Кластеризация и понижение размерности

Clustering & dimensionality reduction

Задача обучения без учителя

Эльбрус буткемп

Цель

> группировать объекты схожей природы по группам (кластерам) так, чтобы объекты в каждом кластере обладали схожими свойствами, т.е. находились близко друг к другу в каком-либо пространстве

На входе:

- признаковое описание объектов X
- матрица расстояний между объектами

Метод K-средних / K-means

Эльбрус буткемп

Алгоритм

1. **Инициализация центроидов**: Инициализируем центроиды кластеров случайным образом или выбираем из наблюдений.
2. **Назначение кластеров**: Для каждой точки данных определяем ближайший кластер по расстоянию до центроидов.
3. **Перемещение центроидов**: Пересчитываем центроиды кластеров, используя средние значения точек данных в каждом кластере.
4. **Повторение шагов 2 и 3**: Повторяем назначение кластеров и перемещение центроидов до сходимости алгоритма.

- `Смотрим демо:` визуализация Kmeans

📊 Демо 1 📊 Демо 2

Иерархические методы

Эльбрус буткемп

Делят выборку не на фиксированное число кластеров, а строят вложенные (иерархические) разбиения.

Алгоритм при *агломеративном* подходе:

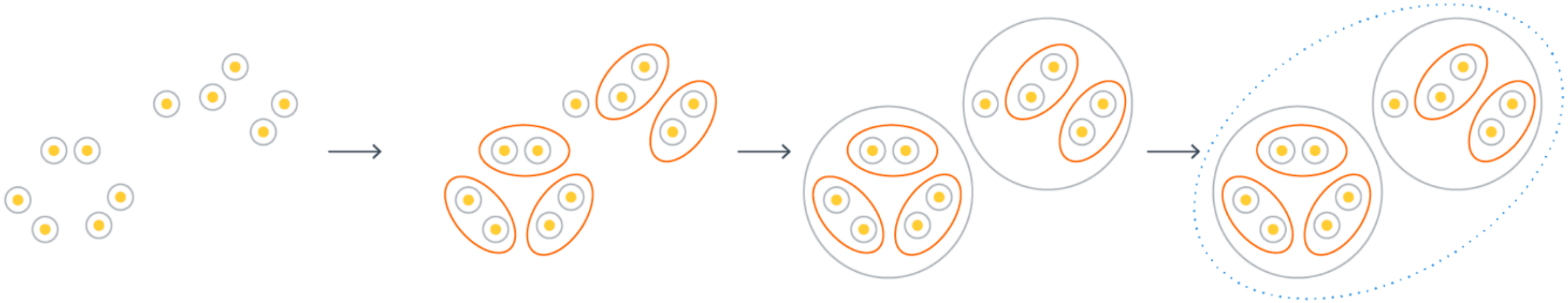
1. Каждый объект считаем отдельным кластером
2. Для каждого объекта находим ближайший, объединяем в новые кластеры
3. Повторяем до тех пор, пока не останется один кластер, содержащий все объекты выборки

Дивизимный подход противоположен: сначала все объекты находятся в одном кластере, далее кластер дробится до тех пор, пока число кластеров не станет равно объему выборки.

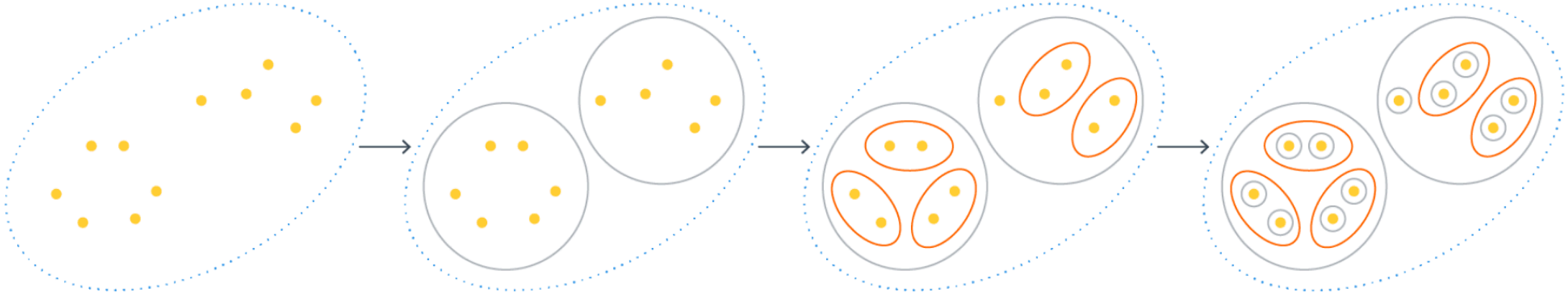
Agglomerative vs. Divisive

Эльбрус буткемп

Agglomerative Hierarchical Clustering



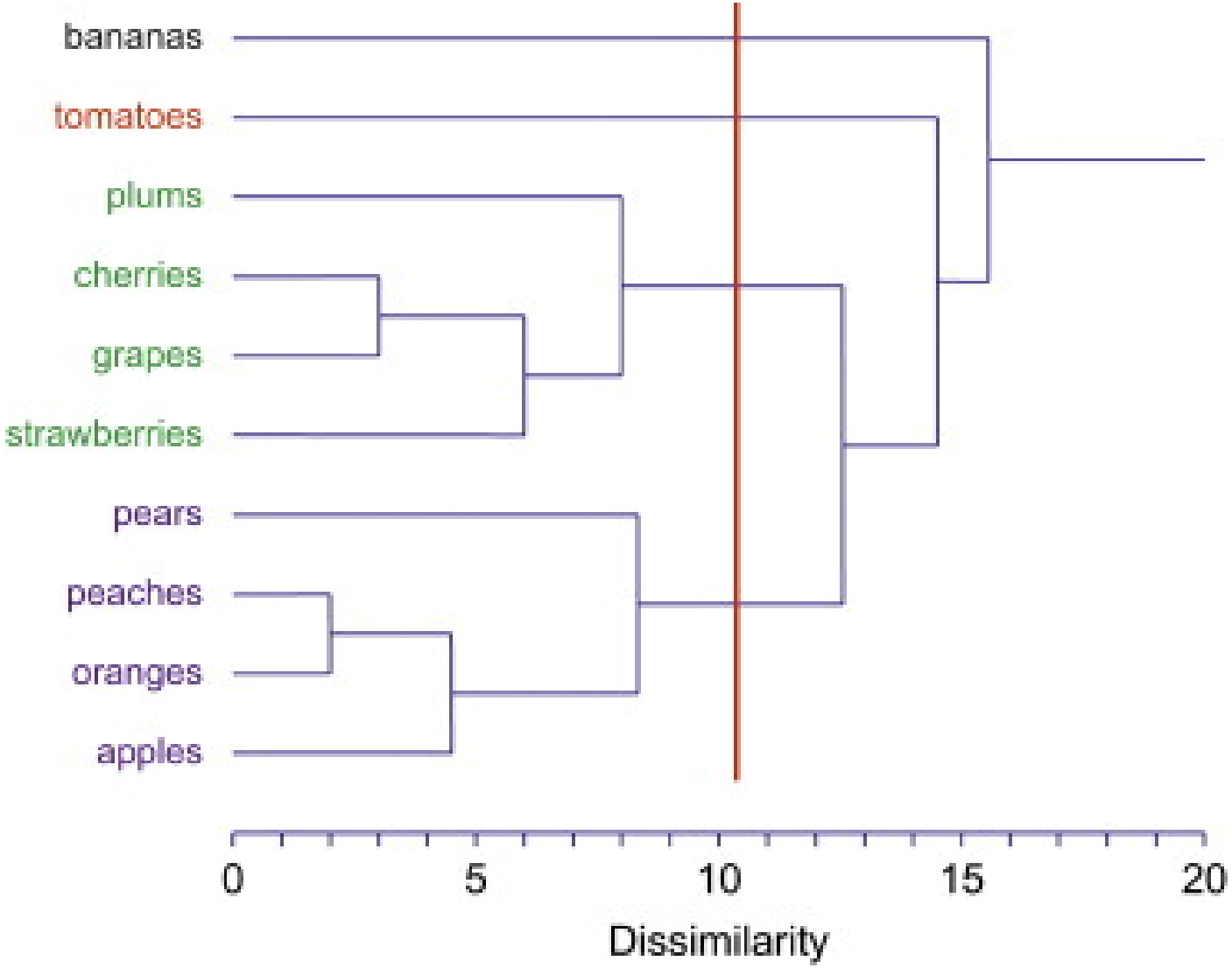
Divisive Hierarchical Clustering



https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html

Дендрограмма

Эльбрус буткемп



Plot Hierarchical Clustering Dendrogram

Плотностной подход / Density based

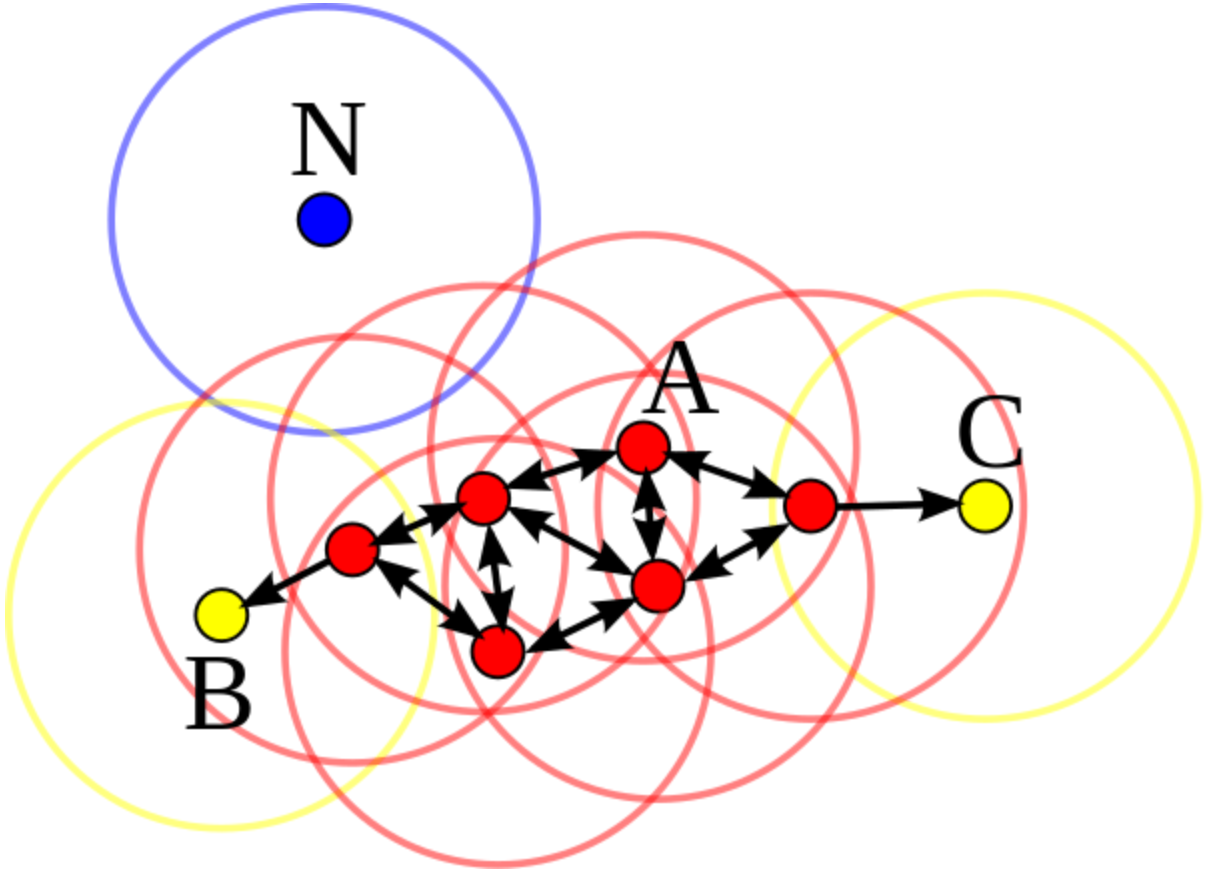
Эльбрус буткемп

> Если объекты расположены близко друг другу и имеют определенное число соседей, то они принадлежат к одному кластеру

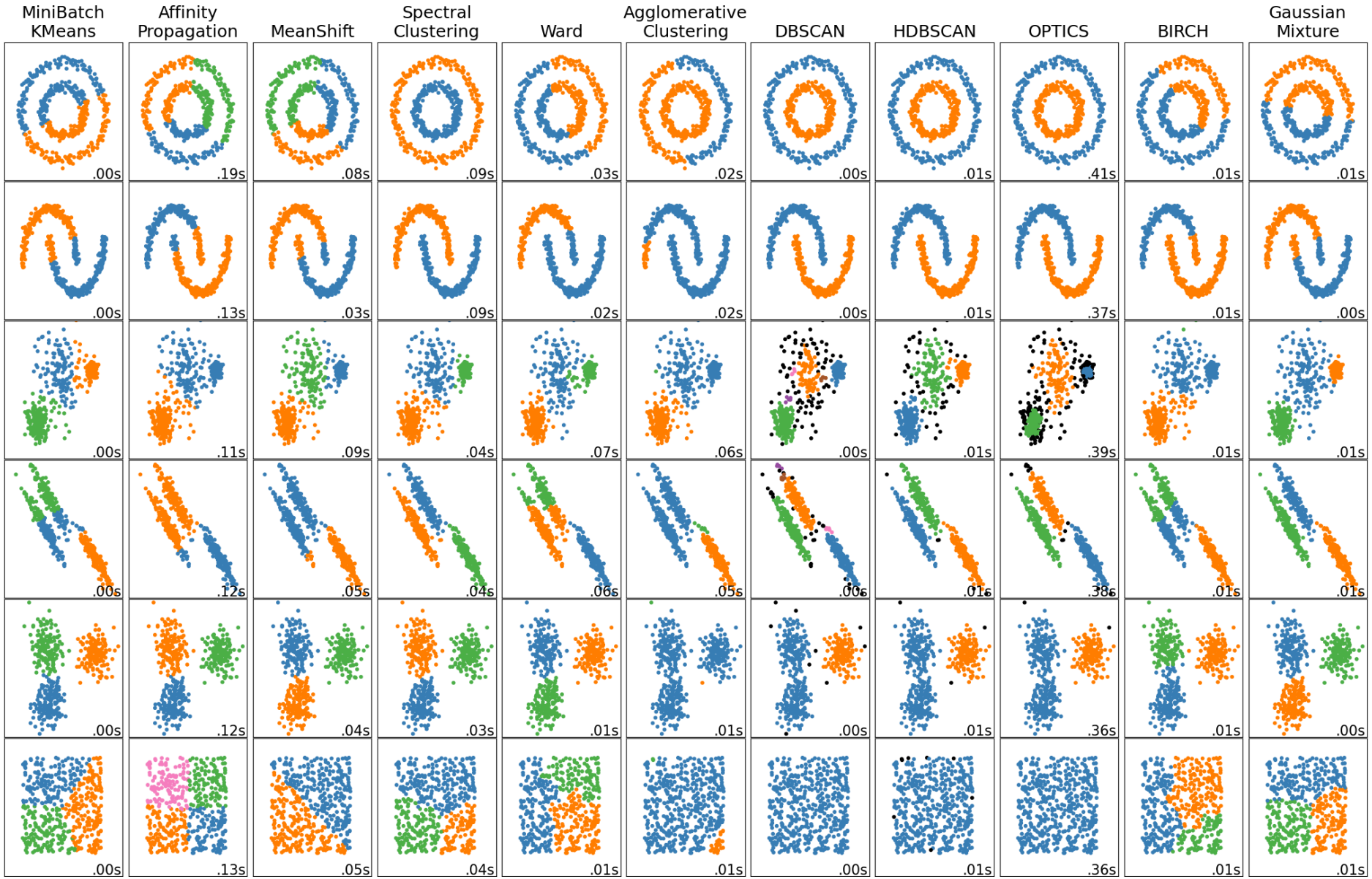
- `sklearn.cluster.DBSCAN`
  - Density-Based Spatial Clustering of Applications with Noise
- sNN
  - Shared Nearest Neighbor

DBSCAN

Эльбрус буткемп



Демо



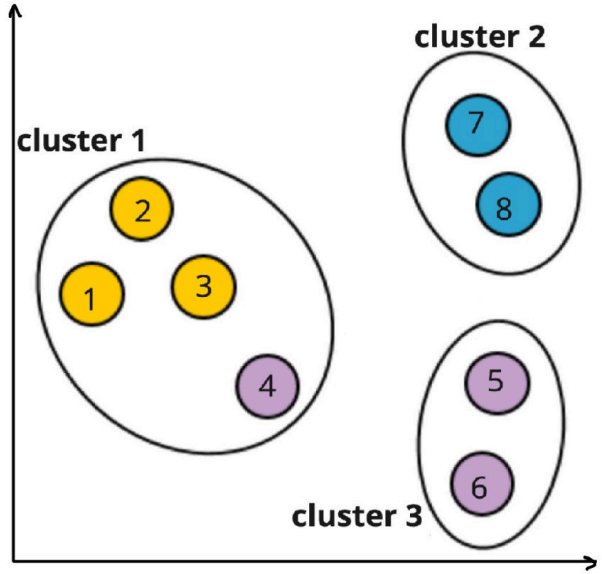
https://scikit-learn.org/stable/modules/clustering.html

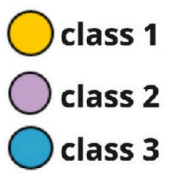
Метрики кластеризации

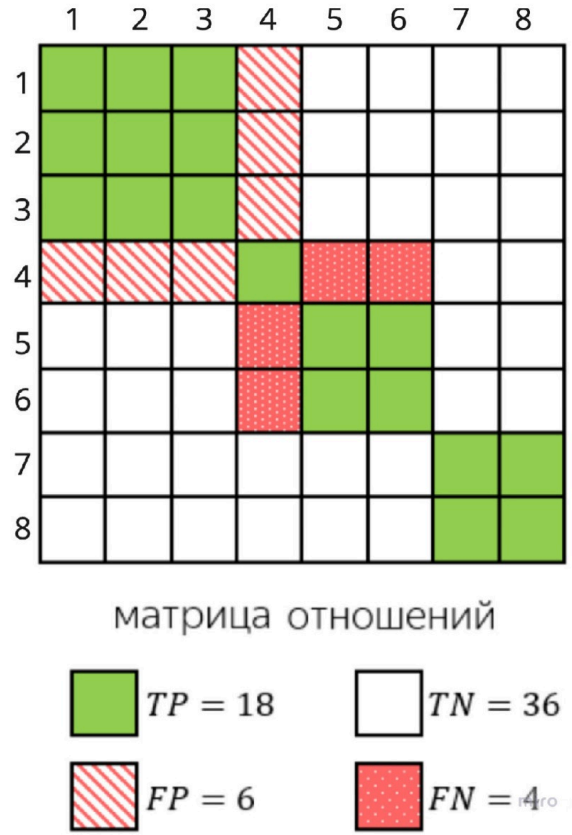
Эльбрус буткемп

- Внешние / External – меры качества, использующие известное распределение по классам
- Внутренние / Internal – меры качества, оценивающие только признаковую информацию об объектах

Внешние метрики кластеризации







- **TP**: кластер и класс совпали
- **FP**: один кластер, но разные классы
- **FN**: разные кластеры, но один класс
- **TN**: разные кластеры и разные классы

Ext Rand Index

Эльбрус буткемп

$$Rand = \frac{TP + TN}{TP + TN + FP + FN}$$

$$0 \leq Rand \leq 1$$

Как называется эта метрика в задачах классификации?

Ext homogeneity(гомогенность)

Эльбрус буткемп

$$homogeneity = \frac{TP}{TP + FP}$$

$$0 \leq homogeneity \leq 1$$

- Что подсвечивает данная метрика, зная, что такое FP
- Как называется эта метрика в задачах классификации?

Ext completeness(полнота)

Эльбрус буткемп

$$completeness = \frac{TP}{TP + FN}$$

$$0 \leq completeness \leq 1$$

- Что подсвечивает данная метрика, зная, что такое FN
- Как называется эта метрика в задачах классификации?

Ext V measure

Эльбрус буткемп

$$v = \frac{(1 + \beta) \times \text{homogeneity} \times \text{completeness}}{(\beta \times \text{homogeneity} + \text{completeness})}$$

- *homogeneity*: каждый кластер содержит элементы только одного класса
- *completeness*: все элементы одного класса были кластеризованы в один кластер

📝V-Measure: A conditional entropy-based external cluster evaluation measure

Int Silhouette Coefficient

Эльбрус буткемп

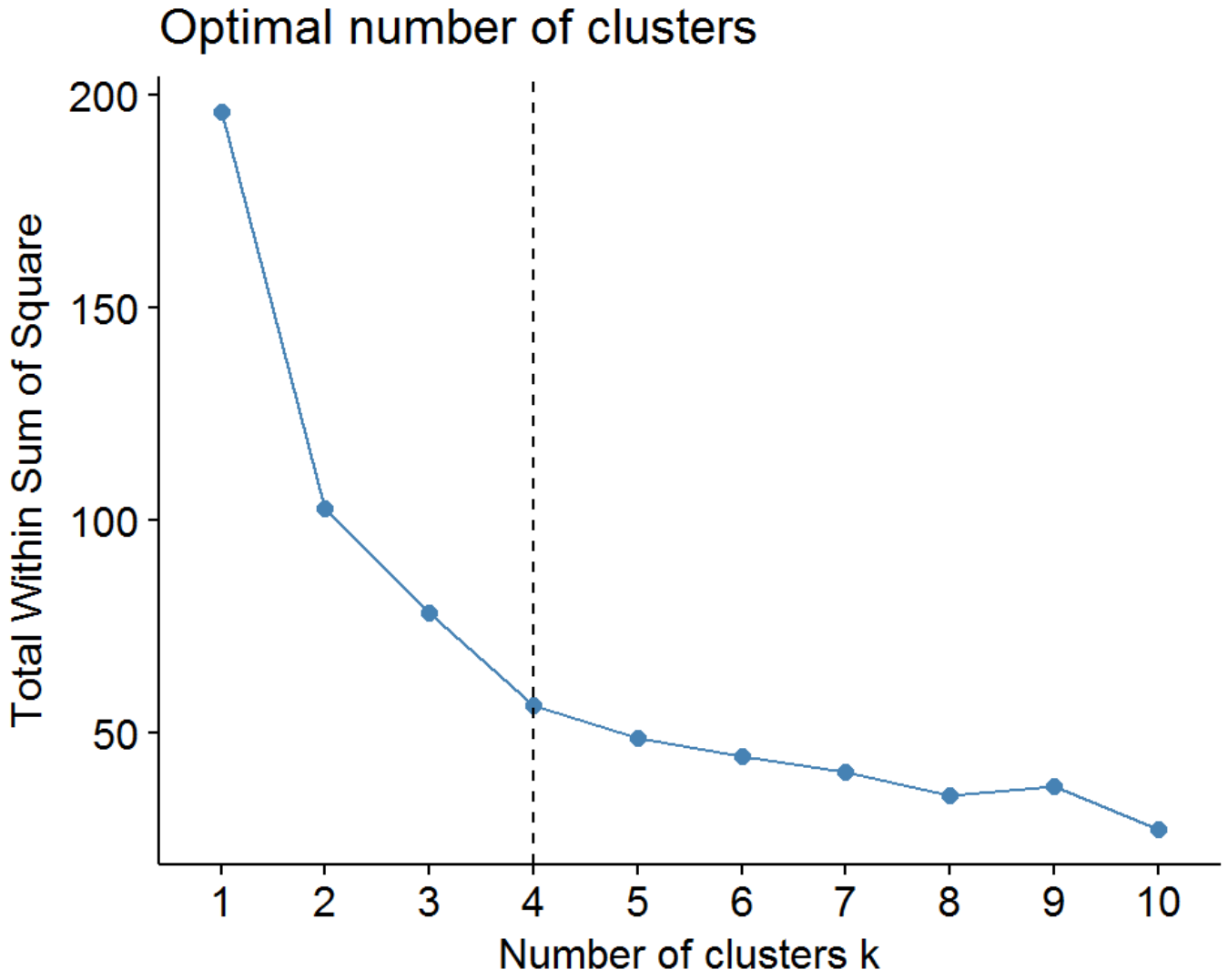
$$s = \frac{b - a}{\max(a, b)}$$

- a – среднее расстояние между фиксированным объектом и остальными объектами в данном кластере
- b – среднее расстояние между фиксированным объектом и объектами другого кластера

sklearn illustration

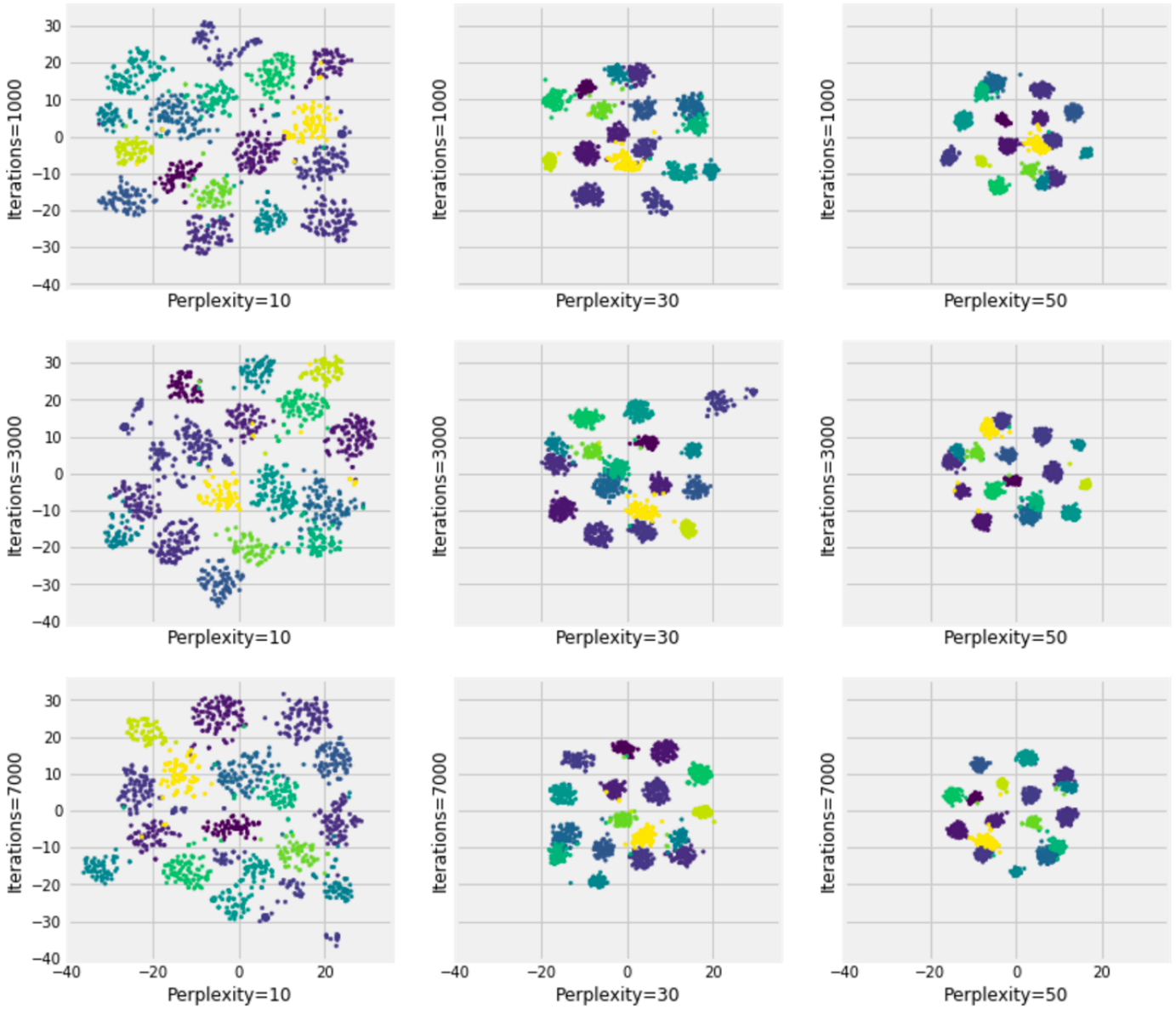
Elbow method

Эльбрус буткемп



📊 Кластеризуем лучше, чем «метод локтя»

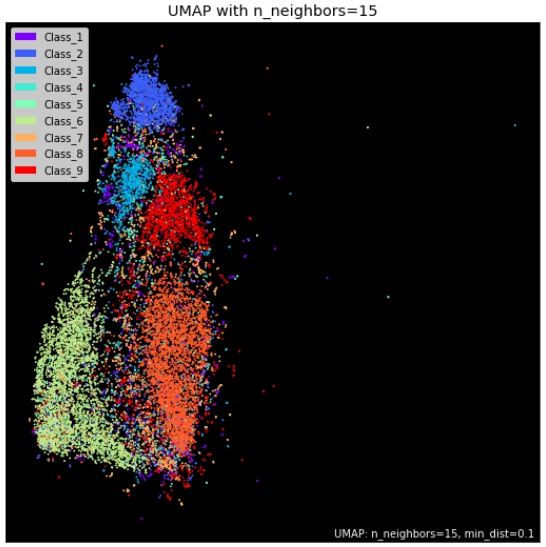
TSNE

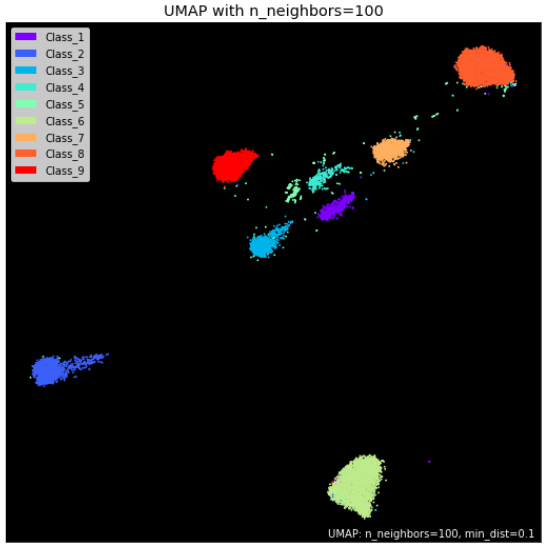


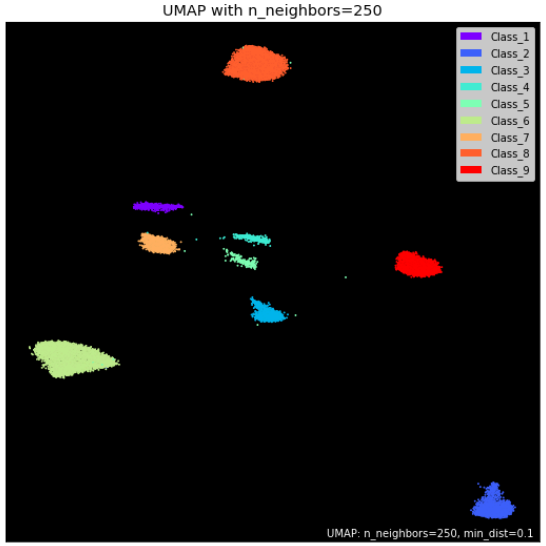
📝 animation tsne 📝Препарируем t-SNE

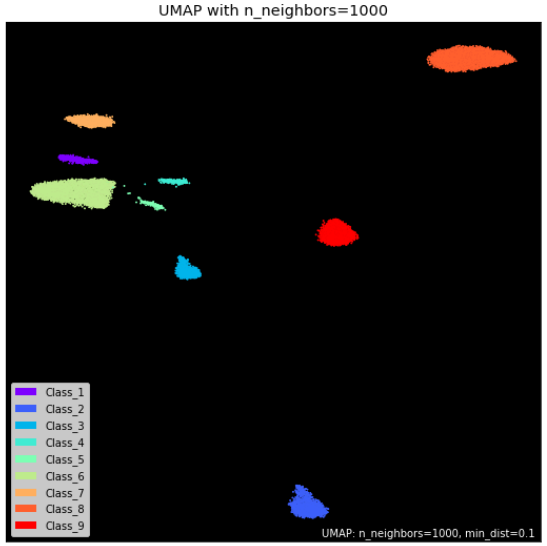
UMAP

Эльбрус буткемп









📝Uniform Manifold Approximation and Projection 📊Потрясающе красиво: как отобразить десятки признаков в данных 📊 Understanding UMAP

Распространенные применения

Эльбрус буткемп

1. Кластеризация табличных данных
2. Кластеризация текстов
3. Кластеризация изображений (см. пример в github)
4. Квантизация изображений (см. пример в github)

Итоги

Эльбрус буткемп

- Кластеризация - способ изучить структуру данных
- В общем виде пайплайн может быть таким
    i. Загрузка данных
    ii. Понижение размерности
    iii. Кластеризация (несколькими разнородными методами)
    iv. Подсчет метрик, интерпретация и визуализация