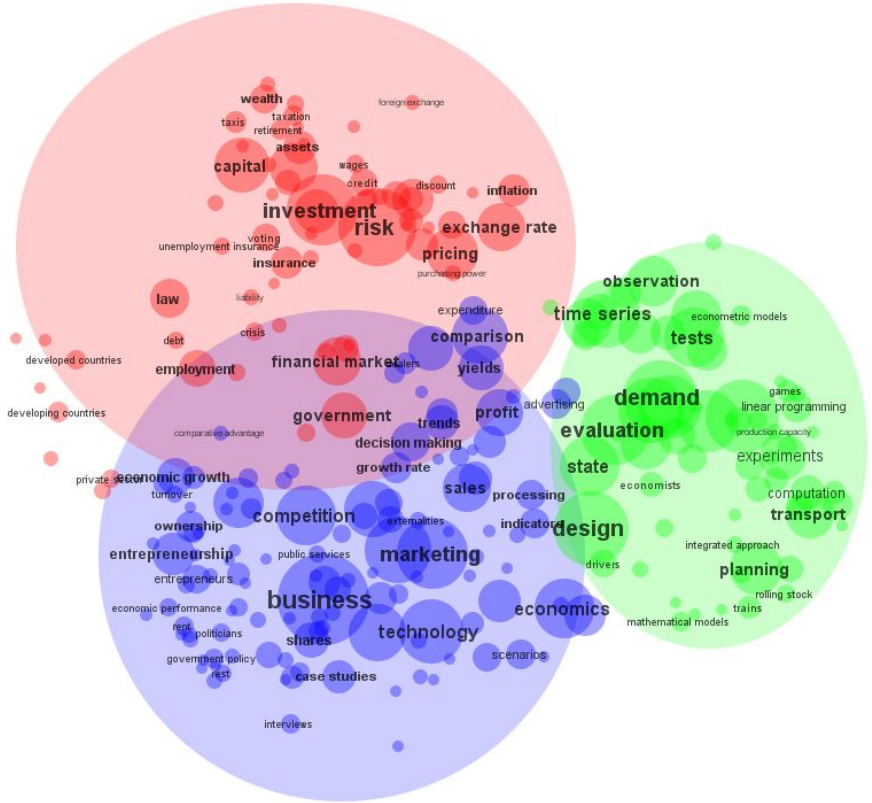
Кластеризация

В основе - презентация Анны Дмитриевой

Задача

- Имея множество **неаннотированных** данных, разделить их на такие группы (кластеры), чтобы элементы внутри каждой группы были похожи друг на друга, а элементы разных групп были разными.
- Кластеризация - задача **обучения без учителя (unsupervised learning)**.



Картинка: N. J. v. Eck, F. Frasincar and D. Chang, "Cluster-Based Visualization of Concept Associations," 2008 12th International Conference Information Visualisation, London, UK, 2008, pp. 409-414, doi: 10.1109/IV.2008.54.

KMeans

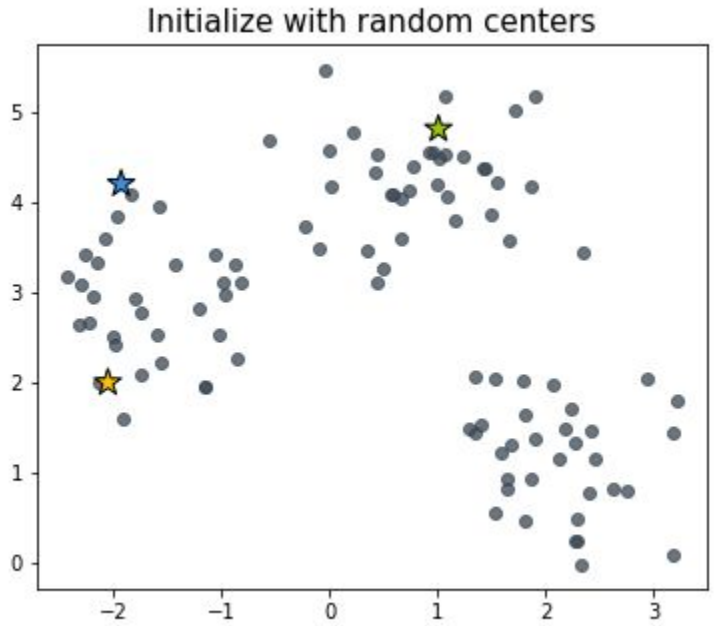
K средних / K means

Алгоритм:

- Выбор K центров кластеров (K задается пользователем);
- Распределение остальных объектов по кластерам по критерию близости к центру;
- Центр масс каждого кластера (среднее арифметическое векторов признаков) становится новым центром кластера;
- Остальные объекты заново распределяются вокруг нового центра.

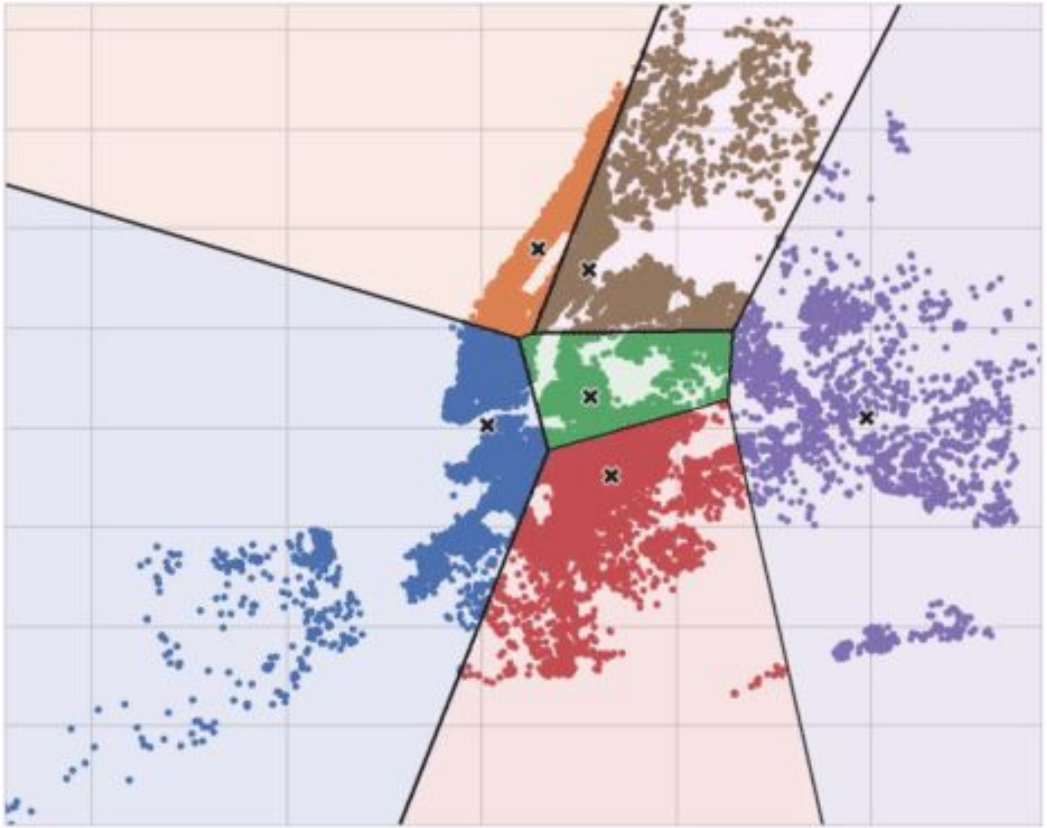
Повторяется, пока кластеры не перестанут меняться.

K средних / K means



Картинка: https://education.yandex.ru/handbook/ml/article/klasterizaciya

K средних /
K means



The K-means clustering algorithm on Airbnb rentals in NYC.

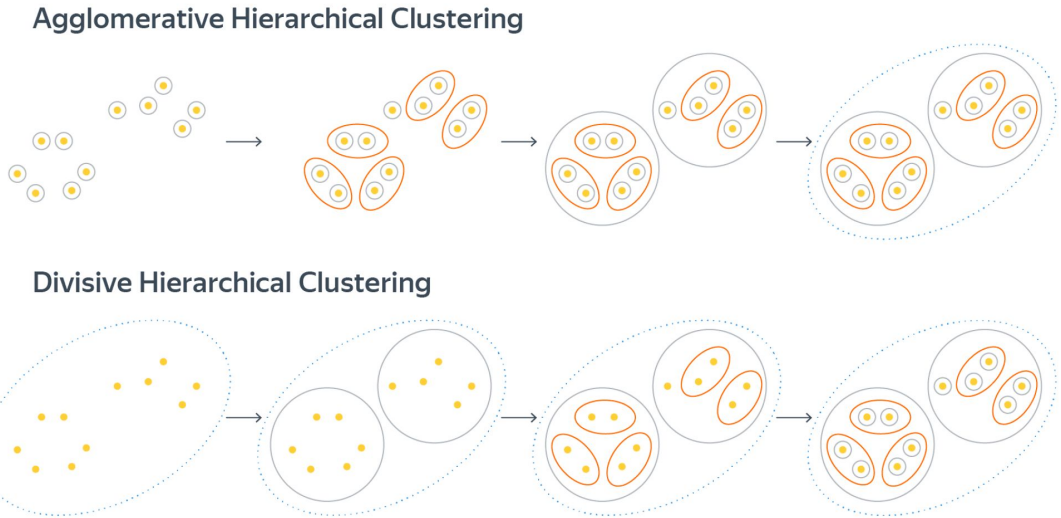
Заметки о выборе центров

- Вначале центры выбираются из объектов выборки, чтобы не попасть в те точки пространства, где нет признаков;
- Начальные положения точек выбираются таким образом, чтобы они были как можно дальше друг от друга.

Иерархическая кластеризация

Иерархическая кластеризация

Может быть агломерационной или дивизионной.



Картинка: https://education.yandex.ru/handbook/ml/article/klasterizaciya

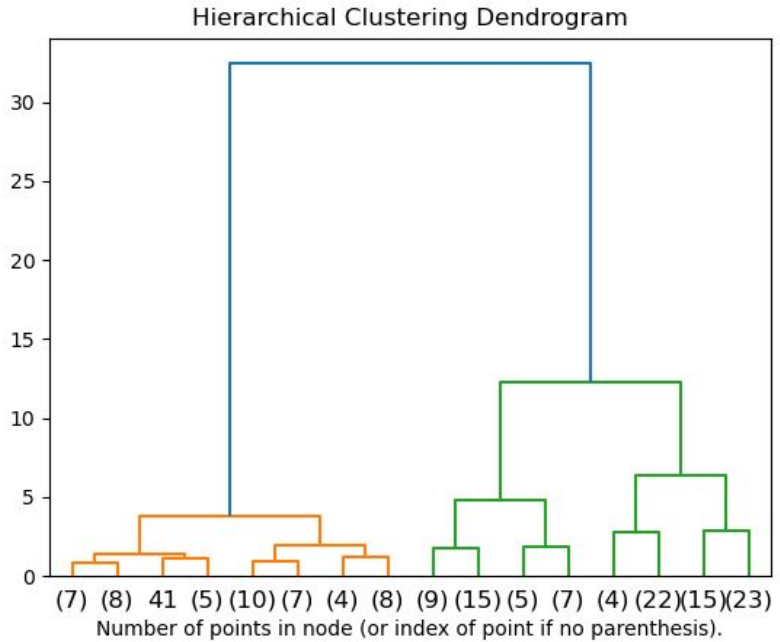
Окончание работы алгоритма

- Все элементы объединены в один кластер;
- Выполняется некое условие расстояния между кластерами: например, расстояние сливаемых кластеров значительно выросло по сравнению с прошлой итерацией.

Как определить, сколько кластеров на самом деле?

По дендрограмме:

- По горизонтали - номера объектов;
- По вертикали - расстояния между кластерами в момент слияния.



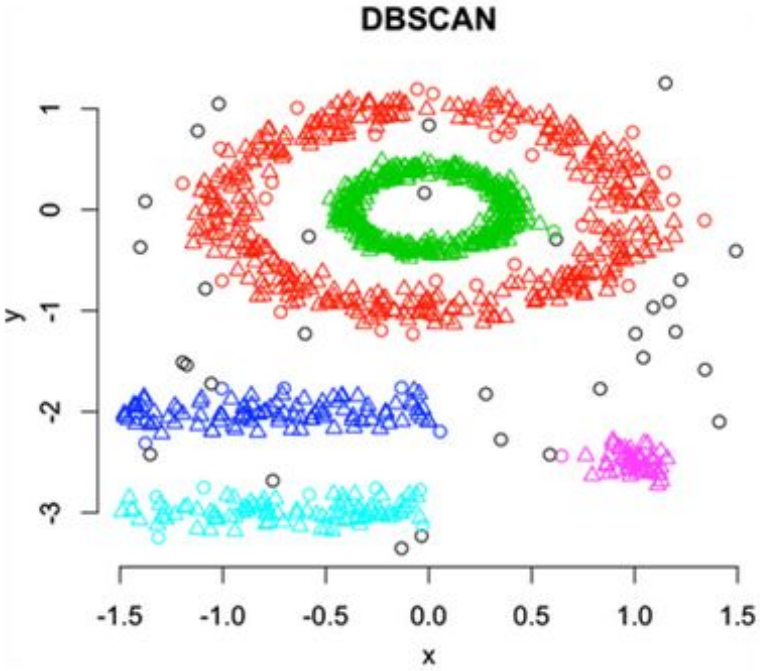
Картинка: https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html

DBSCAN

DBSCAN

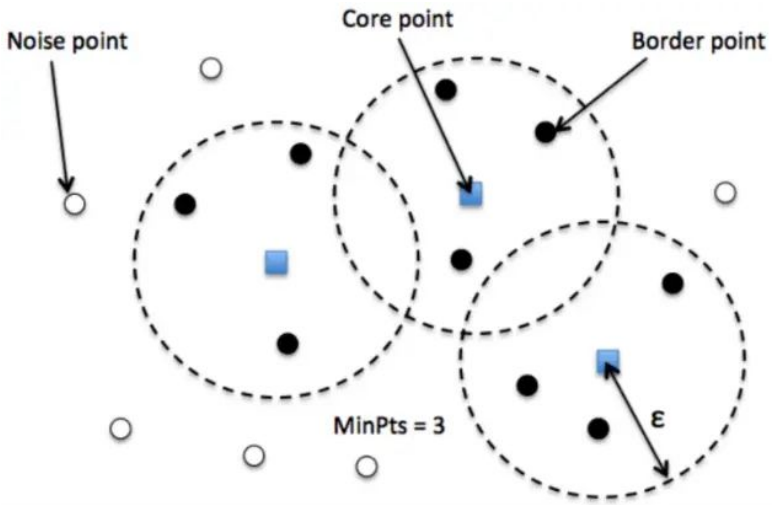
Density-based spatial clustering of applications with noise: кластеризация на основе связных компонент.

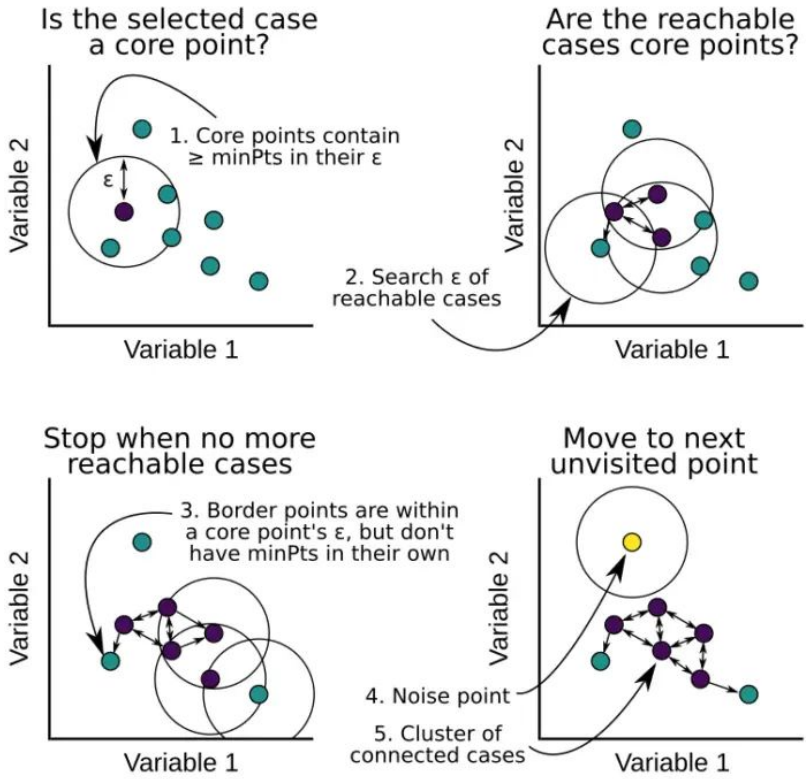
DBSCAN сам определяет количество кластеров и может выявлять также вложенные кластеры.



Картинка: https://education.yandex.ru/handbook/ml/article/klasterizaciya

UPD:





Картинки отсюда
https://machinelearningknowledge.ai/tutorial-for-dbscan-clustering-in-python-sklearn/

DBSCAN: алгоритм

Типы данных:

- **Внутренние/основные** точки: точки, в окрестности радиуса eps которых больше min_samples объектов выборки;
- **Граничные** точки: точки, в окрестности которых есть основные, но общее количество точек в окрестности меньше min_samples;
- **Шумовые** точки: точки, в окрестности которых нет основных точек и в целом содержится менее min_samples объектов выборки.

Гиперпараметры: epsilon/eps, min_samples/min_points.

DBSCAN: алгоритм

1. Шумовые точки убираются из рассмотрения и не приписываются ни к какому кластеру;
2. Основные точки, у которых есть общая окрестность, соединяются ребром;
3. В полученном графе выделяются компоненты связности (максимальные связные подграфы);
4. Каждая граничная точка относится к тому кластеру, в который попала ближайшая к ней основная точка.

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

Оценка качества кластеризации

Оценка качества кластеризации

Без размеченной выборки:

- **Коэффициент силуэта:** мера расстояния между кластерами и их “разделенности”.
- Фиксированная шкала: принимает значения от –1 до +1 и максимизируется, когда кластеры кучные и хорошо отделены друг от друга.
    - Ближе к +1: данная точка находится далеко от ближайших кластеров (к которым она не относится);
    - Ближе к 0: точка находится около границы кластеров;
    - Ближе к -1: точка ближе к объектам ближайших кластеров, чем к объектам своего.

Источники

- https://education.yandex.ru/handbook/ml/article/klasterizaciya
- https://github.com/esokolov/ml-course-hse/blob/master/2018-fall/seminars/sem11_clustering%2Bpca.ipynb
- https://github.com/ischurov/math-ml-hse-2018/blob/master/sem14_clustering/sem14_clustering.ipynb
- https://github.com/nstsj/ML_for_NLP/blob/main/5_clustering/dimred%2Bclustering.ipynb