Data Science with Python

Day 03 - Activity - 01 -1

Linear regression, Multivariate linear regression and Logistic regression

1. Use the given 'boston housing' dataset and attribute information to answer the following questions.

Attribute Information:
These attributes represent different characteristics of the town or neighborhood, such as crime rate, zoning, air pollution, average number of rooms, age of houses, accessibility to highways, property tax rate, pupil-teacher ratio, racial makeup, and socioeconomic status.

- CRIM : per capita crime rate by town. Measures the per capita crime rate by town
- ZN : proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS : proportion of non-retail business acres per town
- CHAS : Charles River dummy variable (= 1 if tract bounds river; 0 otherwise). This variable is a dummy variable that takes a value of 1 if the tract bounds the Charles River, and 0 otherwise. The Charles River is a river that flows through Boston and some of its suburbs.
- NOX : nitric oxides concentration (parts per 10 million)
- RM : average number of rooms per dwelling
- AGE : proportion of owner-occupied units built prior to 1940
- DIS : weighted distances to five Boston employment centres
- RAD : index of accessibility to radial highways
- TAX : full-value property-tax rate per $10,000
- PTRATIO : pupil-teacher ratio by town
- B : 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
- LSTAT : Percentage of lower status of the population. This variable represents the percentage of the population that has a lower socioeconomic status
- MEDV : Median value of owner-occupied homes in $1000's. This variable represents the median value of owner-occupied homes in $1000s

I. Load the dataset into a Pandas DataFrame and explore the data.
II. Create a scatter plot to visualize the relationship between the 'MEDV' and 'LSTAT' columns.
III. Based on the scatter plot, what can you say about the relationship between the 'MEDV' and 'LSTAT' columns?
IV. Calculate the correlation coefficient between the 'MEDV' and 'LSTAT' columns.

V. Fit a multivariate linear regression model to predict the 'MEDV' column using the 'RM', 'LSTAT', and 'PTRATIO' columns. Print the coefficients and intercept of the regression line.

VI. Calculate the R-squared value for the multivariate linear regression model from question 1.

VII. Create a scatter plot to visualize the relationship between the predicted values and the actual 'MEDV' values from the multivariate linear regression model.

VIII. Use the multivariate linear regression model from question 5 to predict the median value of owner-occupied homes in $1000s for a house with 6 rooms, 10% lower status population, and a pupil-teacher ratio of 20.

2. Use the given "Heart Disease" dataset (***heart.csv***) to answer the following questions.

I. Use appropriate visualisation and find the distribution of heart disease in the dataset, and the correlation between different features.

II. Using gradient descent to train a model, implement logistic regression on the heart disease dataset.

III. Evaluate the logistic regression model in terms of accuracy on both the training and testing sets.

IV. Drow a ROC curve to improve the evaluation of the logistic regression model.