Lab -6

PRML
AY 2020-21 Trimester - III

March 10, 2021
Deadline: March 18, 2021, 11:59

Linear Regression

Build a linear regression model for the Medical cost dataset. The dataset consists of age, sex, BMI(body mass index), children, smoker, and region features, and charges. You need to predict individual medical costs billed by health insurance. The target variable here is charges, and the remaining six variables such as age, sex, BMI, children, smoker, region, are the independent variables. The hypothesis function looks like

$$h_\theta(x_i)=\theta_0+\theta_1 age+\theta_2 sex+\theta_3 bmi+\theta_4 children+\theta_5 smoker+\theta_6 region$$

Perform the following tasks for this dataset:-

1. Load the dataset and do exploratory data analysis. [3 Marks]
2. Plot correlation between different variables and analyze whether there is a correlation between any pairs of variables or not. [3 marks]
3. Plot the distribution of the dependent variable and check for skewness (right or left skewed) in the distribution. [3 marks]
4. Convert this distribution into normal by applying natural log and plot it. (If the distribution is normal then skip this). [3 marks]
5. Convert categorical data into numbers. (You may choose one hot encoding or label encoding for that). [3 marks]
6. Split the data into training and testing sets with ratio 0.3. [2 marks]
7. Build a model using linear regression equation $\theta=(X^TX)^{-1}X^Ty$. (First add a feature X_0 =1 to the original dataset). [5 marks]
8. Build a linear regression model using the sklearn library. (No need to add X_0 =1, sklearn will take care of it.) [3 marks]

9. Get the parameters of the models you built in step 7 and 8, compare them, and print comparisons in a tabular form. If the parameters do not match, analyze the reason(s) for this (they should match in the ideal case). [5 marks]
10. Get predictions from both the models (step 7 and step 8). [5 marks]
11. Perform evaluation using the MSE of both models (step 7 and step 8). (Write down the MSE equation for the model in step 7 and use the inbuilt MSE for the model in step 8). [5 marks]
12. Plot the actual and the predicted values to check the relationship between the dependent and independent variables. (for both the models) [5 marks]

Here is the Colab notebook attached for your reference.

Demonstration Colab notebook: Link

Instructions:-

Please Submit the necessary code(s) (Notebook) and a PDF explaining and analyzing the steps in both the questions along with necessary plots/figures.

Note:- No submission will be accepted after the final deadline.