CS 445/545, Machine Learning, Winter, 2014

Homework 2: Support Vector Machines and Boosting

Due Wednesday February 12, 2:00pm.

In this homework you will experiment with classification of dog versus cat images using support vector machines and boosting. The images come from the competition at http://www.kaggle.com/c/dogs-vs-cats

Download data

Download DogsVsCats.zip from the class website:

http://web.cecs.pdx.edu/~mm/MachineLearningWinter2014/DogsVsCats.zip

This archive contains the following files:

- DogsVsCats.train
- DogsVsCats.test

Each instance in these files is a 64-color histogram of the corresponding image. The values of the features are normalized fraction of pixels in the image of a given color bin.

Cats are the negative class and dogs are the positive class.

Download svm_light

Download svm_light from http://svmlight.joachims.org/ Follow the directions (under "Source Code and Binaries" and "Installation") to download and install this package on your computer. Read the "How to Use" section.

Exercise 1: SVM Accuracy and Cross-Validation for Model Selection

Compare the accuracy of the linear kernel versus the polynomial kernel (with degree 5) using 10-fold cross-validation. That is:

Divide the training set into 10 disjoint subsets of approximately equal size. Each subset should have roughly the same number of positive and negative examples.

Do the following 10 times:

- Choose one subset to be the "validation" set. (Choose a different subset at each iteration.) The other 9 subsets put together make up the training data.
- Train an SVM with the linear kernel on the training data , and record the accuracy on the validation set.
- Train an SVM with the polynomial kernel on the training data, and record the

accuracy on the validation set.

Now, for each kernel type, you will have 10 different accuracy measures. Average these 10 measures to get the prediction of test accuracy for each kernel. Call these averages the "**validation accuracy**" for each kernel.

Finally, train the SVM on all the training data (all 10 subsets put together) using the linear kernel. Then do the following:

- Run svm_classify on the training data to get the **training accuracy** of the linear kernel
- Run svm_classify on the test data to get the **test accuracy** of the linear kernel.
- In your report, give the training, validation, and test accuracies of the linear kernel. Which gives a better prediction of the test accuracy, the training accuracy or the validation accuracy?

Repeat the previous step (under "Finally" above), for the polynomial kernel.

Which kernel has the higher test accuracy?

Exercise 2: Boosting SVMs

Implement the Adaboost algorithm described in class to boost SVMs. Use the kernel from Exercise 1 that had the higher test accuracy. (We will discuss how to do this in class.) Set K (the number of boosting iterations) to 10. Run the ensemble classifier H on the test data. In your report, give the test accuracy of your ensemble classifier H. Write a paragraph commenting on your results, and how they differ (if at all) from your results in Exercise 1. Is the accuracy improved by boosting?

Exercise 3: Increasing the number of boosting iterations.

Repeat Exercise 2, but this time set $K = 20$. Give the same information in your report as was requested in Exercise 2 (test accuracy of H and discussion paragraph).

Conclusion

Write a paragraph summarizing all your results, and any interesting observations you have to make on what you learned about SVMs and boosting and this particular task.

Here is what you need to turn in:

- Your **spell-checked**, **double-spaced** report, with all the information requested above.
- Your code for Adaboost.

How to turn it in:

Send these items in electronic format to mm@cs.pdx.edu by 2pm on the due date. **No hard-copy please!**

If there are any questions on this assignment, don't hesitate to ask me, Max (our TA), or e-mail the class mailing list.

Policy on late homework: If you are having trouble completing the assignment on time for any reason, please see me before the due date to find out if you can get an extension. Any homework turned in late without an extension from me will have 5% of the grade subtracted for each day the assignment is late.