Session 20:

Additional Exercise

Session 20: Additional Exercise

Table of Contents

1. Introduction
2. Problem Statement
3. Output

1. Introduction

This assignment will help you to consolidate the concepts learnt in the session.

2. Problem Statement

This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It is in Python *sklearn* library itself, and has been used extensively throughout the literature to benchmark algorithms. The dataset is small in size with only 506 cases.

Dataset Naming

The name for this dataset is simply **Boston**. It has two prototasks: **nox**, in which the nitrous oxide level is to be predicted; and **price**, in which the median value of a home is to be predicted

Miscellaneous Details:

Origin: The origin of the Boston housing data is natural.
Usage: This dataset may be used for Assessment.
Number of Cases: The dataset contains a total of 506 cases.
Order: The order of the cases is mysterious.
Variables: There are 14 attributes in each case of the dataset. They are:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sqft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centers
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

What you should do?

Part 1 :

- Load the data using Python sklearn library.
- Do Exploratory Data Analysis of the data and state your insights.

Part 2 :

- Treat “target” as your Dependent Variable and rest of the columns as Independent Variables.
- Create Scatter Plot of Independent Variable vs Dependent Variable.
- Based on Scatter Plot see if there is any transformation required for Independent Variable.
- Build Multiple Linear Regression model. Use regularization techniques such L1 and L2 models.
- Build a polynomial model and check its accuracy