Практическая работа № 10. Реализация линейной регрессии на Python

Задание:

1. Реализовать задачу линейной регрессии на python с помощью пользовательских функций.

2. В качестве исходных данных использовать датасет Boston.csv.

3. Визуализировать функцию линии регрессии:

- Построить точечную диаграмму: *plt.scatter(x, y)* исходные точки данных отображаются в виде точечной диаграммы.
- Построить линию регрессии: *plt.plot(x, y_pred)* построение линии регрессии с использованием прогнозируемых значений и независимой переменной *x*.
- Вычисление прогнозируемого вектора отклика: *y_pred=b+w*x* вычисляет прогнозируемые значения для *y* на основе рассчитанных коэффициентов.

Теоретический материал

Линейная регрессия - это статистический метод, который используется для прогнозирования непрерывной зависимой переменной (целевой переменной) на основе одной или нескольких независимых переменных (переменных-предикторов).

Метод линейной регрессии предполагает линейную связь между зависимой и независимыми переменными, что означает, что зависимая переменная изменяется пропорционально изменениям независимых переменных.

$$y = wx + b$$

В данном случае представлена простая или парная линейная регрессия, а уравнение вида

$$f_{w,b}(x) = w_0 x_0 + w_1 x_1 + ... w_n x_n + b$$

называется множественной линейной регрессией.

Где,

b - смещение модели,

w - вектор её весов,

x - вектор признаков одного обучающего образца.

Другими словами, линейная регрессия используется для определения степени, в которой одна или несколько переменных могут предсказать значение зависимой переменной.

Основные предположения, которые модель линейной регрессии делает относительно набора данных, к которому она применяется:

- **Линейная связь.**

Связь между откликом и переменными признака должна быть линейной.

Предположение о линейности можно проверить с помощью диаграмм рассеяния (рис.1).

Таким образом, 1-й рисунок даст лучшие прогнозы с использованием линейной регрессии.

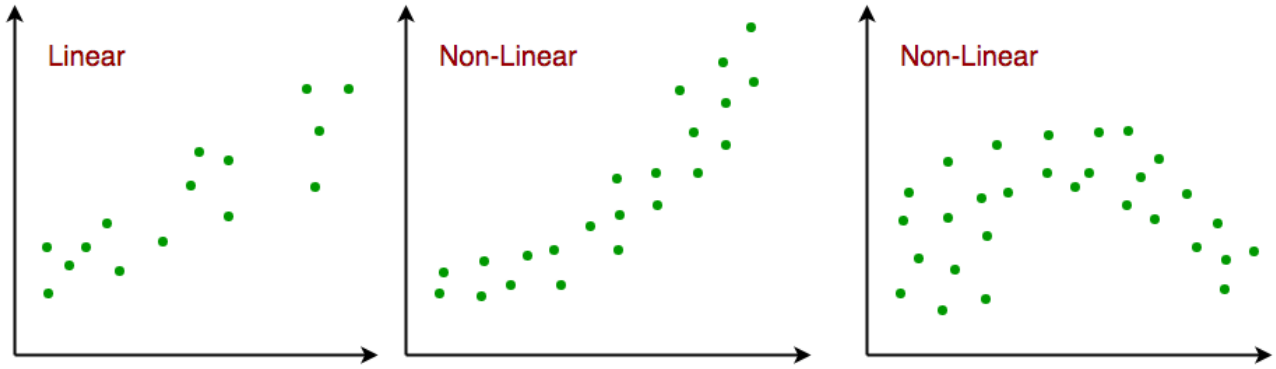

Рисунок 1 - Линейная связь в пространстве признаков

- **Слабая мультиколлинеарность или ее отсутствие,** т.е. заключение о том, что в данных слабая мультиколлинеарность или она отсутствует.

Мультиколлинеарность возникает, когда признаки (или независимые переменные) не являются независимыми друг от друга.

Одномерная линейная регрессия

Одномерная линейная регрессия - это тип регрессии, в котором функция (целевая переменная) зависит только от одной независимой переменной.

Для одномерной регрессии используют одномерные данные. Например, набор данных точек на линии можно рассматривать как одномерные данные, где абсцисса

может рассматриваться как независимая переменная, а ордината - как зависимая переменная.

В линейной регрессии предполагаем, что две переменные, т.е. зависимые и независимые переменные, линейно связаны.

Пример одномерной линейной регрессии

Для функции $y(x) = 2x + 3$ входным объектом будет x, а y -выходным.

x_i	0	1	2	3	4	5	6	7	8	9
y_i	1	3	2	5	7	8	8	9	10	12

Так как в данном примере есть только один входной вектор признаков X, следовательно линия регрессии будет иметь следующий вид $y = wx + b$.

Определим, что:

- X - вектор признаков, т.е. $X = [x_1, x_2, \ldots, x_n]$,
- Y - вектор ответа, т.е. $Y = [y_1, y_2, \ldots, y_n]$
- n – количество наблюдений (в приведенном примере $n=10$).

Диаграмма рассеяния приведенного набора данных представлена на рисунке 2.

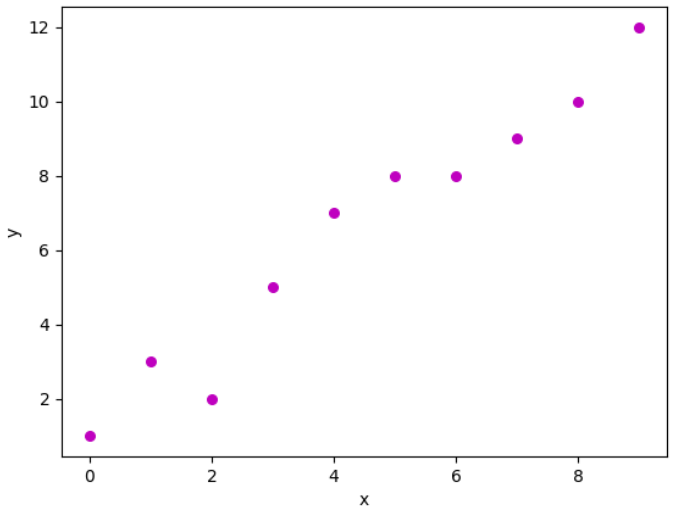

Рисунок 2 - Диаграмма рассеяния для случайно сгенерированных данных

Задача одномерной линейной регрессии состоит в том, чтобы найти линию регрессии, которая лучше всего подходит к приведенной выше диаграмме рассеяния.

Данная задача сводится к нахождению коэффициентов w и b таких, чтобы прогнозируемая (теоретическая) переменная y имела минимальную разницу с фактическим (эмпирическим) y.

Для этого зададим функцию ошибки ε, минимизация которой обеспечит подбор весов w и b, используя метод наименьших квадратов (МНК).

$$y = wx + b + \varepsilon$$

где ε - остаточная ошибка.

В методе МНК (рисунок 3) необходимо выбрать значения w и b таким образом, чтобы общая сумма квадратов разностей между теоретическим и эмпирическим значениями y была минимизирована, т.е. $\varepsilon \to 0$.

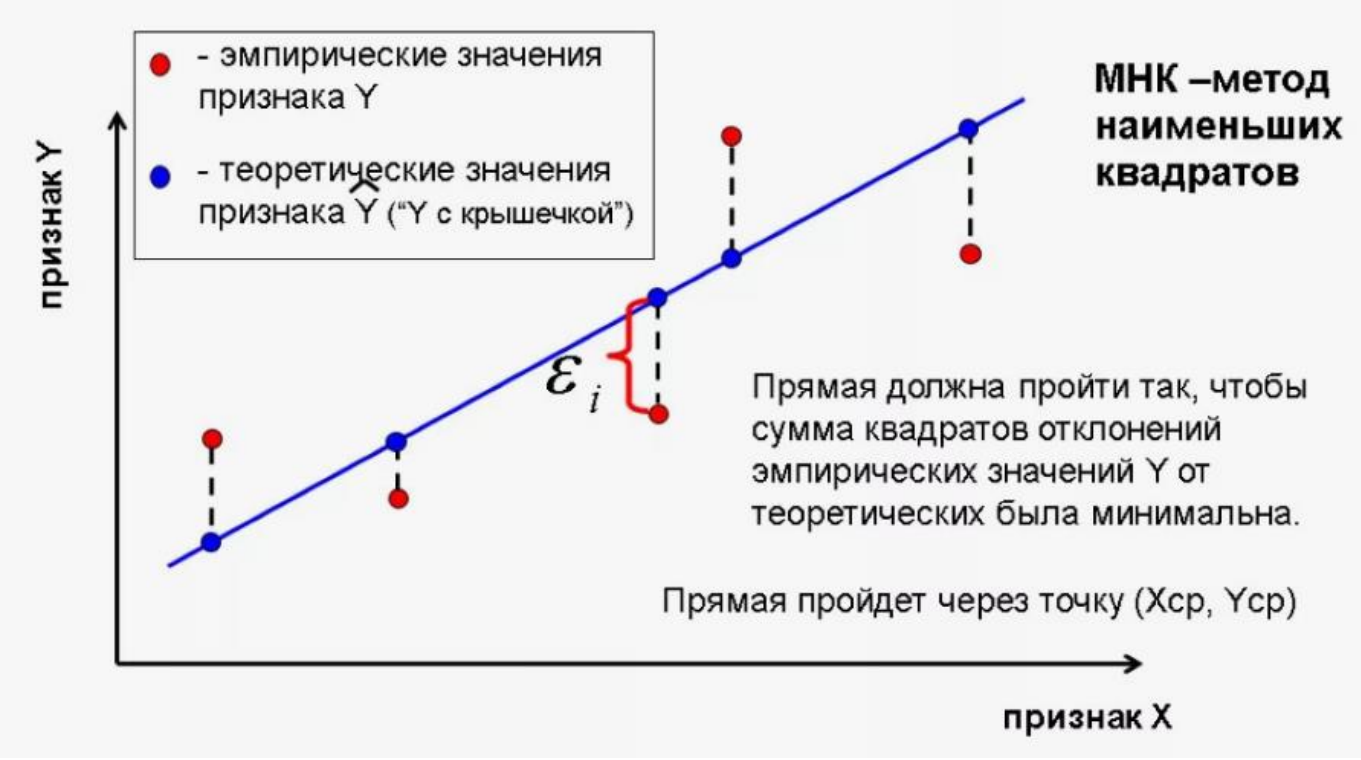

Рисунок 3 - Метод наименьших квадратов

$$w = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},\ b = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

где

i – номер измерения,

x_i и y_i – значения переменных при i-том измерении,

n – число измерений при моделировании системы.

или

$$w = \frac{\overline{xy} - \bar{x} * \bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad b = \frac{\bar{y} * \overline{x^2} - \bar{x} * \overline{xy}}{\overline{x^2} - (\bar{x})^2}$$