# Составление фондового индекса

## Вы научитесь:

- работать с методом главных компонент
- использовать его для вычисления улучшенного индекса Доу-Джонса

## Введение

Метод главных компонент (principal component analysis, PCA) — это один из методов обучения без учителя, который позволяет сформировать новые признаки, являющиеся линейными комбинациями старых. При этом новые признаки строятся так, чтобы сохранить как можно больше дисперсии в данных.Иными словами,метод главных компонент понижает размерность данных оптимальным с точки зрения сохранения дисперсии способом.

Основным параметром метода главных компонент является количество новых признаков. Как и в большинстве методов машинного обучения, нет четких рекомендаций по поводу выбора значения этого параметров. Один из подходов — выбирать минимальное число компонент, при котором объясняется не менее определенной доли дисперсии (это означает, что в выборке сохраняется данная доля от исходной дисперсии).

В этом задании понадобится измерять схожесть двух наборов величин. Если имеется набор пар измерений (например, одна пара — предсказания двух классификаторов для одного и того же объекта), то охарактеризовать их зависимость друг от друга можно с помощью корреляции Пирсона. Она принимает значения от -1 до 1 и показывает, насколько данные величины линейно зависимы. Если корреляция равна -1 или 1, то величины линейно выражаются друг через друга. Если она равна нулю, то линейная зависимость между величинами отсутствует.

## Данные

В этом задании мы будем работать с данными о стоимостях акций 30 крупнейших компаний США. На основе этих данных можно оценить состояние экономики, например, с помощью индекса Доу-Джонса. Со временем состав компаний, по которым строится индекс меняется. Для набора данных был взят период с 23.09.2013 по 18.03.2015, в котором набор компаний был фиксирован (подробнее почитать о составе можно по ссылке из материалов).

Одним из существенных недостатков индекса Доу-Джонса является способ его вычисления — при подсчёте индекса цены входящих в него акций складываются, а потом делятся на поправочный коэффициент. В результате, даже если одна компания заметно меньше по капитализации, чем другая, но стоимость одной её акции выше, то она сильнее влияет на индекс. Даже большое процентное изменение цены относительно дешёвой акции может быть нивелировано незначительным в процентном отношении изменением цены более дорогой акции.

## Реализация в sklearn

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 можно воспользоваться функцией corrcoef из пакета numpy.

## Материалы

- Dow Jones Industrial Average: https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average
- История состава компаний, входящих в индекс Dow Jones: https://en.wikipedia.org/wiki/Historical_components_of_the_Dow_Jones_Industrial_Average

## Инструкция по выполнению

1. Загрузите данные close_prices.csv. В этом 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. Сколько 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?

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