РАБОТА С МЕТОДОМ ГЛАВНЫХ КОМПОНЕНТ

ВЫЧИСЛЕНИЕ ИНДЕКСА ДОУ-ДЖОНСА

Задание основано на применении пакета Python Scikit-learn для работы с методом главных компонент в составе модуля decomposition в классе PCA.

Цели практики:

- овладеть основами работы с методом главных компонент.
- научиться использовать его для вычисления улучшенного индекса Доу-Джонса.

Начало работы

Метод главных компонент реализован в пакете **scikit-learn** в модуле **decomposition** в классе PCA.

Основным параметром является количество компонент **(n_components**).

Для обученного преобразования этот класс позволяет вычислять различные характеристики. Например, поле **explained_variance_ratio_** содержит процент дисперсии, который объясняет каждая компонента.Поле **components_** содержит информацию о том, какой вклад вносят признаки в компоненты.

Чтобы применить обученное преобразование к данным, можно воспользоваться методом **transform**. Для нахождения коэффициента корреляции Пирсона необходимо воспользоваться функцией **corrcoef** из пакета **numpy**.

Задание

В данной задаче мы рассматриваем информацию о ценах на акции 30 крупнейших компаний в США. Эти данные позволяют оценить текущее состояние экономики, основываясь на показателях индекса Доу-Джонса.

Описание:

В течение периода с 23 сентября 2013 года по 18 марта 2015 года состав компаний, включенных в индекс, оставался постоянным (дополнительную информацию о составе можно найти по указанной ссылке в материалах).

Одним из значительных недостатков индекса Доу-Джонса является способ его расчета. При вычислении индекса цены акций включенных компаний суммируются и затем делятся на поправочный коэффициент. В результате даже если капитализация одной компании заметно меньше, чем у другой, но стоимость ее акций выше, она оказывает более сильное воздействие на индекс. Даже значительное процентное изменение стоимости относительно

более дешевых акций может быть скомпенсировано незначительным процентным изменением стоимости более дорогих акций.

1. Загрузите данные **close_prices.csv**. В данном файле приведены цены акций 30 компаний на закрытии торгов за каждый день периода.

2. На загруженных данных обучите преобразование PCA с числом компоненты равным 10. Скольких компонент хватит, чтобы объяснить 90% дисперсии?

3. Примените построенное преобразование к исходным данным и возьмите значения первой компоненты.

4. Загрузите информацию об индексе Доу-Джонса из файла **djia_prices.csv**.

Чему равна корреляция Пирсона между первой компонентой и индексом Доу-Джонса?

5. Какая компания имеет наибольший вес в первой компоненте?

Ответ на каждое задание — текстовый файл, содержащий ответ в первой строчке (#.txt). Обратите внимание, что отправляемые файлы не должны содержать перевод строки в конце.

Утонения по выполнению задания:

Если ответом является нецелое число, то целую и дробную часть необходимо разграничивать точкой, например, 0.42.
При необходимости округляйте дробную часть до двух знаков

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

Метод главных компонент (PCA) представляет собой один из алгоритмов обучения без учителя, который создает новые признаки как линейные комбинации исходных. Эти признаки формируются таким образом, чтобы сохранить максимальную дисперсию в данных. Другими словами, метод главных компонент эффективно уменьшает размерность данных, сохраняя при этом их вариабельность.

Основным параметром метода является *количество новых признаков*. Один из подходов заключается в выборе минимального числа компонент, которое сохраняет определенную долю исходной дисперсии данных. Таким образом, сохраняется определенная часть общей изменчивости.

В данной задаче необходимо измерять схожесть двух наборов данных.

Для оценки взаимосвязи между парами измерений, такими как предсказания двух классификаторов для одного и того же объекта, используется *корреляция Пирсона*. Ее значения находятся в диапазоне *от -1 до 1* и отражают *степень линейной зависимости между величинами*.

Когда корреляция равна *-1 или 1*, это свидетельствует о линейной зависимости, в то время как значение 0 указывает на отсутствие линейной зависимости между величинами.

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html