Experiment No. 4

Aim: Create a Linear Regression model using python to predict home prise using Boston Housing dataset.

Requirement:
- Anaconda Installer
- Windows 10 OS
- Jupyter Notebook
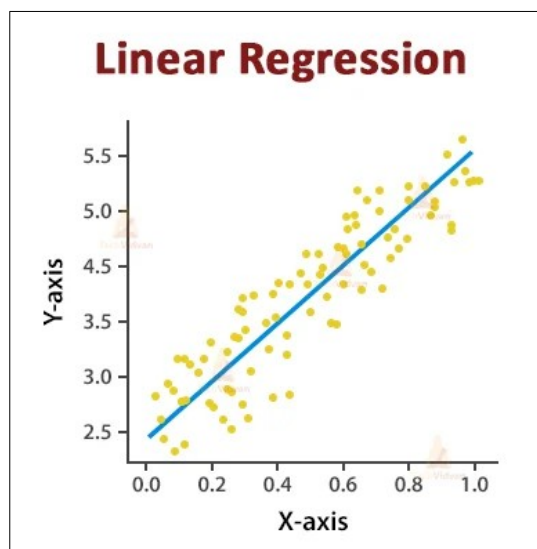
Theory:

Linear Regression in data science:

fig. Linear Regression

The term regression is used when you try to find the relationship between variables. In Machine Learning and in statistical modeling, that relationship is used to predict the outcome of events.

Simple Linear Regression:

Simple linear regression is an approach for predicting a response using a single feature. It is assumed that the two variables are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).

Multiple linear regression:

Multiple linear regression attempts to model the relationship between two or more features and a response by fitting a linear equation to the observed data. Clearly, it is nothing but an extension of simple linear regression.

Assumptions:

Given below are the basic assumptions that a linear regression model makes regarding a dataset on which it is applied:

• Linear relationship: Relationship between response and feature variables should be linear. The linearity assumption can be tested using scatter plots. As shown below, 1st figure represents linearly related variables whereas variables in the 2nd and 3rd figures are most likely non-linear. So, 1st figure will give better predictions using linear regression.
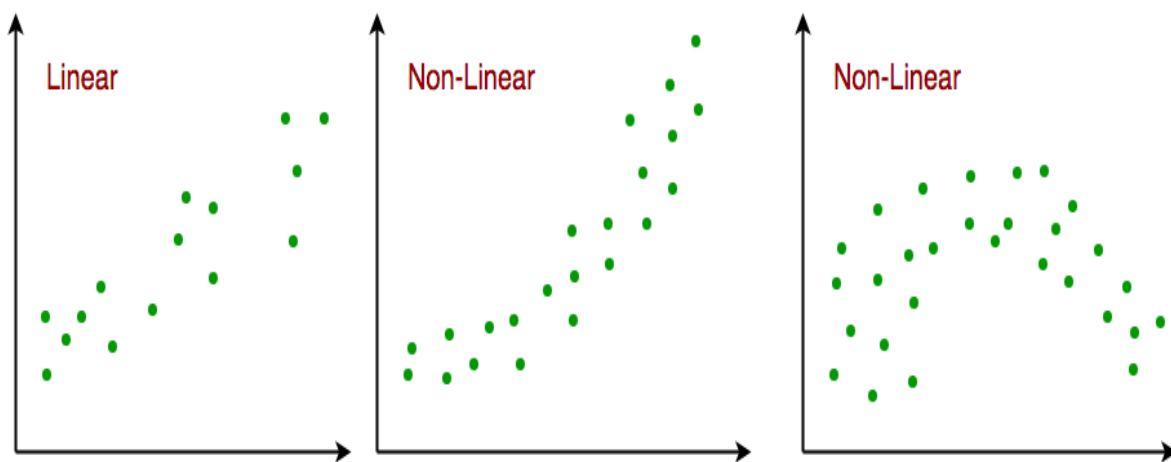
• Little or no multi-collinearity: It is assumed that there is little or no multicollinearity in the data. Multicollinearity occurs when the features (or independent variables) are not independent of each other.

• Little or no auto-correlation: Another assumption is that there is little or no autocorrelation in the data. Autocorrelation occurs when the residual errors are not independent of each other. You can refer here for more insight into this topic.

•Homoscedasticity: Homoscedasticity describes a situation in which the error term (that is, the "noise" or random disturbance in the relationship between the independent variables and the dependent variable) is the same across all values of the independent variables. As shown below, figure 1 has homoscedasticity while figure 2 has heteroscedasticity.

Applications:

1. Trend lines: A trend line represents the variation in quantitative data with the passage of time (like GDP, oil prices, etc.). These trends usually follow a linear relationship. Hence, linear regression can be applied to predict future values. However, this method suffers from a lack of scientific validity in cases where other potential changes can affect the data.

2. Economics: Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumer spending, fixed investment spending, inventory investment, purchases of a country's exports, spending on imports, the demand to hold liquid assets, labor demand, and labor supply.

3. Finance: The capital price asset model uses linear regression to analyze and quantify the systematic risks of an investment. Biology: Linear regression is used to model causal relationships between parameters in biological systems.

Dataset used:

- In this experminet we are going to use the boston housing dataset which contain information about various houses in boston through different parameters.

- There are total 506 samples ad 14 features (columns) in this dataset.

- Our objective is to predict the value of prices of the house using features with the help of linear regression.

Libraries Uesd:

1. Pandas: Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring and manipulating data.

2. Sklearn: It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Conclusion:

In this experiment we have studied abbout linear regression and done house prise prediction using boston housing dataset.