Обучение без учителя: кластеризация. Снижение размерности данных PCA.

Екатерина Кондратьева

Обучение без учителя (unsupervised learning):

Или анализ данных без разметки. Можно условно разделить на три больших направления:

1. кластерный анализ (кластеризация), обнаружение аномалий (anomaly detection);
2. методы снижения размерности (dimensionality reduction), оценка внутренней размерности выборки (component analysis), генерация признаков пониженной размерности (feature engineering);
3. *обучение с подкреплением (reinforcement learning) чаще deep learning, поэтому в этом курсе не рассматривается.

1. Кластерный анализ

Кластеризация

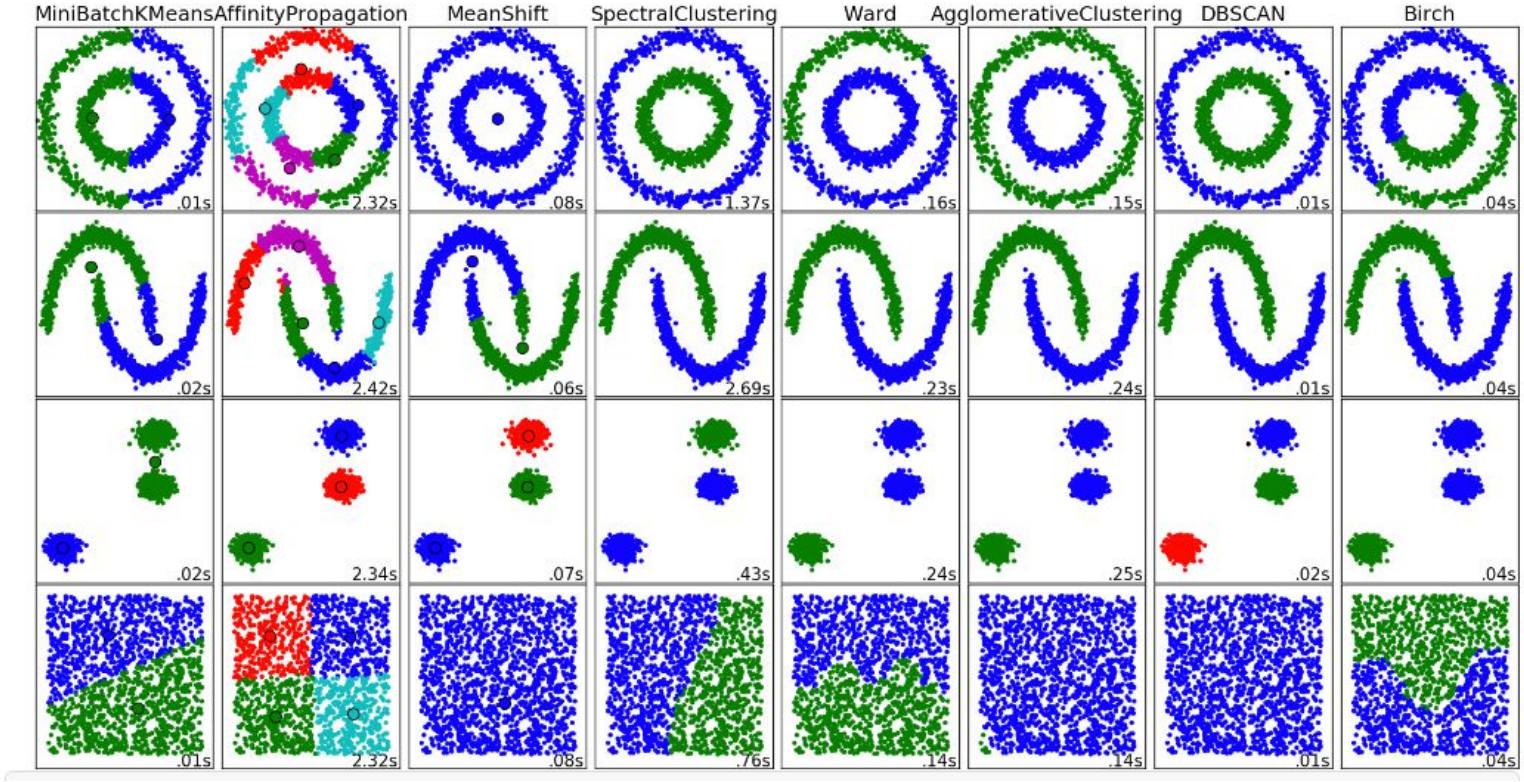
Кластерный анализ (англ. cluster analysis) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Реализации алгоритмов: https://scikit-learn.org/0.18/auto_examples/cluster/plot_cluster_comparison.html, https://scikit-learn.org/stable/modules/clustering.html#k-means

Лекция: https://ru.coursera.org/lecture/unsupervised-learning/vybor-mietoda-klastierizatsii-RZSVo

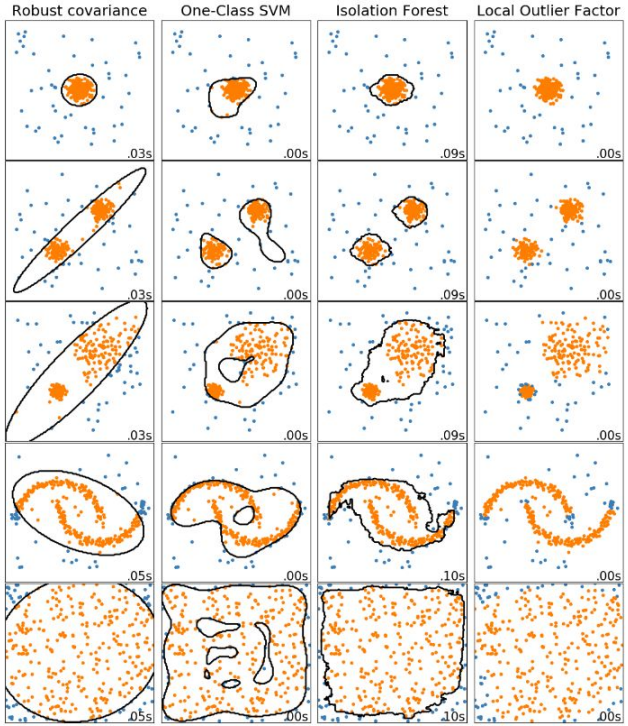
Unsupervised learning: https://ru.coursera.org/learn/unsupervised-learning

Кластерный анализ



https://scikit-learn.org/0.18/auto_examples/cluster/plot_cluster_comparison.html

Обнаружение аномалий (anomaly detection)



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

Аналогия методов классификации (регрессии)

Знакомые нам методы машинного обучения для классификации (регрессии) имеют аналоги (схожие с ними методы) для кластеризации:

- Random Forest Classifier - Isolation Forest -* Agglomerative clustering
- KNN Classifier - KMeans - Local Outlier Factor
- SVC - One-class SVM

Метрики оценивания алгоритмов кластеризации?

Почему не подходят метрики точности классификации?

Метрики оценивания алгоритмов кластеризации

- Полнота (completeness)

all members of a given class are assigned to the same cluster.

- Гомогенность (homogeneity)

each cluster contains only members of a single class

- v_ score

v = 2 * (homogeneity * completeness) / (homogeneity + completeness)

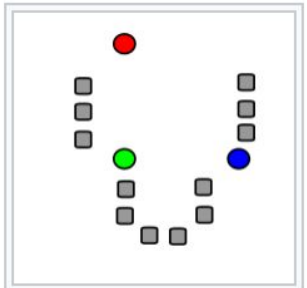
https://scikit-learn.org/0.18/modules/generated/sklearn.metrics.completeness_score.html#sklearn.metrics.completeness_score

Метрики оценивания алгоритмов кластеризации

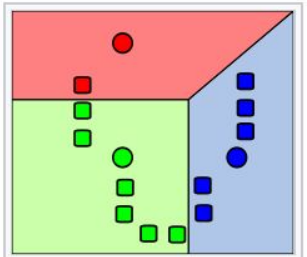
```
>>> from sklearn import metrics
>>> labels true = [0, 0, 0, 1, 1, 1]
>>> labels_pred = [0, 0, 1, 1, 2, 2]

>>> metrics.homogeneity_score(labels_true, labels_pred)
0.66...

>>> metrics.completeness_score(labels_true, labels_pred)
  0.42...
```

https://esigma6.wordpress.com/2018/11/03/2-3-9-3-homogeneity-completeness-and-v-measure/

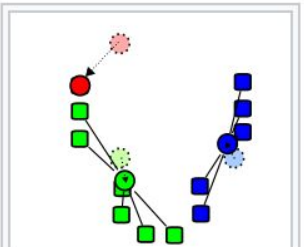
Пример: Метод k- средних



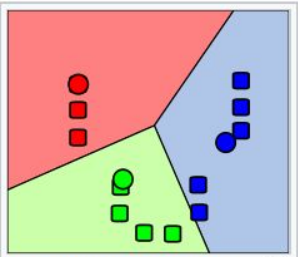
Исходные точки и случайно выбранные начальные точки.



Точки, отнесённые к начальным центрам. Разбиение на плоскости — диаграмма Вороного относительно начальных центров.



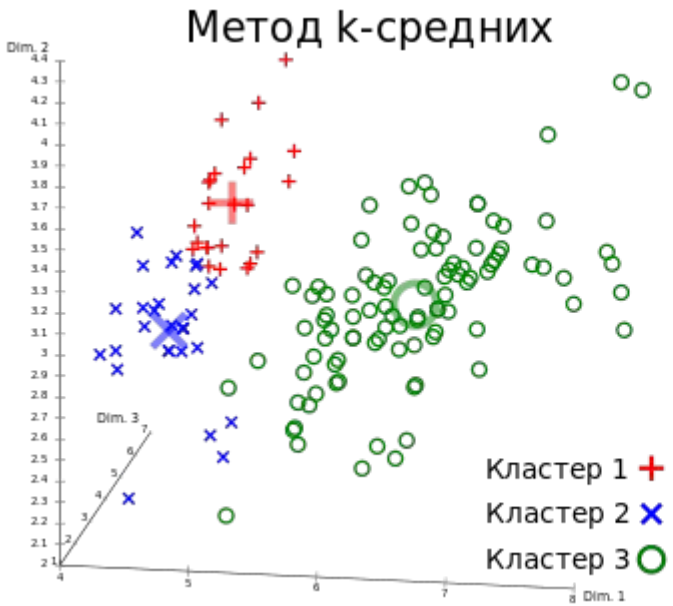
Вычисление новых центров кластеров (Ищется центр масс).

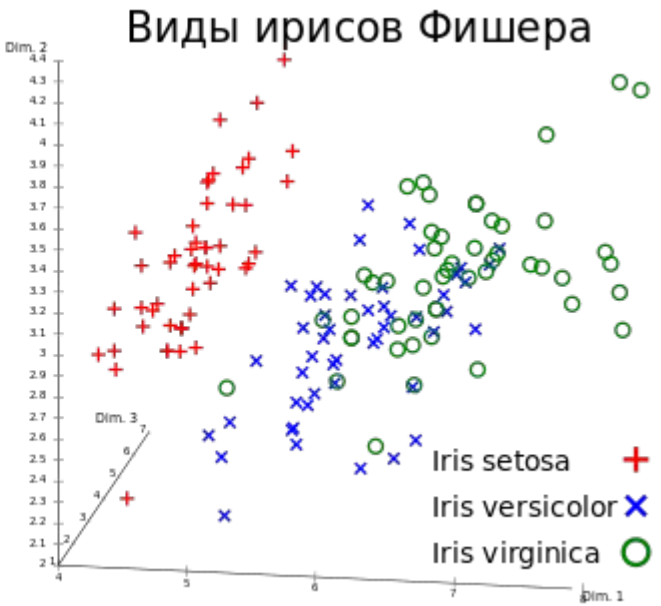


Предыдущие шаги, за исключением первого, повторяются, пока алгоритм не сойдётся.

Минусы метода k-средних

- Не гарантируется достижение глобального минимума суммарного квадратичного отклонения *V*, а только одного из локальных минимумов.
- Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.





2. Методы снижения размерности

Зачем нужно снижать размерность выборки?

Методы снижения размерности

Как уменьшить размерность выборки?

Методы снижения размерности

Как уменьшить размерность выборки?

- удалить неинформативные характеристики объектов (т.е. те, которые вносят наименьший вклад в формирование решающего правила)
- преобразовать имеющиеся характеристики новые, количество которых уменьшит размерность выборки, без потери информации.

Как это сделать?

Методы снижения размерности

Как уменьшить размерность выборки?

- удалить неинформативные характеристики объектов (т.е. те, которые вносят наименьший вклад в формирование решающего правила)
- преобразовать имеющиеся характеристики новые, количество которых уменьшит размерность выборки, без потери информации.

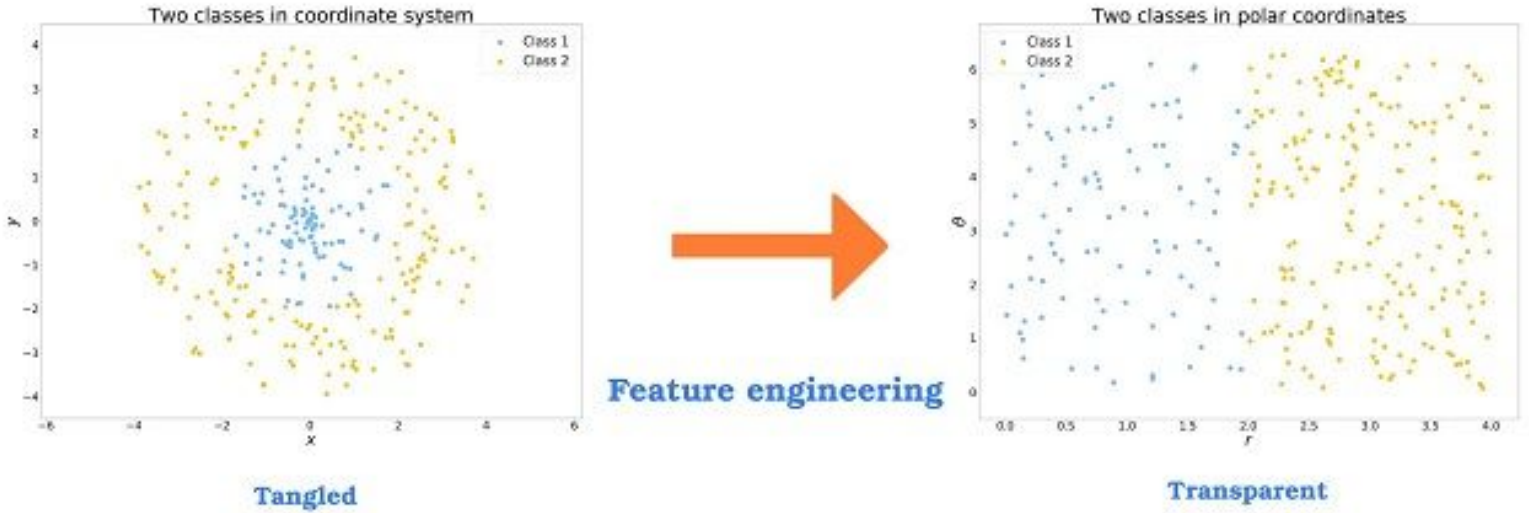
Как это сделать?

- **feature engineering, dimensionality reduction methods** (часто подразумевается manifold learning, или геометрические методы снижения размерности)

Генерация признаков (Feature engineering):

В контексте методов снижения размерности данных и анализа компонент (component analysis), можно говорить о генерации новых признаков, признаков пониженной размерности на многообразии данных.

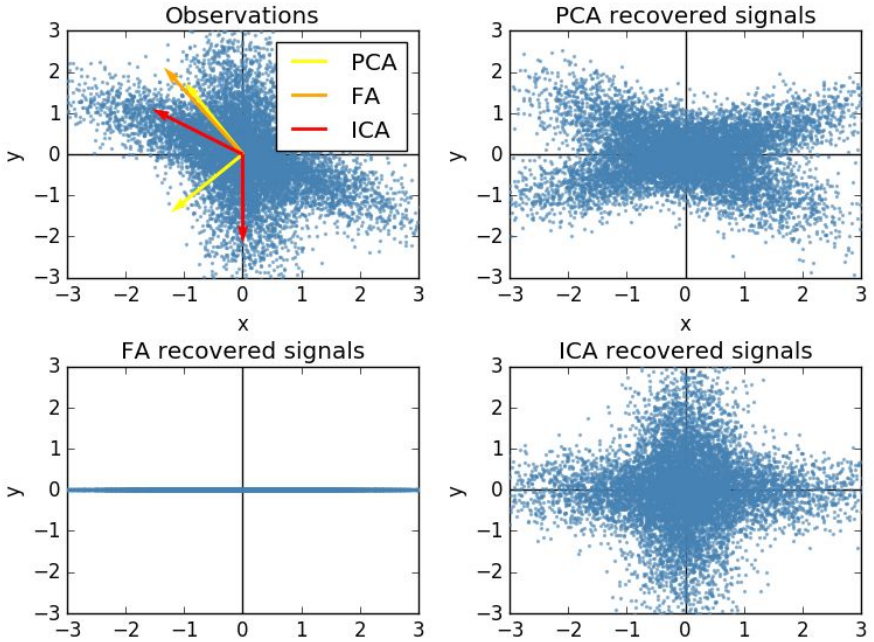
Feature engineering



feature engineering - генерация новых признаков, разделяющих данные

Снижение размерности

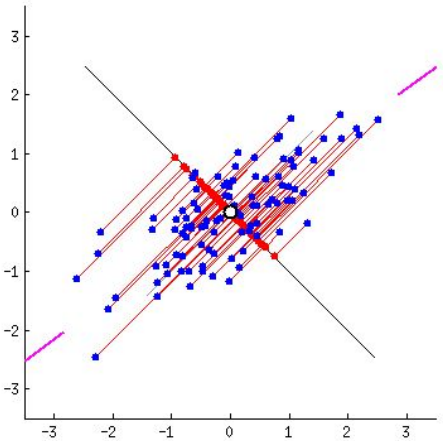
- Линейные (PCA, SVD, ICA и др.)
- Нелинейные (Isomap, tSNE (часто используют как бейзлайн для deep learning) и др.)



https://i.stack.imgur.com/Q0tvQ.png

Снижение размерности данных. PCA

PCA aims to find linearly uncorrelated orthogonal axes, which are also known as principal components (PCs) in the *m* dimensional space to project the data points onto those PCs.



https://towardsdatascience.com/pca-and-svd-explained-with-numpy-5d13b0d2a4d8

Снижение размерности данных. PCA

The PCs can be determined via eigen decomposition of the covariance matrix ***C***. After all, the geometrical meaning of eigen decomposition is to find a new coordinate system of the eigenvectors for ***C*** through rotations.

$$\mathbf{C} = \frac{\mathbf{X}^\top \mathbf{X}}{n-1}$$

Covariance matrix of a 0-centered matrix **X**

$$\mathbf{C} = \mathbf{W}\mathbf{\Lambda}\mathbf{W}^{-1}$$

Eigendecomposition of the covariance matrix **C**

$$\mathbf{X}_k = \mathbf{X}\mathbf{W}_k$$

Project data onto the first k PCs

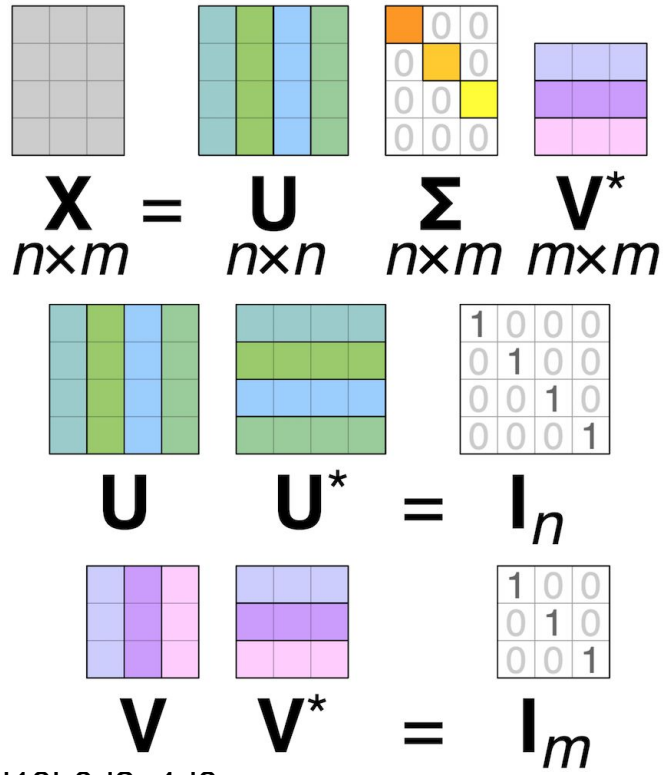
https://towardsdatascience.com/pca-and-svd-explained-with-numpy-5d13b0d2a4d8

Снижение размерности данных. SVD

SVD is another decomposition method for both real and complex matrices. It decomposes a matrix into the product of two unitary matrices ($\boldsymbol{U}$, $\boldsymbol{V}^*$) and a rectangular diagonal matrix of singular values ($\boldsymbol{\Sigma}$):

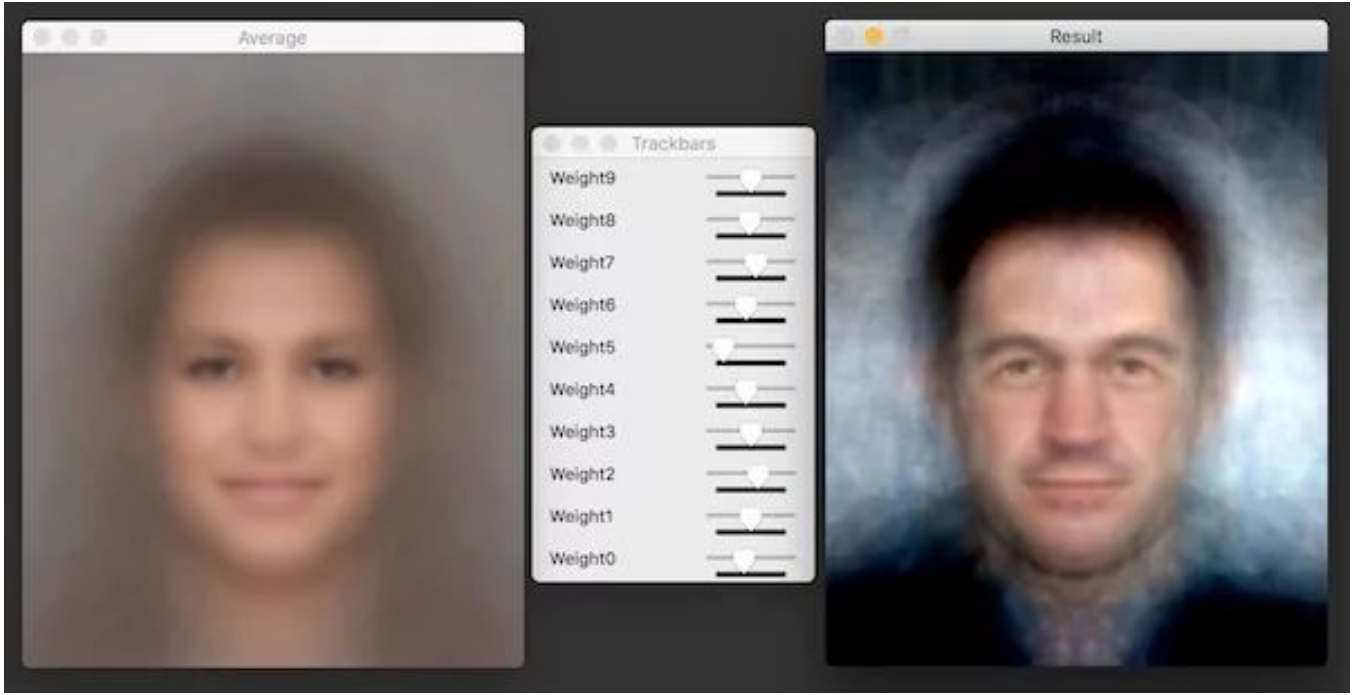
$$\Lambda = \frac{\Sigma^2}{n-1}$$

Relationship between eigenvalue and singular values



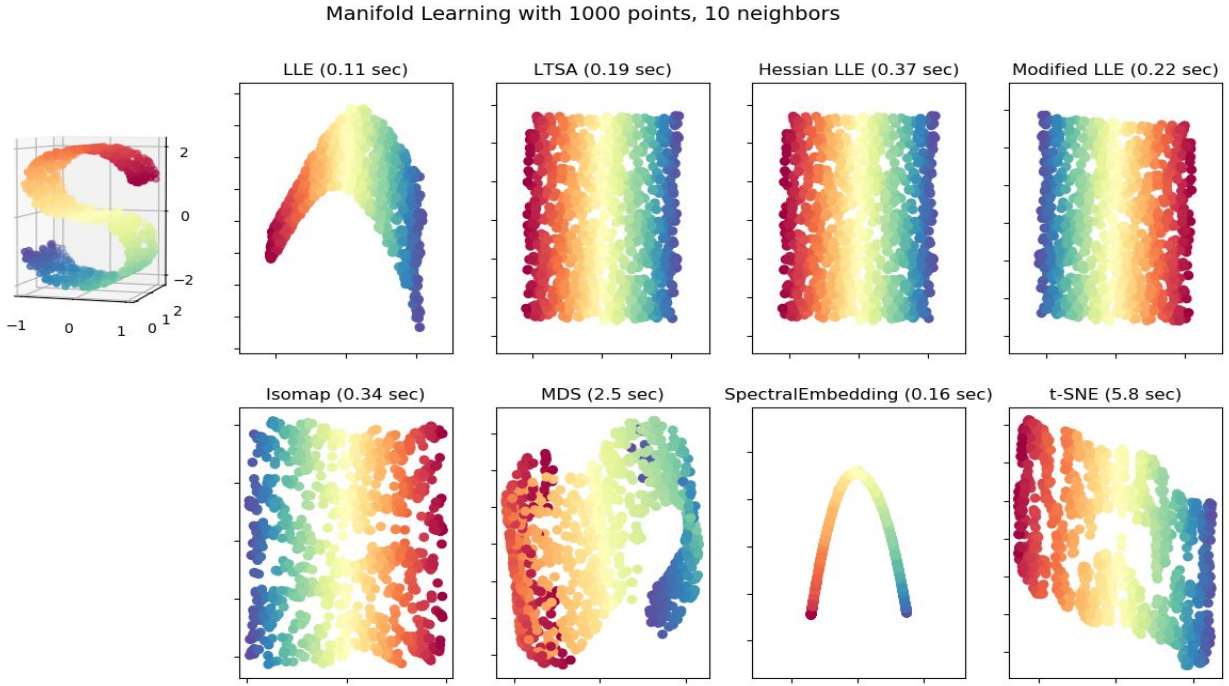
https://towardsdatascience.com/pca-and-svd-explained-with-numpy-5d13b0d2a4d8

Eigenfaces



https://www.learnopencv.com/eigenface-using-opencv-c-python/

Нелинейные методы снижения размерности



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

Источники:

Реализации алгоритмов: https://scikit-learn.org/0.18/auto_examples/cluster/plot_cluster_comparison.html, https://scikit-learn.org/stable/modules/clustering.html#k-means

Кластерный анализ сравнение: https://proglib.io/p/unsupervised-ml-with-python/

Лекция: https://ru.coursera.org/lecture/unsupervised-learning/vybor-mietoda-klastierizatsii-RZSVo

Методы снижения размерности:

Линейные https://ru.coursera.org/lecture/unsupervised-learning/mietod-ghlavnykh-komponient-rieshieniie-e72bH
https://ru.coursera.org/lecture/python-for-data-science/mietod-glavnykh-komponient-principal-component-analysis-X8bem

Нелинейные
https://ru.coursera.org/lecture/vvedenie-mashinnoe-obuchenie/nielinieinyie-mietody-ponizhieniia-razmiernosti-QloeT

*Unsupervised learning: https://ru.coursera.org/learn/unsupervised-learning