Assignment - 2

Due Date: 18th October

Dataset to be used: MNIST dataset (http://yann.lecun.com/exdb/mnist/). The data can be converted into a matlab format by using the codes available here: http://ufldl.stanford.edu/wiki/index.php/Using_the_MNIST_Dataset. You are free to use this or any other available code for converting the data into integer matrices. The data is already divided into training and testing partitions, the same partitions need to be used in the assignment. You are free to use any existing implementation of SVM.

Q1. Two class SVM (50 marks)

A. Using only data pertaining to class '6' and '8', train a linear SVM for a two class problem by performing a grid search for the best value of 'c' (test set has not been used till now). (5)
B. Plot the training set accuracy with respect to the values of 'c' used in your grid search. (5)
C. Using the final model obtained above, report the parameter used and class-wise accuracy on the test set. (5)
D. Depending on the results obtained above, divide the data into two sets, correct classifications and misclassifications. In SVM, each data point has a distance from the hyperplane. Extract the distance for the points pertaining to the two sets created in this part and plot the distances. You are just required to create a single plot, containing the distances of the two sets (classified correctly, classified incorrectly). Report your observations with respect to the distance of the points and the classifications. (10)
E. Apply a RBF kernel and perform grid search on all the parameters to obtain a trained model. Report the final parameters used and class-wise accuracy on the test set for this model. (10)
F. Data pertaining to two classes is given below:
   class1 = [5 12; 5 -12; -5 12; -5 -12; 0 13; 0 -13; 13 0; -13 0];
   class2 = [0 5; 5 0; -5 0; 0 -5; 4 3; -4 3; 4 -3; -4 -3];
   a. Plot the points on the same graph, such that the two classes can clearly be seen. (5)
   b. Train a SVM (with or without kernel) which will be able to give 100% accuracy on the given dataset. If it is possible to train such a SVM, report the parameters, support vectors, and kernel applied. If it is not possible, explain why. (10)

Q2. Multi-class SVM (20 marks)

A. As discussed in class, multiclass svm can be implemented as one-versus-one, half-versus-half and one-versus-all. Libsvm has the one-versus-one approach implemented for multiclass classification. **Implement** either half-versus-half or one-versus-all multiclass svm for the entire MNIST dataset (10 classes). Compute individual class accuracies for each class and report a 10-class confusion matrix for the results obtained on the test set. Analyze the results.

Q3. Text Classification (20 marks)

A. Using the SPAM/HAM dataset in the following link (https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection), train a two class SVM for the problem of SPAM versus HAM. Use the first 200 messages as training for both the classes. Extract meaningful features from the dataset provided (you may use any existing tool for preprocessing and feature extraction **only**, however you should know what the tool is doing). Train a SVM model with a kernel not applied in the above questions on the training set. Report the parameters chosen after grid search and the class-wise accuracy for the remaining data (the test set).

Q4. Theory Questions (30 marks)

A. Give an example of a data (draw on graph) which will be separable by a decision tree and not by a linear SVM. If such an example does not exist, explain why. (10)
B. Give an example (draw on graph) for a two class problem, where the boundary of Linear SVM obtained with c = 0 and c= infinity will almost be the same. If such an example does not exist, explain why. (10)
C. If there is a kernel, $K_a(x,y)$ and another kernel, $K_b(x,y)$, then $K_c(x,y) = K_a(x,y)*K_b(x,y)$ is clearly not a kernel. State true or false with mathematical proof. (10)

You need to submit a report (pdf file) with all your findings.