LINEAR REGRESSION ON IRIS DATA (PROJECT1)

Report by: Thatiparthi Varshini

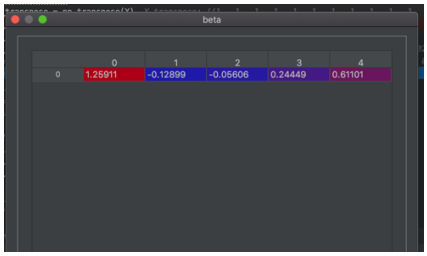
ID: 1001722318

Iris Flower Classification using Linear Regression:

- Goal: The main aim of this project is to classify the iris flower species by training a model using linear regression and perform cross fold validations.
- I have implemented this algorithm using python and the libraries used are numpy and pandas.
- The dataset consists of four independent features namely: sepal length, sepal width, petal length, petal width and one dependent feature is 'species' which consists of 3 groups:
- At first the data is split into independent features(X) and dependent feature(Y)
- By using cross fold validation the data is split into train and test sets.
- By splitting the data into two, the model can be trained and tested on different data
- Model: For finding the model we need to calculate the beta. The method findBeta calculates the beta value using train and test sets.
- findCV is to split the data as well as to calculate the predict the model, find accuracy is to calculate the accuracy score based on the test set and predicted values.

Results: When species values are Iris setosa: 1, Iris versicolor: 2, Iris virgincia: 3

- There are total 149 observations in the dataset.
- When K = 5 the data is divided into 5 folds where each fold consists of 29 observations which is X_fold and Y_fold.
- Then the data is split into train and test set where 4 folds are in train set and 1 fold is in test set.
- The beta value is calculated based on the train and test data. Below picture is the beta value for K = 5



- Later, Prediction values are calculated

  Predicted values = [0.92335488 0.91349236 0.98089589 0.87681954 1.00395313 1.00072949, 0.91248067 0.99345739 0.88109725 0.84406545 0.96272838 0.87515345, 0.86630261 0.70230289 0.88832697 0.90615668 0.93062699 0.90976051, 0.93825808 0.9097817 1.00496482 0.83062029 1.13738781]

  Test values = [1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1., 1. 1. 1. 1. 1. 1.]

- Accuracy score is calculated based on the test set and predicted values
  Accuracy score : 1.034482758620689

- The accuracy is 100% for K = 5 . This model provides better fit.
- Same Steps are repeated for calculating different K values
- Accuracy score for k = 2: 0.5540540540540541 (55%)
- Accuracy score for k = 5 : 1.034482758620689 (100%)
- Accuracy score for k = 10: 1.0714285714285714 (100%)
- Accuracy score for k = 15: 1.1111111111111112
- Accuracy score for k = 20: 1.1428571428571428
- As we can conclude that the model predicts with good accuracy as the K value increases.

When species values are Iris setosa: 2, Iris versicolor: 4, Iris virgincia: 6

- Accuracy score for k = 2: 0.24324324324324326
- Accuracy score for k = 5 : 1.0
- Accuracy score for k = 10: 1.0
- Accuracy score for k = 15: 1.1111111111111112
- Accuracy score for k = 20: 1.1428571428571428
- The model predicts same as before as we can see there is no big difference when we changes the species values.

Conclusion: This model has the best fit when we increase the K values and the accuracy score increases with respect to the partitions.