

## 1. Programming Exercise: Soft Margin SVM.

Dataset: The Spambase folder contains a spam classification data set drawn from the UCI Machine Learning repository (<http://archive.ics.uci.edu/ml/datasets/Spambase>) and normalized to have each feature value in the range  $[0, 1]$ . In the data files provided, each row is a separate training example; the first 57 columns correspond to features, while the last column is the label (+1/ - 1). There are two files train.txt and test.txt containing 250 training instances and 4351 test instances respectively. The CrossValidation folder contains the train and test data for each fold of the 5-fold cross-validation procedure on the training set train.txt. In each fold, there are two files cv-train.txt and cv-test.txt containing 200 training instances and 50 test instances respectively.

Code: You have been given a PYTHON code SVMcode.py to learn an SVM classifier. You need to edit in the PYTHON code “SVMcode.py” as directed.

Instructions:

1. Your program should read the training data, test data, parameter C and print the percentage error of your trained SVM model.
2. Run your implementation of the SVM learning algorithm selecting C from the range  $\{1, 10, 10^2, 10^3, 10^4\}$  through 5-fold cross-validation on the training set. Report the average cross-validation error (over 5 folds) for each value of C in this range. [Note: Use files from CrossValidation folder here.]
3. Take the best value of C (i. e., the one which achieves the least average cross-validation error) and run your SVM implementation with this choice of C. Report the training and test errors achieved by this value of C. [Note: Use train.txt and test.txt files here.]