Assignment # 3

SYDE 675 Winter 2020

This assignment is considered as both the assignment 3 and the final project.

The assignment can be done in Python or Matlab.
You need to submit both your reports and the source codes implementation for all questions (either python or matlab with an additional txt file with all your source code in questions 1 and 2). The report must be a single pdf (one for question 1 and 2 and the second report for question 3 and 4 as your final project) and the source code must be a single .py or .m file for each question.

For question 4 please upload a separate .py or .m file and a separate .txt file.

Please include brief comments in your code. Be sure to label all figures and include a legend where appropriate.
The due date for this assignment is **April 20th, 2020**, due to the firm deadline of final grades, the deadline will not be extended.

Part A
In this homework, you will be using support vector machines to gain an intuition of how SVMs work. You are allowed to use any existing implementations of SVM including MATLAB's built-in functions, OSU-SVM, LibSVM and etc. As a suggestion, you can use the Lib-SVM toolbox.

Question 1 (7)
Linear SVM for Two-class Problem

Use the **hw3_dataset1** and **hw3_dataset2** for this question. In this part, you try different values of the C parameter of SVM. Informally, the C parameter is a positive value that controls the penalty for misclassified training examples. A large C parameter tells the SVM to try to classify all the examples correctly. Use the whole set for training purposes. Answer the questions below based on the two datasets separately and compare the results. Answer the questions for both datasets:

1. Load data of two classes and plot to visualize the dataset on a figure.
2. Train a linear SVM on the dataset. Try to use different values of C and see how the decision boundary varies. Use C={0.001, 0.01, 0.1, 1}.
3. Plot different decision boundaries with different C and compare them beside each other on one figure in your report.
4. Which value of C is the best value for this dataset? Explain the effect of C in training of SVM.

Question 2 (8)

Adaboost

In this part you will create an adaboost classifier based on linear SVM to classify the dataset in Question 2.

1. Load and plot **'classA.csv'** and **'classB.csv'** and visualize them on the same figure.
2. Train a linear SVM with proper C value from the set {0.1, 1 , 10 , 100} and visualize the decision boundary and report the accuracy based on 10-times-10-fold cross validation.
3. Create an ensemble of classifiers based on the Adaboost-M1 approach to classify the dataset again. Use a linear SVM with the selected C in part 2 as your weak learner classifier. Use T = 50 as the max number of weak learners.

   Note:

   I) For each iteration draw only 100 samples from the dataset to train each classifier.

   II) If the training error is higher than 50% in one iteration, discard the classifier and re-sample the training set and train a new classifier. Continue until you have trained 50 unique SVMs.
4. Report the mean and variance of accuracy for 10-times-10-fold cross validation approach.
5. Visualize the decision boundary of the ensemble model on the plot in part 1.

Part B

The answers to the following questions should be submitted as a separate report. Name your report as "Report#2.pdf". This question and the next question are considered as your course project. You can use any public library to do the next two questions. However, you must cite all the papers and libraries you use to answer the question. It is recommended to structure your report for the next two questions like a research paper with 2-3 pages and try to answer the asked questions through your report.

Question 3 (10)

Write a one-page paper review (two-column format) on an advanced topic based on topics we discussed in the class. You can select any arbitrary topic which you like to learn more about. However you need to use your method to solve the problem in question 4 as well.

Question 4 (15)

Use the method you described in question 3 to solve the following problem. Please download the dataset from the link:
https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

The dataset is one of the UCI datasets.

You are going to solve Human Activity Recognition based on the sensory data. The task is to predict six activities (WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING) based on signals extracted by wearing a smartphone on the waist.

The dataset is a combination of several signals extracted from an accelerometer and gyroscope when with different human activities. The goal of this question is to create a model (you are free to use any machine learning model) to be able to predict the human activity based on the input signals.

1. Talk about your methodology, justification why you use that method and explain how to design your method to address the problem in detail.
2. Use the training data to train your model and report the 10-times 10-folds accuracy of the model you designed based on the training data.
3. Report the best set of parameters and model setup which you found based on the cross-validation.
4. Report the test accuracy and show the confusion matrix. Explain your findings and how it is possible to improve the model as possible future work.