Assignment 4

V Manaswini - EE22B065

Import the required libraries `numpy` and `csv`(to read the files with extension .csv)

Predicting Chance of Admission

There are various parameters which affects the chance of getting admission in top rated university. Here, we'll deal with the parameters such as GRE Score, TOEFL Score, Rating, SOP, LOR, CGPA, Research.

The goal here is to predict the chance of admit as a function of these parameters. The python file `ee22b065_Assignment4.py` contains the code which predict the chance of admit as a linear function of these parameters.

Let these parameters are denoted by `GRE, TOEFL, Rat, SOP, LOR, CGPA, Res` and their corresponding coefficients are denoted by `p1,p2,p3,p4,p5,p6,p7`, Chance of Admit is denoted by `Chance`, Number of sets of Data points is represented as `N`.

- Chance of Admit = p1(GRE) + p2(TOEFL) + p3(Rat) + p4(SOP) + p5(LOR) + p6(CGPA) + p7(Res)

Description of the Code

- Initially it extracts the data from the file `Admission_Predict_Ver1.1.csv` and creates 8 different types of lists which contains the data corresponding to the parameter.
- Theses lists get updated to a new list by removing the header and then converted to `int()` or `float()` data types depending upon the parameter.

Construction of Matrix M

The matrix **M** consists of 8 different columns; first column filled with all GRE Scores, second column with TOEFL Scores and so on upto the CGPA, the eight column is filled with ones.

$$M = \begin{pmatrix} GRE[1] & TROEF[1] & Rat[1] & SOP[1] & SOP[1] & LOR[1] & CGPA[1] & Res[1] & 1 \\ GRE[2] & TROEF[2] & Rat[2] & SOP[2] & SOP[2] & LOR[2] & CGPA[2] & Res[2] & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ GRE[N] & TROEF[N] & Rat[N] & SOP[N] & SOP[N] & LOR[N] & CGPA[N] & Res[N] & 1 \end{pmatrix}$$

The matrix **M** can be constructed by using the function `column_stack()`.

Solving for the coefficients

The coefficients can be estimated by using least squares regression. `np.linalg.lstsq()` would take the matrix **M**, **Chance** as its arguments and estimates the coefficients `p1,p2,p3,p4,p5,p6,p7`.

- **linalg.lstsq()** solves for the coefficients based on the equation $Mx = Chance$, where the column matrix `x` contains the coefficients.

Estimated Coefficients:

GRE Score: 0.0018585064850102188

TOEFL Score: 0.002777972391419563

University Rating: 0.005941368040176877

SOP: 0.0015861374557667321

LOR: 0.016858742352418608

CGPA: 0.11838505345773793

Research: 0.024307478582165632

Constant term in Linear Function: -1.2757250829969884

- To find the parameter which will affect the chance of admit by large amount can be found out by multiplying the coefficients of a parameter with their corresponding maximum values. The resultant values are said to be Parameter importances, whose values are 0.669, 0.333, 0.029, 0.008, 0.084, 1.183, 0.024 respectively.

From the above values we can say that **CGPA** will affect more on the chance of admit, **SOP** doesn't affect more on the chance of admit.

- You should focus more on CGPA to get admitted to a top rated university.

Errors

Root Mean Square Error(RMSE): 0.059504208777649545

Mean Absolute Error(MAE): 0.042572390149733345

- Those two errros follows the basic inequality: RMSE > MAE.

Conclusions

Since there is no huge amount of error(approximately (4 to 6)%) while predicting the chance of admit, therefore, Linear Regression Model can predict more accurate values as comapared to other Non-linear Models.

- Average Ratio of Estimated Chance to the given chance is 0.9998971962833112, which is approximately equal to 1. Hence, the predicted values by linear model are accurate.