

CS 410/510
Machine Learning
Spring, 2007

Homework 5:
Linear Discrimination and Support Vector Machines

Due Tuesday, May 15.

In this assignment, you will experiment with using a support vector machine on a data set of your choosing.

The support vector machine system you will use is called SVM_light. It was written by Thorston Joachims. You can download the code from <http://svmlight.joachims.org/> This web site also gives instructions on how to compile and run the code.

Here is what you need to do for this assignment:

1. Choose a data set to experiment on. You can use one of the UC Irvine data sets (<http://www.ics.uci.edu/~mllearn/MLRepository.html>) or find a different one on the web. You need to find a data set that contains a reasonably large number of training examples (say, > 1000), so you can play with different size training sets. SVM_light does only binary classification, so if your data set has multiple classes, you need to change the classification problem to a binary one (e.g., for a hand-written character recognition problem, “letter A” versus “not letter A”). Make sure that your training and test data have approximately balanced numbers of positive and negative examples.
2. Download the SVM_light code from the URL given above, and read the instructions on how to run it. You can simply use all the default parameters, which means you don’t need to specify any options (except for the experiments below on varying the kernel type.)
3. Put your data set in the format required by SVM_light.
4. Run svm_learn on the training data. Then run svm_classify on the test data. Report the following (you can get all this information from the output of the SVM code and from the “model” file it creates):
 - The classification accuracy (i.e., fraction of correct classifications) on the training data.
 - The classification accuracy (i.e., fraction of correct classifications) on the test data.
 - The precision and the recall on the test data.
 - The number of support vectors that were used in the model created by the SVM.
 - The estimated VC-Dimension of the model (or an upper bound on it).

Repeat two more times, using different sizes of training sets (you can choose the sizes—if you make the training set larger, simply move some examples from the test set to the training set). How does changing the size of the training set affect the above measures?

5. In step 4 you used the default *linear* kernel. Repeat step 4 with the original-size training and test data using two other kernels (specified with the `-t` option): *polynomial* (1) and *radial basis function* (2). Report the same results for each of these that you reported in step 3.

6. Run the C4.5 system and your Naive Bayes system on this same data. Compare the error of these two and the SVM system on the test data. Use 5-fold cross-validation to compute the confidence in the hypothesis that the observed best method actually has true error less than the observed second best method. In the writeup, show your work on this.

7. Summarize. Write a few paragraphs or a page or two describing your data set, and summarizing all of your results from the above steps.

What to hand in

Hand in a hard-copy of your (computer formatted, not handwritten) write-up as described above. That's it!