



Машинное обучение

DS-поток

Лекция 1

Кластеризация



Обучение без учителя

Обучение с учителем (supervised learning):

x_1, \dots, x_n — объекты.

Y_1, \dots, Y_n — таргет.

Требуется научиться предсказывать таргет по объектам.

- ▶ Задаем множество моделей и функционал ошибки.
- ▶ Обучение — выбор лучшей модели с точки зрения функционала.

Обучение без учителя (unsupervised learning):

x_1, \dots, x_n — объекты.

Отсутствует таргет.

Требуется исследовать данные на наличие внутренней структуры.



Кластеризация

Задача кластеризации

Работа в командах

Определяемся с требованиями

Метрики качества

K-means

DBSCAN

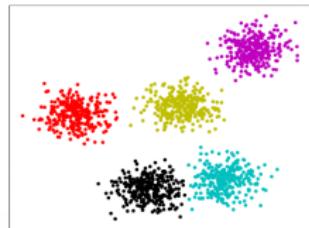
Иерархическая кластеризация

Постановка задачи кластеризации

Дана выборка объектов $X = (x_1, \dots, x_n)$.

Задача кластеризации:

выявить в данных K кластеров.



Кластер может быть:

▶ *Подвыборкой*

Построить правило $f : X \rightarrow \{1, \dots, K\}$,
определяющее номер кластера только для объектов выборки.

▶ *Областью пространства*

Построить правило $f : \mathcal{X} \rightarrow \{1, \dots, K\}$,
определяющее номер кластера для любых объектов пр-ва \mathcal{X} .

▶ *Нежестким*

Построить правило $f(x) = (p_1, \dots, p_K)$,
определяющее распределение объекта по кластерам,
где $p_k : \mathcal{X} \rightarrow [0, 1]$ — вероятность принадлежности x к класт. k .

Число K может быть известно заранее, т.е. гиперпараметр.

Цели кластеризации

- ▶ Упростить дальнейшую обработку данных.
Разбить выборку X на группы схожих объектов и работать с каждой группой отдельно.
- ▶ Сократить объем хранимых данных.
Например, оставить лишь по одному представителю из каждого кластера.
- ▶ Выделить нетипичные объекты.
Объекты, которые не подходят ни к одному из кластеров.
- ▶ Использовать для разбиения данных на группы.
*Аналог классификации в случае, если нет целевых меток.
Например, можно кластеризовать клиентов и разным группам предлагать разные услуги.*

Пример: кластеризация пользователей

Цель: выделить кластеры схожих по поведению пользователей.

Данные:

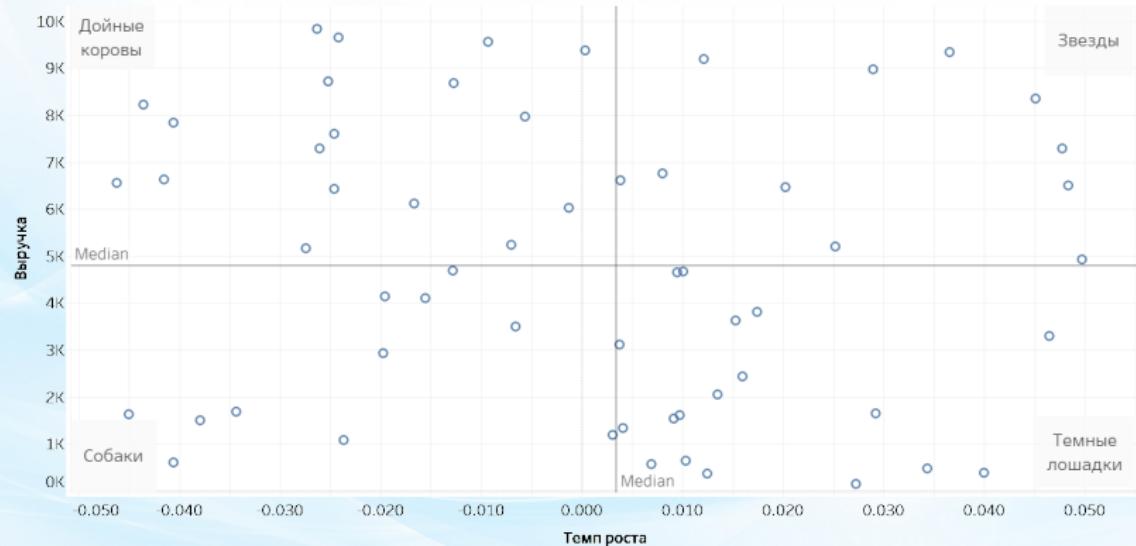
- ▶ Номер карты лояльности
- ▶ Фамилия Имя
- ▶ Пол
- ▶ Возраст
- ▶ Город, регион
- ▶ Номер телефона
- ▶ Вид деятельности
- ▶ Дата регистрации в программе лояльности
- ▶ Сумма покупок
- ▶ Частота посещений
- ▶ Средний чек
- ▶ Часто покупаемые категории продуктов

Полученные кластеры можно проинтерпретировать, проанализовав отличие по признакам. Например, *мужчины в возрасте от 30 до 40, посещают в среднем 5 раз в неделю, чаще всего покупают готовую еду*. Этот кластер скорее всего характеризует офисных работников, которые посещают супермаркет в обеденный перерыв.

Пример: кластеризация пользователей

В качестве бейзлайна можно использовать простую сегментацию пользователей по порогам нескольких признаков.

Матрица BSG





Пример: кластеризация пользователей

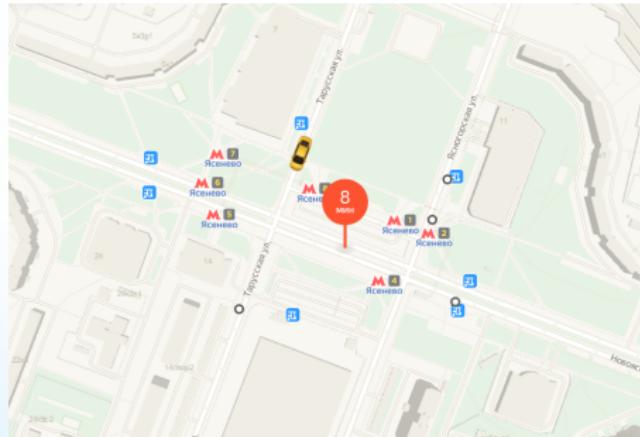
Эти 4 сегмента обычно интерпретируют следующим образом:

- ▶ **Звезды** — крупнейшие клиенты с высоким темпом роста выручки. Это сильная точка роста, и им нужно уделять наибольшее внимание.
- ▶ **Дойные коровы** — крупные клиенты, с низкими или отрицательными темпами выручки. Эти клиенты будут формировать ядро вашей текущей выручки. Проглядите коров и потеряете бизнес.
- ▶ **Темные лошадки** — пока мелкие клиенты, но с большим темпом роста. Это группы клиентов, на кого надо обращать внимание, т.к. они могут вырасти до звезд или дойных коров.
- ▶ **Собаки** — мелкие клиенты с низкими или отрицательными темпами роста. Это клиенты, кому можно уделять наименьшее внимание и применять к ним массовые методы обслуживания.

Пример: точки посадки в такси

Цель:

определить наиболее удобные точки, где пассажир садится в такси.



Возможные проблемы:

- ▶ Неточность GPS-сигнала — в некоторых случаях погрешность может составлять 100 метров и более.
- ▶ Водитель может отметить начало поездки в приложении не сразу.

Работа в командах

Задача № 0. Кластеризоваться в команды по 4-5 человек.

Задача: время 10 минут

Имеется несколько тысяч супермаркетов.

Цель: кластеризовать магазины по похожести,
внутри кластеров выявить успешные магазины,
определить для остальных магазинов в кластере точки роста.

Данные:

- ▶ Продажи и выручка за каждый день, в т.ч. по отдельным товарам и категориям
- ▶ Потери от списаний
- ▶ Площадь торгового зала
- ▶ Количество сотрудников
- ▶ Торговые центры рядом
- ▶ Другие продуктовые магазины рядом
- ▶ Плотность и кол-во населения
- ▶ Городской трафик
- ▶ Образов. учреждения рядом

Вопросы: как кластеризовать, какие требования к модели, как измерить качество?

Типы кластерных структур



Шарообразные



Ленточные



Соединяются перемычками



Разреженный фон
из шумовых объектов

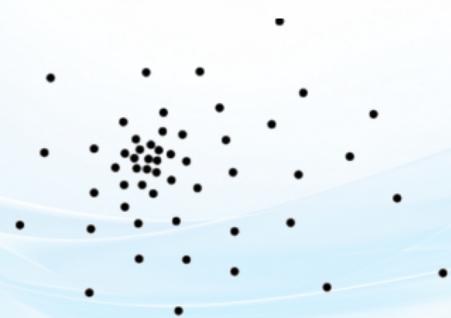
Типы кластерных структур



Могут перекрываться



Сходство не по расстоянию



Кластеры отсутствуют

Неоднозначность задачи кластеризации

Решение задачи кластеризации неоднозначно:

- ▶ Точной постановки задачи кластеризации нет.
- ▶ Существует множество критериев качества.
- ▶ В реальных задачах метрикой качества кластеризации не редко оказывается моральная удовлетворенность заказчика
- ▶ Число кластеров K обычно не известно заранее.
- ▶ Результат зависит от выбора метрики расстояния (схожести) между объектами.

Пример: Сколько здесь кластеров?



Задача: время еще 10 минут

Имеется несколько тысяч супермаркетов.

Цель: кластеризовать магазины по похожести,
внутри кластеров выявить успешные магазины,
определить для остальных магазинов в кластере точки роста.

Данные:

- ▶ Продажи и выручка за каждый день, в т.ч. по отдельным товарам и категориям
- ▶ Потери от списаний
- ▶ Площадь торгового зала
- ▶ Количество сотрудников
- ▶ Торговые центры рядом
- ▶ Другие продуктовые магазины рядом
- ▶ Плотность и кол-во населения
- ▶ Городской трафик
- ▶ Образов. учреждения рядом

Вопросы: как кластеризовать, как измерить качество?

Презентация результатов



Кластеризация

Задача кластеризации

Работа в командах

Определяемся с требованиями

Метрики качества

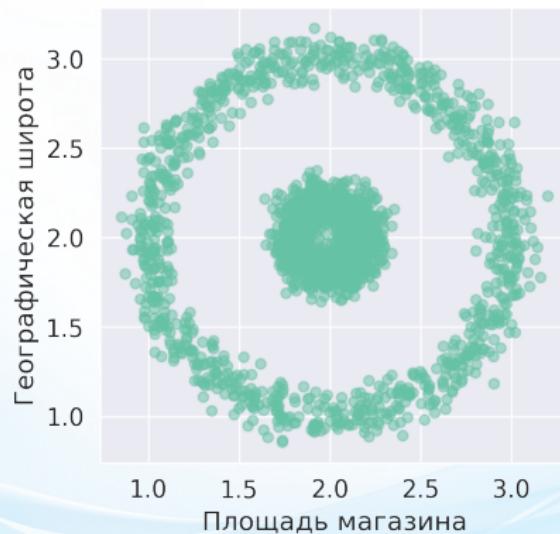
K-means

DBSCAN

Иерархическая кластеризация

Какие требования к форме кластеров?

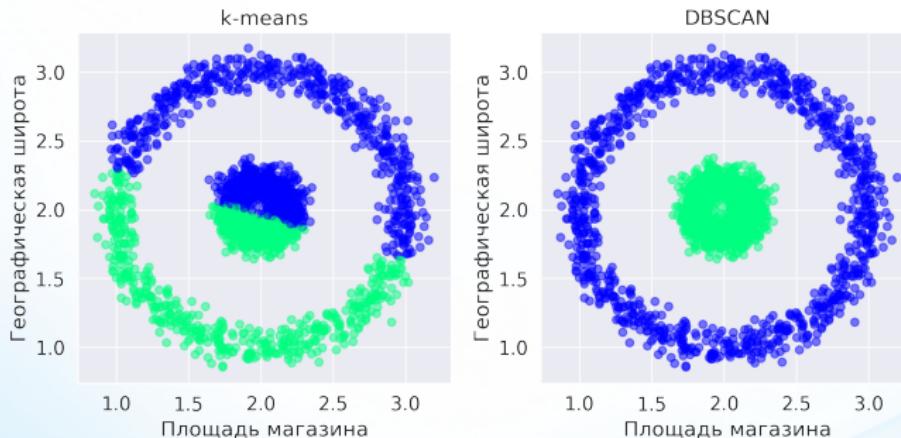
Пусть данные выглядят так:



Сколько здесь кластеров? Какие они?

Какие требования к форме кластеров?

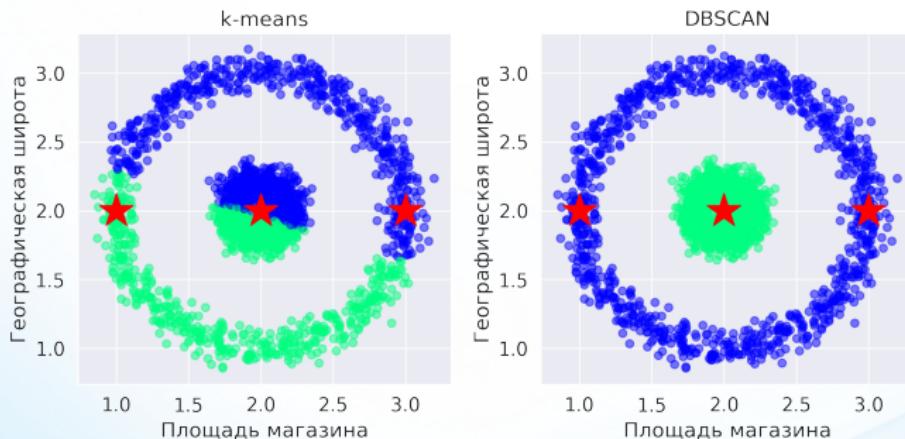
Применим два популярных метода для кластеризации на 2 кластера:



Какой лучше и почему?

Какие требования к форме кластеров?

Применим два популярных метода для кластеризации на 2 кластера:



Какой лучше и почему?

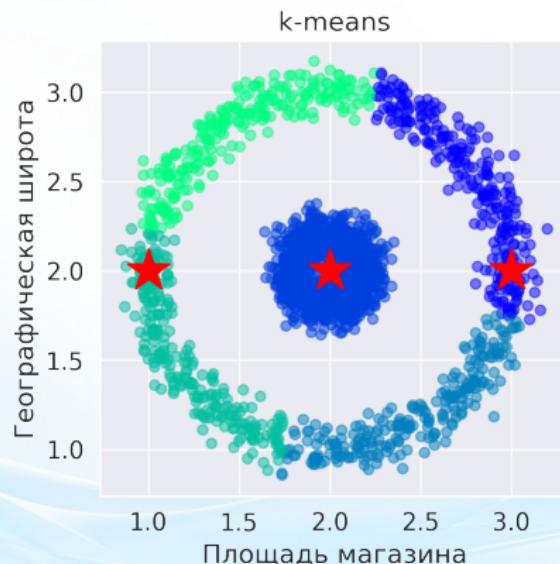
Для ответа на этот вопрос рассмотрим три магазина разной площади на одной широте.

Видим, что второй метод дает неинтерпретируемый результат: магазины с площадью 1 и 3 лежат в одном кластере, а магазин с площадью 2 — в другом.

Какие требования к форме кластеров?

Достаточное ли количество кластеров?

Возьмем больше для первого метода:



Кажется, так лучше.

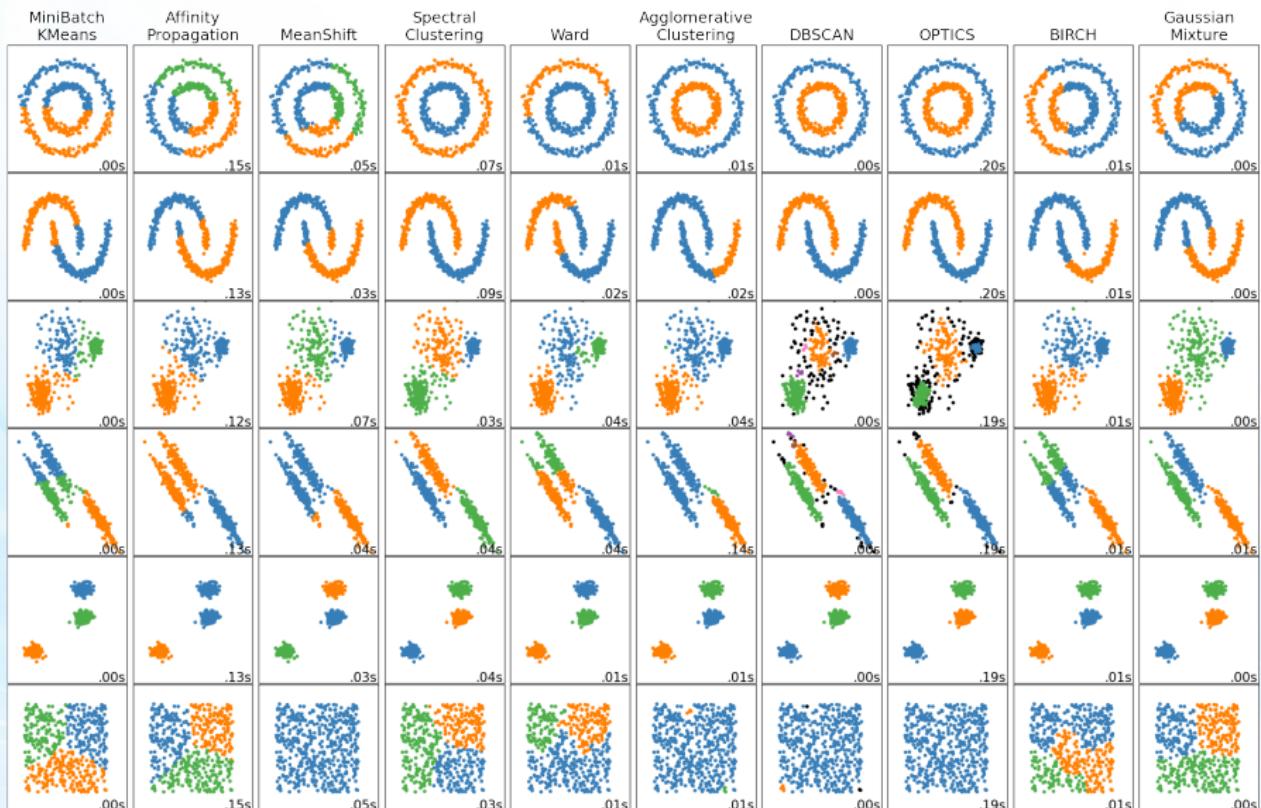


Какие требования к форме кластеров?

Итог:

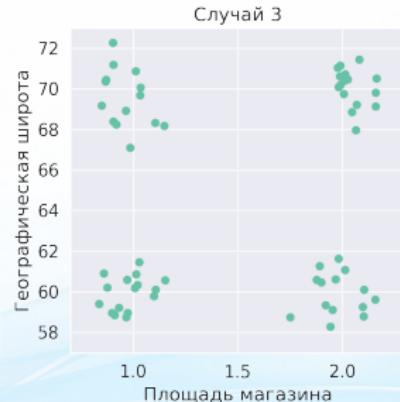
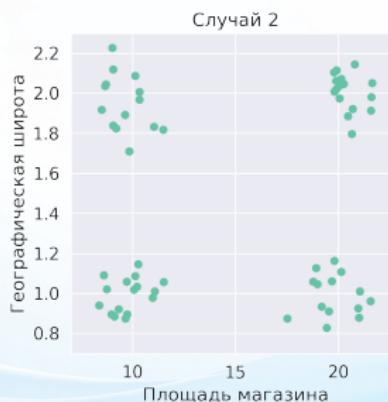
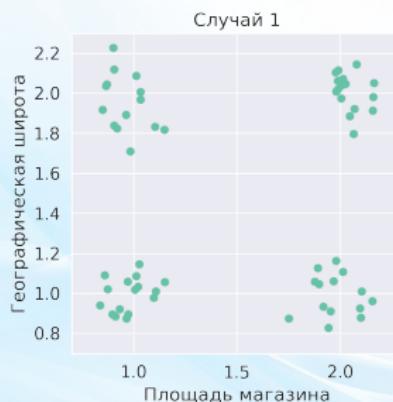
- ▶ Нужно понимать, какую форму кластера можно получить в зависимости от применяемого метода.
- ▶ Для решаемой задачи нужно понять, каким свойством должен обладать кластер.
- ▶ Если нужны интерпретируемые для бизнеса кластеры, то они обязательно должны быть выпуклыми.
Желательно также получать более компактные кластеры, нежели сильно растянутые.
- ▶ Если же кластеризация используется какой-либо ML-моделью, то форма кластера может быть не сильно важна, стоит следить за качеством ML-модели.

Сравнение разных методов



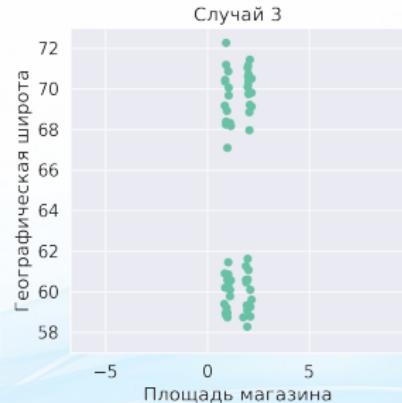
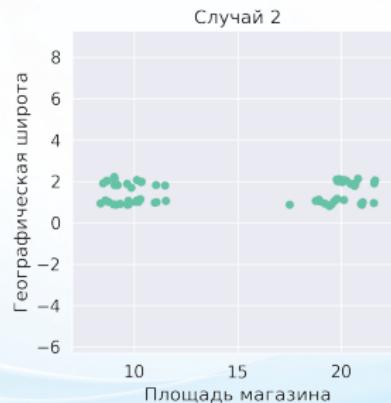
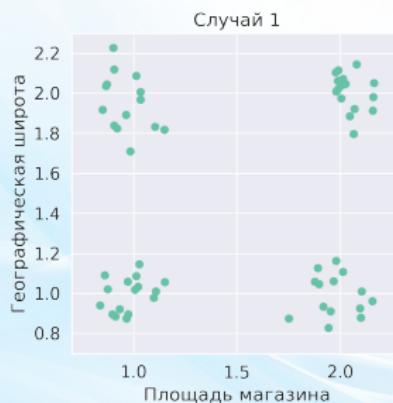
Подумаем еще

Посмотрим на три набора данных. В чем их отличие?



Подумаем еще

Как на них посмотрит метод кластеризации?



Масштабирования и расстояния

Итог:

- ▶ Результат кластеризации сильно зависит от используемой метрики (функции расстояния).
- ▶ Если предполагается ручной анализ кластеров, то не стоит выбирать неинтерпретируемые метрики и выполнять неинтерпретируемые преобразования признаков.
- ▶ В простом случае стоит выполнять стандартизацию признаков. Иначе результат кластеризации во многом будет определяться признаком, который имеет самый большой диапазон значений.
- ▶ В идеале стоит подумать, какие признаки более значимы. Например, потребовать, чтобы изменения были сопоставимыми:
 1. Площадь магазина увеличилась на $10\ m^2$ при той же широте
 2. Широта магазина увеличилась на x при неизменной площади

Исходя из желаемого значения x расставить веса признакам после стандартизации.



Кластеризация

Задача кластеризации

Работа в командах

Определяемся с требованиями

Метрики качества

K-means

DBSCAN

Иерархическая кластеризация

Расстояния внутри и между кластерами

Пусть задана функция расстояния между объектами — $\rho(x_1, x_2)$ и f — построенный метод кластеризации.

1. Среднее внутрикластерное расстояние:

$$F_0(f) = \frac{\sum_{i < \ell} I\{f(x_i) = f(x_\ell)\} \cdot \rho(x_i, x_\ell)}{\sum_{i < \ell} I\{f(x_i) = f(x_\ell)\}} \rightarrow \min_f$$

2. Среднее межкластерное расстояние:

$$F_1(f) = \frac{\sum_{i < \ell} I\{f(x_i) \neq f(x_\ell)\} \cdot \rho(x_i, x_\ell)}{\sum_{i < \ell} I\{f(x_i) \neq f(x_\ell)\}} \rightarrow \max_f$$

3. $F_0(f)/F_1(f) \rightarrow \min_f$

Метрика 1 не подходит для выбора количества кластеров:
её оптимум достигается, если все кластеры — одноэлементные.

Расстояния внутри и между кластерами

- Среднее расстояние до центра кластера.

Будем считать, что каждый кластер характеризуется своим центром μ_k .

$$\sum_{k=1}^K \sum_{i=1}^n I\{f(x_i) = k\} \cdot \rho(x_i, \mu_k)$$

- Индекс Данна (Dunn Index):

$$\frac{\min_{1 \leq k' < k \leq K} d(k', k)}{\max_{1 \leq k \leq K} d(k)} \longrightarrow \max_f$$

$d(k', k)$ — расстояние между кластерами k' и k .

Например, евклидово расстояние между центрами кластеров.

$d(k)$ — внутрикластерное расстояние для кластера k .

Например, сумма расстояний от всех объектов кластера k до его центра.

Силуэт

Пусть точка x лежит в кластере C_k .

a_x — среднее расст. от x до всех других объектов из его же кластера:

$$a_x = \frac{1}{|C_k| - 1} \sum_{z \in C_k, z \neq x} \rho(x, z)$$

b_x — среднее расстояние от x до всех объектов из ближайшего другого кластера:

$$b_x = \min_{\ell \neq k} \frac{1}{|C_\ell|} \sum_{z \in C_\ell} \rho(x, z)$$

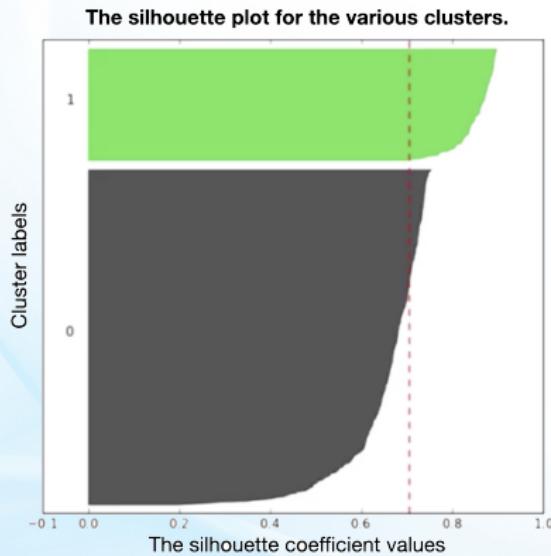
Силуэт для точки x : $s_x = \frac{b_x - a_x}{\max(b_x, a_x)}$

Если $b_x \gg a_x$ — хороший случай, то s_x около 1.

Если $b_x \ll a_x$ — плохой случай, то s_x около -1 .

Средний коэффициент силуэта по выборке: $s = \frac{1}{n} \sum_{i=1}^n s_{x_i}$

Силуэт

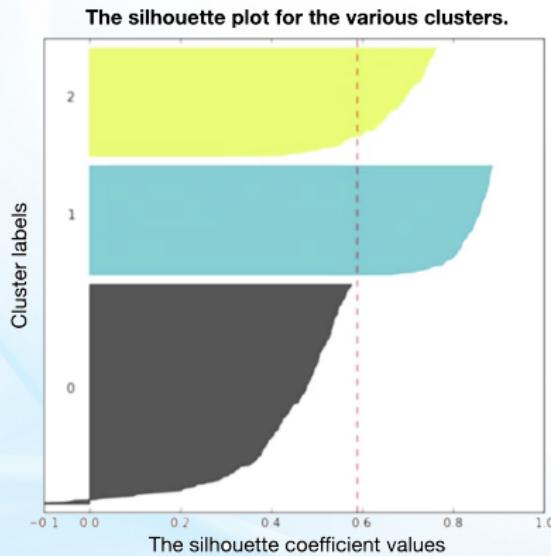


Пунктирная линия – среднее значение силуэта по выборке.

Разброс значений силуэта между кластерами не очень большой.

Кластеризация считается хорошей.

Силуэт



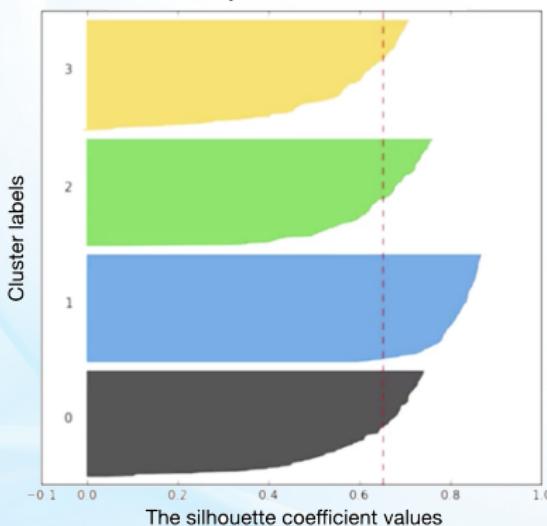
Разброс значений силуэта между кластерами большой.

Значения силуэта для кластера 0 ниже среднего значения.

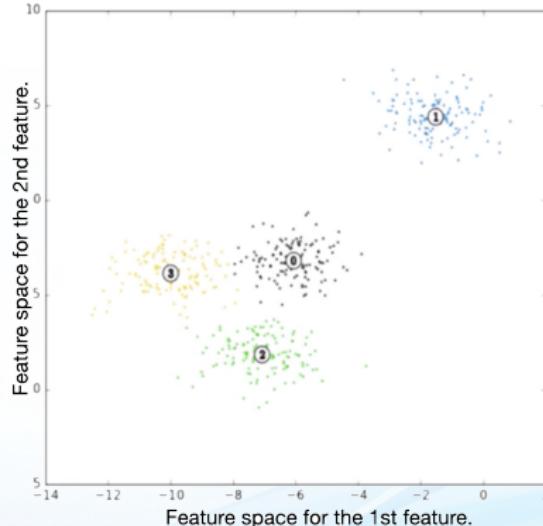
Кластеризация считается плохой.

Силуэт

The silhouette plot for the various clusters.



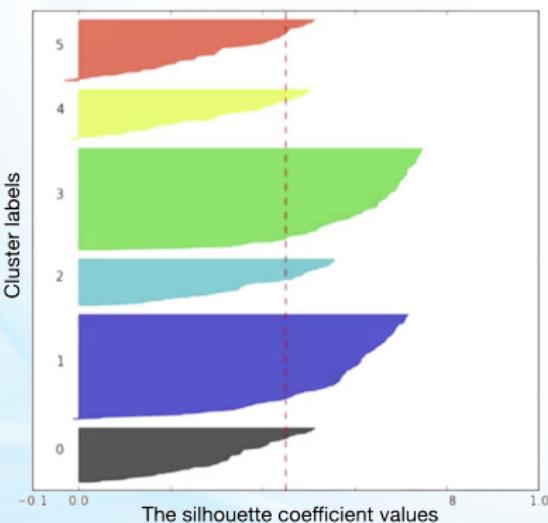
The visualization of the clustered data



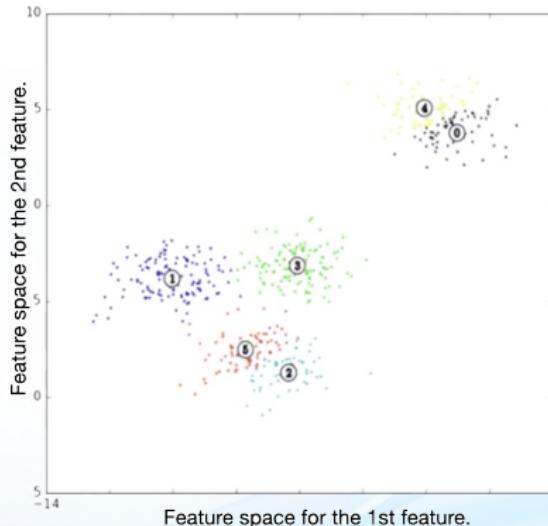
Разброс значений силуэта между кластерами не очень большой.
Кластеризация считается хорошей.

Силуэт

The silhouette plot for the various clusters.

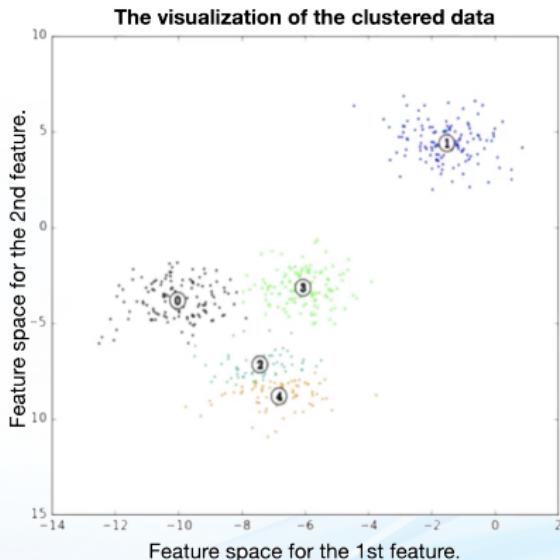
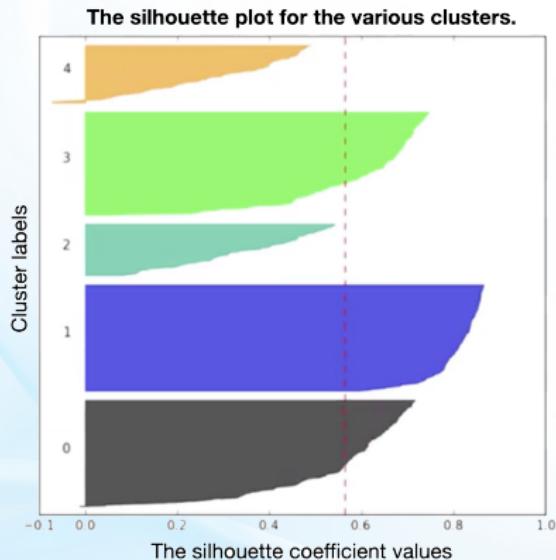


The visualization of the clustered data



Разброс значений силуэта между кластерами большой.
Кластеризация считается плохой.

Силуэт



Разброс значений силуэта между кластерами большой.

Значения силуэта для кластеров 2 и 4 ниже среднего значения.

Кластеризация считается плохой.

Как поступать на практике

Часто на практике:

- ▶ "Внутренние" метрики не интерпретируемые и не согласуются с желаниями заказчика.
- ▶ Объекты плохо разбиваются на кластеры или кластеров нет вообще. А заказчик хочет кластеры.



- ▶ Необходимо просто и наглядно убедить заказчика, что результат кластеризации хороший.



Как поступать на практике

- ▶ Заказчика может устроить подробная интерпретация каждого кластера по совокупности признаков.
- ▶ Возможно, есть разметка объектов, которые должны лежать в одном или разных кластерах. В таком случае стоит посчитать на них метрики качества классификации.
- ▶ Обычно кластеризация — промежуточная задача для решения другой задачи. В таком случае можно использовать такой критерий качества, который согласуется с качеством целевой задачи.
 - ▶ Если цель — на кластерах построить разные регрессионные модели, то можно оценивать по MSE весь пайплайн предобработка -> кластеризация -> регрессия.
 - ▶ Если цель — выявить успешные магазины в каждом кластере, то можно максимизировать среднюю внутрикластерную дисперсию выручки среди "хороших" кластеризаций.



Кластеризация

Задача кластеризации

Работа в командах

Определяемся с требованиями

Метрики качества

K-means

DBSCAN

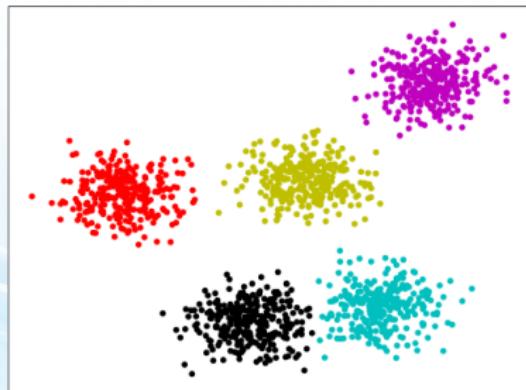
Иерархическая кластеризация

K-means (метод K-средних)

Дана выборка $X = (x_1, \dots, x_n)$.

Задано число кластеров K .

Нужно построить отображение $f : \mathcal{X} \longrightarrow \{1, \dots, K\}$,
то есть отнести каждый объект к одному из кластеров.



K-means (метод K-средних)

В качестве метрики расстояния между объектами обычно используется евклидова метрика: $\rho(x, z) = \|x - z\|^2$

Процедура:

1. Задать начальное приближение центров кластеров $\mu_1, \mu_2, \dots, \mu_K$.
2. Повторять

- 2.1 Отнести каждый объект к ближайшему центру:

$$f(x_i) = \arg \min_k \|x_i - \mu_k\|^2.$$

- 2.2 Вычислить новые положения центров:

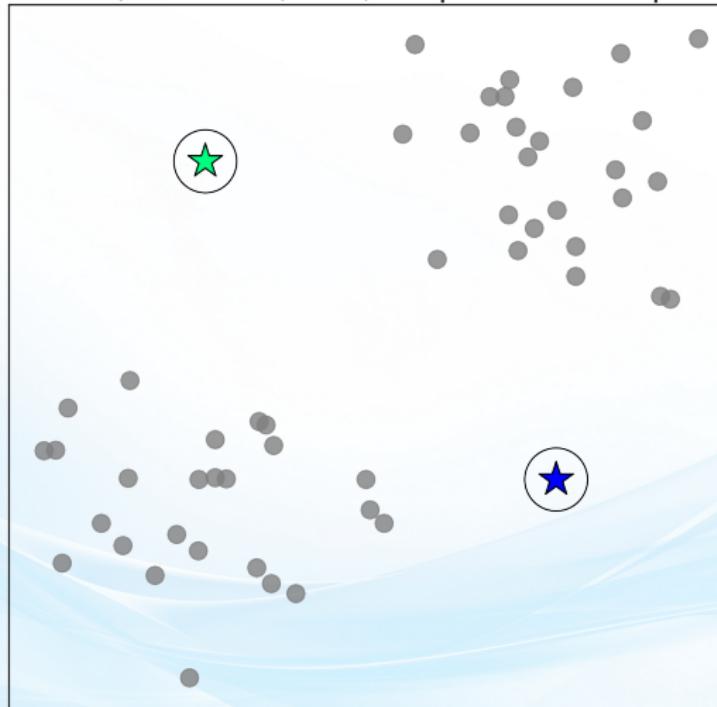
$$\mu_k = \frac{\sum_{i=1}^n I\{f(x_i) = k\} x_i}{\sum_{i=1}^n I\{f(x_i) = k\}}$$

- 2.3 Пока $f(x_i)$ не перестанут изменяться.

Метод применим и к новым данным: берем ближайший к точке кластер.

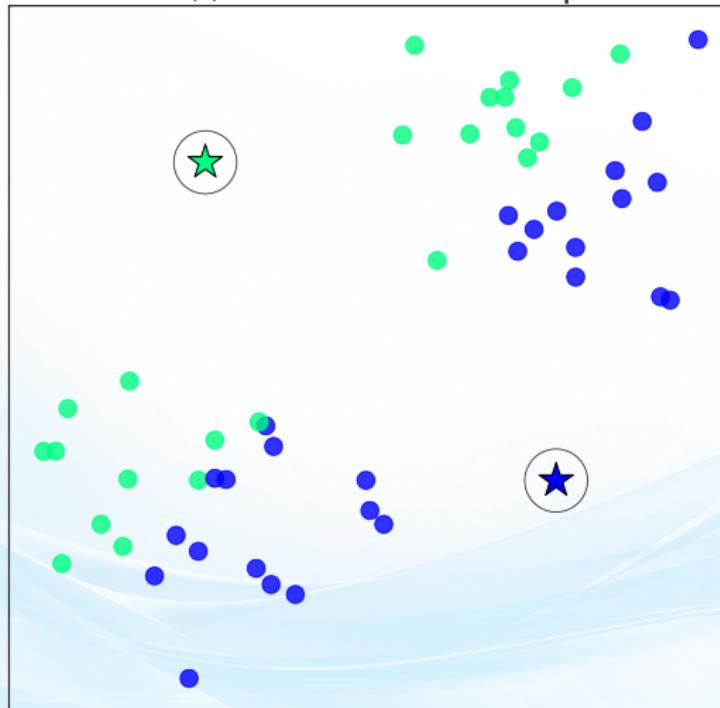
K-means: пример работы

Инициализация центров кластеров



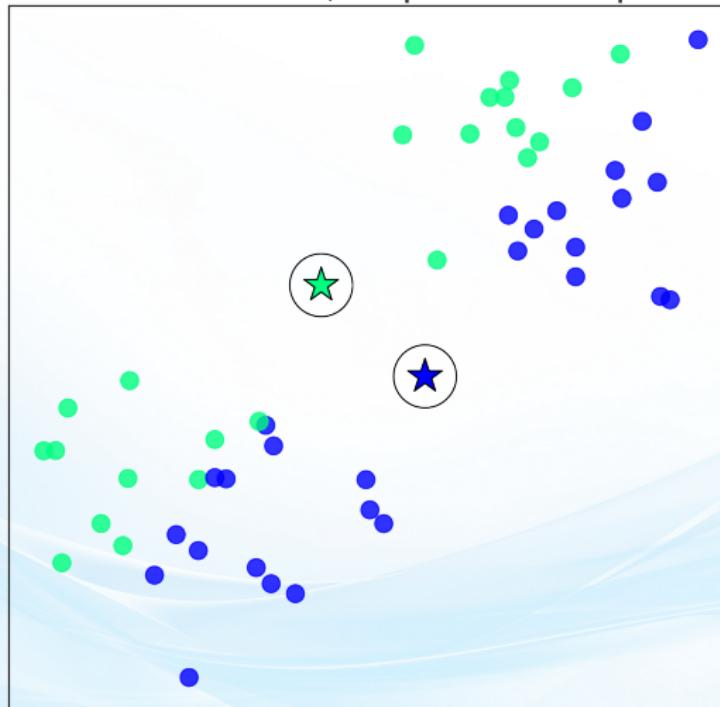
K-means: пример работы

Разделение на кластеры



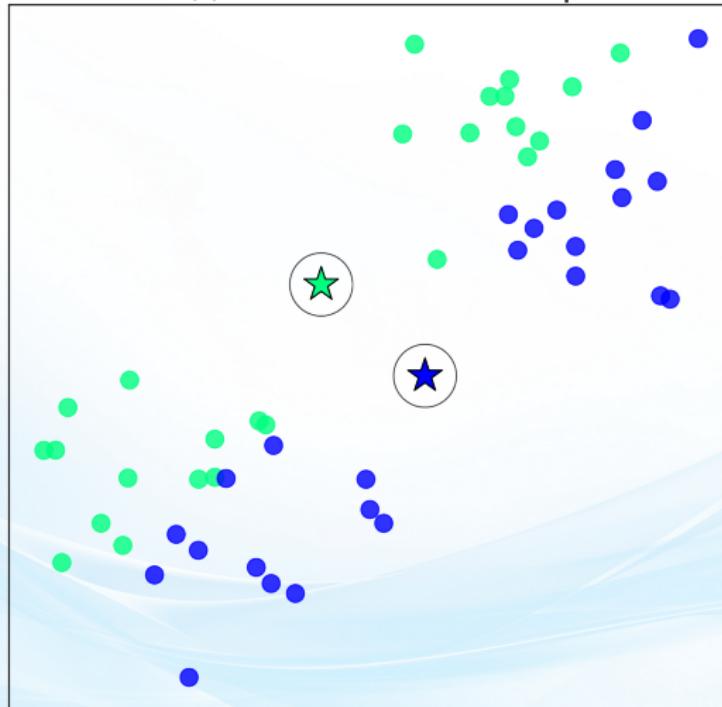
K-means: пример работы

Обновление центров кластеров



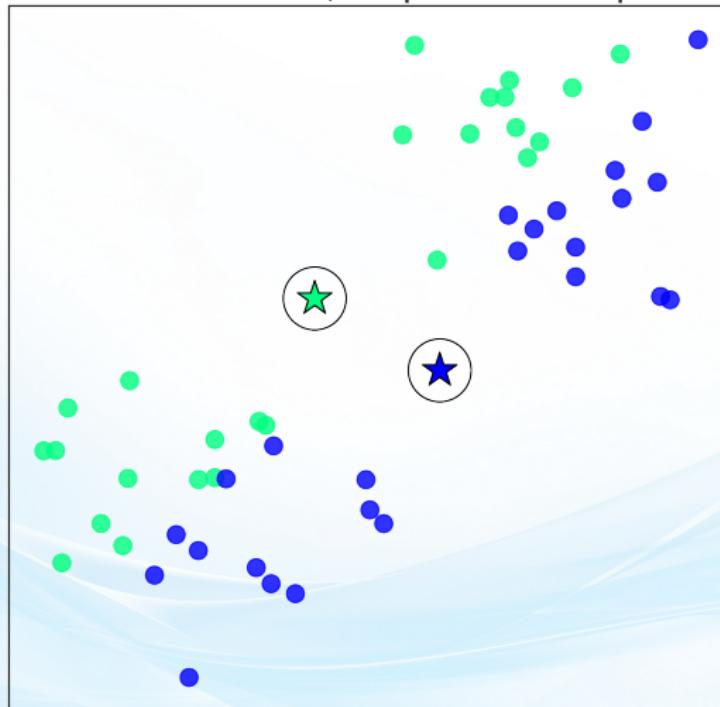
K-means: пример работы

Разделение на кластеры



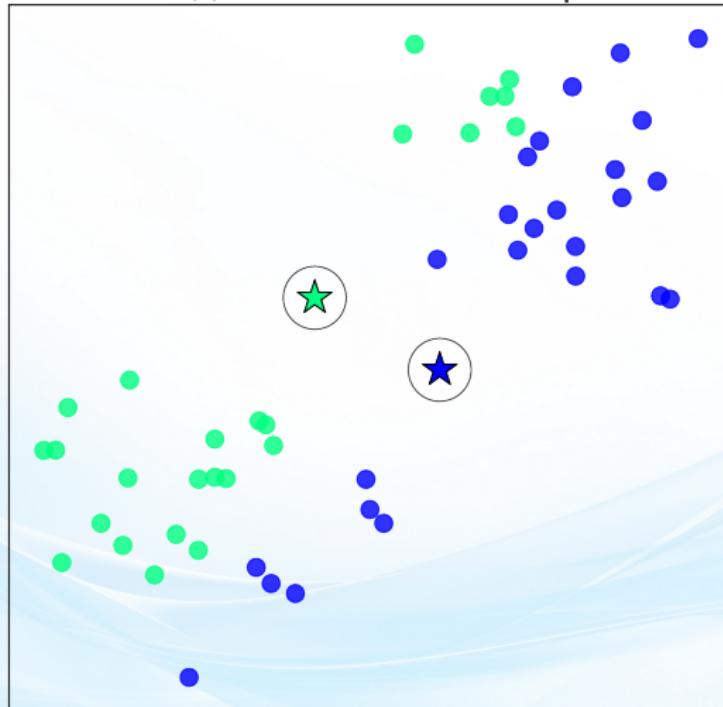
K-means: пример работы

Обновление центров кластеров



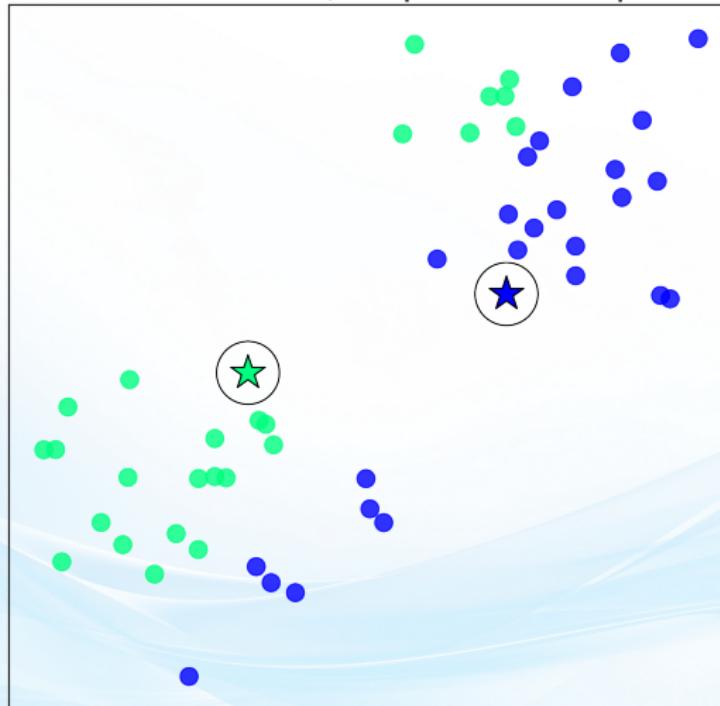
K-means: пример работы

Разделение на кластеры



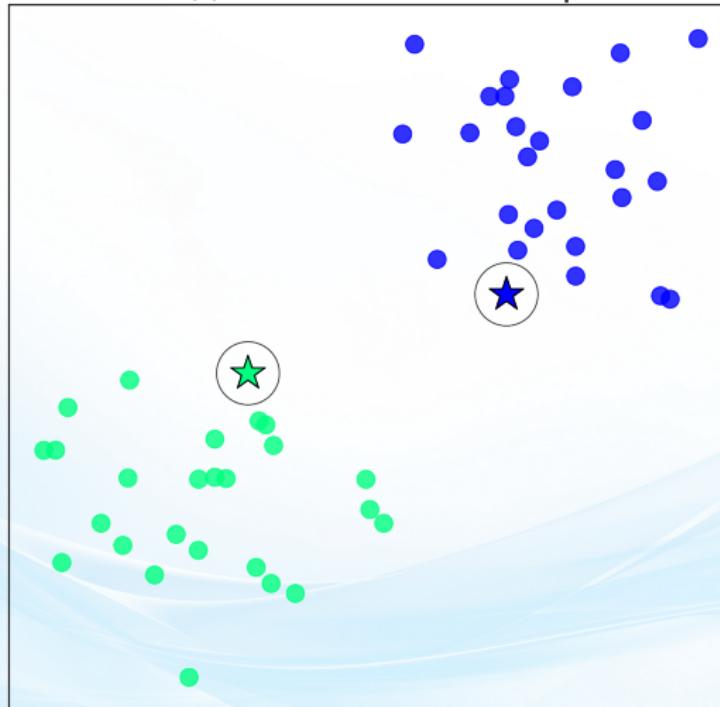
K-means: пример работы

Обновление центров кластеров



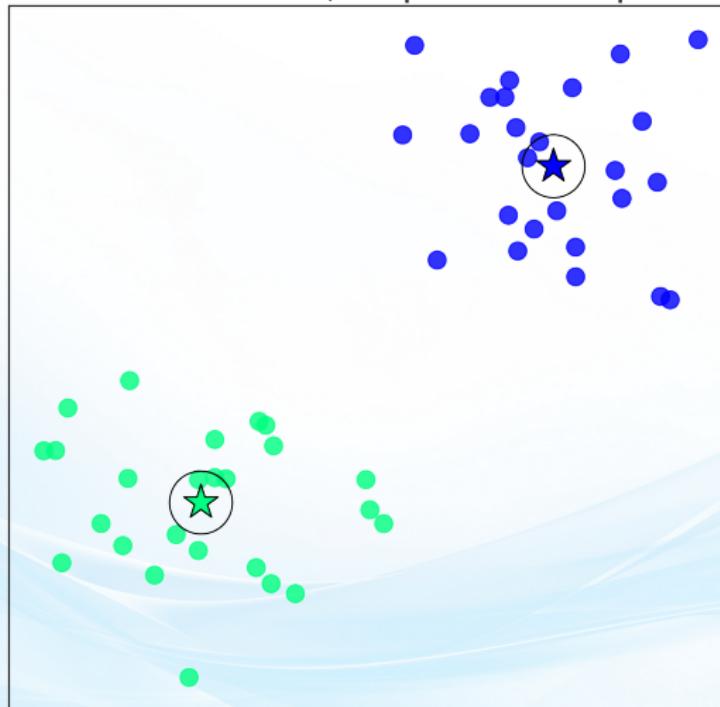
K-means: пример работы

Разделение на кластеры



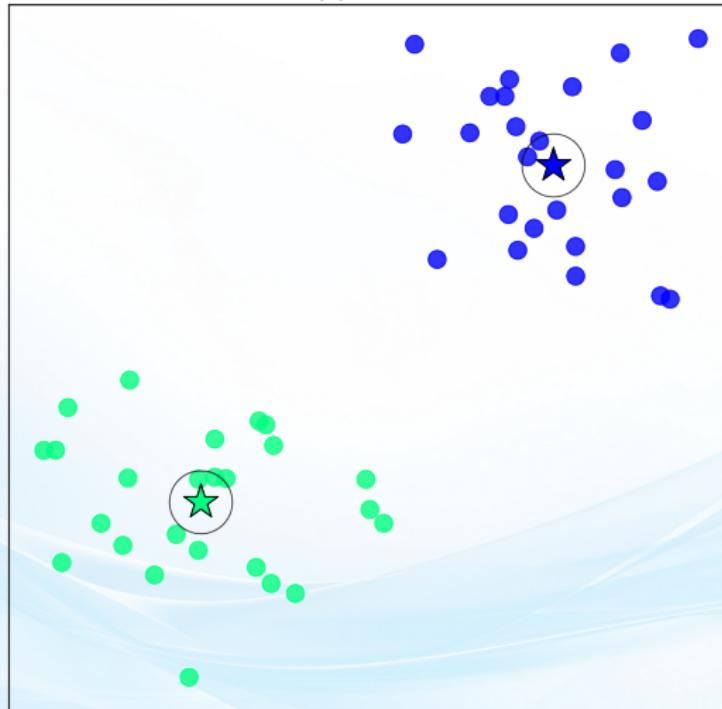
K-means: пример работы

Обновление центров кластеров



K-means: пример работы

Метод сошелся



Что оптимизирует K-means

Утверждение.

K-means оптимизирует сумму квадратов внутрикластерных расстояний до центра кластера:

$$Q(f) = \sum_{i=1}^n \|x_i - \mu_{f(x_i)}\|^2 \rightarrow \min_{f, \mu}$$

Доказательство

Покажем, что на каждом шаге $Q(f)$ убывает или не изменяется:

- ▶ **Шаг 1:** Отнести каждый объект к ближайшему центру.
Центры — фиксированные точки. При отнесении точки x_i к ближайшему центру не возрастает значение $\|x_i - \mu_{f(x_i)}\|^2$.
 \Rightarrow функционал $Q(f)$ не возрастает.
- ▶ **Шаг 2:** Вычислить новые положения центров.
Кластеры фиксированы. Функционал $Q(f)$ разбивается на K независимых слагаемых, для минимизации которых в качестве центра нужно взять среднее: $\frac{1}{|\{i: f(x_i)=k\}|} \sum_{i: f(x_i)=k} x_i$.
 \Rightarrow функционал $Q(f)$ не возрастает.

Что оптимизирует K-means

Утверждение.

K-means оптимизирует сумму квадратов внутрикластерных расстояний до центра кластера:

$$Q(f) = \sum_{i=1}^n \|x_i - \mu_{f(x_i)}\|^2 \rightarrow \min_{f, \mu}$$

Доказательство

Доказали, что на каждой итерации $Q(f)$ не возрастает.

Равенство $Q(f)$ между итерациями достигается либо если отображение объектов в кластеры не меняется, либо в редких случаях симметрии.

Метод найдет лишь локальный минимум функционала $Q(f)$. Нахождение глобального минимума — NP-полнная задача.

Особенности

1. Сходится к локальному оптимуму,
имеет смысл запускать из разных начальных приближений.
2. Кластеры представляют собой выпуклые множества.
3. **Mini-batch k-means**

На каждом шаге:

- ▶ Выбираем случайное подмножество объектов.
 - ▶ Распределяем это подмножество объектов по кластерам.
 - ▶ Считаем центры кластеров по данным объектам.
-
4. K-means — метрический метод, имеет смысл делать снижение размерности: отбор признаков, PCA, UMAP, t-SNE и прочее.

K-means++

В зависимости от начального приближения центров кластеров

- ▶ может потребоваться разное время для сходимости;
- ▶ результаты могут получиться разными.

Решение: брать центры подальше друг от друга.

Как?

1. Первый центр выбираем случайно из равномерного распределения на выборке.
2. Каждый следующий центр выбираем случайно по некоторой вероятности из оставшихся точек.

При этом вероятность выбрать каждую точку пропорциональна квадрату расстояния от нее до ближайшего выбранного центра.

EM-алгоритм

Нежесткая кластеризация

Пусть объект x не строго принадлежит одному кластеру, а имеет некую вероятность принадлежности к каждому кластеру.

Предположим, что имеется K кластеров.

Причем объекты подчиняются модели смеси распределений с плотностью:

$$p(x) = \sum_{k=1}^K \pi_k p_{\theta_k}(x), \quad \pi_k \geq 0, \quad \sum_{k=1}^K \pi_k = 1$$

где π_k — вероятность получить объект из кластера k ,

$p_{\theta_k}(x)$ — плотность объекта внутри кластера k .

Смысл параметров:

- ▶ π_k отвечают за соотношение кластеров,
- ▶ θ_k за положения кластеров в пространстве.

EM-алгоритм

Рассмотрим смесь гауссовских распределений.

$$p_{\theta_k} = \mathcal{N}(\mu_k, \Sigma_k)$$

E-шаг: Оцениваем вероятности принадлежности объектов к кластерам, при фиксированных параметрах распределений.

$$\gamma_{ik} = P(x_i \text{ в кластере } k) = \frac{\pi_k \cdot p_{\mu_k \Sigma_k}(x_i)}{\sum_{k=1}^K \pi_k \cdot p_{\mu_k \Sigma_k}(x_i)}$$

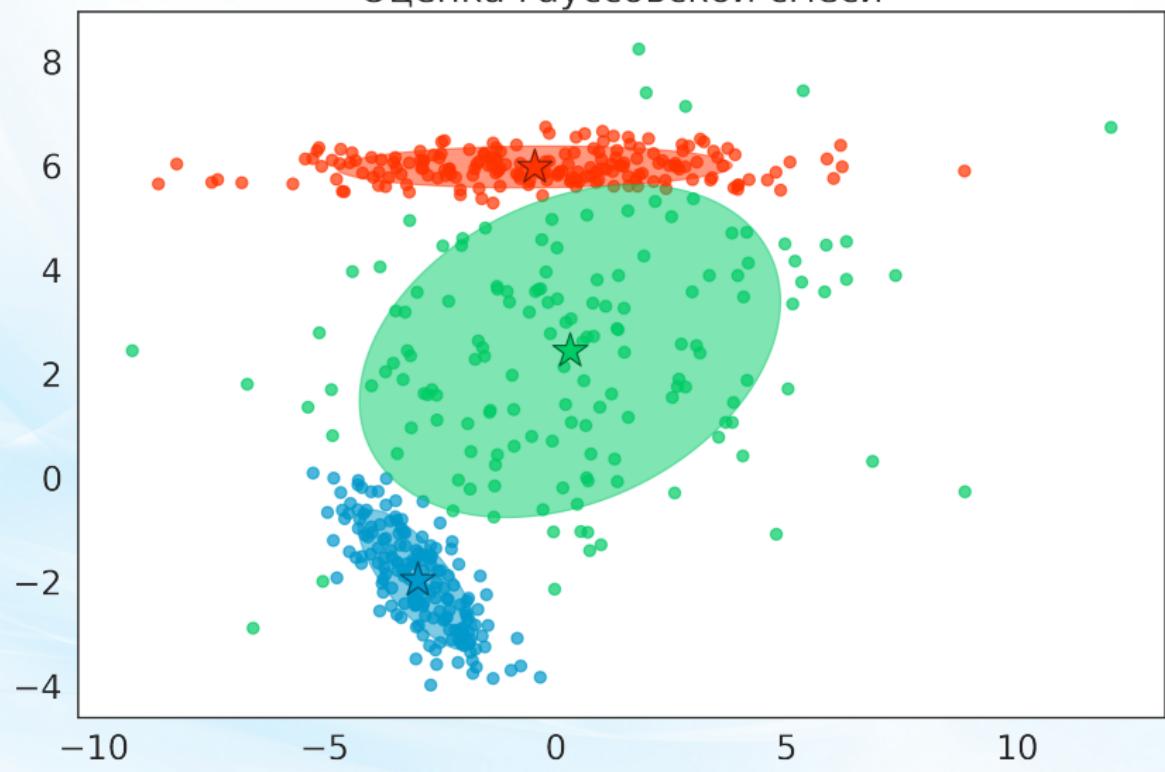
M-шаг: Оцениваем параметры распределений при фиксированных вероятностях принадлежности к кластерам.

$$\mu_k = \frac{\sum_{i=1}^n \gamma_{ik} x_i}{\sum_{i=1}^n \gamma_{ik}} \quad \Sigma_j = \frac{\sum_{i=1}^n \gamma_{ik} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{i=1}^n \gamma_{ik}} \quad \pi_k = \frac{1}{n} \sum_{i=1}^n \gamma_{ik}$$

Повторяем до сходимости

ЕМ-алгоритм: Пример

Оценка гауссовой смеси





Сравним K-means и EM

EM

E-шаг:

Для каждого объекта
оцениваем вероятность
принадлежности к кластерам.

M-шаг:

Оцениваем параметры,
задающие распределения.

K-means

E-шаг:

Для каждого объекта
находим ближайший кластер.

M-шаг:

Оцениваем центры кластеров.

⇒ K-means — упрощенный вариант EM алгоритма
для разделения смеси гауссовских распределений.



Кластеризация

Задача кластеризации

Работа в командах

Определяемся с требованиями

Метрики качества

K-means

DBSCAN

Иерархическая кластеризация

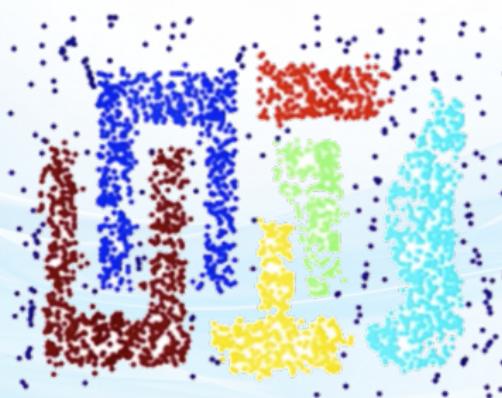
Density Based методы

$X = (x_1, \dots, x_n)$ — выборка в d -мерном пространстве.

Требуется разбить выборку на какое-то число кластеров.

Density Based методы работают по следующей идеологии:

*Кластеры — непрерывные области большой плотности,
разделенные от др. кластеров областями с низкой плотностью.*



DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

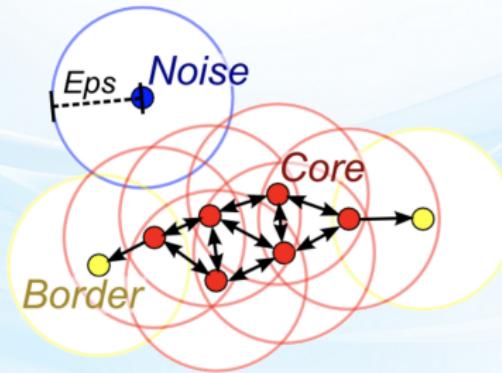
Задаются 2 гиперпараметра: ε , m .

Для каждой точки рассматривается ее ε -окрестность:

$$U_\varepsilon(x) = \{z \in X \mid \rho(x, z) \leq \varepsilon\}.$$

Каждый объект может быть одного из **трех типов**:

- ▶ *Основной*: $|U_\varepsilon(x)| \geq m$.
- ▶ *Пограничный*: Не является основным, но находится в ε -окрестности основного объекта.
- ▶ *Шумовой*: Не является основным или пограничным.



DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Процесс:

1. Берем еще не помеченный объект x .
2. Если $|U_\epsilon(x)| < m$, то помечаем x как возможно шумовой.
Позже x может стать частью какого-то кластера.
3. Если $|U_\epsilon(x)| \geq m$, то создаем кластер $K = U_\epsilon(x)$.

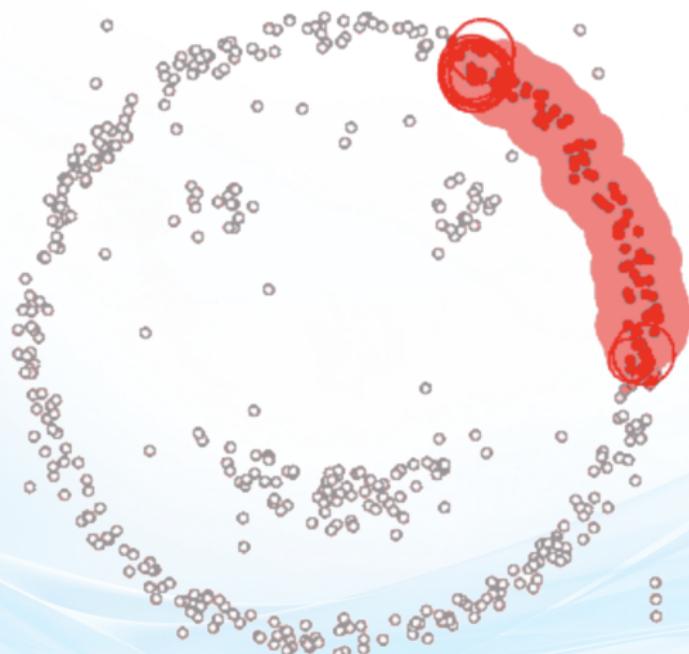
Находим все точки достижимые из x по основным точкам в т.ч. те, что были помечены шумовыми.

Помечаем найденные точки как точки кластера K .

Примечание. Точка y **достижима** из основной точки x , если существует путь по основным точкам из x в y .

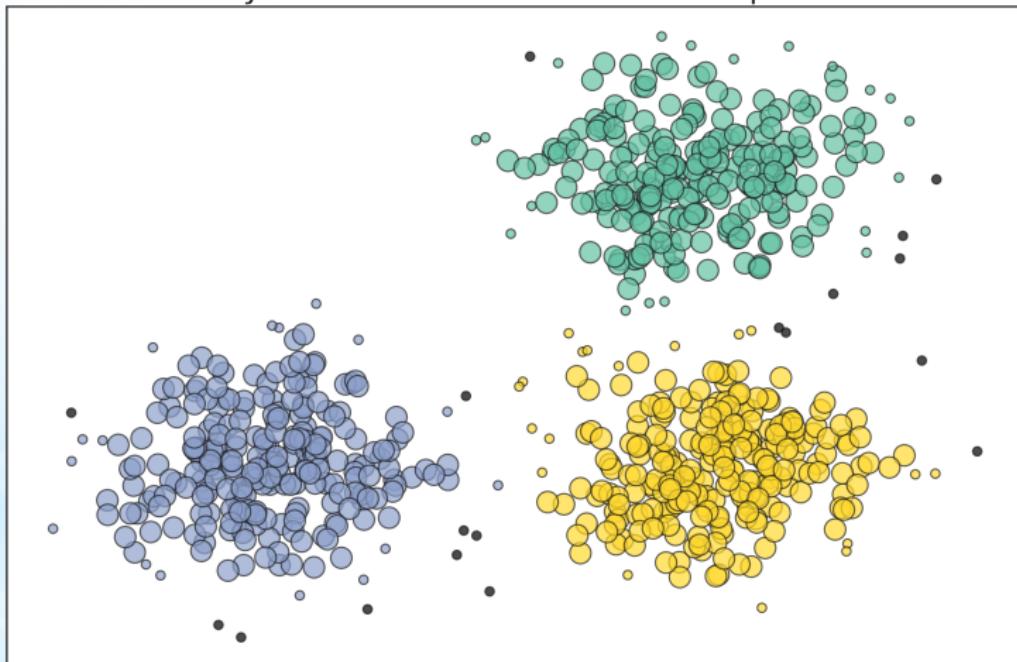
4. Повторяем 1-3 пока есть неразмеченные объекты.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)



DBSCAN: Примеры

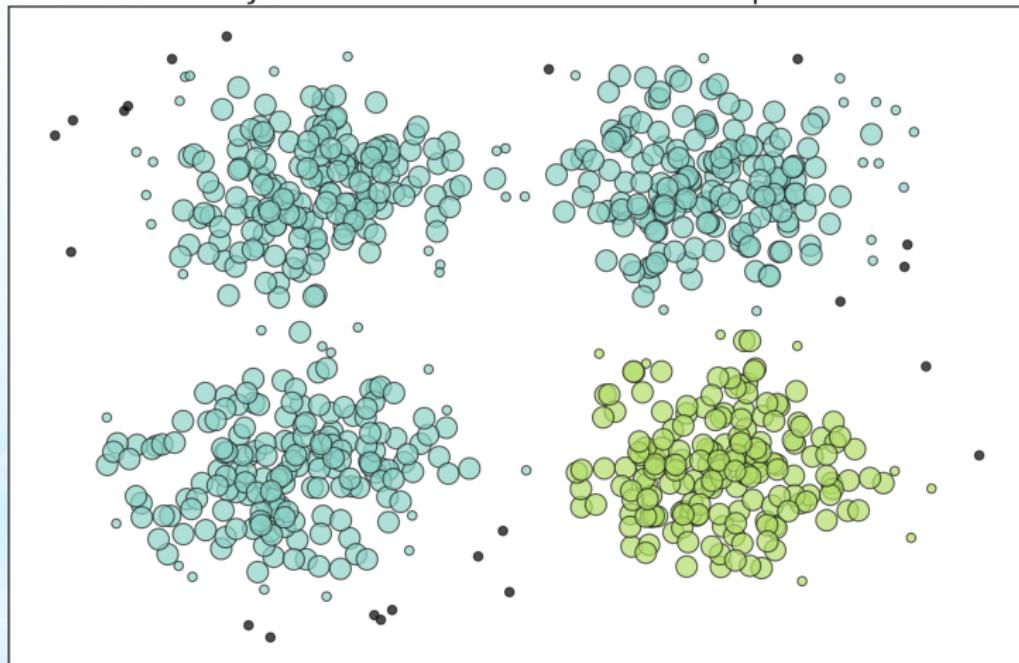
Полученное количество кластеров: 3



Выделяет шумовые точки.

DBSCAN: Примеры

Полученное количество кластеров: 2



Расположенные близко кластеры могут сливаться в один.

DBSCAN

Плюсы:

- ▶ Кластеры могут быть произвольной формы.
- ▶ Деление объектов на основные, пограничные и шумовые.
 ⇒ Можем выкинуть шум.
- ▶ Быстрая кластеризация.

В худшем случае работает $O(n^2)$.

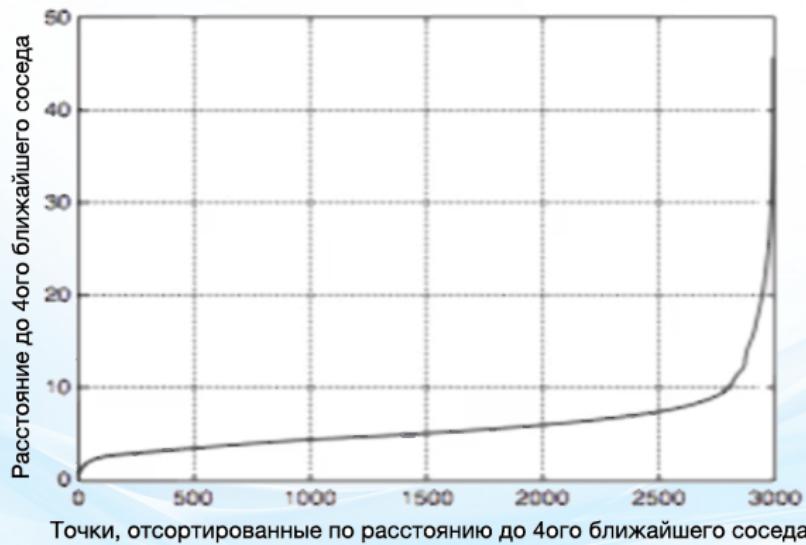
Минусы:

- ▶ Чуть-чуть сливающиеся кластеры определяются как один.
- ▶ Плохо работает, разные кластеры имеют разную плотность.
- ▶ Нужно подбирать гиперпараметры.

DBSCAN: Подбор гиперпараметров

Какое ε и m выбрать?

Нарисуем график:



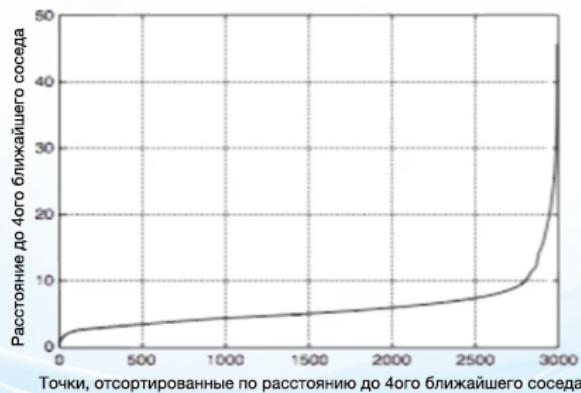
Ось OX — точки, отсортированные по расстоянию до k -ого соседа.

Ось OY — расстояние до k -ого соседа.

DBSCAN: Подбор гиперпараметров

У хвоста точек есть резкий рост расстояния.
Это в основном шумовые точки.

Выберем ε как расстояние в начале резкого роста расстояния,
 m выберем равным k .



Перебрав разные k получим разные графики
и решим сколько точек готовы сделать шумовыми.



Кластеризация

Задача кластеризации

Работа в командах

Определяемся с требованиями

Метрики качества

K-means

DBSCAN

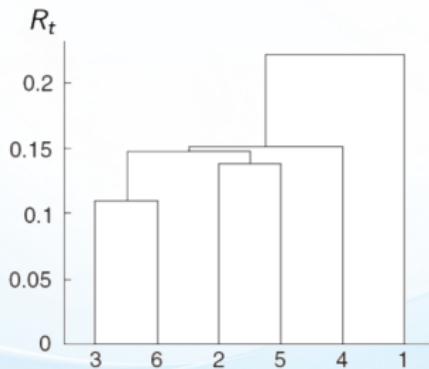
Иерархическая кластеризация

Иерархическая кластеризация

Работает по следующему принципу:

1. Вводим функцию расстояния между кластерами.
2. Выстраиваем иерархию вложенных друг в друга кластеров.
3. Получаем дерево, вершины в котором — кластеры.

Результат — иерархия: список вложенных друг в друга кластеров.

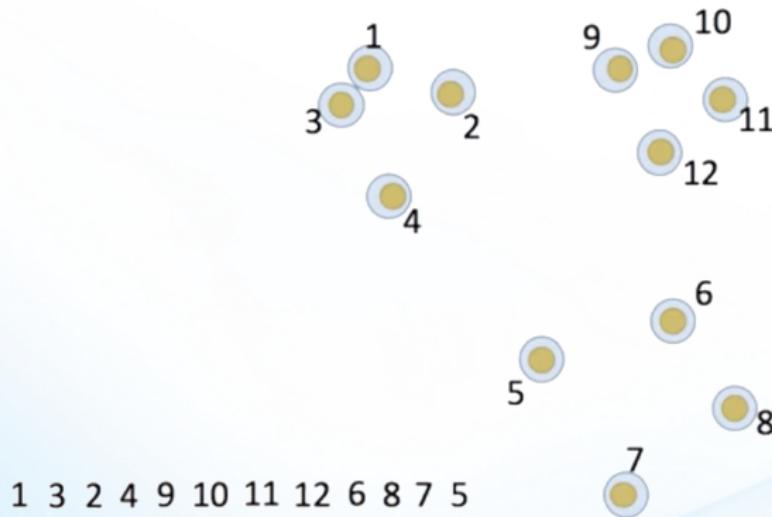


Дендрограмма — график расстояний между кластерами в момент слияния.

Ось OX — номера объектов.

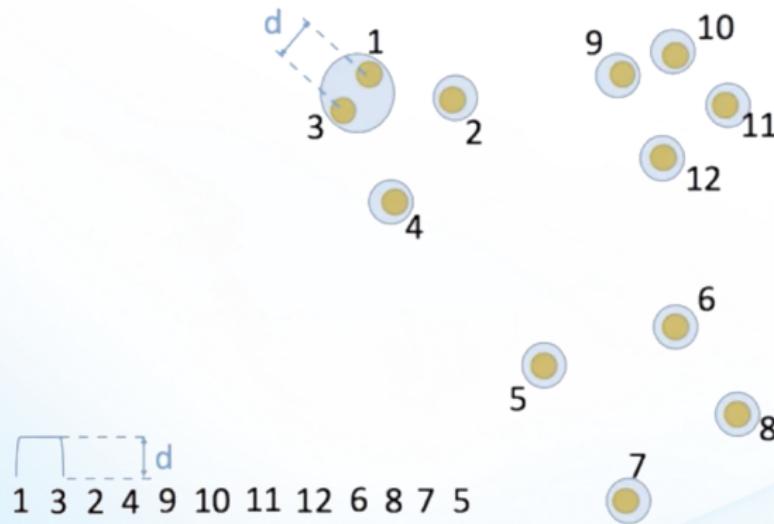
Ось OY — расстояние при слиянии двух кластеров.

Дендрограмма



Изначально каждый объект лежит в своем отдельном кластере.

Дендрограмма

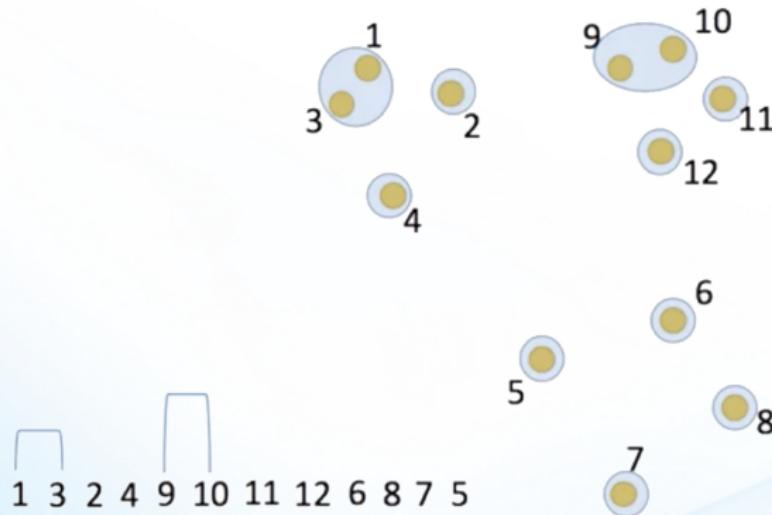


Соединяем самые близкие к друг другу кластеры.

Здесь это {1} и {3}. Пусть расстояние между ними равно d .

Отложим стобик высоты d соединяющий эти 2 кластера.

Дендрограмма

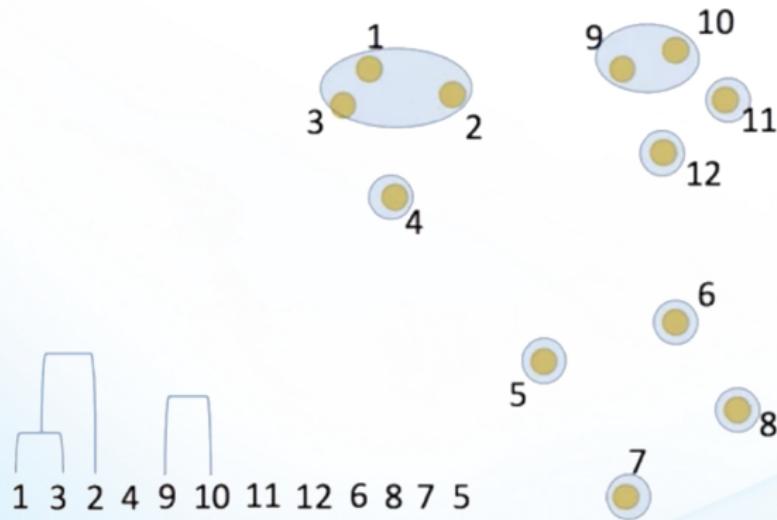


Снова находим самые близкие к друг другу кластеры.

Это кластеры $\{9\}$ и $\{10\}$.

Откладываем расстояние между ними на дендрограмме.

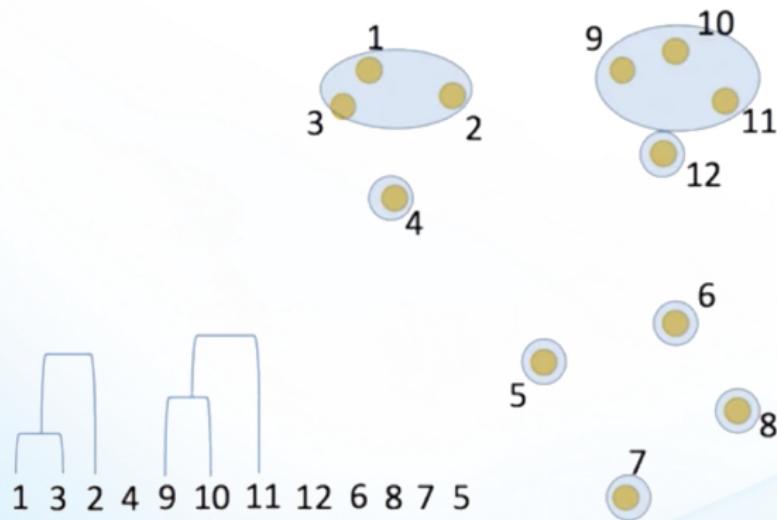
Дендрограмма



Объединяем в один кластер кластеры $\{1, 3\}$ и $\{2\}$.

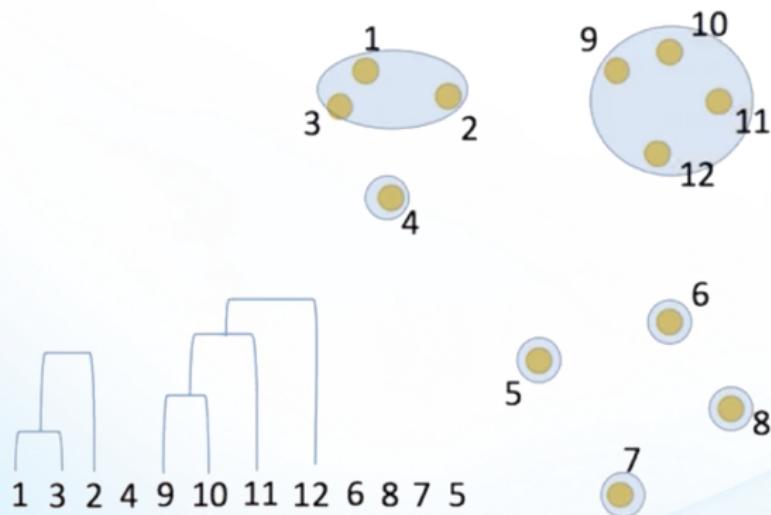
Откладываем расстояние между этими кластерами на дендрограмме.

Дендрограмма



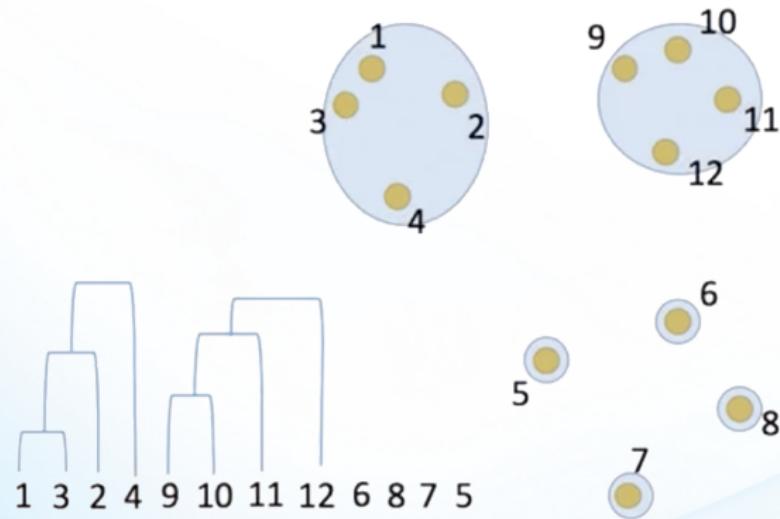
И так далее ...

Дендрограмма



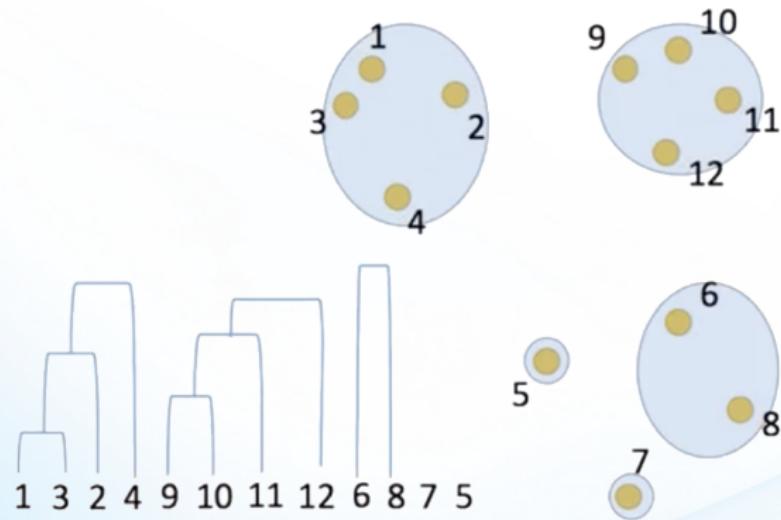
И так далее ...

Дендрограмма



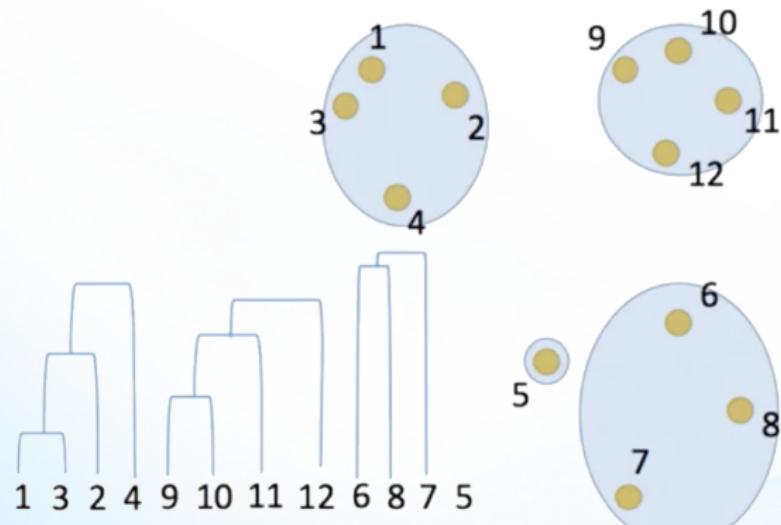
И так далее ...

Дендрограма



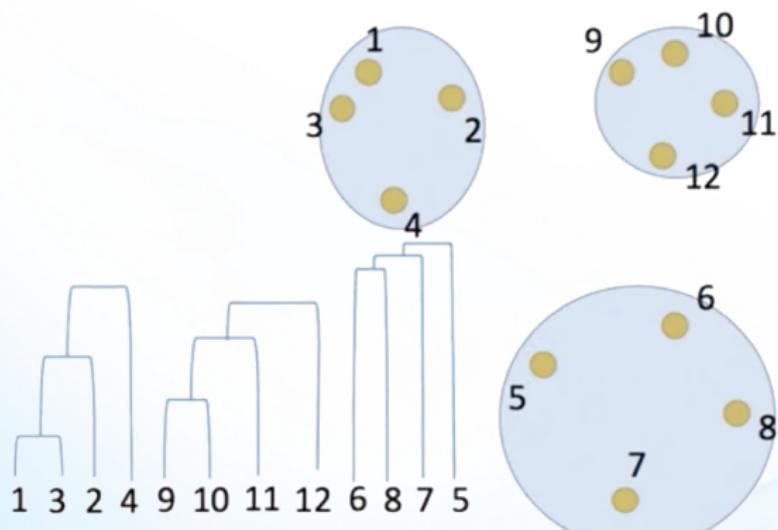
И так далее ...

Дендрограмма



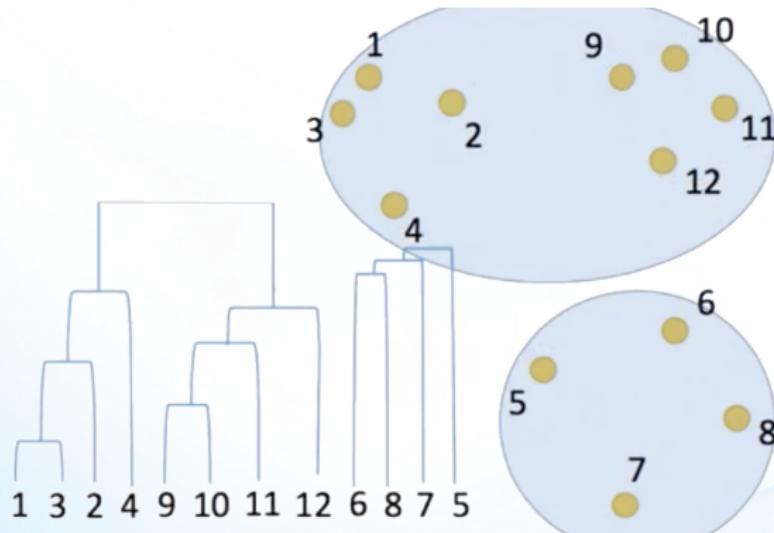
И так далее ...

Дендрограмма



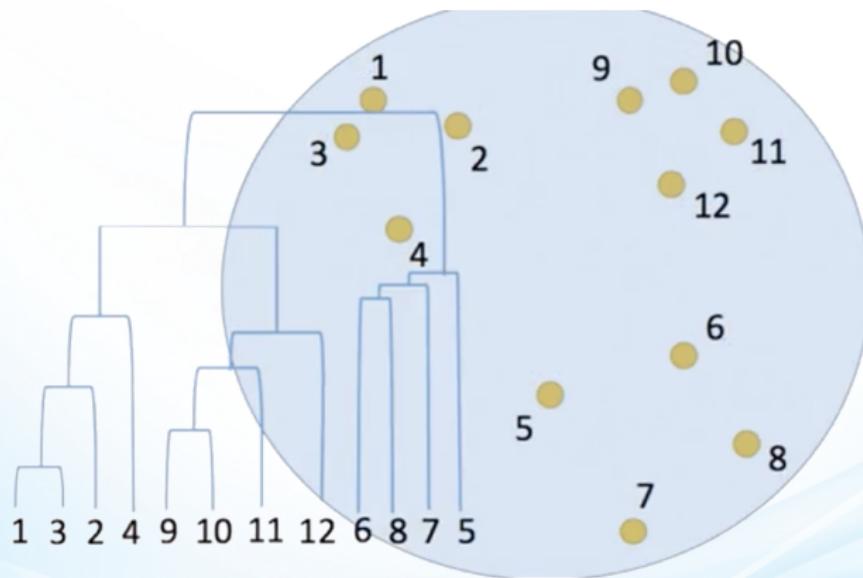
И так далее ...

Дендрограмма



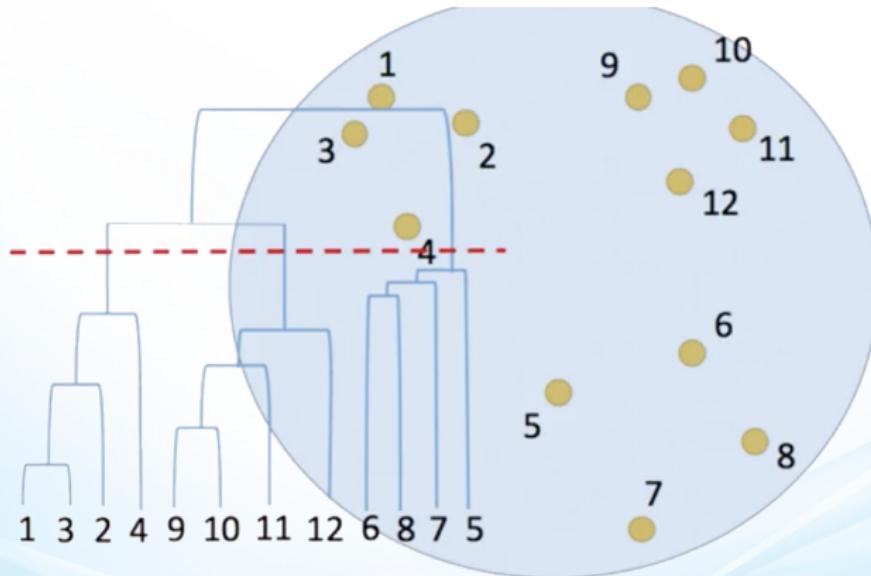
И так далее ...

Дендрограмма



Получили один кластер, в котором лежат все объекты.

Дендрограмма



Можем обрезать дендрограмму по какой-то итерации если

- ▶ расстояния между сливаемыми кластерами стали большими
- ▶ разница в расстоянии для соседних слияний большая

Агломеративная иерархическая кластеризация

Пусть $C_t = \{K_1, K_2, \dots, K_s\}$ — набор кластеров на итерации t .

R_{K_1, K_2} — некоторое расстояние между кластерами K_1 и K_2 .

1. Все кластеры одноэлементные: $C_1 = \{\{x_1\}, \{x_2\}, \dots, \{x_n\}\}$
 $R_{\{x\}, \{z\}} = \rho(x, z)$ — расстояния между кластерами.
2. Итерации: для всех $t \in \{2, \dots, n\}$:
3. Найти в C_{t-1} пару кластеров K_1, K_2 с минимальным R_{K_1, K_2} .
4. Слить их в один кластер:
$$K' = K_1 \cup K_2$$

$$C_t = C_{t-1} \cup \{K'\} \setminus \{K_1, K_2\}$$
5. Для всех $K \in C_t$:
6. Вычислить расстояние $R_{K', K}$.

Агломеративная иерархическая кластеризация

Как выбрать расстояние между кластерами?

- ▶ Среднее расстояние между объектами из разных кластеров.
- ▶ Макс. расстояние между объектами из разных кластеров.
- ▶ Мин. расстояние между объектами из разных кластеров.
- ▶ Расстояние между центрами кластеров.

Какая сложность пересчета расстояния на каждой итерации?

- ▶ Сложность подсчета среднего, максимального и минимального расстояний $R_{K',K} = O(|K'| \cdot |K|) \cdot O(d)$.
- ▶ Сложность подсчета расстояния между центрами кластеров — $O(|K'| + |K|) \cdot O(d)$.

На каждой итерации t необходимо найти $R_{K',K}$ для каждого $K \in C_t$.

⇒ Итоговая сложность на одной итерации большая.

Агломеративная иерархическая кластеризация

Как сделать подсчет расстояний более эффективным?

Пусть на итерации t решили объединить кластеры K_1 и K_2 в один кластер K' .

Из предыдущей итерации $t - 1$ уже знаем расстояния $R_{K_1, K}$ и $R_{K_2, K}$ для $\forall K \in C_{t-1} = C_t \cup \{K_1, K_2\} \setminus K'$.

Может ли это нам помочь?

Рассмотрим случай минимального расстояния между кластерами:

$$\begin{aligned} R_{K', K} &= \min_{x \in K', z \in K} \rho(x, z) = \\ &= \min \left[\min_{x \in K_1, z \in K} \rho(x, z), \min_{x \in K_2, z \in K} \rho(x, z) \right] = \min(R_{K_1, K}, R_{K_2, K}) \end{aligned}$$

\implies можем пересчитать расстояние $R_{K', K}$ за $O(1)$.

Формула Ланса-Уильямса

- Рекурсивная формула пересчета расстояний за $O(1)$:

$$R_{K_1 \cup K_2, K} = \alpha_{K_1} R_{K_1, K} + \alpha_{K_2} R_{K_2, K} + \beta R_{K_1, K_2} + \gamma |R_{K_1, K} - R_{K_2, K}|.$$

- Обобщает множество разных способов ввести расстояние:

- Минимальное расстояние: $\alpha_{K_1} = \alpha_{K_2} = \frac{1}{2}$, $\beta = 0$, $\gamma = -\frac{1}{2}$
- Максимальное расстояние: $\alpha_{K_1} = \alpha_{K_2} = \frac{1}{2}$, $\beta = 0$, $\gamma = \frac{1}{2}$
- Среднее расстояние: $\alpha_{K_1} = \frac{|K_1|}{|K_1 \cup K_2|}$, $\alpha_{K_2} = \frac{|K_2|}{|K_1 \cup K_2|}$, $\beta = \gamma = 0$
- Расстояние между центрами:

$$R_{K', K}^c = \rho \left(\sum_{x_i \in K'} \frac{x_i}{|K'|}, \sum_{x_j \in K} \frac{x_j}{|K|} \right)$$

$$\alpha_{K_1} = \frac{|K_1|}{|K_1 \cup K_2|}, \quad \alpha_{K_2} = \frac{|K_2|}{|K_1 \cup K_2|}, \quad \beta = -\alpha_{K_1} \alpha_{K_2}, \quad \gamma = 0$$

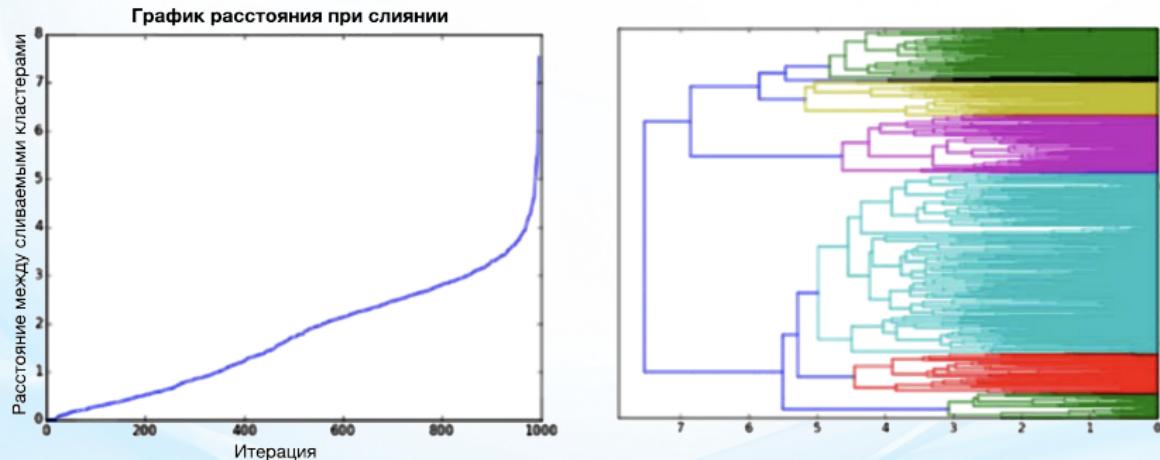
- Расстояние Уорда:

$$R_{K', K}^Y = \frac{|K'| \cdot |K|}{|K'| + |K|} \cdot R_{K', K}^c$$

$$\alpha_{K_1} = \frac{|K| + |K_1|}{|K| + |K_1 \cup K_2|}, \quad \alpha_{K_2} = \frac{|K| + |K_2|}{|K| + |K_1 \cup K_2|}, \quad \beta = \frac{|K|}{|K| + |K_1 \cup K_2|}, \quad \gamma = 0$$

Агломеративная иерархическая кластеризация

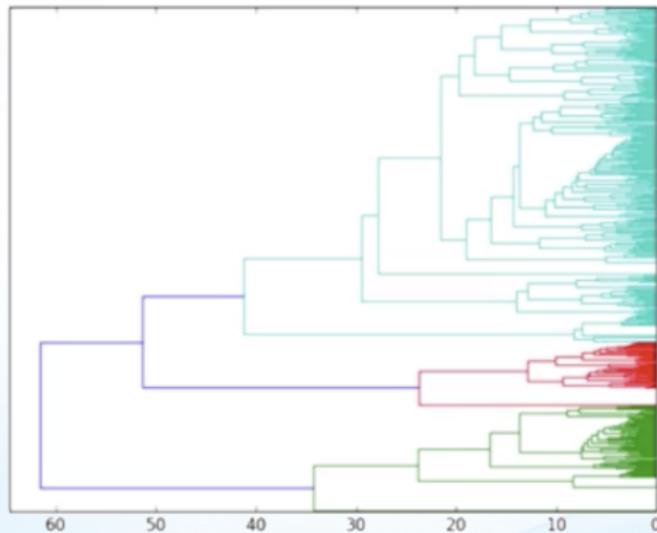
Полезно посмотреть на график расстояния между сливаемыми кластерами от номера итерации.



На последних итерациях расстояния резко увеличиваются.

⇒ существует некоторое разумное число кластеров, которые более менее друг от друга удалены.

Агломеративная иерархическая кластеризация



Особенность иерархической кластеризации —
часто возникает один большой кластер и несколько маленьких.
Но иногда хочется кластеры примерно одинакового размера.

Агломеративная иерархическая кластеризация

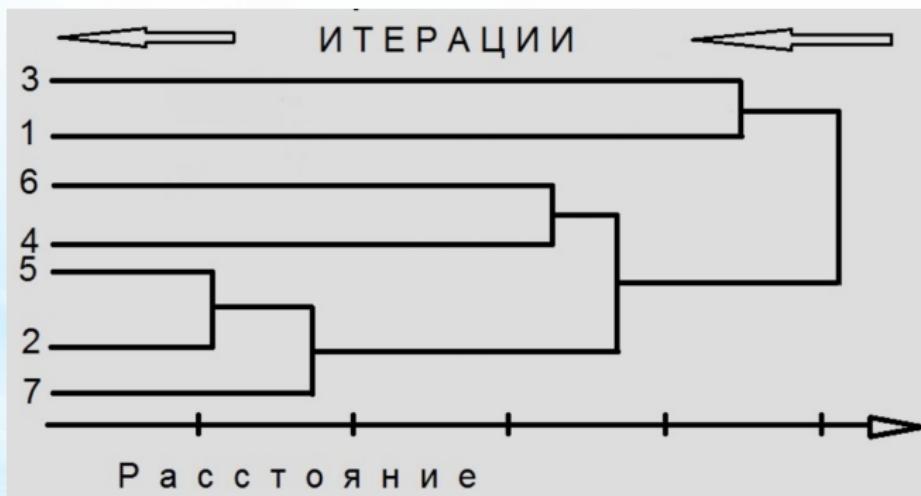
Рекомендации

- ▶ Пользоваться расстоянием Уорда R^Y .
- ▶ Обычно строят несколько вариантов и выбирают лучший визуально по дендограмме.
- ▶ Определение числа кластеров:
 1. Если число кластеров известно заранее, то обрезать дендограмму так, чтобы получить требуемое число кластеров.
 2. Иначе обрезать если разница расстояний соседних объединений $|R_{t+1} - R_t|$ стала больше некоторого порога.
Результирующее множество кластеров — C_t .

Дивизионная иерархическая кластеризация

Является логической противоположности агломеративным методам.

В начале работы все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры.



Дивизионная иерархическая кластеризация

Пример дивизионного метода

Пусть $C_t = \{K_1, K_2, \dots, K_s\}$ — набор кластеров на итерации t .

1. Все элементы входят в один кластер: $C_1 = \{x_1, x_2, \dots, x_n\}$
2. Итерации: для всех $t \in \{2, \dots, n\}$
3. Найти в C_{t-1} кластер K и элементы $x, z \in K$ такие, что расстояние ρ между ними максимально среди всех пар элементов из одного кластера.
4. Делим кластер K на 2 кластера K_1 и K_2 с элементами x и z соответственно.

Относим остальные элементы из K к двух кластерам:

$x' \in K_1$ если $\rho(x', x) \leq \rho(x', z)$.

$x' \in K_2$ если $\rho(x', x) > \rho(x', z)$.

$$C_t = C_{t-1} \cup \{K_1\} \cup \{K_2\} \setminus \{K\}$$

Theta



BCE !