




ОНЛАЙН-ОБРАЗОВАНИЕ

Онлайн-образование



Меня хорошо видно && слышно?

Ставьте , если все хорошо
Напишите в чат, если есть проблемы

Обучения без учителя

Иерархическая кластеризация, DBSCAN



Андрей Канашов
Senior Data Scientist
BestDoctor

Проверить, идет ли запись!



Преподаватель



Андрей Канашов

- Senior Data Scientist в BestDoctor
 - Прогнозирование ключевых метрик
 - Мэтчинг медицинских услуг
 - Рекомендации клиник
 - Тарификация и ценообразование

Правила вебинара



Активно участвуем



Задаем вопрос в чат или голосом



Off-topic обсуждаем в Slack #канал группы или #general



Вопросы вижу в чате, могу ответить не сразу

Цели вебинара | После занятия вы узнаете

1

Иерархическая кластеризация

2

DBSCAN

3

Силуэтный коэффициент

4

Применить алгоритмы на практике



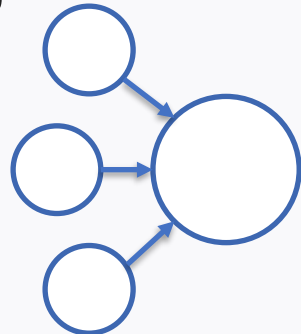
Иерархическая кластеризация

Иерархическая кластеризация

Agglomerative clustering

Агломеративная кластеризация

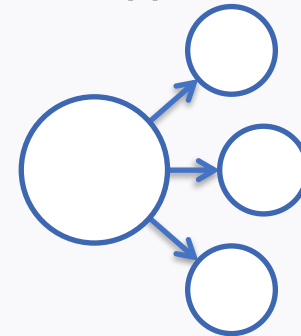
- Каждая точка – отдельный кластер
- На каждом шаге объединяются два ближайших кластера
- Объединение продолжается до тех пор, пока данные не сольются в один кластер



Divise clustering

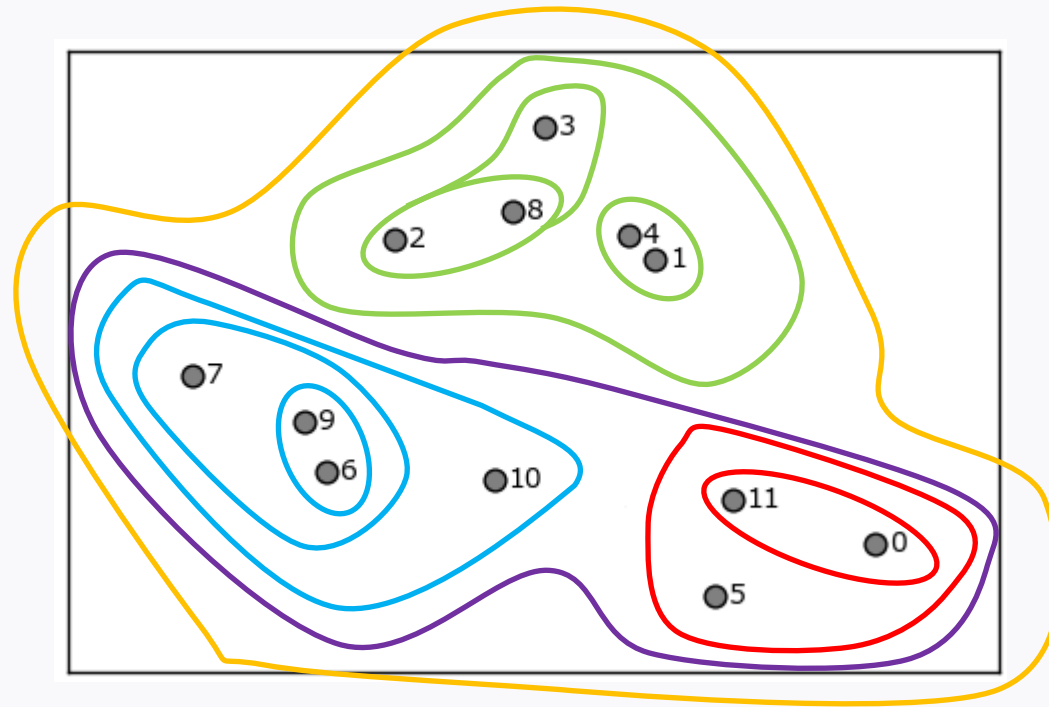
Дивизивная кластеризация

- Все данные – один большой кластер
- На каждом шаге разделяется один из кластеров на две части
- Разделение продолжается до тех пор, пока кластеры не будут состоять из одной точки



Agglomerative clustering

Объединение данных



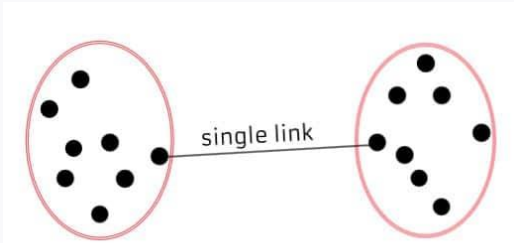
Как определить расстояние между кластерами?



Agglomerative clustering

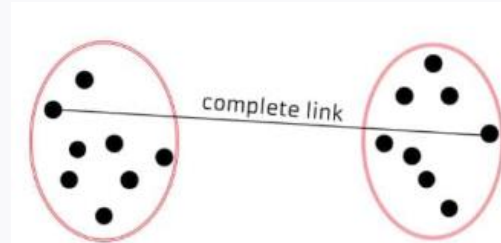
Критерии связи (linkage)

Single linkage



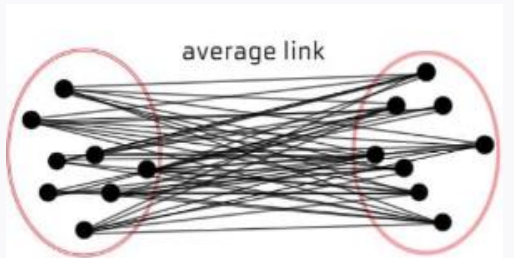
$$D(X, Y) = \min_{x \in X, y \in Y} d(x, y)$$

Complete linkage



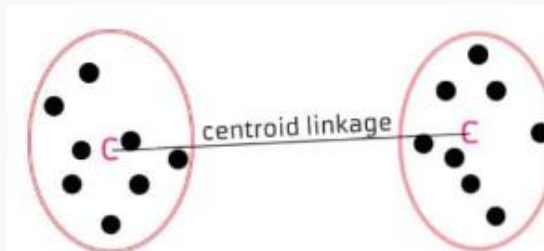
$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y)$$

Average linkage



$$D(X, Y) = \frac{1}{|X| \cdot |Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y)$$

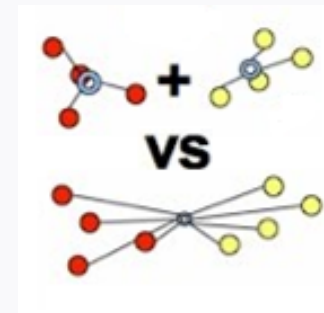
Centroid linkage



$$D(X, Y) = \|c_x - c_y\|$$

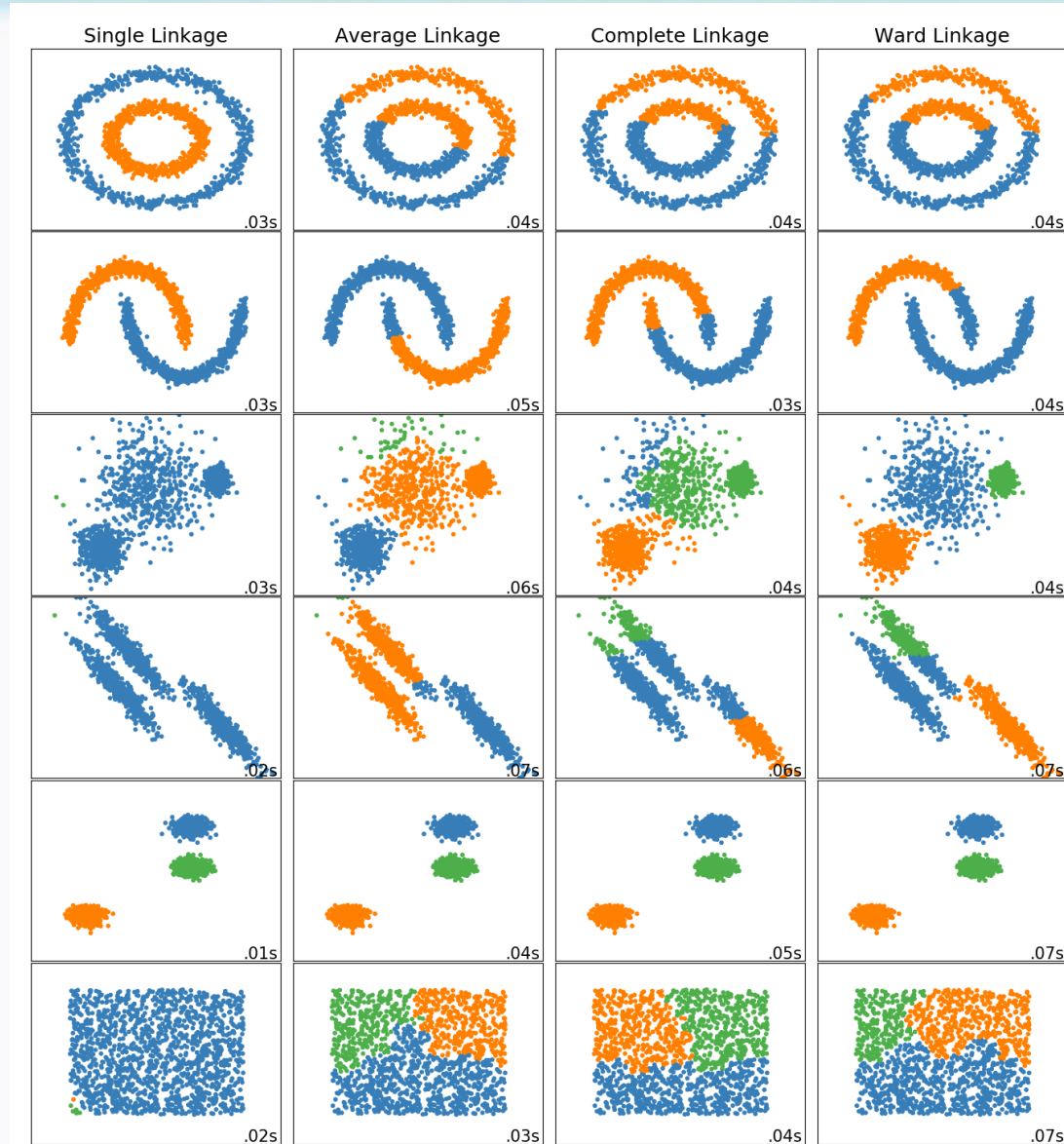
Ward linkage

В качестве расстояния между кластерами берётся прирост дисперсии, т.е. суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения



$$L_{Ward}(X, Y) = \sum_{x \in X} \sum_{y \in Y} \|x - y\|^2$$

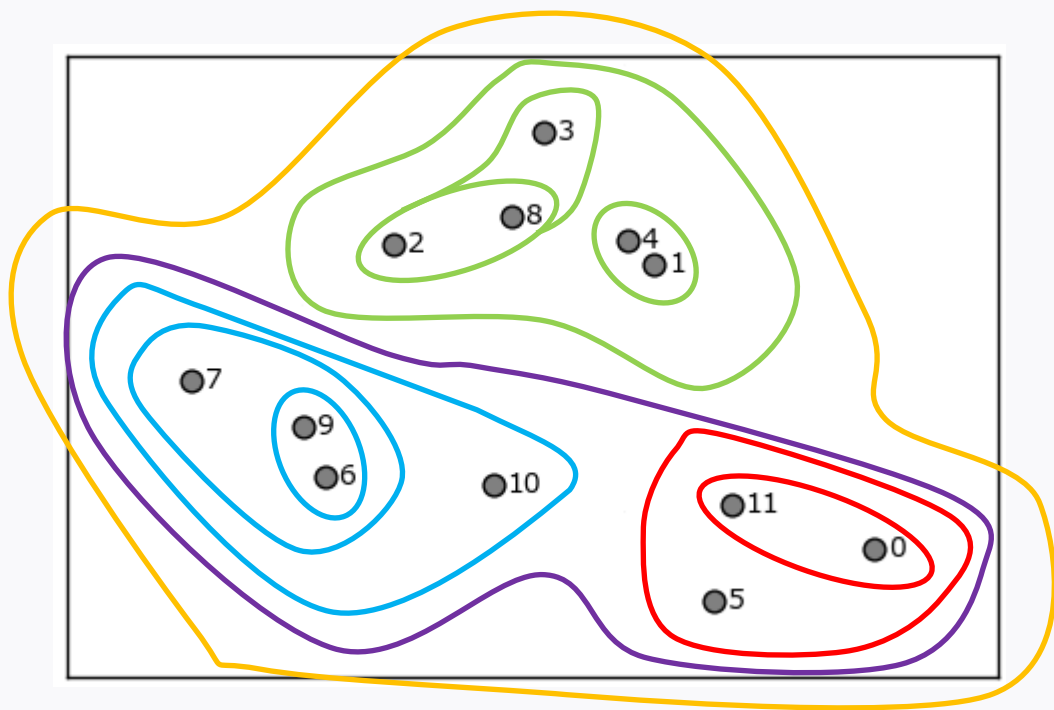
Agglomerative clustering



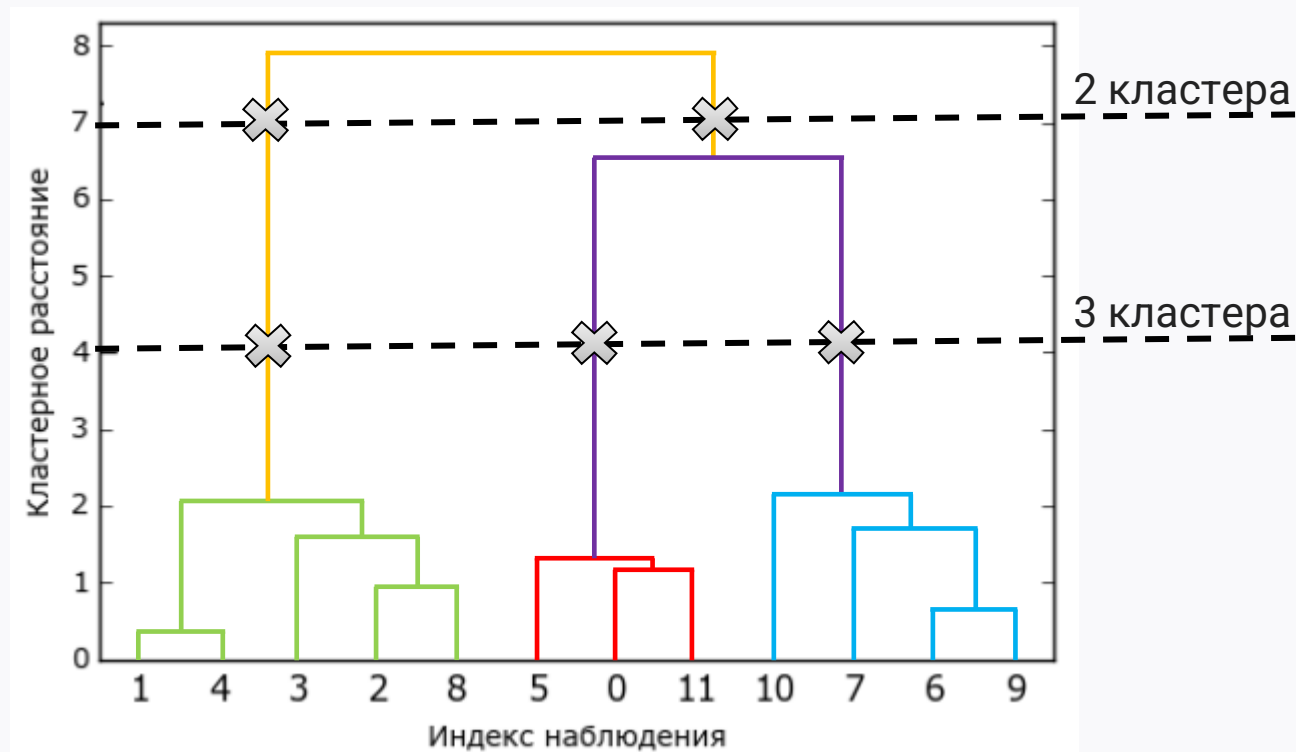
Agglomerative clustering

Выбор числа кластеров

Объединение данных



Дендрограмма



Плюсы и минусы иерархической кластеризации

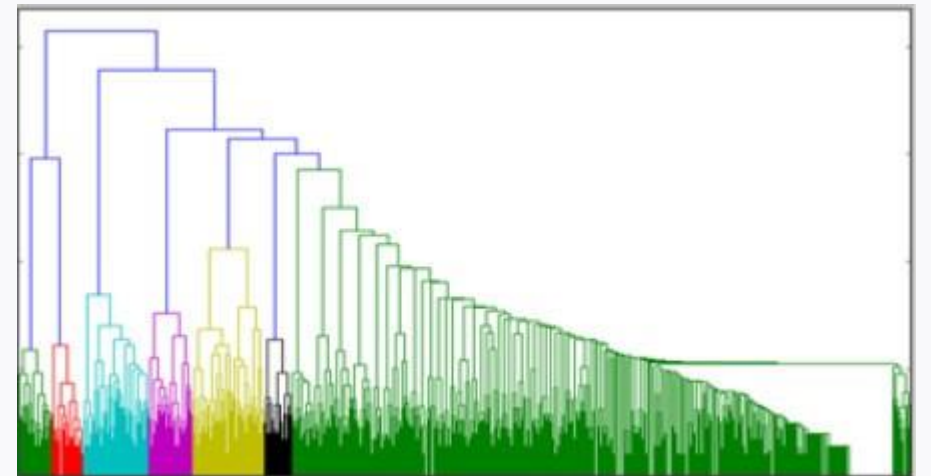
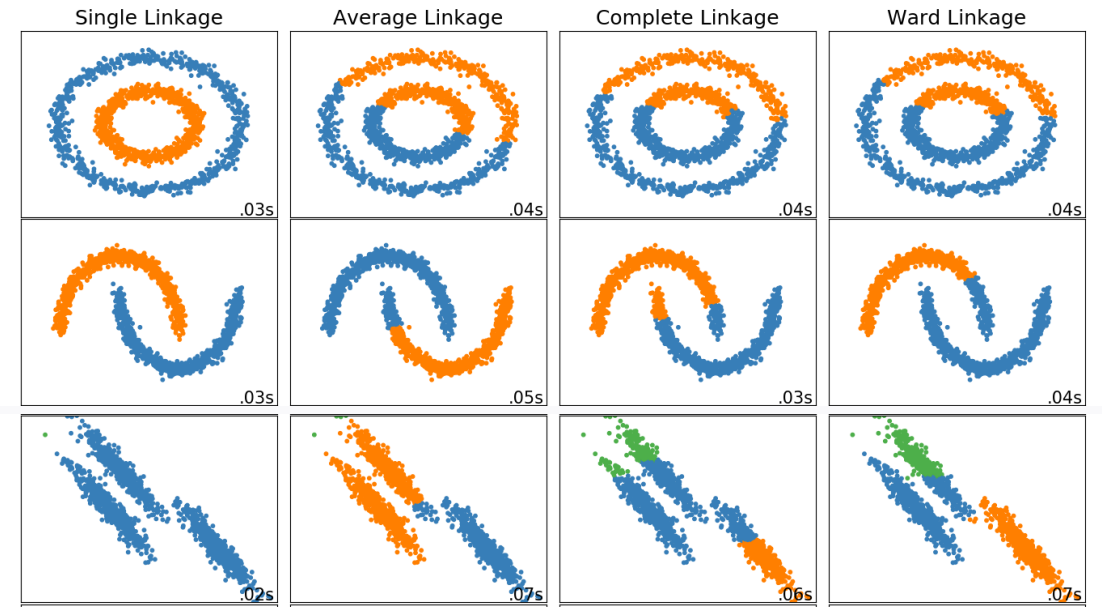
Плюсы:

- + простая в реализации
- + иерархия кластеров



Минусы:

- вычислительные ресурсы и память
- не всегда может выделить сложные формы кластеров
- дендрограмма для небольшого объема данных
- выбор критериев связи



The background of the slide is a high-angle, blue-tinted aerial photograph of a dense urban skyline, likely New York City. Overlaid on this image is a semi-transparent network diagram consisting of numerous small blue dots connected by thin, light-blue lines, creating a web-like pattern across the center of the slide. The text 'DBSCAN' is centered within this network pattern.

DBSCAN

DBSCAN

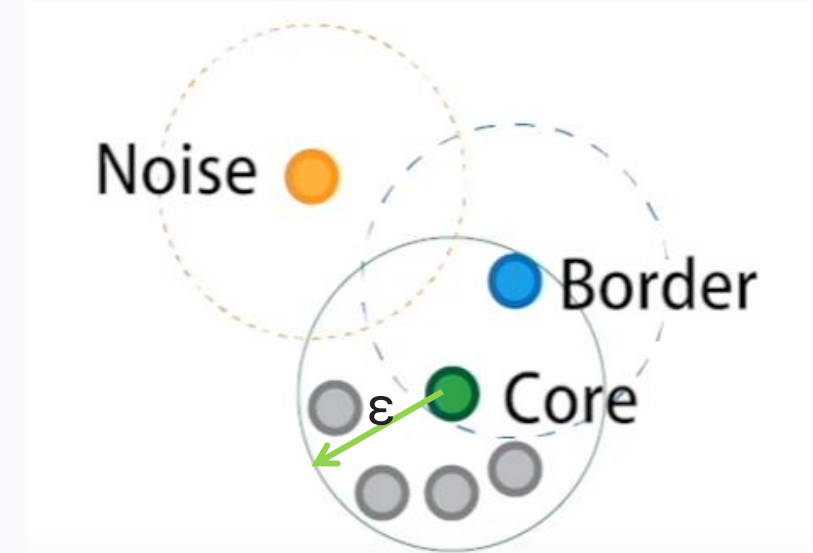
DBSCAN (density-based spatial clustering of applications with noise) - плотностный алгоритм кластеризации пространственных данных с присутствием шума

Параметры:

- **eps** – радиус окрестности
- **min_samples** – минимальное количество точек, которое должно находиться в окрестности *eps*

Типы точек:

- **core points (ядровые точки)** – в радиусе *eps* находится не менее *min_samples* точек
- **border points (пограничные точки)** – находятся в пределах радиуса окрестности ядровых точек, при этом в радиусе своей окрестности имеют меньше *min_samples* точек
- **noise points (шумовые точки)** – в радиусе окрестности *eps* меньше *min_samples* точек и не попадают в окрестность ядровых точек

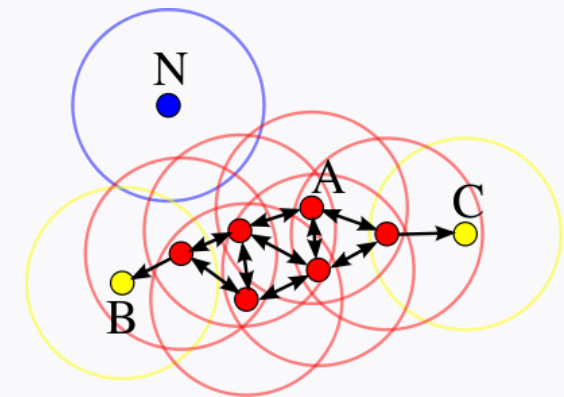
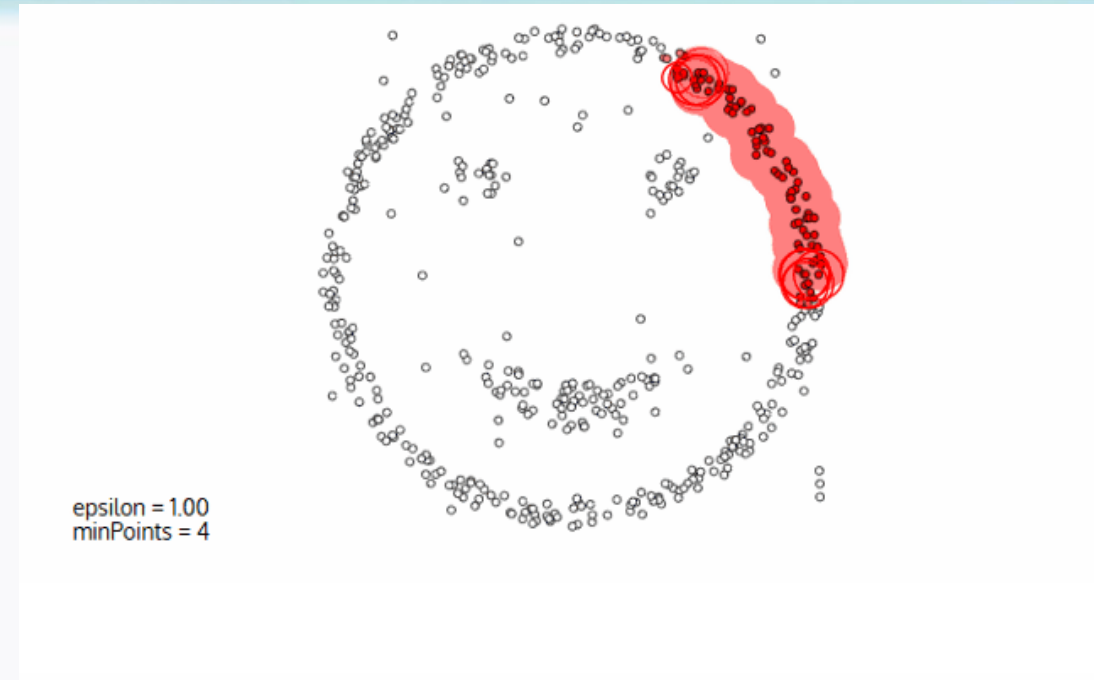


DBSCAN

Выбирается произвольная точка

- Находятся все точки, удаленные от стартовой точки на расстоянии, не превышающем радиуса окрестности ϵ .
- Если множество точек, находящихся в пределах радиуса окрестности ϵ , меньше значения min_samples , стартовая точка помечается как шум (noise).
- Если это множество точек больше значения min_samples , стартовая точка помечается как ядровая и ей назначается метка нового кластера.
- Затем посещаются все соседи этой точки (находящиеся в пределах ϵ). Если они еще не были присвоены кластеру, им присваивается метка только что созданного кластера. Если они являются ядровыми точками, поочередно посещаются их соседи и т.д.

Кластер растет до тех пор, пока не останется ни одной ядерной точки в пределах радиуса окрестности ϵ . Затем выбирается другая точка, которая еще не была посещена, и повторяется та же самая процедура.



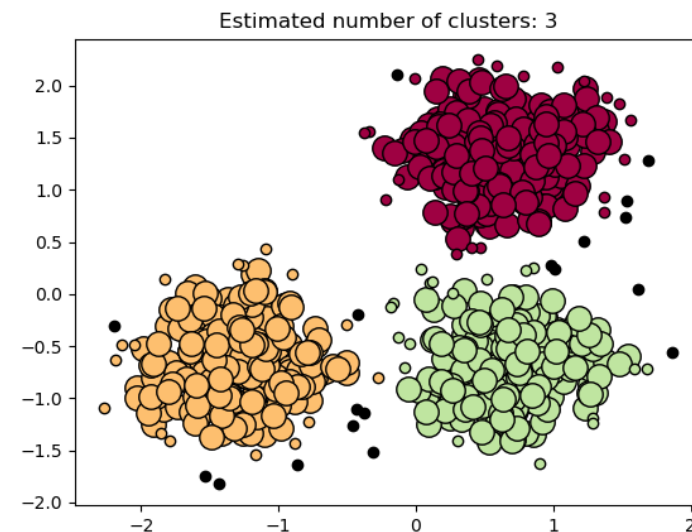
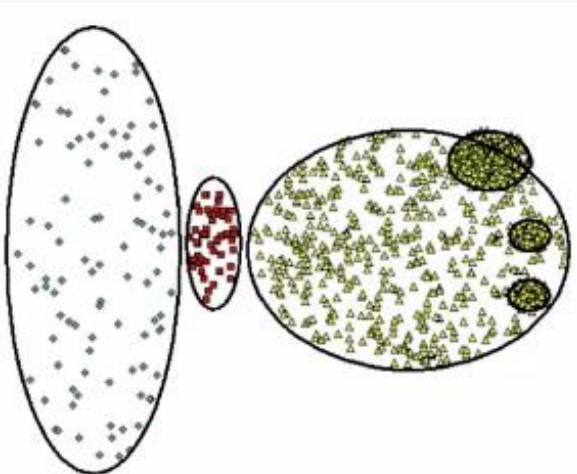
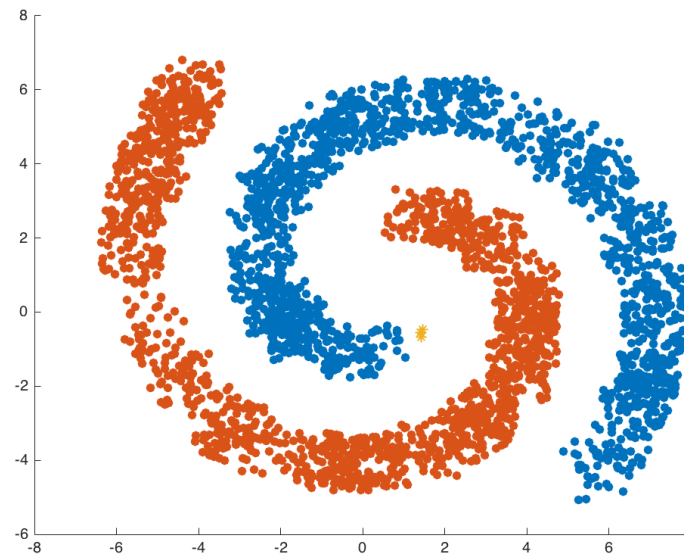
Плюсы и минусы DBSCAN

Плюсы:

- + сам определяет количество кластеров
- + выделяет сложные формы кластеров
- + находит выбросы

Минусы:

- выбор параметров
- плохо выделяет кластеры с разной плотностью



The background of the entire image is an aerial photograph of a dense city skyline, likely New York City, with numerous skyscrapers. A semi-transparent blue overlay covers the entire image. In the center, there is a network of thin, light blue lines connecting small dots, creating a web-like pattern. The text "Силуэтный коэффициент" is written in a large, white, sans-serif font across the middle of the image.

Силуэтный коэффициент

Силуэтный коэффициент

Показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров.

$$s = \frac{b - a}{\max(a, b)}$$

a — среднее расстояние от данного объекта до объектов из того же кластера

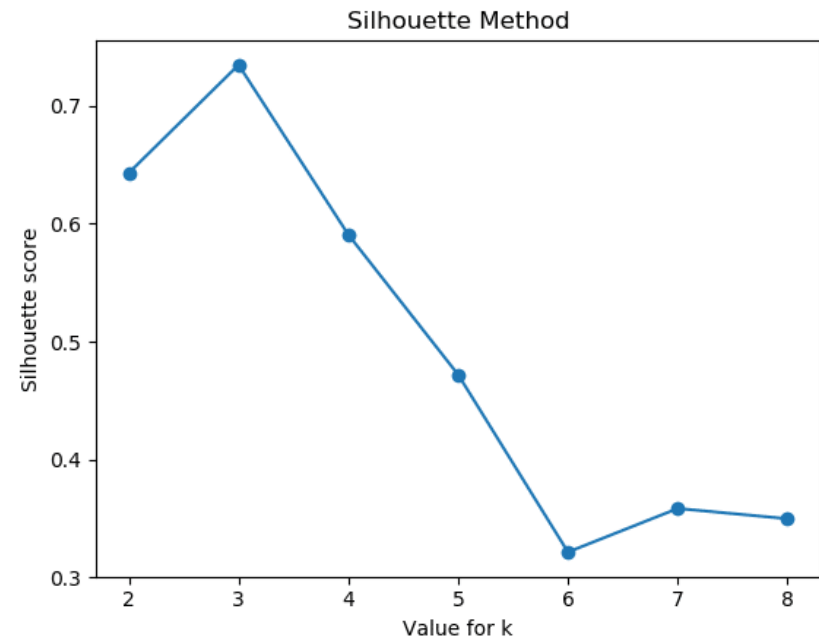
b — среднее расстояние от данного объекта до объектов из другого ближайшего кластера

Диапазон коэффициента: $[-1, 1]$

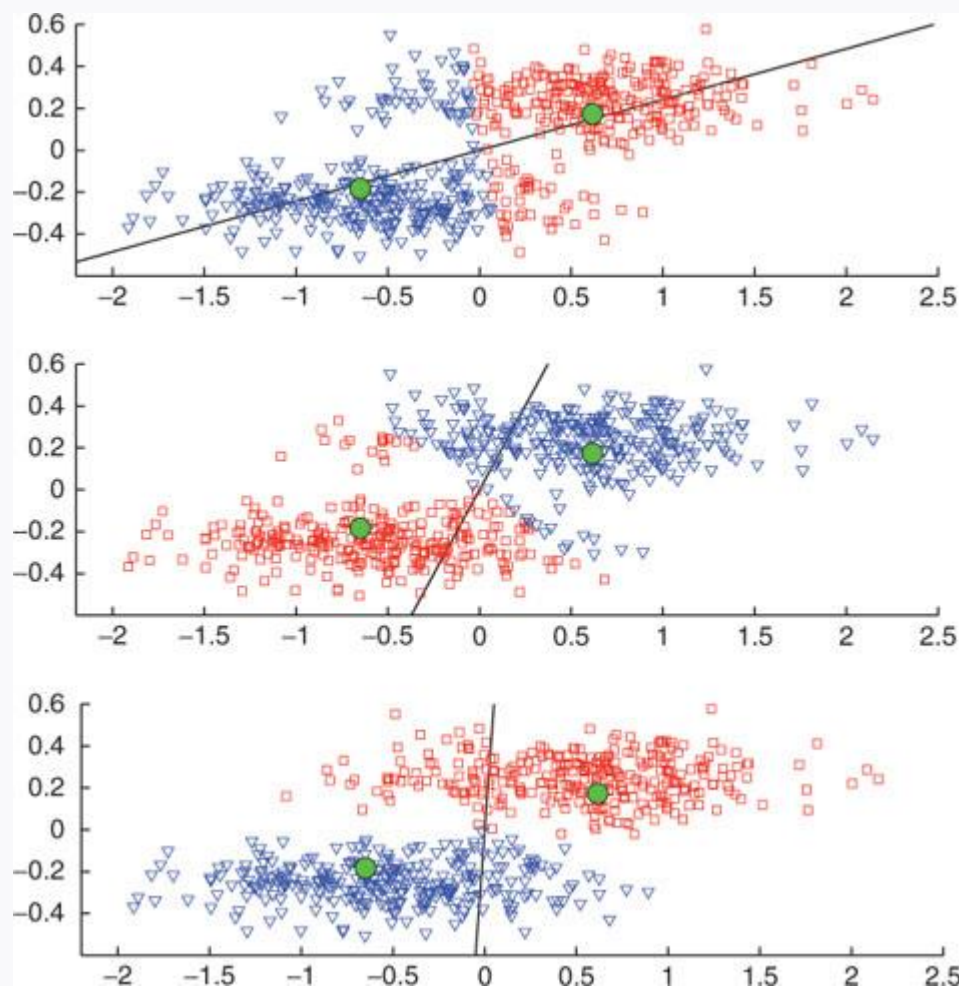
-1 — худший вариант, разрозненная кластеризация


0 — кластеры пересекаются и накладываются

1 — лучший вариант, четко выделенные кластеры



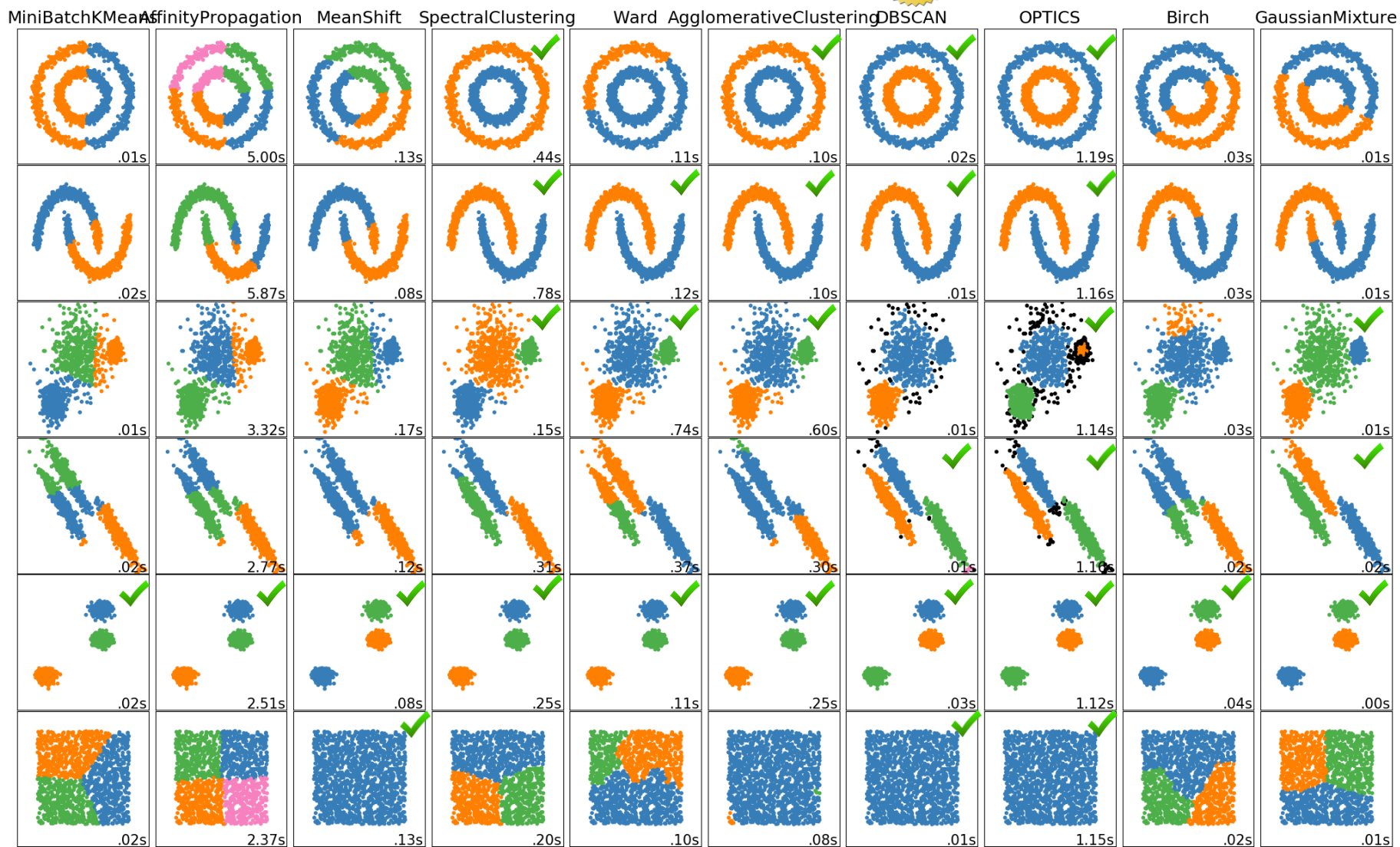
Силуэтный коэффициент





Сравнение алгоритмов кластеризации

Сравнение кластеризации





LIVE



Дополнительные материалы

Обзор кластеризации из библиотеки sklearn:

<https://scikit-learn.org/stable/modules/clustering.html>

<https://scikit-learn.org/stable/modules/clustering.html#hierarchical-clustering>

<https://scikit-learn.org/stable/modules/clustering.html#dbscan>

Силуэтный коэффициент из библиотеки sklearn:

<https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient>

Визуализация DBSCAN:

<https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>

Дополнительные материалы

OPTICS:

<https://scikit-learn.org/stable/modules/clustering.html#optics>

https://en.wikipedia.org/wiki/OPTICS_algorithm

<https://towardsdatascience.com/clustering-using-optics-cac1d10ed7a7>

HDBSCAN:

https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html

Spectral clustering:

<https://scikit-learn.org/stable/modules/clustering.html#spectral-clustering>

https://en.wikipedia.org/wiki/Spectral_clustering

Домашнее задание

Сегментация клиентов банка

- **EDA и Preprocessing**

1. Скачайте данные по клиентам немецкого банка: <https://www.kaggle.com/uciml/german-credit>
2. Проведите EDA, чтобы познакомиться с признаками
3. Преобразуйте все признаки в числовые подходящими методами
4. Приведите все данные к одному масштабу (а заодно поясните, почему это необходимая операция при кластеризации)

- **Моделирование**

1. Постройте три варианта кластеризации: k-means, hierarchical и DBSCAN, выберите оптимальное количество кластеров для каждого метода при помощи Elbow method и Silhouette plot
2. Также воспользуйтесь различными вариантами сжатия признакового пространства (PCA, UMAP, tSNE) и визуализируйте результаты кластеризации на двумерной плоскости

- **Интерпретация**

1. Теперь ваша задача - попытаться проинтерпретировать получившиеся кластеры, начните с простого расчета средних значений признаков для каждого из кластеров, есть ли интересные закономерности?
2. Теперь постройте boxplot-ы для каждого признака, сгруппировав значения по кластерам, по каким признакам заметно наибольшее отличие кластеров друг от друга? Можно ли их интерпретировать?

The background of the slide is a high-angle, blue-tinted aerial photograph of a dense urban skyline, likely New York City. Overlaid on this image is a semi-transparent blue band that contains a white network diagram. This diagram consists of numerous small dots connected by thin white lines, creating a web-like structure that spans the width of the slide. Centered within this band is the main title in a large, white, sans-serif font.

Проверка достижения целей

Цели вебинара | Проверка достижения целей

1 Как работает иерархическая кластеризация?

2 Как работает DBSCAN?

3 Какие метрики качества кластеризации?

4 Какой алгоритм кластеризации лучше всех?


Рефлексия



Достигли ли вы цели вебинара?



С какими основными мыслями и инсайтами уходите с вебинара?

The background of the image is an aerial photograph of a city with many skyscrapers, overlaid with a semi-transparent blue layer. A network of thin, light-blue lines connects various points across the blue area, creating a digital or technological aesthetic. The text is centered within this blue area.

Заполните, пожалуйста,
опрос о занятии по ссылке в чате

Спасибо за внимание!
Приходите на следующие вебинары



Андрей Канашов
Senior Data Scientist
BestDoctor