Project 3

Total Points: 25

Support Vector Machine

The objective of this part is to use SVM on Sonar data set available at UCI Machine Learning Dataset Repository.

Dataset information

Information for this dataset is available at
http://archive.ics.uci.edu/ml/datasets/connectionist+bench+(sonar,+mines+vs.+rocks)

Deliverables:

- A single code file which should run on a click when the sonar.mat file is placed in the working directory. (Functions are to be written at the end of file in MATLAB)
- A pdf file of all the results

Following tasks are to be performed:

1. Loading the data in MATLAB

   ```
   load('sonar.mat')
   ```

   After loading the data, you will have 2 variables in your workspace: X and Y. X is the data matrix where rows are observations. Y contains the class labels, either -1 or +1. There are 208 observations in this dataset.

   Select random 25% of the data for testing. Remember to select respective data labels from Y matrix. The remaining data is training data.

   You should have following matrices after this step,

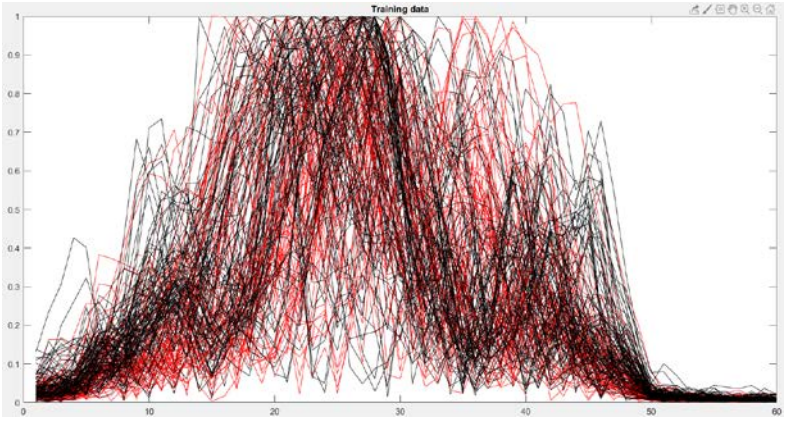
   trainfeatures (156*60)

   trainlabels (156*1)

   testfeatures (52*60)

   testlabels(52*1)

2. Plot training data and test data in **2 different plots**. Use different colors for different colors. Here is my training data with red color for label==-1 and black color for label==+1. Don't forget to add titles to the plot. Legend to the plot can be in the plot or can be added as a caption.



3. Apply a hard margin SVM and report the testing accuracy. You can use inbuilt function for this, or you can code it on your own.
4. Apply a soft margin SVM and report the testing accuracy. Soft margin SVM has a tunable parameter C. Optimum value of C can be found out using k-fold cross validation.
    a. Use 3 fold cross validation. Report value of obtained C and report the value of testing accuracy at this C (Search for optimum value of C with C=0.01, 0.02, 0.03 … 1.00).
    b. Use 5 fold cross validation. Report value of obtained C and report the value of testing accuracy at this C. (Search for optimum value of C with C=0.01, 0.02, 0.03 … 1.00).
    c. Plot Mean accuracy vs Value of C for 3-fold and 5-fold cross validation in 2 different plots.
5. Aggregate all results in the following table

Method	**Testing Accuracy**
Hard Margin SVM	
Soft Margin SVM with C = … (C value obtained using 3-fold CV)	
Soft Margin SVM with C = … (C value obtained using 5-fold CV)	

How to use cross validation for tuning the hyper parameter C?

Step 1: Set a value of C (say C=0.01).

Step 2: Split the dataset into k equal partitions.

Step 3: Use first partition as testing data and union of other partitions as training data and calculate testing accuracy.

Step 4: Repeat step 2 and step 3. Use different set as test data different times. That is if we are dividing the dataset into k folds. On the first iteration, 1st fold will be test data and union of rest will be training data. Then we will calculate the testing accuracy. Then on next iteration 2nd fold will be test data and union of rest will be training data. Likewise, we will do for all folds.

Step 5: Take the average of these test accuracy as the accuracy of the sample. Example, for k=5 you will have 5 test accuracies and you need to take mean of those 5 accuracies. This would be the mean testing accuracy at the selected C value.

Step 6: Now update the C value (say C=0.02) and repeat step 2 to step 5.

Remember once you obtain the optimum value of C, the train and test data split would be similar to what you used for hard margin SVM.

Note: Shuffle the data before you obtain k partitions of the data.

Point distribution:

Hard Margin SVM: 5points

Soft Margin SVM: 10 points

Implementation of cross validation: 10 points

Acknowledgements:

The sonar data was obtained from UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.