Лабораторная работа № 7 (последняя)

Тема: Метод главных компонент (PCA).

Цель работы – научиться применять алгоритм PCA и анализировать результаты его работы.

По ссылке вы найдете несколько примеров, доказывающих важность нормализации (масштабирования) данных перед использованием метода PCA:

https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html#sphx-glr-auto-examples-preprocessing-plot-scaling-importance-py

Ваша задача разобраться в этих примерах и реализовать код, который:

1. Загружает данные, вместо рассмотренного в примере датасета используйте любой другой датасет для классификации с большим количеством параметров, например *breast_cancer , он тоже встроен в scikit-learn и его можно загрузить:*

```
from sklearn.datasets import load_breast_cancer
X, y =load_breast_cancer(return_X_y=True, as_frame=True)
```

2. Строит гистограмму весов первых компонент без нормализации и после нормализации параметров.
3. Визуализирует при помощи метода PCA распределения классов без нормализации и с нормализацией.
4. Далее обучите на выбранном в предыдущем задании датасете модель опорных векторов (*как вы делали это раньше, без применения PCA*). Выведите точность полученной модели и время обучения.
5. Теперь обучите модель опорных векторов с использованием метода PCA, но для этого определите количество главных компонент, которые обеспечат сохранение 90% дисперсии данных. Определить количество главных компонент можно из графика процента дисперсии от количества компонент, пример графика в лекции на стр. 10. Выведите точность полученной модели и время обучения.

Вопросы:

1. Какие преимущества дает уменьшения размерности данных?
2. Расскажите про методы понижения размерности данных Feature Selection и Feature Extraction.
3. Расскажите принцип работы метода PCA.
4. Что означает понятие главная компонента?