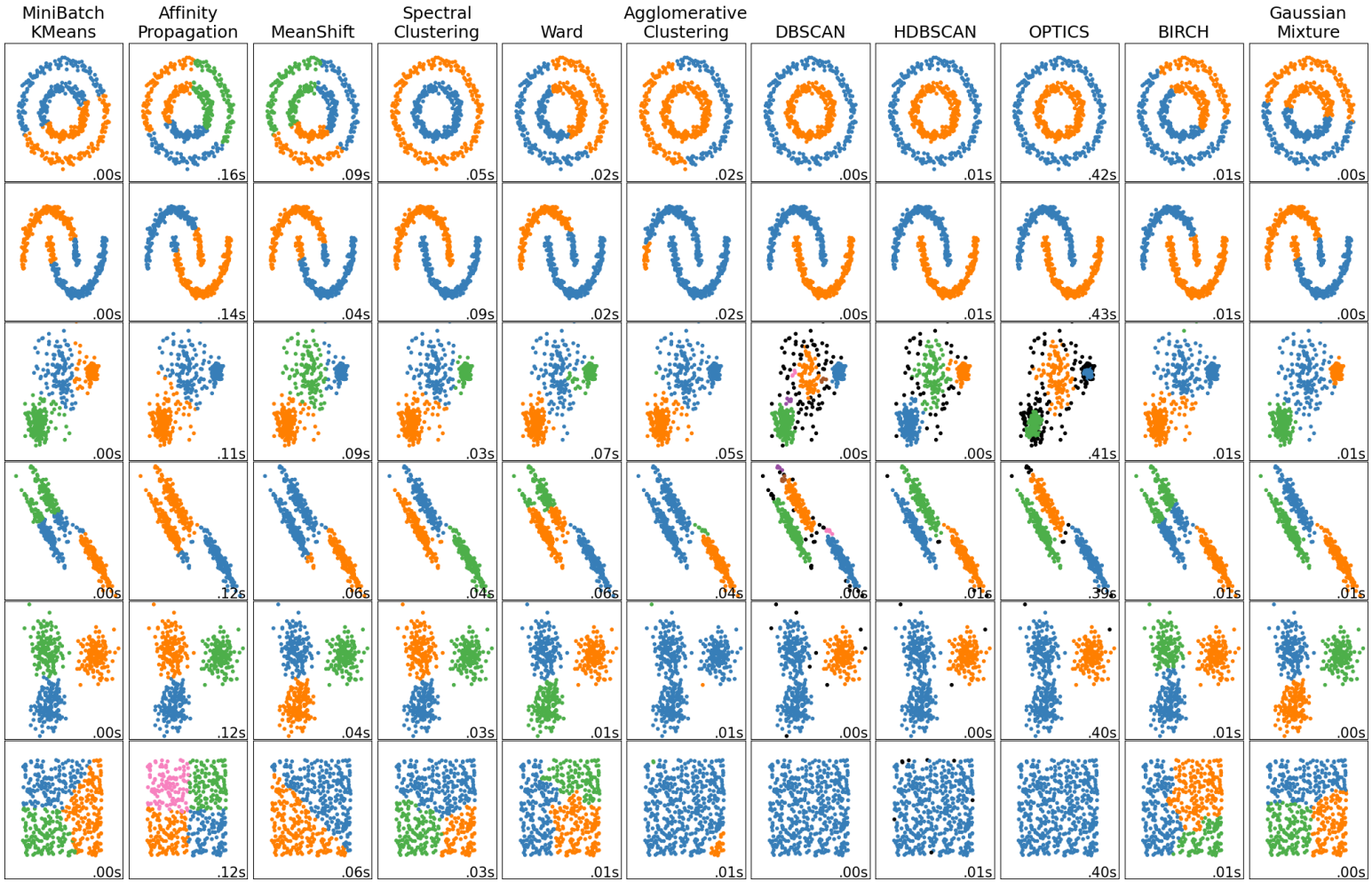
Кластеризация и понижение размерности

Что такое кластеризация

Задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждыи кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.
Теорема невозможности Клейнберга.

Картинка


MiniBatch KMeans
Affinity Propagation
MeanShift
Spectral Clustering
Ward
Agglomerative Clustering
DBSCAN
HDBSCAN
OPTICS
BIRCH
Gaussian Mixture
.00s .16s .09s .05s .02s .02s .00s .01s .42s .01s .00s
.00s .14s .04s .09s .02s .02s .00s .01s .43s .01s .00s
.00s .11s .09s .03s .07s .05s .00s .00s .41s .01s .01s
.00s .12s .06s .04s .06s .04s .00s .01s .39s .01s .01s
.00s .12s .04s .03s .01s .01s .00s .00s .40s .01s .00s
.00s .12s .06s .03s .01s .01s .00s .01s .40s .01s .00s

Виды кластеризации

- Centroid-based
- Connectivity-based
- Distribution-based
- Constraint-based

K-Means

1. Выбираем K точек и объявляем их центрами масс
2. Для каждой точки из выборки ищем ближайший центр масс и относим её к кластеру этого центра масс
3. Вычисляем новые центры масс

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Нет гарантий сходимости.

Метрики качества кластеризации

- Внешние — используют дополнительные знания о кластеризуемом множестве: распределение по кластерам, количество кластеров и т.д.
- Внутренние — оценивают качество структуры кластеров опираясь только непосредственно на нее, не используя внешней информации.

Внутренние метрики

- Среднее внутрикластерное расстояние (компактность)
- Среднее межкластерное расстояние (отделимость)
- Силуэт — насколько объект похож на свой кластер по сравнению с другими кластерами:

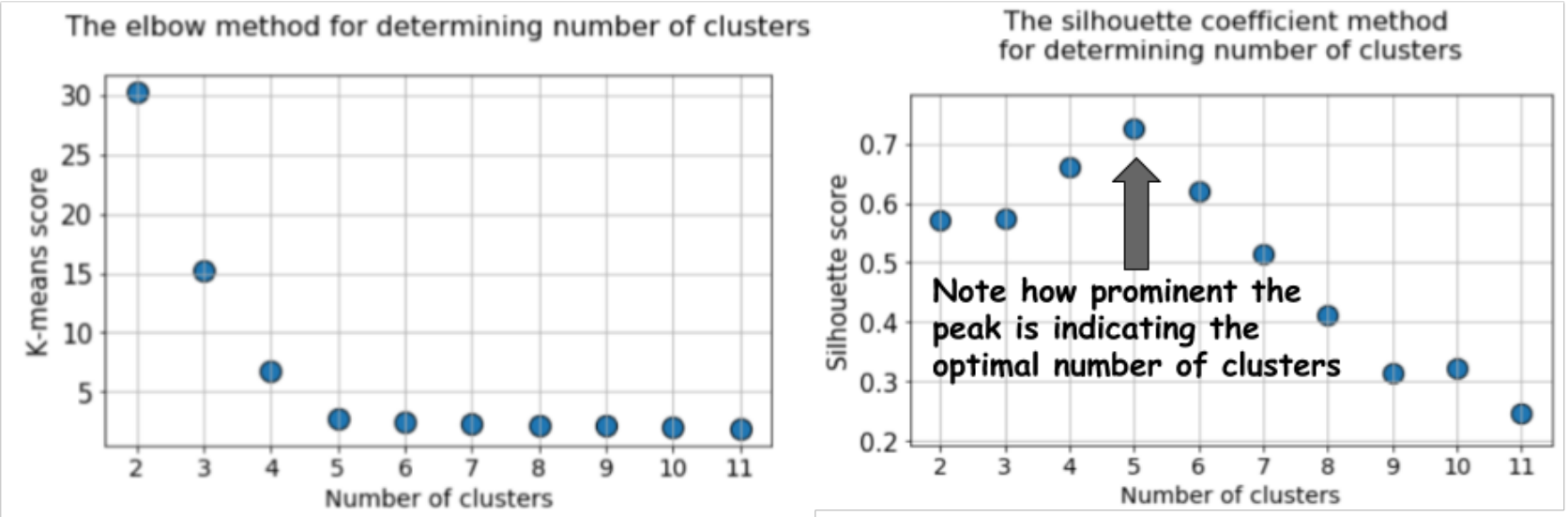
$$S(x_i) = \frac{B(x_i) - A(x_i)}{\max(B(x_i), A(x_i)},$$

где $A(x_i)$ — среднее расстояние между x_i и объектами того же кластера, $B(x_i)$ — среднее расстояние между x_i и объектами ближайшего другого кластера.

- Индекс Дэвиса-Болдуина

Выбор числа кластеров

1. Вычисляем метрику
2. Рисуем график
3. Выбираем k



Односвязный

1. Выбираем случайную точку и кладём её в кластер
2. Ищем соседей с расстоянием ближе, чем ρ и добавляем их в кластер
3. Если соседи кончились, то берём следующую точку и проделываем п. 2
4. Если больше не добавить точек, то переходим к п. 1

DBSCAN

1. Выбираем случайную непосещённую точку
2. Если у неё больше, чем m соседей в радиусе r, то кладём её в K и создаём новый кластер, который включает все точки из окрестности, иначе помечаем как шум.
3. Если точка из окрестности уже является частью другого кластера Cj, то все точки данного кластера добавляются в кластер K.

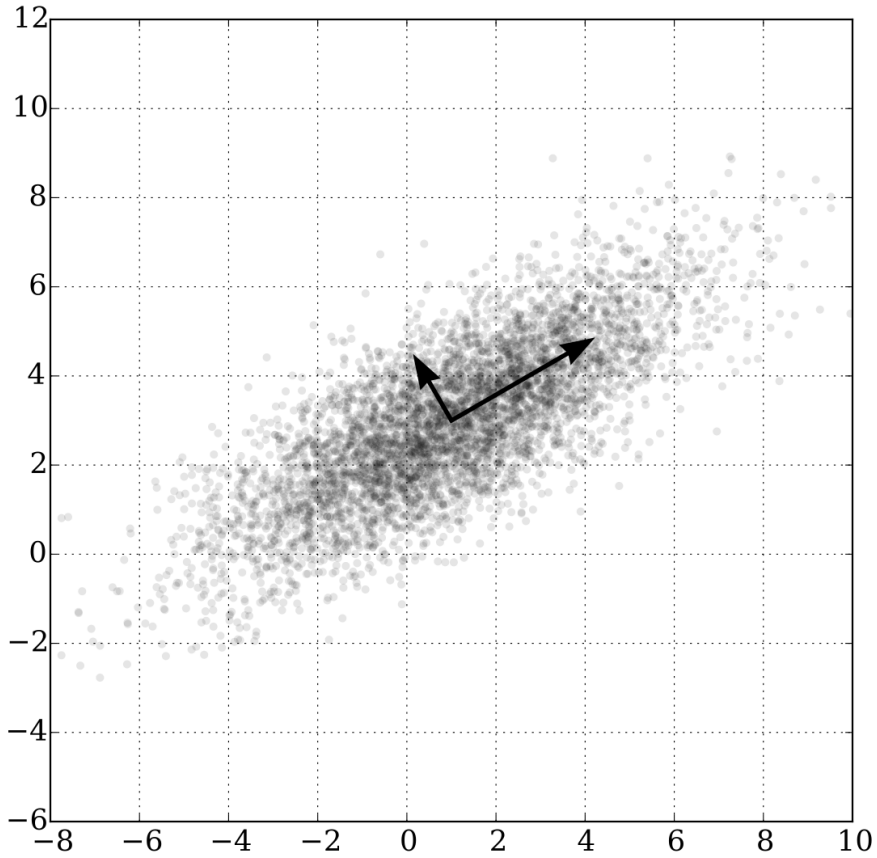
https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Понижение размерности

1. Избавляемся от шума и корреляций
2. Ускоряем вычисления
3. Визуализация

Метод главных компонент

Хотим аппроксимировать данные линейным многообразием меньшей размерности.



Метод главных компонент

Пусть $X \in \mathbb{R}^{n \times m}$, $W \in \mathbb{R}^{m \times k}$ и

$$T = XW$$

- Минимизируем ошибку проецирования
- Максимизируем дисперсию (больше дисперсия — больше информации)

t-SNE и UMAP

Свойства \ Алгоритм	**PCA**	**UMAP**	**t-SNE**
Линейность	Да	Нет	Нет
Детерминированность	Да (условно)	Нет	Нет
Обучаемость	Нет (условно)	Да	Да
Выявляет структуру	Только глобальную	Локальную, отчасти глобальную	Локальную, отчасти глобальную
Скорость работы	Быстрый	Средний	Медленный