AI Community

7. Кластеризация

План лекции

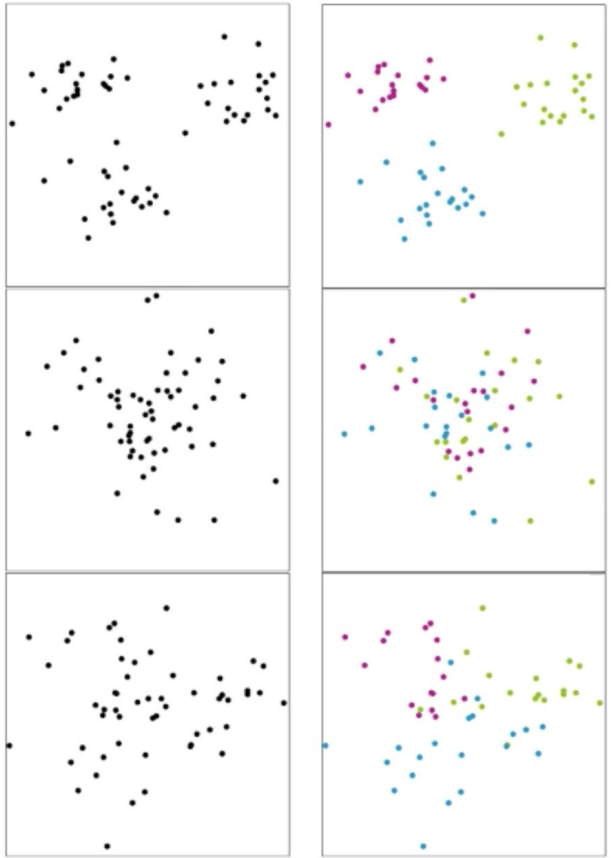
1. Определение
2. K-means
3. K-means++
4. Иерархическая кластеризация
    a. Дендограммы
    b. Разделительный подход
    c. Агломеративный подход
5. DBSCAN

Кластеризация

Кластеризация

Кластеризация является примером обучения без учителя (unsupervised learning). Задача похожа на классификацию с тем отличием, что у нас нет информации какой объект какому кластеру принадлежит и сколько кластеров у нас вообще.

$$\mathbb{D} = \{(x_i)|x_i \in \mathbb{R}^p\}_{i=1}^m$$



Цель кластеризации

Мы не можем делать предсказания, так как нет меток, которые можно было предсказать.

Но вместо этого мы можем найти скрытые паттерны и структуры в данных, что даст возможность в дальнейшем обрабатывать кластеры определенным образом, а также находить различные аномалии в данных.

Кластеризация пользователей

SPORTS

WORLD NEWS

Метрики оценки разности кластеров

1. Евклидово расстояние (L2 норма)
2. Манхэттенское расстояние (L1 норма)
3. Косинусная мера (нормированное скалярное произведение)
4. Метрики основанные на корреляции

K-means

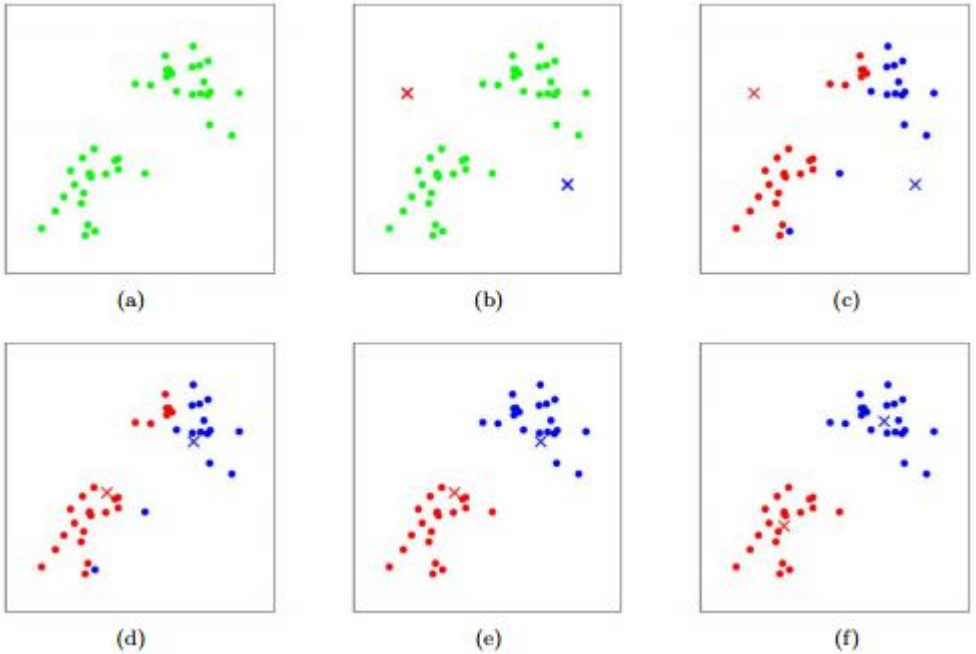
K-means

Простой алгоритм, позволяющий разделить датасет на ***K*** непересекающихся кластеров.

Но необходимо выбрать, сколько будет кластеров (число ***K***) и как инициализировать центры кластеров (случайно, ручной выбор, наиболее дальние точки)

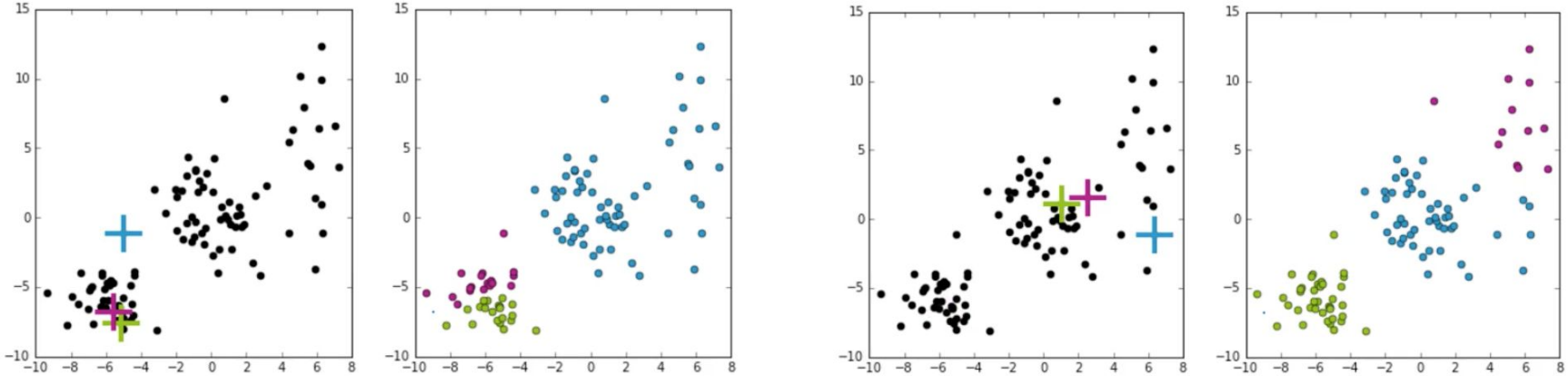
K-means

1. Выбираем центры кластеров
2. Добавляем точки в кластер, центр которого ближе всего.
3. Пересчитываем центры как среднее среди всех точек этого кластера
4. Возвращаемся к шагу 2, пока центры кластеров не перестанут менятся



K-means

Алгоритм очень чувствителен к начальной инициализации



Необходим более умный подход к начальной инициализации

K-means++

K-means++

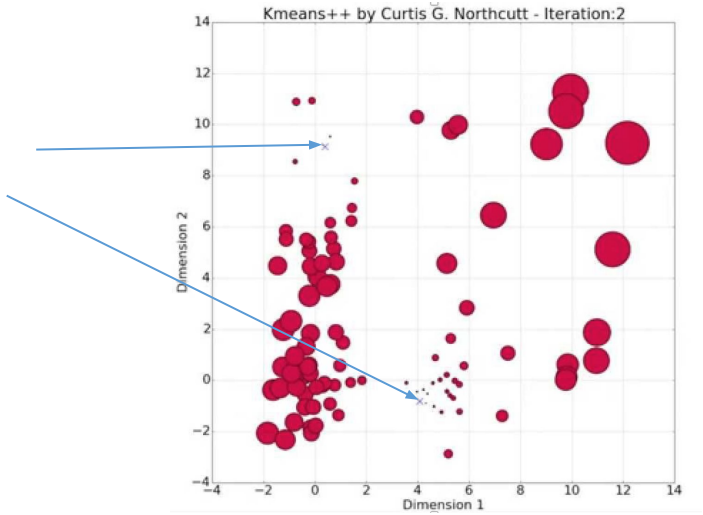
Основная идея алгоритма в том, что мы делаем начальную инициализации более умным способом.

Первый центр выбирается случайно с равной вероятностью среди всех точек.

Все последующие центры выбираются с вероятностью пропорционально квадрату расстояния до ближайшего центра кластера

K-means++

Уже выбранные центры



Размер точки пропорционален квадрату расстояния до ближайшего центра и определяет вероятность выбора этой точки в качестве нового центра

Как выбрать K?

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x$$

Центр кластера

$$RSS_i = \sum_{x \in C_i} (x - \mu_i)^2$$

Дисперсия кластера

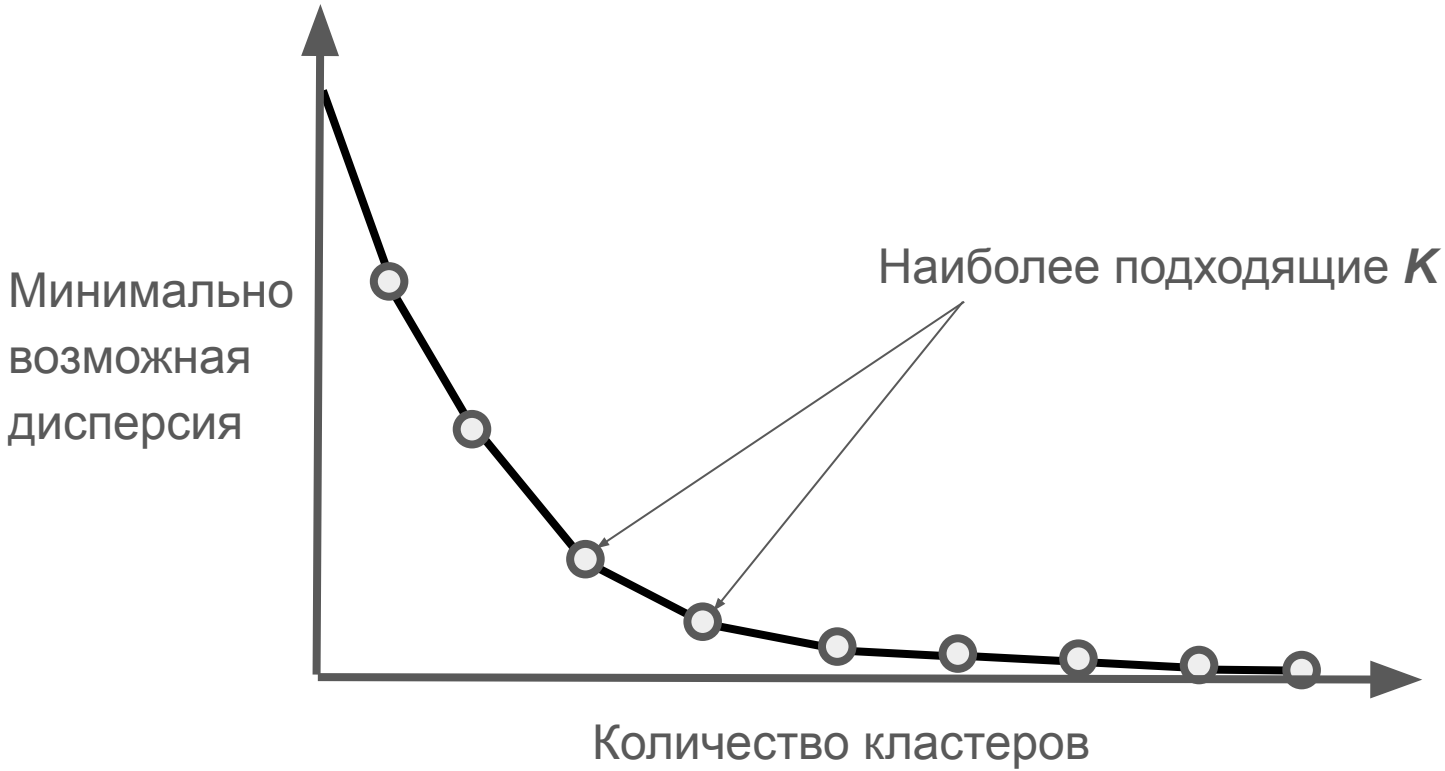
$$RSS = \sum_{i=1}^{k} RSS_i$$

Общая дисперсия

Мы хотим, чтобы общая дисперсия была как можно меньше. Но не всё так просто.

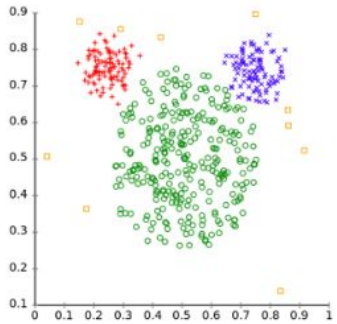
При ***K*** = 1 у нас будет максимальная дисперсия. При ***K*** = ***N*** дисперсия будет равна 0, но такая кластеризация бесполезна.

Как выбрать K?

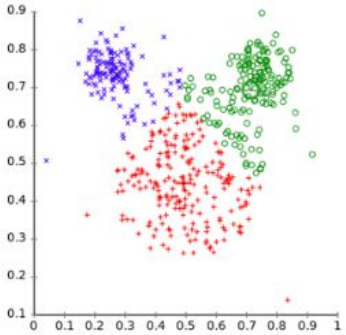


Проблемы K-means

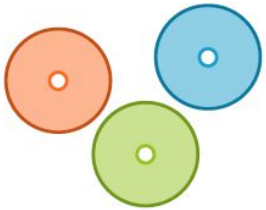
Исходные данные



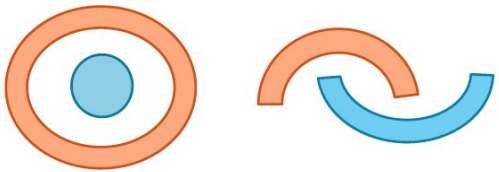
K-means



Хорошо на сферических кластерах



Но что будет на таких?



1. Необходимость выбора оптимального ***K***
2. Работает только с кластерами одинакового размера
3. Плохо работает для кластеров, имеющих не сферическую форму
4. Не учитывает выбросы в данных

Иерархическая Кластеризация

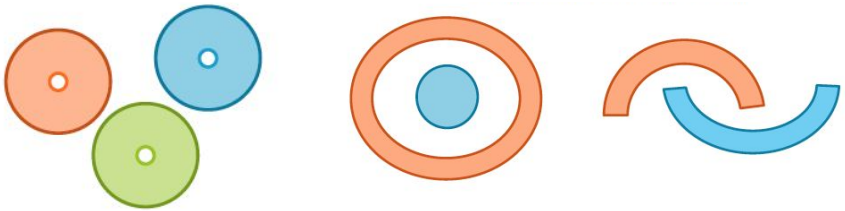
Иерархическая кластеризация

Способна находить более сложные формы, по сравнению с K-Means.

Представлена двумя классами методов:

- Агломеративные - подход **снизу-вверх**
- Разделительные (дивизионные) - подход **сверху-вниз**

Примеры, на которых может работать иерархическая кластеризация



Агломеративная кластеризация

Начинаем с того, что каждую точку в датасете объявляем кластером, состоящим из одного элемента.

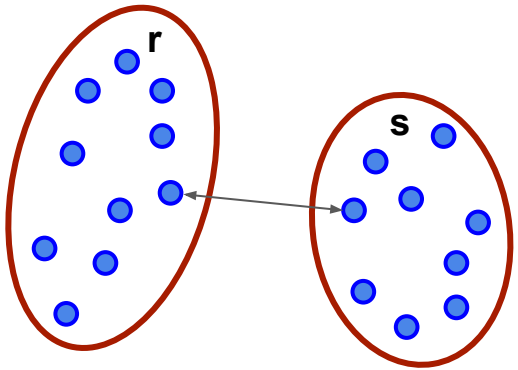
Затем на каждом шаге **объединяем два наиболее близких кластера** в один до тех пор, пока не останется один кластер, содержащий в себе все точки.

Необходимо определить, как измерять близость кластеров.

Дистанция между кластерами

Используется несколько подходов к оценке расстояния:

1. Метод одиночной связи - расстояние между двумя наиболее близкими точками из разных кластеров.

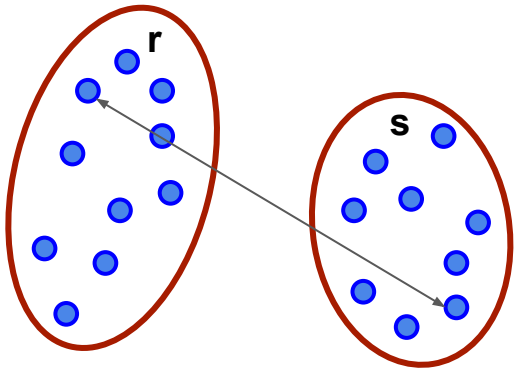


$$L(r,s) = \min\left(D\left(x_{ri}, x_{si}\right)\right)$$

Дистанция между кластерами

Используется несколько подходов к оценке расстояния:

1. Метод одиночной связи - расстояние между двумя наиболее близкими точками из разных кластеров.
2. Метод полной связи - расстояние между двумя наиболее далекими точками из разных кластеров.

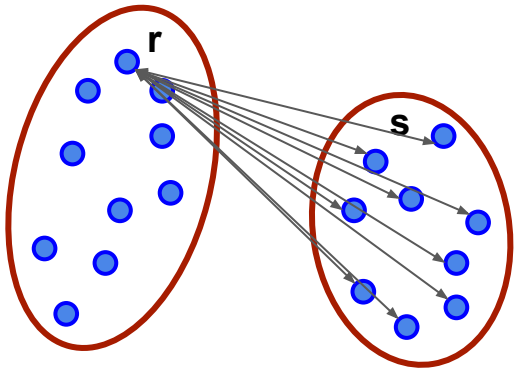


$$L(r, s) = \max\left(D\left(x_{ri}, x_{si}\right)\right)$$

Дистанция между кластерами

Используется несколько подходов к оценке расстояния:

1. Метод одиночной связи - расстояние между двумя наиболее близкими точками из разных кластеров.
2. Метод полной связи - расстояние между двумя наиболее далекими точками из разных кластеров.
3. Метод средней связи - среднее расстояние между всеми точками двух кластеров.

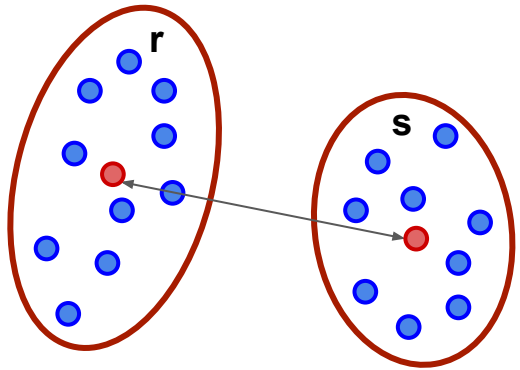


$$L(r,s) = \frac{1}{|r||s|} \sum_{i=1}^{|r|} \sum_{j=1}^{|s|} \left(D\left(x_{ri}, x_{sj}\right) \right)$$

Дистанция между кластерами

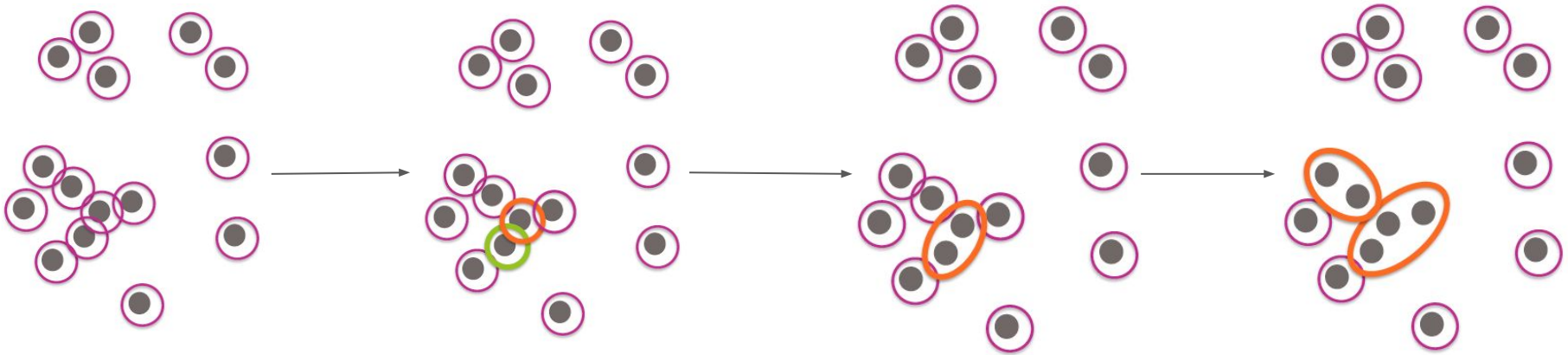
Используется несколько подходов к оценке расстояния:

1. Метод одиночной связи - расстояние между двумя наиболее близкими точками из разных кластеров.
2. Метод полной связи - расстояние между двумя наиболее далекими точками из разных кластеров.
3. Метод средней связи - среднее расстояние между всеми точками двух кластеров.
4. Центроидный метод - расстояние между центроидами кластеров.

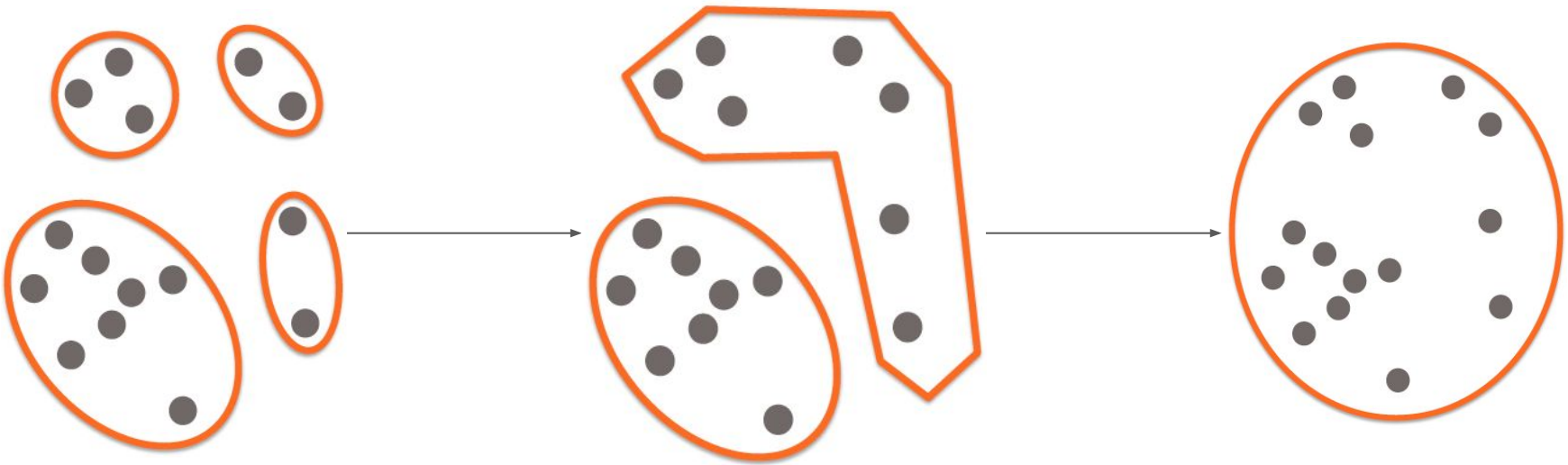


$$L(r,s) = D(c_r, c_s)$$

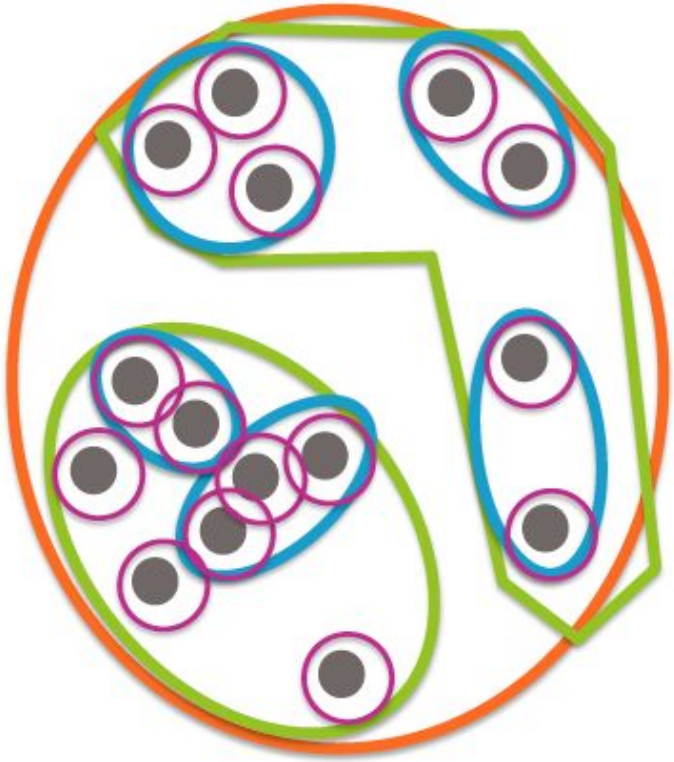
Агломеративная кластеризация



Агломеративная кластеризация



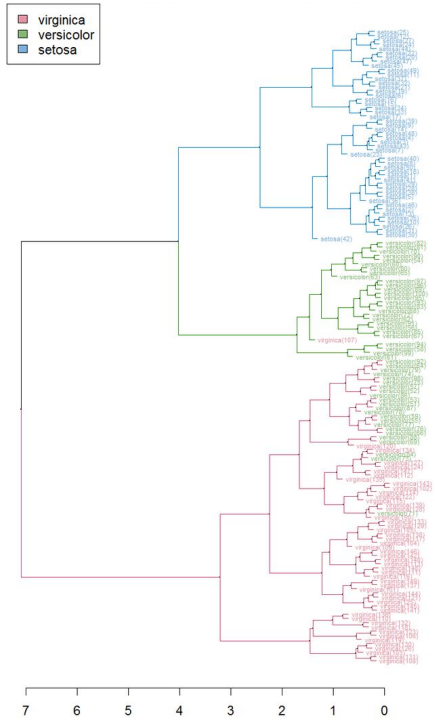
Агломеративная кластеризация



Дендограммы

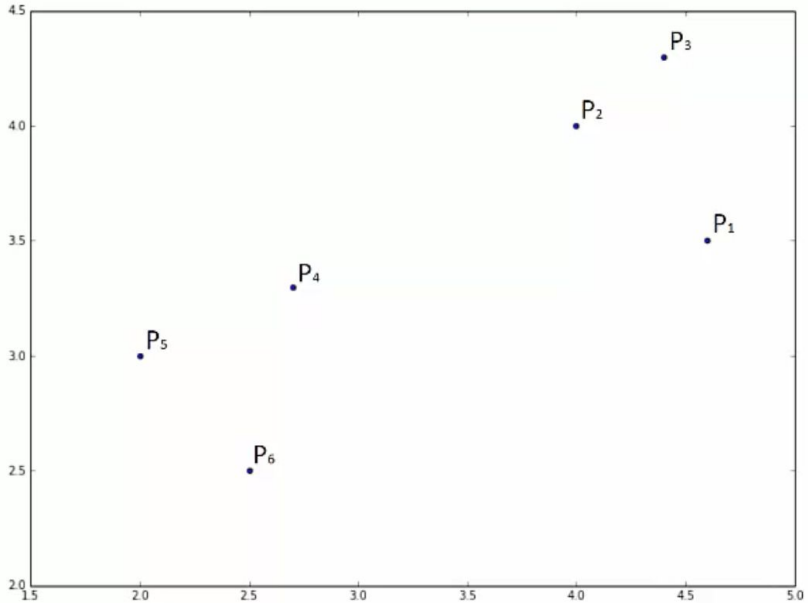
Иерархическая кластеризация позволяет строить **дендограммы** - дерево вложенных кластеров.

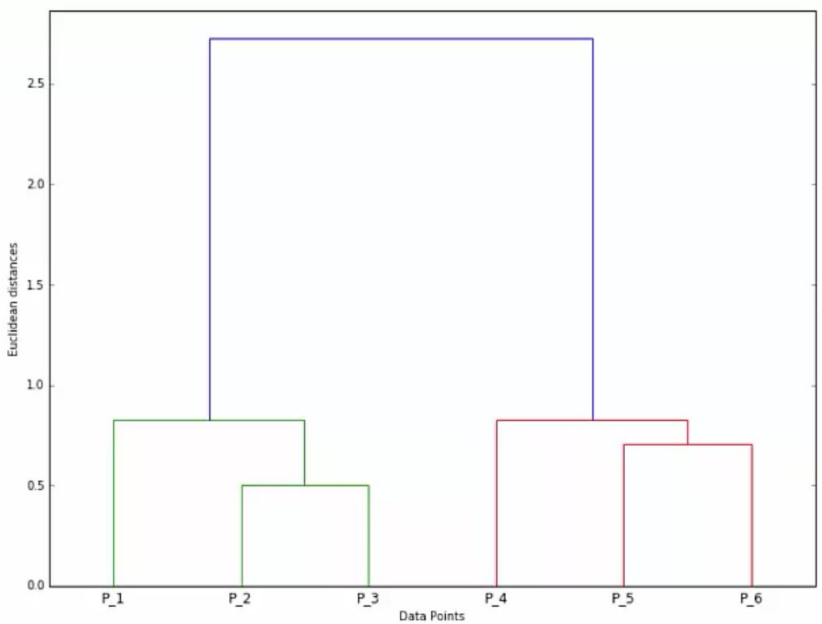
По ось X показывает дистанцию между парой кластеров в точке разветвления



Дендограмма, полученная после иерархической кластеризации датасета **Iris**

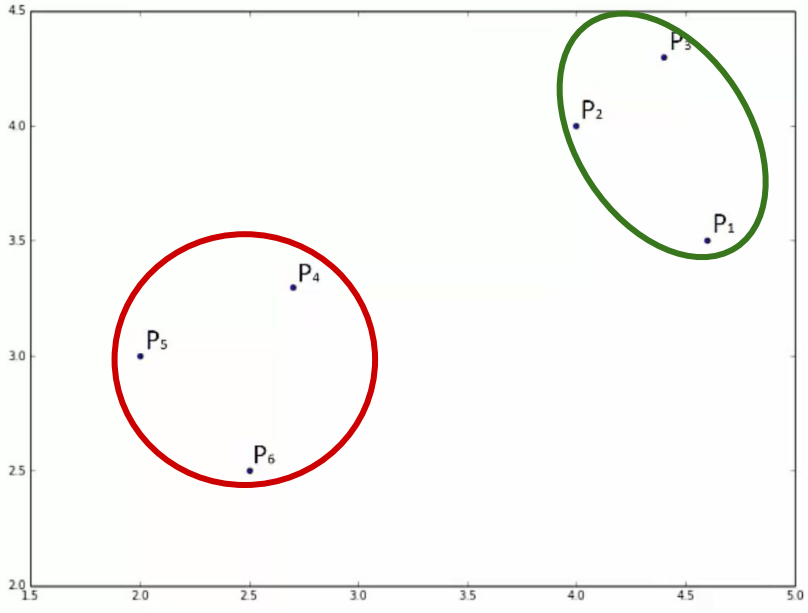
Оптимальное разделение

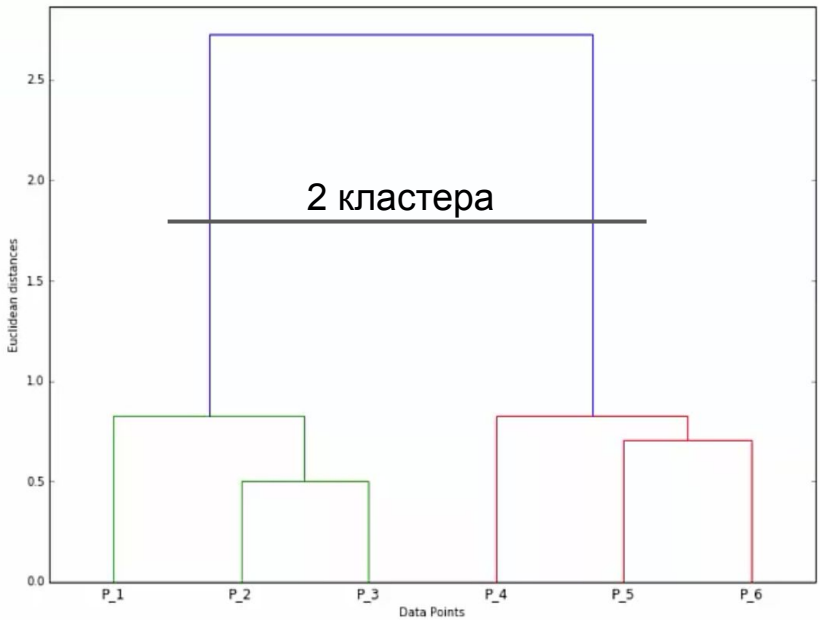




Оптимальное разделение

Половина наиболее длинной вертикальной линии





DBSCAN

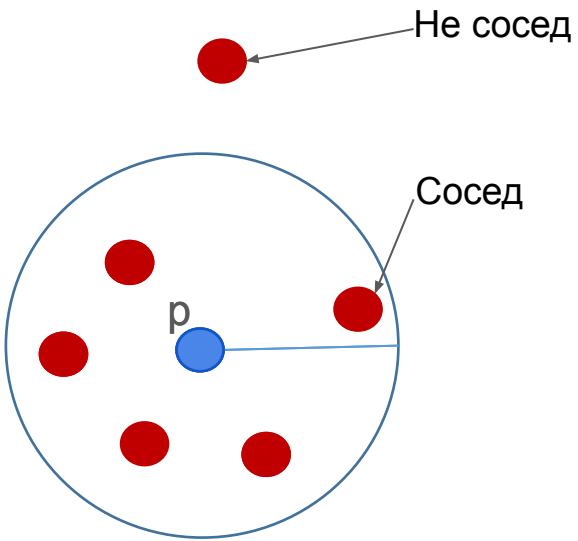
DBSCAN

DBSCAN (Density-based spatial clustering of applications with noise) - имеет схожий принцип работы с иерархической кластеризацией, но при этом позволяет находить выбросы в данных и не включать их итоговые кластеры.

Строит кластеры, основываясь на плотности распределения точек, отделяя области высокой плотности друг от друга.

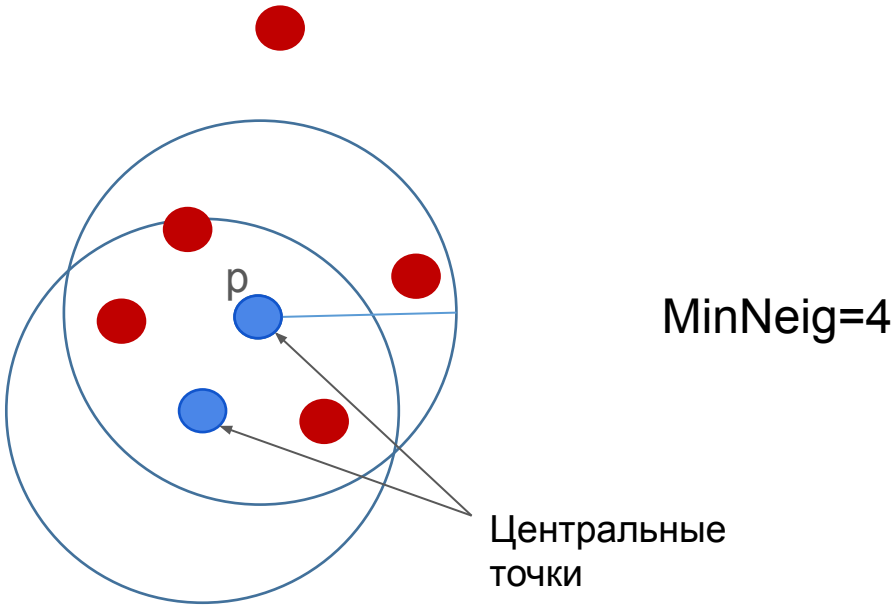
DBSCAN

Соседями точки ***p*** являются точки, удалённые от неё на расстояние не большее, чем ***r***.



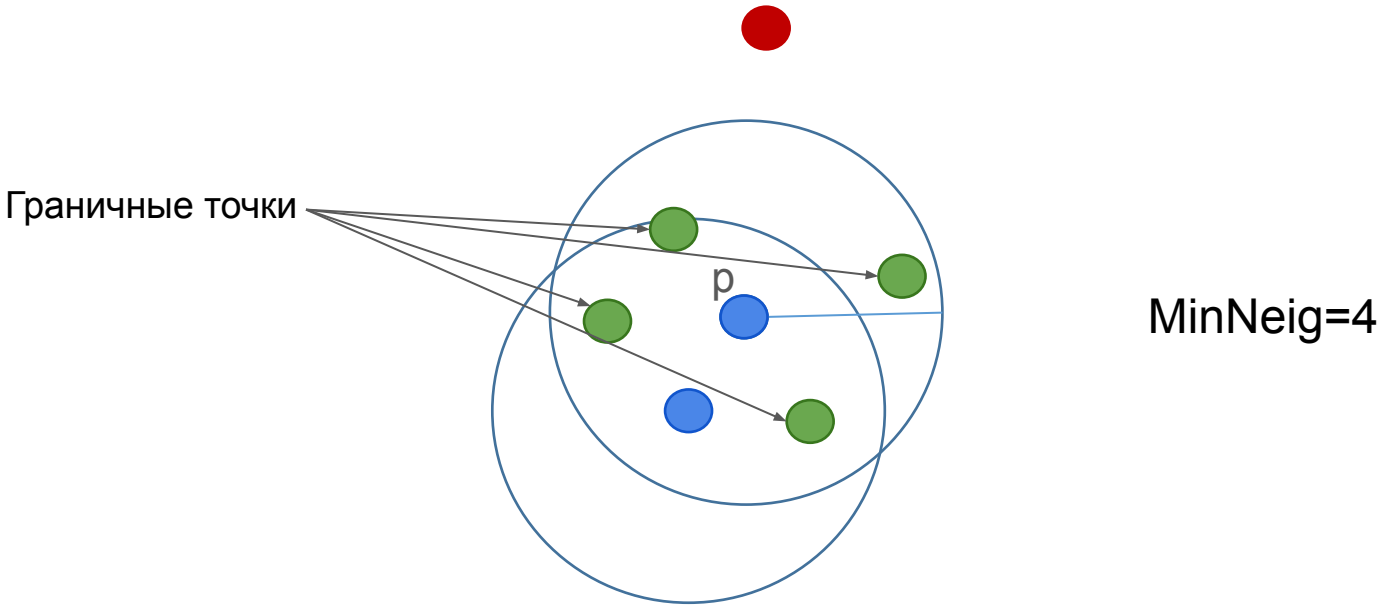
DBSCAN

Центральной точкой является точка, количество соседей которых равно или превышает некоторый порог ***MinNeig***.



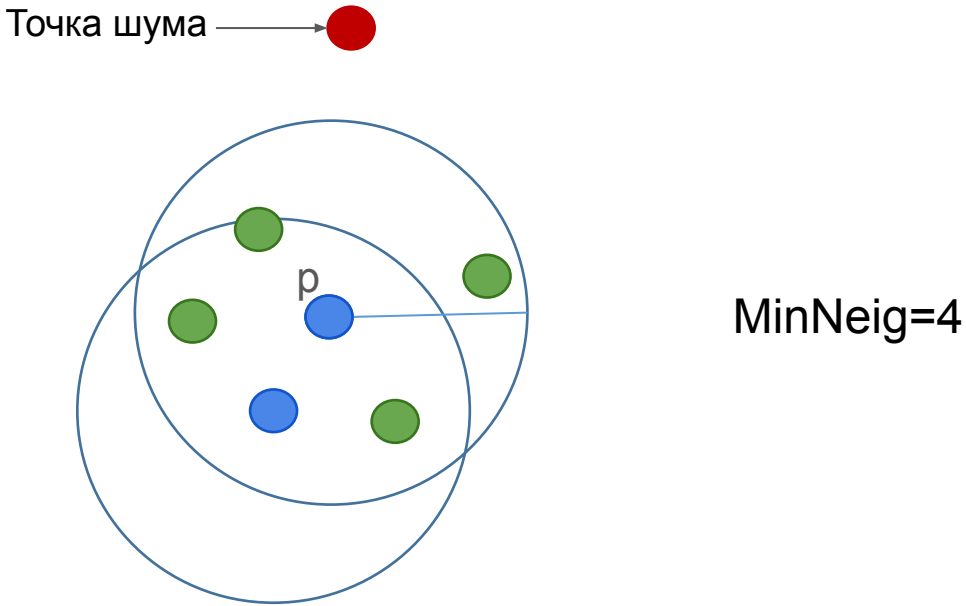
DBSCAN

Граничной точкой является точка, которая не является центральной точкой но имеет среди соседей центральную точку.



DBSCAN

Точкой шума является точка, которая не является ни центральной точкой, ни граничной точкой.



Достижимость и Связанность

Точка ***q*** **прямо достижима** из точки ***p*** если ***p*** - центральная точка и точка ***q*** является соседом точки ***p***.

Точка ***q*** **достижима** из точки ***p*** если найдётся такая цепь точек ***p_1, p_2, … p_n***, что ***p = p_1***, ***q = p_n***, и ***p_(i+1)*** прямо достижима из точки ***p_i***.

Точки ***q*** и ***p*** связаны, если найдётся такая точка ***c***, что обе точки ***q*** и ***p*** достижимы из ***c***.

Кластер в DBSCAN

Кластером ***D*** называется такое множество точек, в котором для всех точек ***q*** и ***p*** принадлежащих ***D*** выполняется условие связанности ***q*** и ***p.***

При этом, если точка ***p*** принадлежит ***D*** и точка ***q*** достижима из точки ***p***, то ***q*** принадлежит ***D.***

Пример