Name: Sai Rohit Kalyan Gandham

ID: 1002070724

Project1 – Report

Instructions given to complete the Project: Linear Regression

- The Data to be modeled is Iris data from the fallowing link
  Link: http://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
- Need to split the data more than 50-50 for the Training and Testing
- Should not use the Packages for the Linear Regression Specially "Sklearn"

Introduction

- **Linear Regression:** It is ML algorithm which is based on the Supervised Learning.
  It performs a regression tasks. This models a target prediction value based on the independent variables.
  When given a independent variable say (X), linear regression predicts the value of the dependent variable (Y), Based on the relation with X and Y it will draw a regression line which is the best fit line for our model.
- I have split the data to train = 70% and for the testing = 30% for generating a linaer regression model in order to generate the beta value.
- Formula to generate the linear regression using training data is
  $$\widehat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}$$
- I used the 4-fold cross validation method to validate and check the Accuracy of the model

Code

- Packages Used:
  1. Numpy
  2. Pandas
- I have used the pandas to read the csv file from the given link
- Prepared a map function which maps flower names to integer column
  {'Iris-setosa':0,'Iris-versicolor':1,'Iris-virginica':2}
- Creating a extra column with encode(Map)
- After creating a new column Shuffling the data so that I can use the 70% for the training
  And remaining 30% for testing the data
- Making sure the shapes of the trained and tested data are same
- Now defined the cost function and train function which return the train parameters(b)
- Once the training is completed we starting to predict the Y values
- Using for loop to check the Y_ predict = Y_test incrementing a variable called "correct"
- As the final step of the project used the 4-fold cross validation.
- In the 4-fold, each fold differs the proportion of the observation with a given categorical value

Which leads to a different accuracy values

ScreenShots:

The output of my code with cross validation information is seen in the screenshot.