CSE/ECE 343/543: Machine Learning
Assignment-1 Linear Regression, Logistic Regression & Naive Bayes
Max Marks: 100 (Programming:90, Theory:10) **Due Date**: 20/09/2021, 9.00 PM

Instructions

- You are allowed to discuss but the final answer should be your own. 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 **AdmissionNo_HW1.zip**. Including only the **relevant files** arranged with proper names. A single **.pdf report** explaining your codes with details of EDA and pre-processing, relevant graphs, visualization and solution to theory questions. Anything not in the report will not be considered for evaluation. The structure of submission should follow:

  AdmissionNo_HW1
  |− Q1.py
  |− Q2.py
  |− Q3.py
  |− Q4.py
  |− Report.pdf
  |− Weights (folder)
  |− plots (folder
- Restrict to using only Python for coding assignments.
- You are free to use math libraries: Numpy, Pandas; and use Matplotlib library for plotting.
- Use of inbuilt function for any evaluation metric is not allowed. Each of the metric needs to be implemented from scratch.
- 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 **two days before the deadline.**
- **Document** your code. Lack of comments and documentation would result in loss of 20% of the *obtained* score.
- The assessment will be done on basis of the following components:
  - Working codes
  - Analysis and clarity of results (drawing comparisons across different parts) clarity of the report

– Understanding the theoretical concepts/viva

QUESTIONS

1. (30 points) **Linear Regression**

   Download the **Abalone dataset** , **Readme** dataset for this dataset. It contains 9 variables in which the last column is the output variable and the other 8 are input variables.

   1. Perform Linear regression (implemented from scratch) on this dataset, use 8:2 train:test split, set random seed to 0. Report the RMSE on training and testing data (10 marks)
   2. Apply regularization techniques Ridge Regression and Lasso Regression (sklearn's implementation can be used) here. You have a hyperparameter alpha that you can modify to regulate how much the coefficients are restricted.
      (a) Plot a graph to show the effect of alpha value (at least 10 different alphas has to be used) on the testing data's RMSEand report best model's coefficients / parameters for both Ridge and Lasso Regression respectively. (10 marks)
      (b) Use Sklearn's Grid search function to find the best alpha value and report the best model coefficient for both Ridge and Lasso Regression respectively? Compare the best model coefficients with reported in Q1.2.a. (10 marks)

2. (30 points) **Logistic Regression**

   Download the **Diabetes Dataset**, **Readme** for this data. This dataset contains 9 columns, use the "Outcome" column as the target value and the other columns as the features.

   You need to implement gradient descent (both SGD and BGD) from scratch (You may use numpy, but other libraries like sklearn, keras are not allowed)

   1. Perform Logistic Regression (implemented from scratch) on this dataset, use a 7:2:1 train:val:test split. Show all your preprocessing steps, and mention them in your report
      (a) Include loss plots between training loss v/s iterations and validation loss v/s iterations. Comment on the convergence of the model. Compare and give your analysis between the plots. (7 marks)
      (b) Re-Run your implementation for different learning rates 0.01, 0.0001, 10. Compare and give your analysis. (8 marks).
      (c) Make the confusion matrix and report the accuracy, precision, recall and f1 score obtained. (7 marks)
   2. Choose an appropriate learning rate, number of epochs from part Q2. part 1. Using the same learning rate, run sklearn's Logistic Regression on the dataset above and compare the following:

(a) Loss plots of sklearn and your implementation. (3 marks)

(b) Number epochs to converge of sklearn and your implementation. (2 marks)

(c) Performance sklearn's implementation and your implementation. Report accuracy, precision, recall and f1 score. (3 marks)

3. (30 points) **Naive Bayes**
   Use the FMNIST dataset which has 60K training data points and 10K testing points. The dataset has 10 different classes of clothes and the image pixels are in the range 0-255, you can binarize the images to 0,255.

   1. Implement (from scratch) a machine learning algorithm using pixel values as the feature and Naive Bayes Classifier to differentiate between two classes : Trouser and Pullover. (10 marks)
   2. You will need to perform K-Fold cross-validation in this exercise (implement from scratch). Choose an appropriate value of K and justify it in your report along with the preprocessing strategy. (10 marks)
   3. Plot the curves, sklearn library can be used
      (a) Make the Confusion Matrix (3 marks)
      (b) Plot the ROC curve (5 marks)
      (c) Find the Accuracy, Precision, Recall (2 marks)

4. (10 points) **Theory**

   1. Assume you want to estimate the wage equation $W_i = \beta_0 + \beta_1 X_i + u_i$, where W_i is the wage of worker i and X_i is the labour market experience of worker i.
      (a) Assume you suspect that the intercept in the equations is different for men and women. Explain how you will change the above model to test this suspicion? (2 marks)
      (b) Assume you suspect that the slope ("return to experience") is different for men and women. Explain how you will change the above model to test this suspicion? (2 marks)
      (c) Assume you suspect that the relationship between wage and experience resembles an upward slope. Explain how you will change the above model to test this suspicion? (2 marks)
   2. L2 regularisation (also known as RIDGE regression) promotes smaller coefficients. Briefly explain the intuition behind regularisation. (1 marks)
   3. Take a Gaussian prior on parameters, and derive the equation for L2 regularisation using Maximum A Posterior Inference. (3 marks)