	4	5
18	Richard M	4	5	10	4	10	4	10	4	4	10

Иерархический КА ДЕНДРОГРАММА



4
кластера

Этапы кластерного анализа

Формулировка проблемы

Выбор метода кластеризации

Интерпретация и профилирование кластеров

Оценка достоверности кластеризации

Методы кластеризации

1 Статистические методы кластеризации

- EM-алгоритм
- Метод k -средних

2 Сети Кохонена

- Модели конкурентного обучения
- Карты Кохонена

3 Иерархическая кластеризация (таксономия)

- Агломеративная иерархическая кластеризация
- Дендрограмма

Выбор метода кластеризации

Существует два основных класса методов кластеризации – иерархические и итерационные.

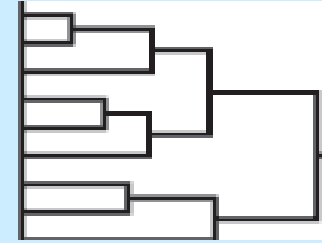
- ❑ Иерархический кластерный анализ (древовидная кластеризация).

- ❑ Исходное количество кластеров равно количеству объектов.

- ❑ Метод К-средних - итерационный.

- ❑ Исходное количество кластеров равно К.

Иерархические методы



Главное различие между иерархическими методами заключается в том, как они определяют расстояние между кластерами, т.е. в стратегии процесса объединения объектов в кластеры.

В зависимости от способа измерения **расстояний между кластерами** методы иерархического кластерного анализа можно разбить 7 групп.

Типичным результатом иерархической кластеризации является иерархическое дерево, или **дендрограмма**.

Методы иерархического кластерного анализа

Расстояния между кластерами

Известно 12 методов присоединения к кластеру нового объекта.
НАИБОЛЕЕ РАСПРОСТРАНЕННЫЕ:

- **Метод одиночной связи** (Single Linkage, Nearest Neighbor), или **Расстояние ближнего соседа**.
- **Метод полной связи** (Complete Linkage, Furthest Neighbor), или **Расстояние дальнего соседа**.
- **Метод невзвешенного (или взвешенного) попарного среднего** (Unweighted (или Weighted) pair-group average, Between-groups linkage) или **Групповое среднее расстояние**.
- **Метод Уорда** (Варда) (*Ward's method*).

Невзвешенный центроидный метод (Unweighted), измеряется расстояние между *центрами тяжести*.

- **Медианный метод – взвешенный центроидный метод** (с учетом числа объектов в кластере в качестве весов).

Расстояния между кластерами

- **Метод одиночной связи** (Single Linkage, Nearest Neighbor), или **Расстояние ближнего соседа**.
- Объект должен иметь наибольшее сходство (по сравнению с прочими «кандидатами на присоединение») с **одним** из членов кластера. Результатом такого метода являются большие **продолговатые кластеры («гребенка»)**, длинные **"цепочки"**.
- Нечувствителен к наличию в данных выбросов, к наличию совпадений в данных, не зависит от преобразования данных.

Расстояния между кластерами

- **Метод полной связи** (Complete Linkage, Furthest Neighbor), или **Расстояние дальнего соседа**.
- Сходство между новым объектом и **всеми** членами кластера должно превышать некоторое пороговое значение (вычисляемое программой).
- Этот метод дает **компактные кластеры** и хорошо работает с группами разного размера.

Расстояния между кластерами

- **Метод невзвешенного (или взвешенного) попарного среднего или Групповое среднее расстояние.**
- Своеобразный компромисс между двумя предыдущими методами, расстояние между новым объектом и кластером определяется как среднее арифметическое расстояний между этим объектом и всеми членами кластера.
- Кластеры обычно получаются довольно продолговатыми.
- Хорошо работает с группами разного размера, эффективно выделяет структуру, «скрытую» случайной изменчивостью признаков.

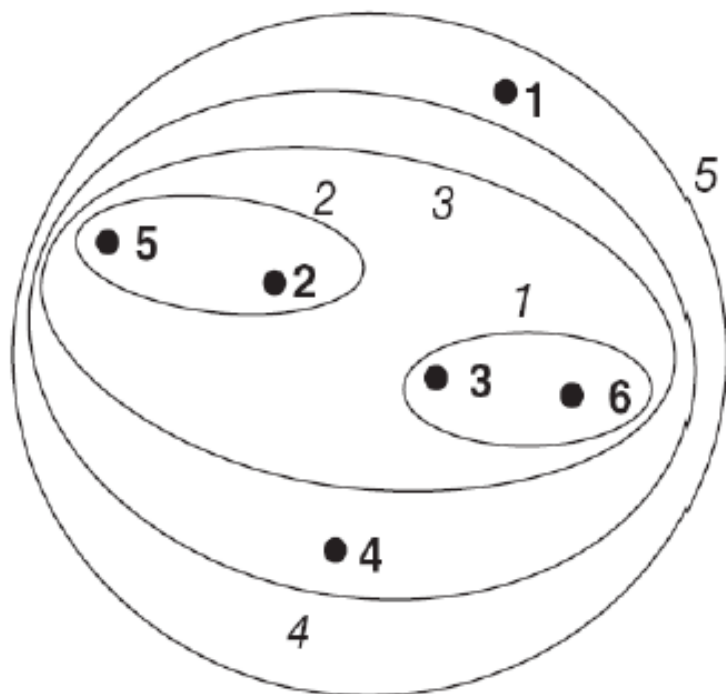
Расстояния между кластерами

- **Метод Уорда** (Варда) (*Ward's method*)
- Минимизирует внутрикластерный разброс объектов. Минимизирует минимальную дисперсию внутри кластеров.
- Позволяет получить компактные хорошо выраженные кластеры. Имеет тенденцию к нахождению кластеров приблизительно равного размера и имеющих гиперсферическую форму.
- Хорошо работает с группами сходных размеров, эффективно выделяет структуру, «скрытую» случайной изменчивостью признаков.

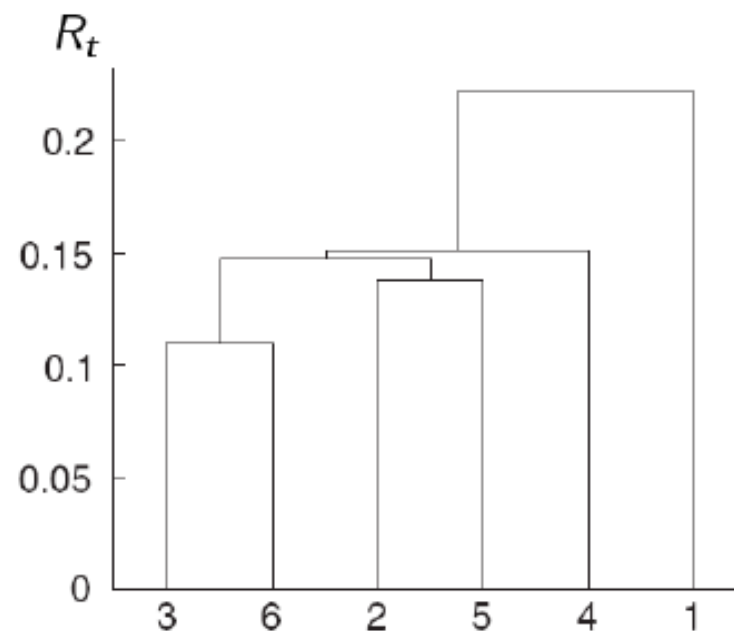
Визуализация кластерной структуры

1. Расстояние ближнего соседа:

Диаграмма вложения

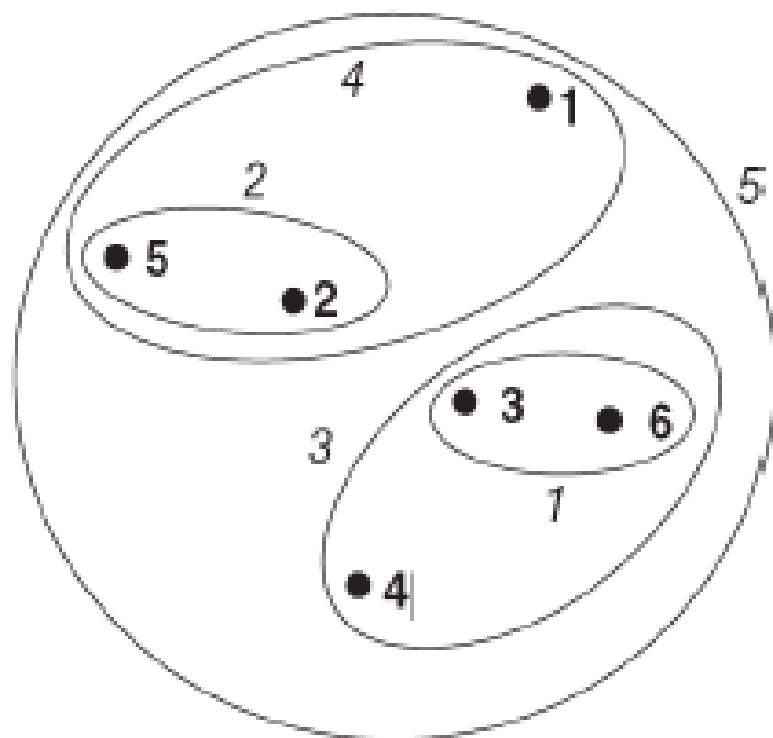


Дендрограмма

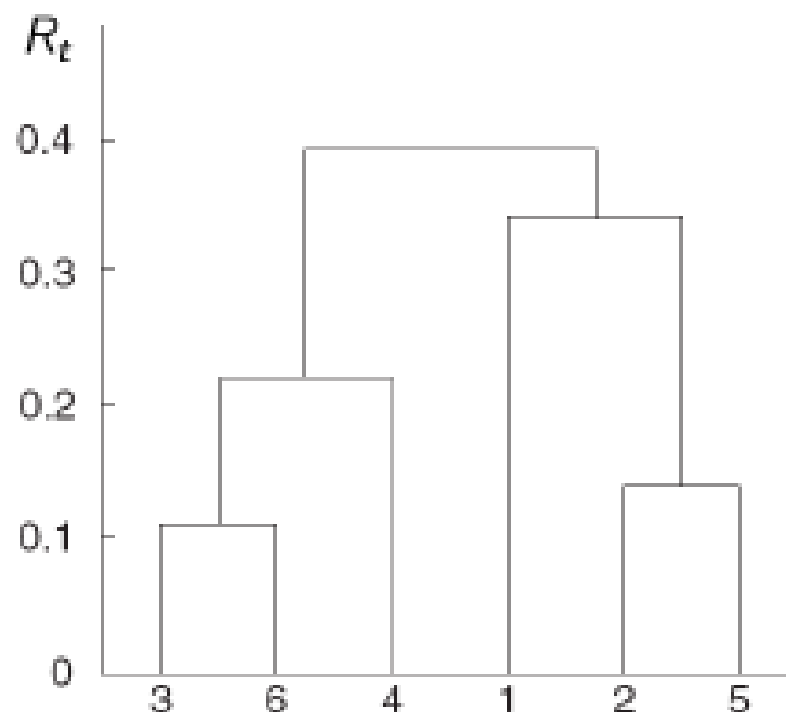


2. Расстояние дальнего соседа:

Диаграмма вложения

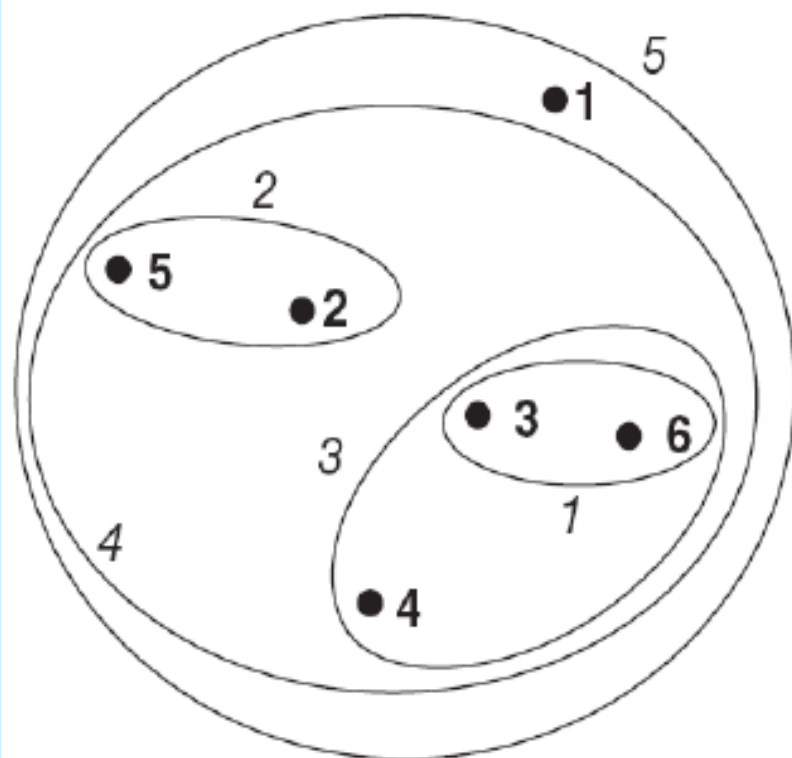


Дендрограмма

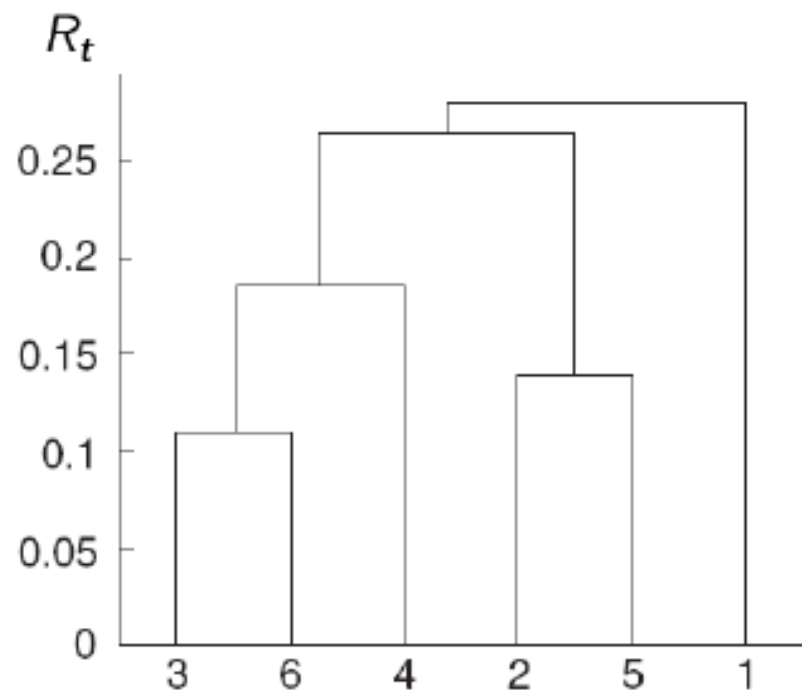


3. Групповое среднее расстояние:

Диаграмма вложения

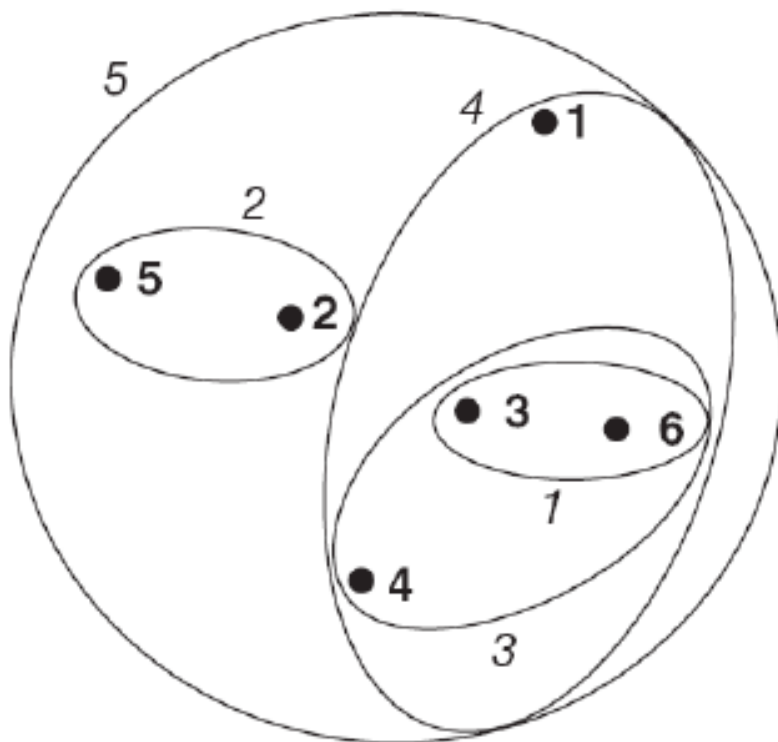


Дендрограмма

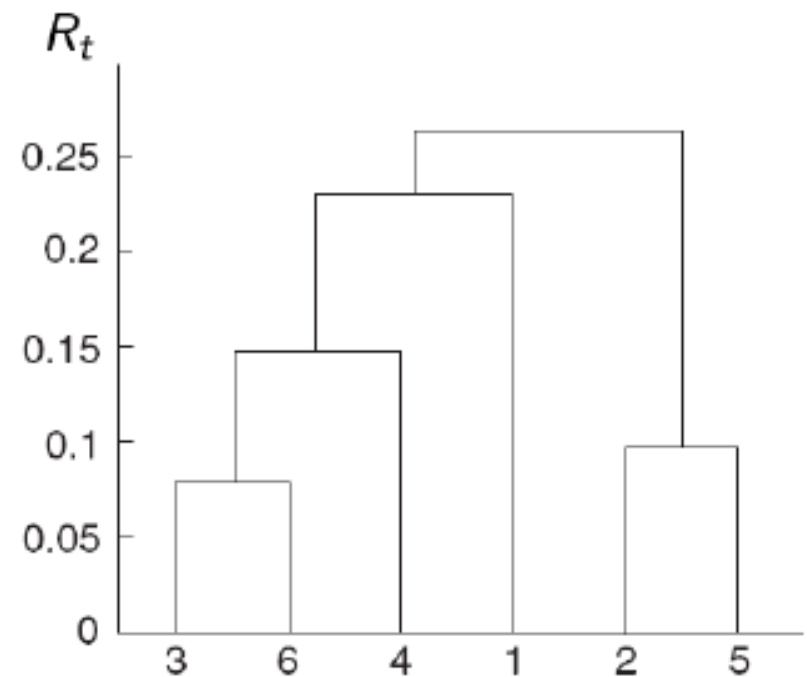


Расстояние Уорда:

Диаграмма вложения



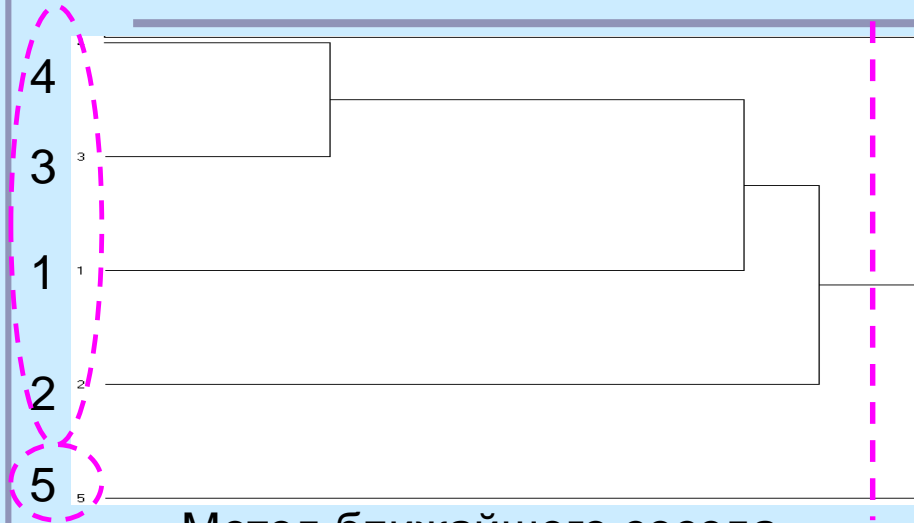
Дендрограмма



Советы специалистов

- Использовать метод Уорда (Варда)

ПРИМЕР



Метод ближайшего соседа

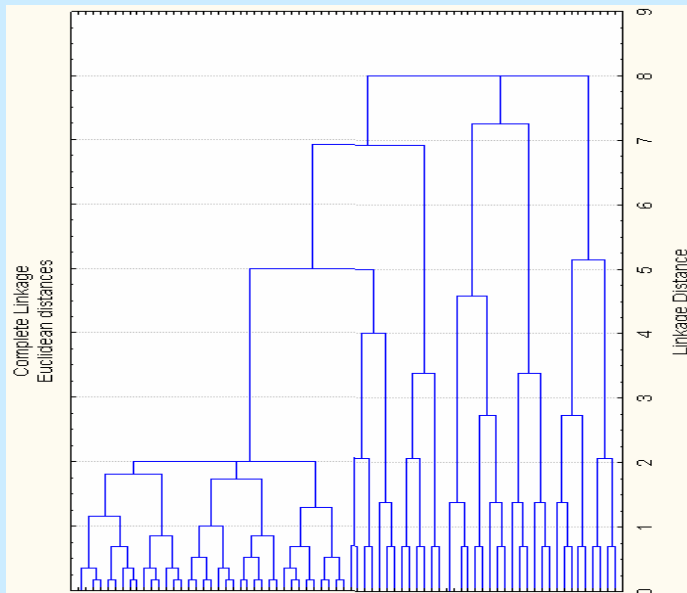
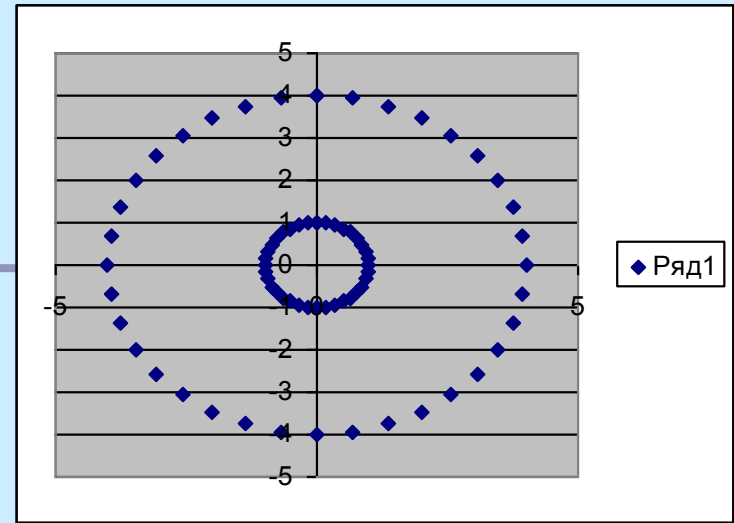
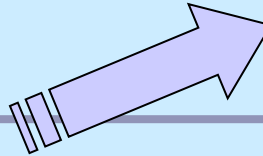


Метод дальнего соседа

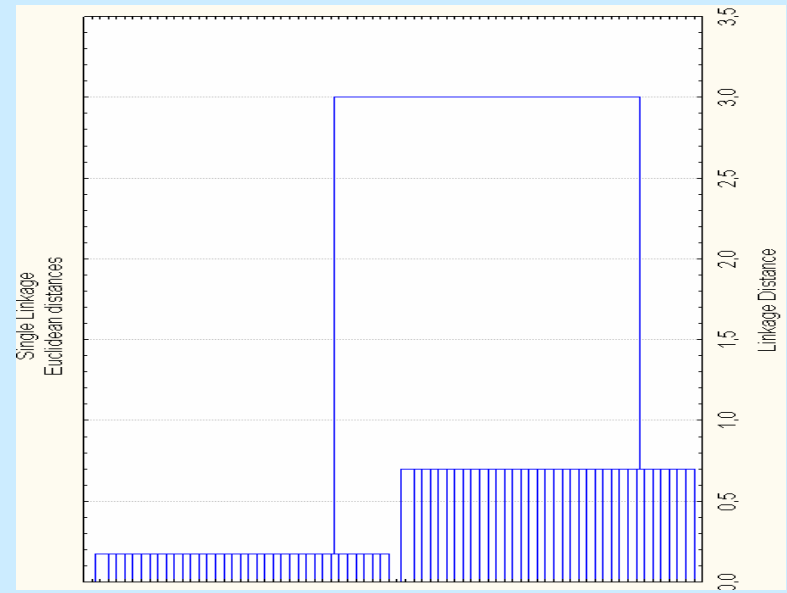
ДЕНДРОГРАММЫ

ПРИМЕР

ИСХОДНЫЕ ДАННЫЕ



МЕТОД ДАЛЬНОГО СОСЕДА



МЕТОД БЛИЖАЙШЕГО СОСЕДА

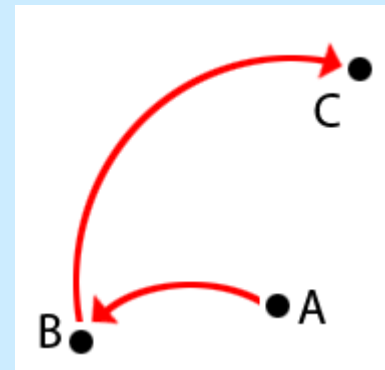
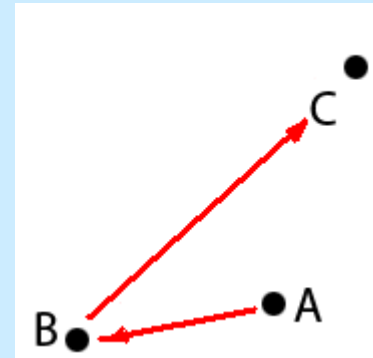
Выбор меры расстояния

РАССТОЯНИЕ - МЕТРИКА

Свойства расстояния

- $\rho(X_i, X_j) = 0 \Leftrightarrow X_i = X_j$
- $\rho(X_i, X_j) = \rho(X_j, X_i)$
- $\rho(X_i, X_j) \leq \rho(X_i, X_k) + \rho(X_k, X_j)$

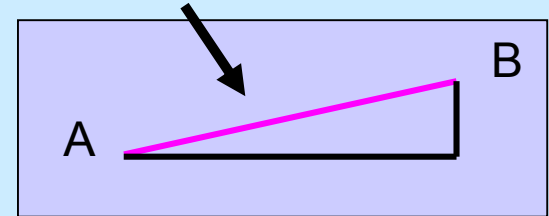
$$X_i = (X_{i1}, X_{i2}, \dots, X_{im})$$



Расстояние между объектами

- Евклидова метрика

$$\rho(X_i, X_j) = \sqrt{\sum_{k=1}^m (X_{ik} - X_{jk})^2}$$



X_{ik} – значение k -того признака i -того объекта

- Евклидова дистанция между двумя точками x и y — это наименьшее расстояние между ними. В двух- или трёхмерном случае — это прямая, соединяющая данные точки.

- Взвешенная евклидова метрика

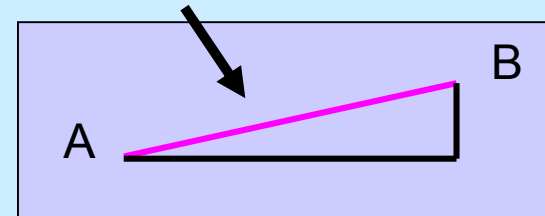
$$\rho(X_i, X_j) = \sqrt{\sum_{k=1}^m W_k (X_{ik} - X_{jk})^2}$$

W_k - вес k -того фактора

Расстояние между объектами

- Квадрат Евклидовой метрики

$$\rho(X_i, X_j) = \sum_{k=1}^m (X_{ik} - X_{jk})^2$$



X_{ik} – значение k -того признака i -того объекта

Используется, когда требуется придать большие веса более отдаленным друг от друга объектам.

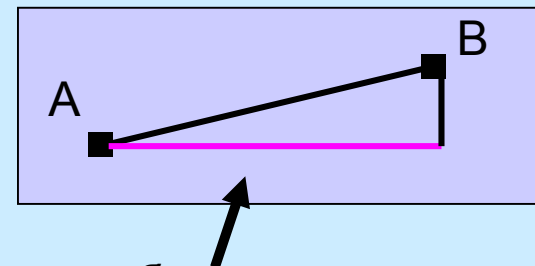
Эта мера должна всегда использоваться при построении кластеров при помощи центроидного и медианного методов, а также метода Варда.

Расстояние между объектами

- Расстояние Чебышева

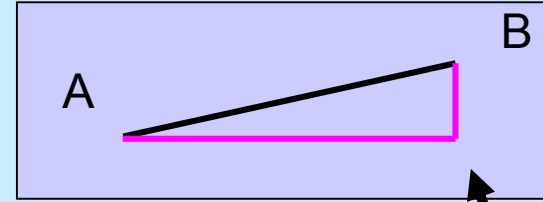
$$\rho(X_i, X_j) = \max_{k=1, \dots, m} |X_{ik} - X_{jk}|$$

Это расстояние используют, когда хотят определить два объекта как «различные», если они различаются по какой-либо одной координате (каким-либо одним измерением).



- Степенное расстояние

$$\rho(X_i, X_j) = \left[\sum_{k=1}^m |X_{ik} - X_{jk}|^p \right]^{1/r}$$



- **Хемингово расстояние, (городских кварталов, манхэттенское расстояние).**

$$\rho(X_i, X_j) = \sum_{k=1}^m |X_{ik} - X_{jk}|$$

В шутку называется “дистанцией таксиста”.

Для этой меры влияние отдельных больших разностей (выбросов) уменьшается (так как они не возводятся в квадрат).

Расстояние Махаланобиса — обобщение предыдущих расстояний

$$d_0 = (\mathbf{x}_k - \mathbf{x}_l)' \Sigma^{-1} (\mathbf{x}_k - \mathbf{x}_l)$$

Здесь k и l — номера объектов, $\mathbf{x}_k, \mathbf{x}_l$ — их векторы признаков, Σ — ковариационная матрица признаков

Основные характеристики

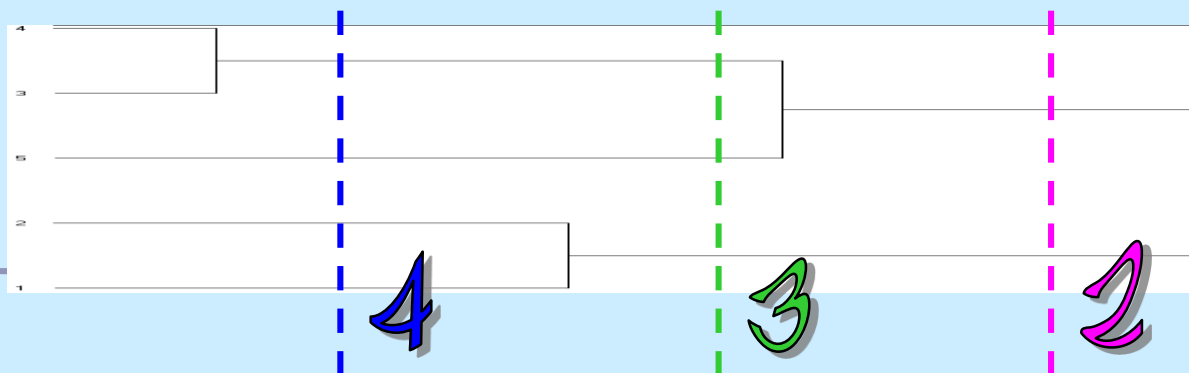
- Учитывает возможную корреляцию между переменными
- Если корреляция между переменными отсутствует, то расстояние Махаланобиса равно расстоянию Евклида



Prasanta Chandra Mahalanobis
1893 - 1972

Принятие решения о количестве кластеров

- Таким образом, в результате иерархического анализа получаем систему вложенных кластеров.
- Когда закончить разбиение на кластеры?
- Если число кластеров заранее известно, то классификацию заканчивают как только будет сформировано разбиение с этим числом кластеров. При неизвестном числе кластеров правило остановки связывают с понятием *порога*— это некоторое расстояние, определяемое условиями конкретной задачи.



ПРИМЕР

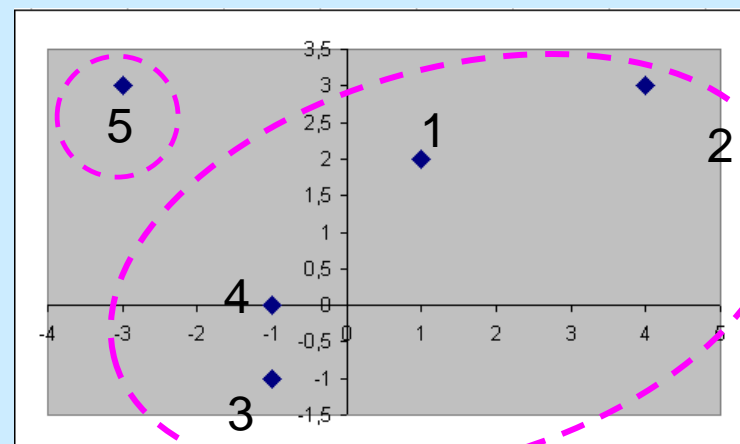
ФАКТОРЫ

ПЕРЕМЕННЫЕ

НАБЛЮДЕНИЯ
ОБЪЕКТЫ

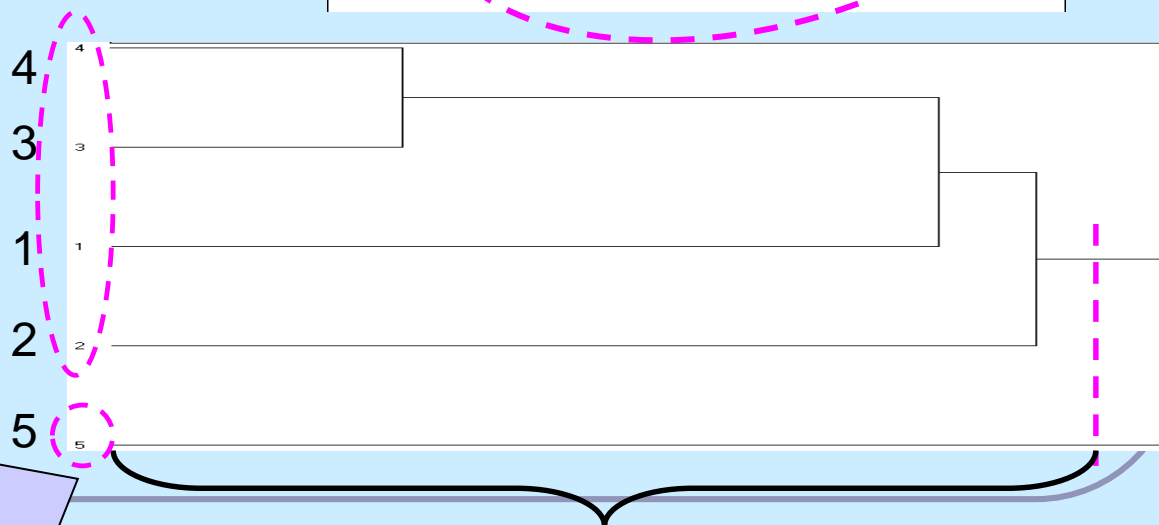
Data		
	COLUMN 1	COLUMN 2
1	1	2
2	4	3
3	-1	-1
4	-1	0
5	-3	3

ГРАФИЧЕСКАЯ
ИЛЛЮСТРАЦИЯ



Результат кластерного
анализа

ДЕНДРОГРАММА



РАССТОЯНИЕ

ПРИМЕР

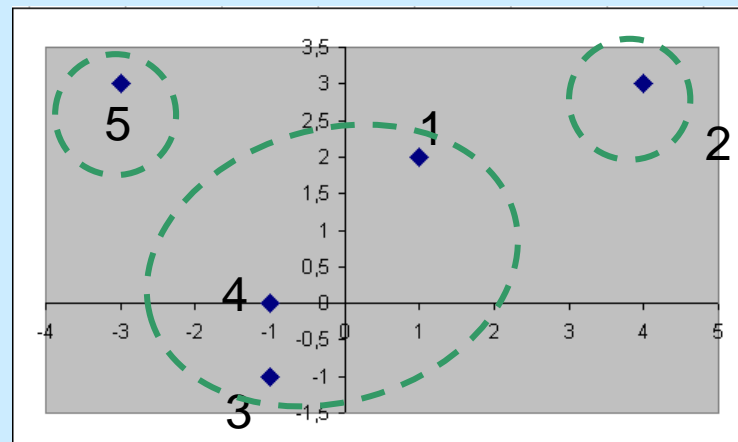
ФАКТОРЫ

ПЕРЕМЕННЫЕ

НАБЛЮДЕНИЯ
ОБЪЕКТЫ

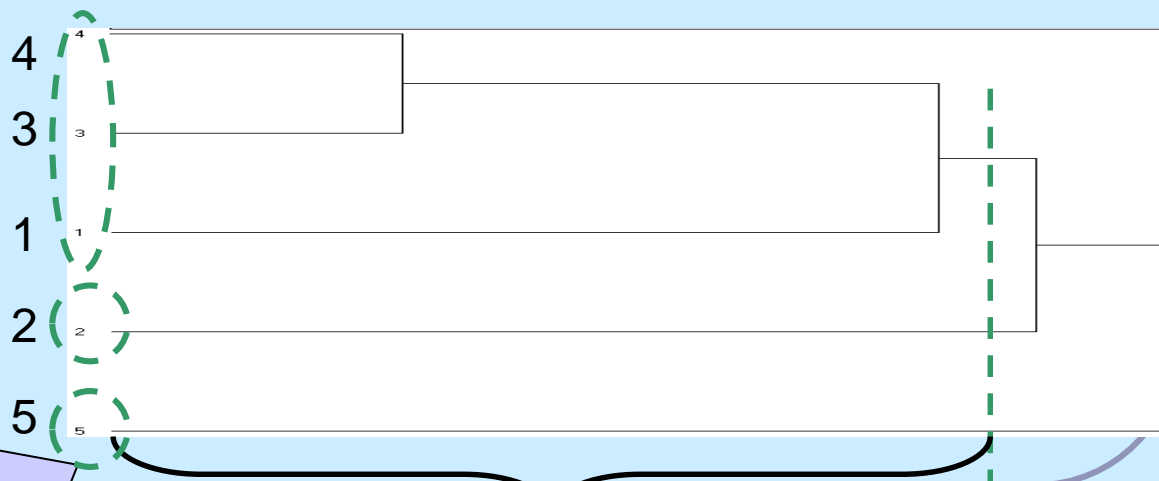
Data		
	COLUMN 1	COLUMN 2
1	1	2
2	4	3
3	-1	-1
4	-1	0
5	-3	3

ГРАФИЧЕСКАЯ
ИЛЛЮСТРАЦИЯ



Результат кластерного
анализа

ДЕНДРОГРАММА



РАССТОЯНИЕ

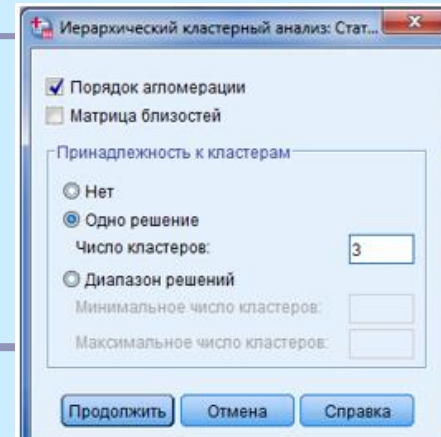
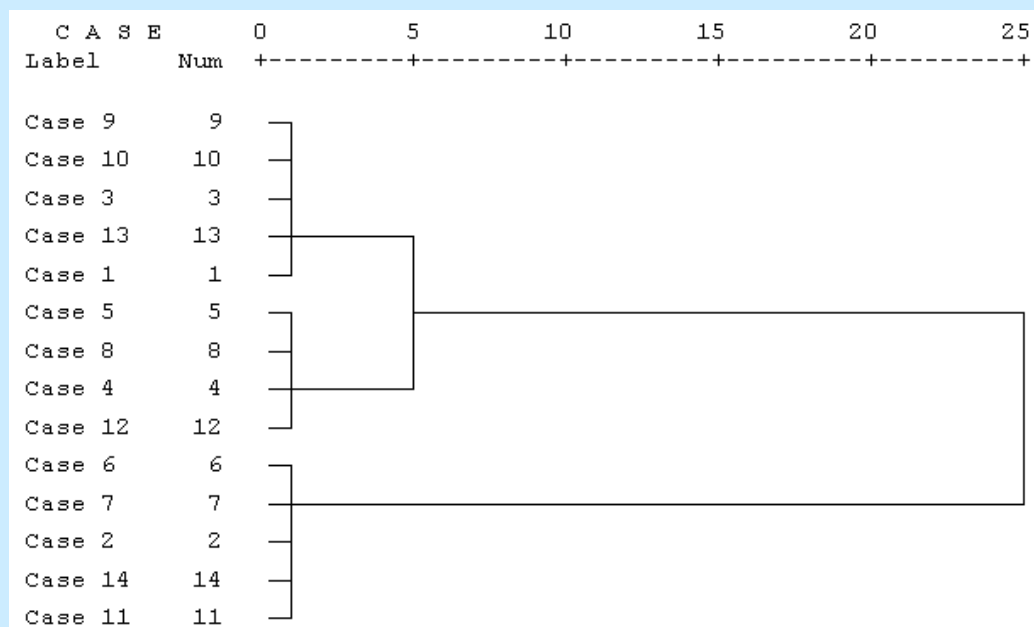
Таблица агломерации

- Оптимальное число кластеров =
число наблюдений – количество шагов до скачка

Таблица 13.2. Порядок агломерации

	Cluster Combined		Coefficients
	Cluster 1	Cluster 2	
1	9	10	,000
2	2	14	1,461E-02
3	3	9	1,461E-02
4	5	8	1,461E-02
5	6	7	1,461E-02
6	3	13	3,490E-02
7	2	11	3,651E-02
8	4	5	4,144E-02
9	2	6	5,118E-02
10	4	12	,105
11	1	3	,120
12	1	4	1,217
13	1	2	7,516

Пример: $14 - 12 = 2$

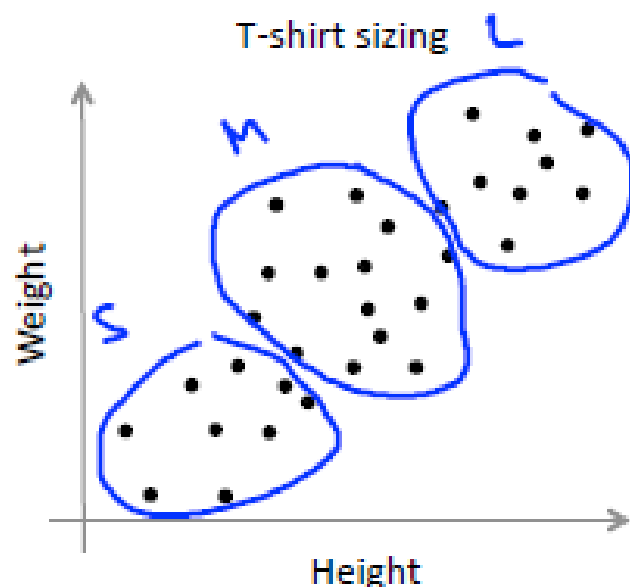


Choosing the value of K

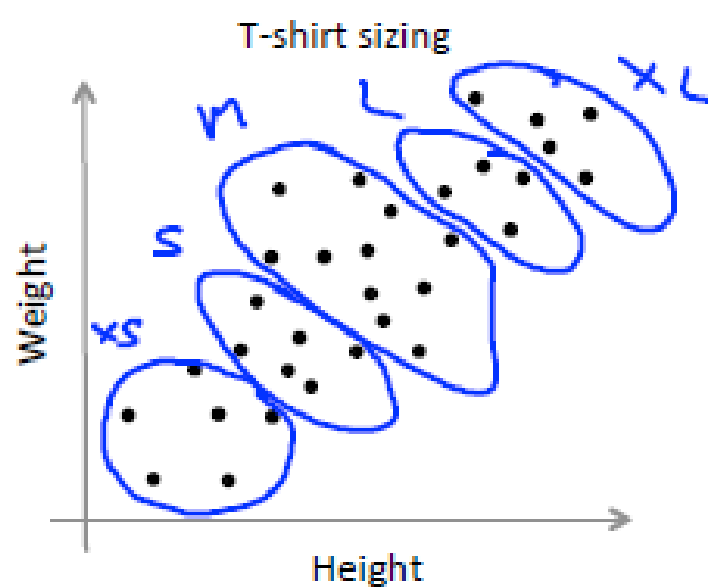
Sometimes, you're running K-means to get clusters to use for some later/downstream purpose. Evaluate K-means based on a metric for how well it performs for that later purpose.

$K=3$ S, M, L

E.g.



$K=5$ XS, S, M, L, XL



Andrew Ng

-
- Иерархические методы используются обычно в таких задачах классификации небольшого числа объектов (порядка нескольких десятков), где больший интерес представляет не число кластеров, а анализ структуры множества этих объектов и наглядная интерпретация проведенного анализа в виде дендрограммы.
 - Если же число кластеров заранее задано или подлежит определению, то для классификации чаще всего используют *параллельные* кластер-процедуры
 - *итерационные алгоритмы, на каждом шаге которых используется одновременно (параллельно) все наблюдения*
 - *Метод K-средних («K-Means Clustering») с заранее заданным числом классов.*


МЕТОД K -средних

Дана случайная выборка из N точек (наблюдений, Cases), каждая из которых имеет m признаков (переменных, Variables).

Требуется найти K центров, представляющих кластеры в N точках ($K < N$) так, чтобы каждая из N точек относилась ровно к одному из K кластеров и центр каждого кластера совпадал с центром тяжести относящихся к нему точек.

K – задано!

Метод К-средних – итерационный метод

- 
- Шаг1. За центры искоемых кластеров $Z_1(1)$, $Z_2(1)$,... $Z_k(1)$, принимают случайно выбранные наблюдения, обычно это K первых точек.
- Шаг2. Для каждой из оставшихся точек находят ее расстояние до центров кластеров и точку относят к тому кластеру, расстояние до которого минимально.
- Шаг3. Рассчитывают новые центры тяжести кластеров, так чтобы сумма квадратов расстояний между всеми элементами, принадлежащими кластеру и новым центром кластера должна быть минимальна.
- Шаг4. Если пересчет центров тяжести практически не приводит к изменению кластеров, процедуру заканчивают, иначе повторяют процедуру, начиная с Шага 2.

Метод k -средних (k -means)

$$X = \mathbb{R}^n.$$

1: начальное приближение центров μ_y , $y \in Y$;

2: **повторять**

3: отнести каждый x_i к ближайшему центру:

$$y_i := \arg \min_{y \in Y} \rho(x_i, \mu_y), \quad i = 1, \dots, \ell;$$

4: вычислить новые положения центров:

$$\mu_{yj} := \frac{\sum_{i=1}^{\ell} [y_i = y] f_j(x_i)}{\sum_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad j = 1, \dots, n;$$

5: **пока** y_i не перестанут изменяться;

Критерии качества классификации.

1. Сумма квадратов расстояний до центров классов:

$$F_1 = \sum_l \sum_i d^2(X_i, \bar{X}_l),$$

где l - номер кластера;

\bar{X} - центр l -го кластера;

X_i - вектор значений переменных для i -го объекта в l -ом кластере;

$d(X_i, \bar{X}_l)$ - расстояние между i -ом объектом и центром l -го кластера.

2. Сумма внутриклассовых расстояний между объектами:

$$F_2 = \sum_l \sum_{\bar{ij}} d^2_{\bar{ij}}$$

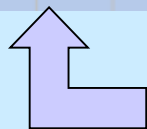
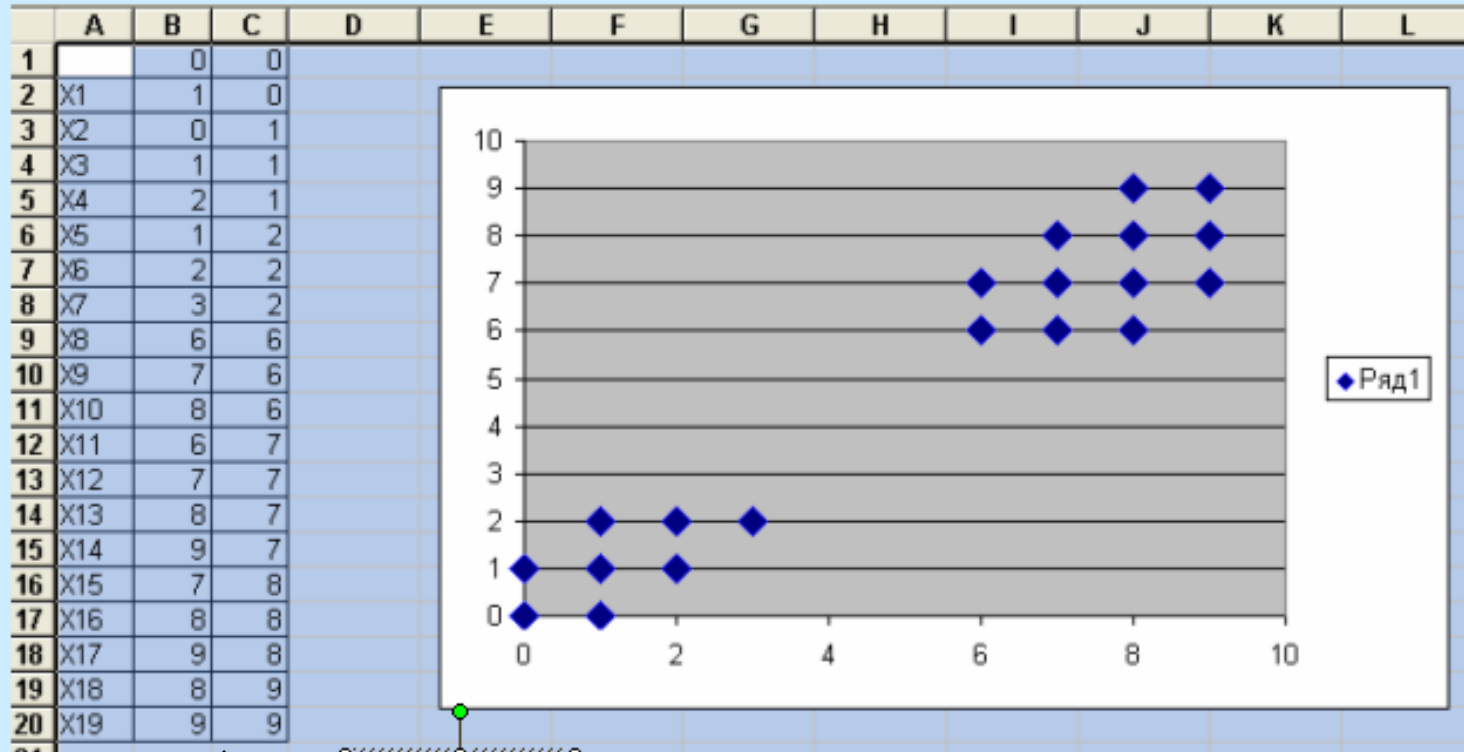
3. Суммарная внутриклассовая дисперсия:

$$F_3 = \sum_l \sum_j \sigma_{lj}^2,$$

где σ_{lj}^2 - дисперсия j -ой переменной в кластере S_l

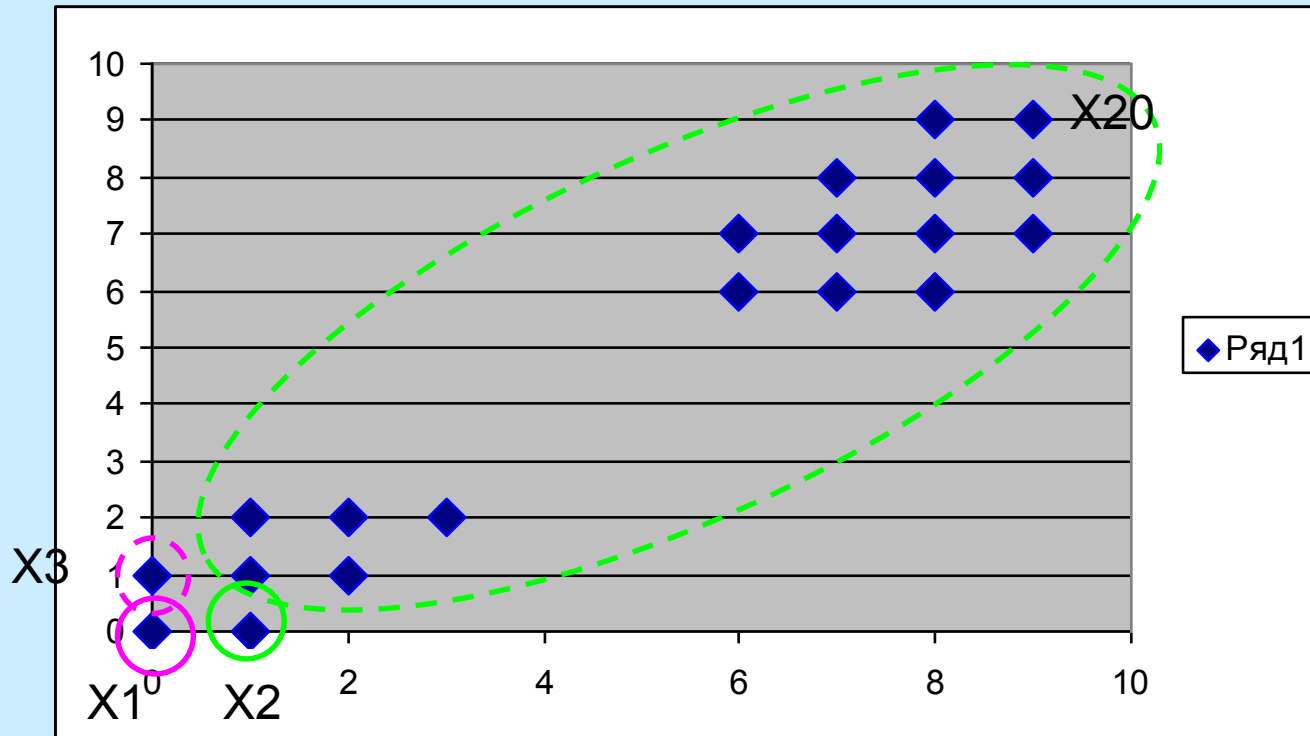
ПРИМЕР

(20 объектов, 2 переменные)



КООРДИНАТЫ КЛАСТЕРОВ

Пусть $K=2$

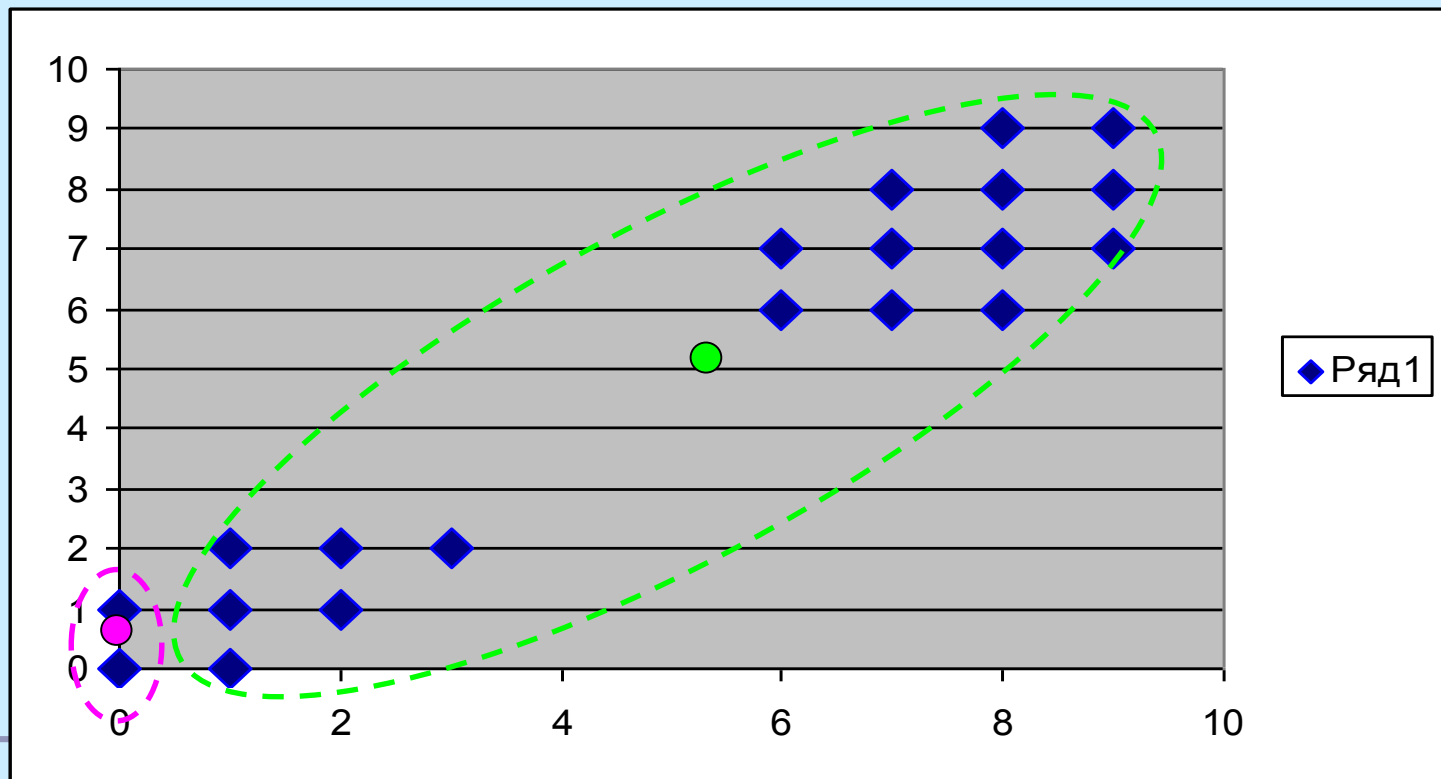


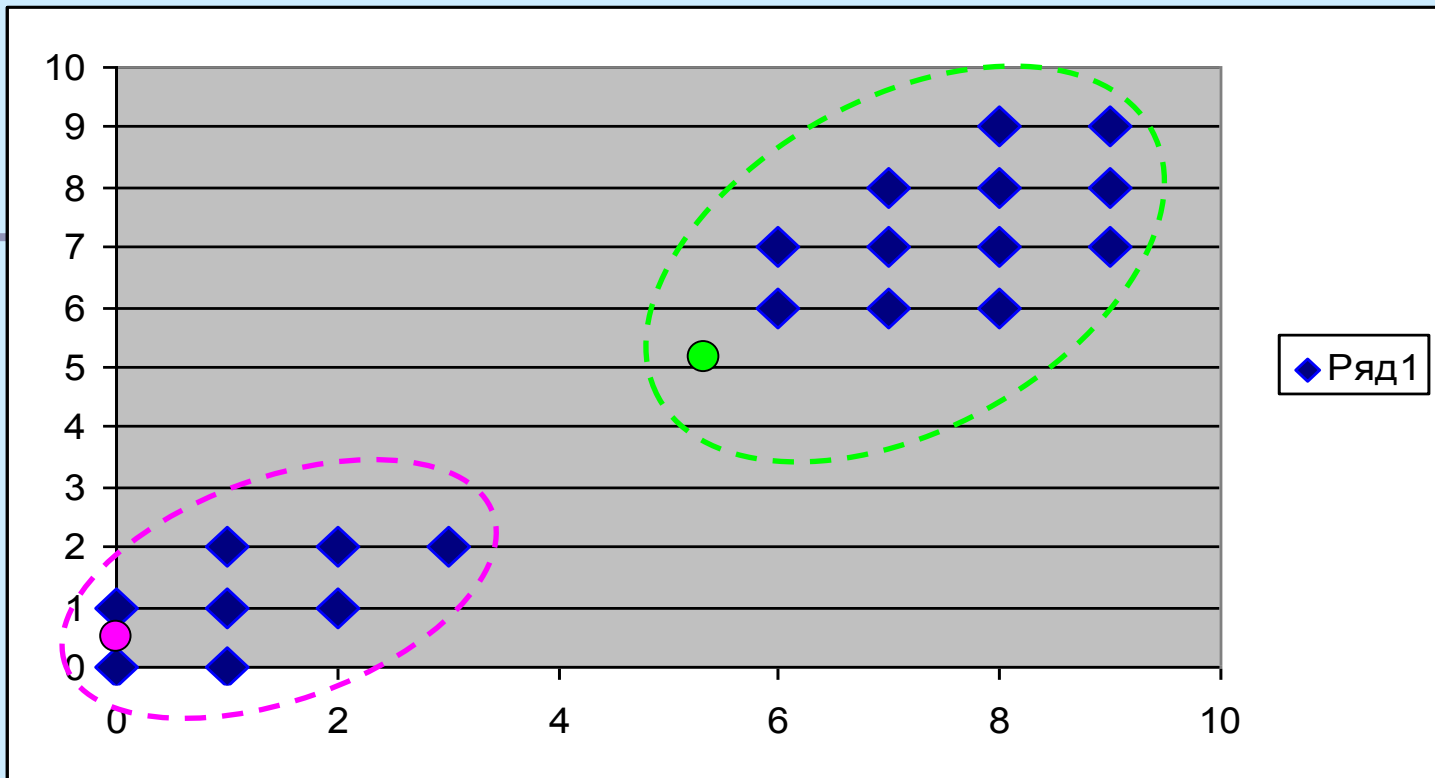
На Шаг1 в качестве кластеров берутся два первых элемента.

На Шаг 2 ищутся ближайшие к ним. $\{X1, X3\}$ и $\{X2, X4, X5, \dots, X20\}$

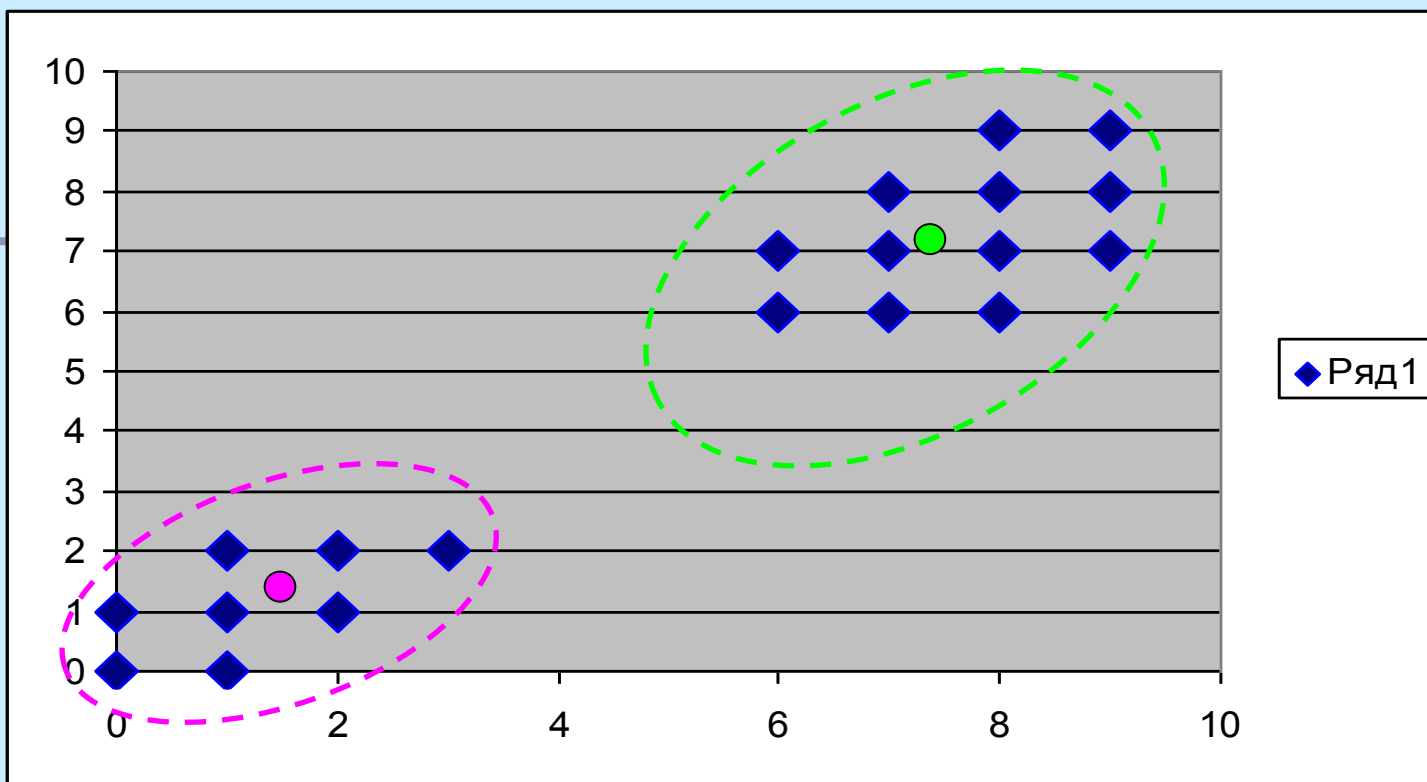
Шаг 3. Коррекция центров кластеров.

- Центр первого $Z1(2)=(0, 0.5)$,
- Центр второго $Z2(2)=(X2+X4+..)/18=(5.67, 5.33)$





- На следующей итерации формируем новые кластеры, группируя их по близости к новым центрам.
- Получим два кластера $\{x_1, x_2, \dots, x_8\}$ и $\{x_9, x_{10}, \dots, x_{20}\}$
- Рассчитываем новые центры кластеров.



- На следующей итерации сформируем новые кластеры, группируя их по близости к новым центрам.
- Новый пересчет расстояний дает те же результаты, процесс можно закончить.

K – means (k – средних)

- Недостатки:
 - Чувствительность к выбору начального приближения
 - Необходимость задавать k
- Как преодолеть:
 - Провести несколько случайных кластеризаций
 - Постепенно наращивать k.

Интерпретация и профилирование кластеров

- Проводиться исследование полученных кластеров
- Построение портрета прецедентов
- Исследования кластера методами дисперсионного анализа

Оценка достоверности кластеризации

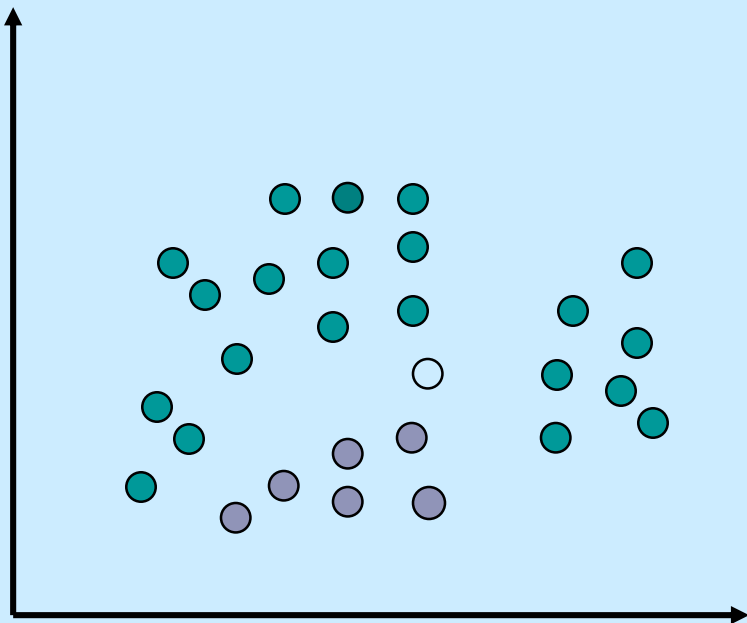
- Многие методы кластерного анализа – довольно простые эвристические процедуры, как правило не имеют достаточного статистического обоснования.
- Различные методы могут порождать различные решения для одних и тех же данных.
- Формальные процедуры оценки надежности и достоверности результатов достаточно сложны.

ЧТО ДЕЛАТЬ?

- Использовать различные способы измерения расстояния.
- Использовать различные методы кластеризации.
- Выполнить кластерный анализ по сокращенному набору переменных.
- Разбить данные на две части случайным образом.

СРАВНИТЬ РЕЗУЛЬТАТЫ

Проблемы



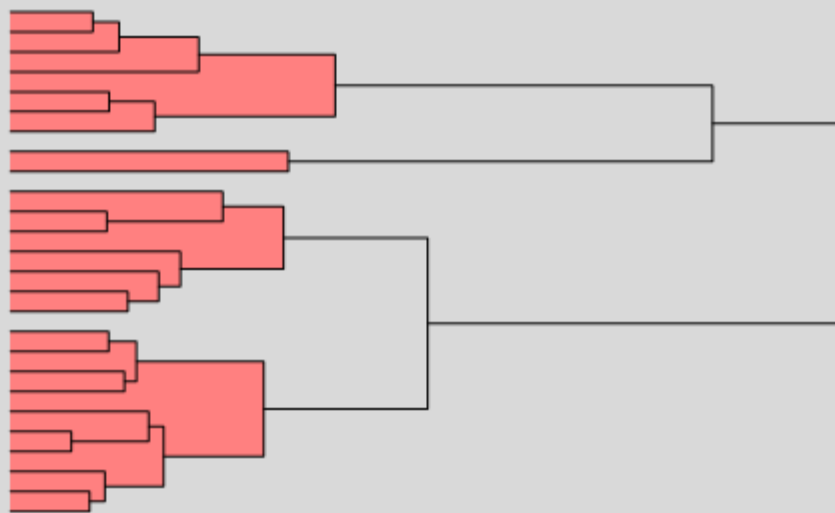
ПРИМЕР WM

КЛАСТЕРНЫЙ АНАЛИЗ

Иерархический кластерный анализ

```
Clear[distanceMatrix]
Needs["HierarchicalClustering`"]
(distanceMatrix = DistanceMatrix[stdata, DistanceFunction → EuclideanDistance]) //
  MatrixForm;
Clear[clust]
clust = DirectAgglomerate[distanceMatrix, countries, Linkage → "Ward"];
```

```
plot = DendrogramPlot[clust, Orientation → Right, HighlightLevel → 4,
  HighlightStyle → Pink]
```



Получили 4 кластера, но сразу можно увидеть, что они не очень однородные по количеству стран - есть один кластер, который довольно маленький (2 страны).

Метод к - средних 4 или 3 кластера (хочется убрать кластер, который состоит из 2 стран).

```
clustKmeans = FindClusters[stdata, 4, Method -> "KMeans"];  
clustK1 = clustKmeans[[1];  
clustK2 = clustKmeans[[2];  
clustK3 = clustKmeans[[3];  
clustK4 = clustKmeans[[4];  
Map[Length, {clustK1, clustK2, clustK3, clustK4}]  
  
{14, 5, 1, 6}
```

Если разбивать на 4 кластера, то группы получаются достаточно неоднородные по количеству, попробуем 3.

```
clustKmeans2 = FindClusters[stdata, 3, Method -> "KMeans"];  
clustK12 = clustKmeans2[[1];  
clustK22 = clustKmeans2[[2];  
clustK32 = clustKmeans2[[3];  
Map[Length, {clustK12, clustK22, clustK32}]  
  
{17, 2, 7}
```

Сначала 3 кластера:

```
pos1 = Flatten@Table[Position[stdata, i], {i, clustK12}];  
pos2 = Flatten@Table[Position[stdata, i], {i, clustK22}];  
pos3 = Flatten@Table[Position[stdata, i], {i, clustK32}];  
country1 = Table[countries[[i]], {i, pos1}]  
country2 = Table[countries[[i]], {i, pos2}]  
country3 = Table[countries[[i]], {i, pos3}]
```

```
{Belgium, Denmark, France, W_Germany, Ireland,  
Italy, Luxembourg, Netherlands, United_Kingdom, Austria,  
Finland, Greece, Norway, Portugal, Spain, Sweden, Switzerland}
```

```
{Turkey, Yugoslavia}
```

```
{Bulgaria, Czechoslovakia, E_Germany, Hungary, Poland, Rumania, USSR}
```


1 кластер : средние уровни занятости по большинству отраслей. Низкая занятость в сельском хозяйстве и в горнодобывающей промышленности, самая высокая занятость в сфере услуг и соц. служб.

2 кластер: самый высокий уровень занятости в сельском хозяйстве и финансовом секторе (такое бывает?). Самая низкая занятость в таких отраслях, как производство, энергетика, строительство, услуги, соц. службы и транспорт и связь.

3 кластер: страны, у которых наблюдается наибольший уровень занятости в таких отраслях, как: горнодобывающая, производство, энергетика, строительство, транспорт и связь. По остальным отраслям страны занимают срединное значение.

```
ListLinePlot[Mean /@ Table[clustKmeans2[[i]], {i, 1, 3}],  
  PlotLegends → {"1 кластер", "2 кластер", "3 кластер"},  
  PlotLabel → "Средние значения кластеров по всем признакам"]
```

