CSE343/ECE343: Machine Learning
Assignment-1 Linear Regression, Logistic Regression & Naive Bayes
Max Marks: 25 (Programming: 15, Theory: 10) **Due Date**: 21/9/2022, 11:59 PM

Instructions

- Keep collaborations at high level discussions. Copying/Plagiarism will be dealt with strictly.
- Late submission penalty: As per course policy.
- Your submission should be a single zip file **2020xxx_HW1.zip** (Where *2020xxx* is your roll number). Include **all the files (code and report with theory questions)** arranged with proper names. A single **.pdf report** explaining your codes with results, relevant graphs, visualization and solution to theory questions should be there. The structure of submission should follow:

  2020xxx_HW1
  |− code_rollno.py/.ipynb
  |− report_rollno.pdf
  |− (All other files for submission)

- Anything not in the report will **not** be graded.
- Remember to **turn in** after uploading on Google Classroom. No excuses or issues would be taken regarding this after the deadline.
- Start the assignment early. Resolve all your doubts from TAs in their office hours at-least **two days before the deadline.**
- Your code should be neat and well commented.
- **You have to do either Section B or C.**
- **Section A is mandatory.**

---

1. (10 points) **Section A (Theoretical)**

   (a) (2 marks) For simple linear regression, the least square fit line always passes through the point $(\bar{X}, \bar{Y})$, where $\bar{X}$ and $\bar{Y}$ represent the arithmetic mean of the independent variables and dependent variables respectively. Prove it.

   (b) (3 marks) Let us suppose if two variables have a high correlation with a third variable, does this convey that they will also be highly correlated? Justify your answer with the help of an example.

   (c) (2 marks) Provide proof of the weak law of large numbers (LLN). Provide a pseudo-code to illustrate the weak LLN assuming some distribution for the random variable.

(d) (3 marks) Derive the Maximum A Posteriori (MAP) solution for linear regression. (assuming gaussian prior distribution of the wieghts).

2. (15 points) **Section B (Scratch Implementation)**

   Linear Regression

   Implement Linear Regression on the given Dataset. You need to implement gradient descent from scratch i.e. you cannot use any libraries for training the model (You may use numpy, but libraries like sklearn are not allowed).

   Dataset: **Housing Price Prediction Dataset**

   (a) (6 marks) You will need to perform K-Fold cross-validation (K=2-5) in this exercise (implement from scratch). What is the optimal value of K? Justify it in your report along with the table for the mean RMSE of K-values and K-value.

   (b) (3 marks) Plot the RMSE V/s iteration graph for all models trained with optimal value of K for K-Fold cross-validation. RMSE should be reported on the train and val set.

   (c) (4 marks) Modify your Regression implementation by including L1 (LASSO) and L2 (Ridge Regression) regularization. Implement both regularization functions from scratch and train the model again. Try different values of the regularization parameter and report the best one. Plot similar RMSE V/s iteration graph as before (train and val loss).

   (d) (2 marks) Implement the normal equation (closed form) for linear regression and get the optimal parameters directly for each fold (optimal K). Report the RMSE on respective validation sets.

OR

3. (15 points) **Section C (Algorithm implementation using packages)**

   In this question, you are expected to understand and run Naive Bayes Algorithm.

   Dataset: **Dry Bean Dataset**

   (a) (1 marks) For the given dataset, plot the class distribution and analyze.

   (b) (2 marks) Perform EDA (histograms, box plots, scatterplots, etc.) and give at least five insights on the data. Check the missing values in the dataset.

   (c) (3 marks) Use TSNE (t-distributed stochastic neighbor embedding) algorithm to reduce data dimensions to 2 and plot the resulting data as a scatter plot. Comment on the separability of the data.

   (d) (2 marks) Run the sklearn's implementation of Naive Bayes (Any 2 of your choice - refer **here**). Report Accuracy, Recall, and Precision. Comment on the results and their differences from the two implementations of Naive Bayes. (80:20 train test split)

(e) (3 marks) Use Principal Component Analysis (PCA) to reduce the number of features and use the reduced data set for model training. Use values 4,6,8,10 and 12 for the number of components. Compare results (Accuracy, Precision, Recall, and F-1 score). (80:20 train test split)

(f) (2 marks) Use Scikit-learn to plot the ROC-AUC curves and comment on the output.

(g) (2 marks) Train your model using Sklearn's implementation of Logistic Regression, choose appropriate parameters, and comment on your choice. Compare the results with the ones obtained from Naive Bayes models. (80:20 train test split)