DMML
Assignment

November 18, 2014

This exercise is from Andrew Ngs notes on classifying text as spam and non spam using SVM.

Here you use a library called libsvm available at: `http://www.csie.ntu.edu.tw/~cjlin/libsvm/#java`

Please download the appropriate library depending upon your choice of implementation language/package. The readme file on the downloaded file has the instruction to compile the library on your machine. (If you face any difficulty in getting the library compiled please let me know)

You don't have to write the SVM but you need to play around with the library. The input files are formatted in such a way that it is easy to read with the library mentioned above. The library provides a train method to train the model, and a predict method to test the model. Check the library documentation for details.

There is a twofeature.txt which has 2 features. In this case, you can visually see the input by plotting and check how well your classifier performs, by drawing $w^T x + b = 0$ line. Try with various values of C and plot the input along with the model ie. the line $w^T x + b = 0$.

For email classification, train your model with different training sets available:
(email_train-50.txt, email_train-100.txt, email_train-400.txt and email_train-all.txt). So you will have 4 models, one for each training set. Do a prediction using the models obtained, on the test set (email_test.txt) and tabulate the results ie. the training set vs accuracy. Compare these results with the previous exercise's result (Naive Bayesian prediction) and explain your observation.