ES 647 - Pattern Recognition & Machine Learning

Assignment 2

Release Date : March 29, 2018
Due Date: April 8, 2018

Instructions

1. Follow the honor code of the institute while doing any assignment. Any violation in that would be taken quite seriously.
2. You can consult/discuss with any of your friend to develop the solution strategy. You can also take help of your friend in setting up your machine. However, the final solution and code should be written by you from scratch and you should not copy even a single bit of it from others. You should acknowledge the help taken from your friend(s) in your code at the top part (in comments).
3. You will be required to submit one single **.py** file for the entire Assignment 2. The submission needs to be done via Canvas only.
4. You should name the file as follows: *RollNumber_assignment2.py* . Files not following this naming convention will not be evaluated.
5. The submission should be done by 11:59 PM on the due date. Late submissions will be penalized.
6. All the plots should be properly titled. The axes should have proper title and markers. In any plot, the width of curves and markers (if any) should be chosen sufficiently so that the plot is visible properly. Further, highlight gridlines or additional lines wherever it make sense and wherever it adds more value to the plot.
7. For any kind of clarification on the problem definition and what you need to do in this assignment, you can contact our TAs via email communication in Canvas. You can also post your queries in the announcement section of Canvas and let your friends or TAs answer that eventually. You also feel free to answer the queries of others on canvas (but don't provide the solution).

Problem Set for Assignment 2

In Assignment 2 you are required to implement 4 well known Classification Algorithms on following datasets.

Datasets

1. MNIST: Digit Images Dataset - Multiclass classification (Available in Scikit Learn)
2. Default of Credit Card Clients - Binary Classification Kaggle

The Credit Card dataset has categorical features. For some of the algorithms you require to convert these features to numerical values. One of the ways is onehot encoding. You are free to use any other method/idea of your choice to do the task. However you must be able to justify your choice.

Classification Algorithms

You will apply following classification algorithms on each of these datasets:

1. K-nearest neighbors
2. Decision Tree
3. SVM
4. Logistic Regression

As problem 1(a,b,c,d) implement all the algorithms in above order on MNIST dataset and as problem 2(a,b,c,d) do the same for the Credit Card dataset

Reporting Metrics

For both the data sets and all classification algorithms you have to show the confusion matrices as a compulsory metric for both training and testing data. Other than that you can report the following metrics:

1. For MNIST: Top - 1 Accuracy, Top -3 Accuracy etc.
2. For Credit Card : Accuracy, F1- score, precision and recall score, ROC curve etc.

You might need to read about some of these metrics. There are plenty of resources available on the net. Wikipedia should be a good starting point. Modules might be available on sklearn too. Be sure you understand the metric that you are reporting properly.

This is a much open ended assignment than the previous one. You are free to choose the size of your training and test sets. You are also free to tune the hyper parameters of your classifiers by yourself. You can use cross validation method, brute force etc. You are encouraged to try different combinations and report the best ones. You can also decide on how to report the metrics using graphs, figures etc. You can also use metrics not mentioned here. You are encouraged to think as an ML practitioner. The evaluation will be based on your understanding, efforts, and presentation of your results.