Introduction to Machine Learning

IML-IBM

Assignment-1

Max Marks: 10
Due Date: 10/April/2021

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 **OPDxxxxx_ASG1.zip** (Where OPDxxxxx is your roll number).
- Include only the **relevant files** arranged with proper names. A **.pdf report** explaining your codes with relevant graphs and visualization and theory questions.
- Do **NOT** include data files in your submission. It makes your files unnecessarily big while downloading.
- Ensure that everything required for a particular question is present in their respective files in terms of functions (not comments). Failure to do so would result in a penalty. Follow the following file structure for submission:

  OPDxxxxx_ASG1

  |− Q1.py

  |− Q2.py

  |− Q3.py

  |− Report.pdf

  |− Plots (folder)

- Remember to **turn in** after uploading on google classroom.
- Resolve all your doubts from TA in their office hours **two days before the deadline.**
- **Document** your code. Lack of comments and documentation or improper file names would result in loss of 20% of the *obtained* score.

1. (3.5 points) **Linear Regression**

Download the datasets **Dataset 1** (README). Perform Linear Regression on the given dataset. Also perform K-Fold cross-validation in this exercise.

Analysis to be included in your report:

(a) Choose an appropriate value of K and justify it in your report along with the preprocessing strategy. (0.5 point)

(b) Include plots between training loss v/s iterations and validation loss v/s iterations. (1.5 points)

(c) Implement gradient descent using two losses - RMSE loss and MAE loss. Include the best RMSE and MAE value achieved in your report. (1.5 points)

2. (1.5 points) Use the **Real_time_dataset** for this. Apply different types of loss functions to predict class 'gender' and report their performances.

3. (5 points) Use Decision Tree(DT) and Gaussian Naive Bayes (GNB) classifier to train **Dataset 1**. Split the data into 60-20-20 train-val-test splits. Implement K-Fold cross validation for both GNB and DT.

(a) Save the best model, load the saved model to predict the results on the test data (2 points).

(b) Evaluate testing data on the basis of accuracy, precision, recall, F1-Score, plot ROC-curve and confusion matrix (1 point).

(c) Find optimal depth as a parameter in-case of DT using Grid Search and use K-Fold cross validation to validate it (1 point).

(d) For DT plot training and validation accuracy plot with respect to tree depth and write your analysis(1 point).