CS 4661: Introduction to Data Science
Homework3
Due Date: Fri, Nov 4

Up to 3 students can team up to work on this homework. One of the team members should submit the homework on behalf of the team. Make sure to include the name/CIN of everyone on every submitted file.

Question1: Write and submit your python codes in "Jupyter Notebook" to perform the following tasks. Make sure to provide proper descriptions as MarkDown for each section of your code (each section of the code must have a short meaningful description right before that, describing what this part of the code is supposed to do!).

In this question, we work with another dataset from the textbook of "An Introduction to Statistical Learning."

a- Read the dataset file "Credit.csv" (you should download it from CSNS), and assign it to a Pandas DataFrame.

b- Check out the dataset. The "Credit" dataset includes "balance" column (average credit card debt for a number of individuals) as target as well as several features: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), marital status, and rating (credit rating).

c- Generate the feature matrix and target vector (target is "balance" in this dataset). Then, normalize (scale) the features (**note**: don't normalize the target vector!).

d- Split the dataset into testing and training sets with the following parameters: test_size=0.2, random_state=2.

e- Use Linear Regression to train a linear model on the training set. Check the coefficients of the linear regression model. Which feature is the most important? Which feature is the least important?

f- Predict "balance" for the users in testing set. Then, compare the predicted balance with the actual balance by calculating and reporting the **RMSE** (as we saw in lab tutorial 4).

g- Now, use 10-fold Cross-Validation to evaluate the performance of a linear regression in predicting the balance. Thus, rather than splitting the dataset into testing and training, use Cross-Validation to evaluate the regression performance. What is the **RMSE** when you use cross validation?

Question2 (no need for coding for Question2): Suppose we have a dataset with 3 features: X_1 = GPA, X_2 = Age, X_3 = Type of Position (1 for Technical positions, and 0 for Non-Technical positions), and build a non-linear regression model as:

$$\text{Target} = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \theta_4 X_1 X_2 + \theta_5 X_1 X_3$$

The target is "starting salary after graduation" (in thousands of dollars). Suppose we train (fit) the model, and get θ_0 = 30, θ_1 = 20, θ_2 = 0.07, θ_3 = -30, θ_4 = 0.01, θ_5 = 10.

(a) Which answer is correct, and why?

i. For a fixed value of Age and GPA, Technical positions earn more on average than non-technical positions.

ii. For a fixed value of Age and GPA, Non-Technical positions earn more on average than Technical positions.

iii. For a fixed value of Age and GPA, Technical positions earn more on average than Non-Technical positions when the GPA is high enough.

iv. For a fixed value of Age and GPA, Non-Technical positions earn more on average than Technical positions when the GPA is high enough.

(b) Predict the salary of a Technical and a Non-Technical positions with Age of 27, GPA of 4.0.

Question3 (we will cover plotting ROC curves in python and computing AUC in next session of class. So, you can wait until then before starting this problem):

Write and submit your python codes in "Jupyter Notebook" to perform the following tasks. Make sure to provide proper descriptions as MarkDown for each section of your code.

a- In this question, we work with a simplified version of Heart dataset (remember that this dataset is a little different from what you have used in HW2). Read the dataset file "Hearts_short.csv" (you should download it from CSNS), and assign it to a Pandas DataFrame.

b- Generate the feature matrix and label vector (AHD). Then, normalize (scale) the features.

c- Split the dataset into testing and training sets with the following parameters: test_size=0.2, random_state=3.

d- Use Logistic Regression Classifier to **predict** Heart Disease occurrence based on the training/testing datasets that you built in part(c). Then, compute and report the **Accuracy**.

e- Now, Use Logistic Regression Classifier to **predict the probability** of Heart Disease based on the training/testing datasets that you built in part (c) (you have to use "my_logreg.**predict_proba**" method rather than "my_logreg.**predict**"). Then, Plot **Roc Curve** for this classifier, and also Compute the **AUC** (Area Under Curve for ROC).