

COMPUTER PROGRAMMING IN JAVA

COLUMBIA UNIVERSITY HIGH SCHOOL SCIENCE HONORS PROGRAM

Machine Learning: Naïve Bayes Assignments 2008 Apr 26 Sat

This week's assignment will not involve much (or any) coding. Instead, you will be using the provided code to perform experiments on a given corpus. Read BOTH tasks first before starting, as one does NOT depend on the other, and it would be nice if the class split itself between the two tasks so that we can share and compare results.

Preparation

Begin by downloading “virus-classifier.zip” from the webpage. From inside CUNIX, execute the following command:

```
wget http://www.cs.columbia.edu/~mwc2110/shp/virus-classifier.zip
```

You will see a new file if you do an `ls -l`. Extract the file as follows:

```
unzip virus-classifier.zip
```

After you expand it, change into the new directory (`cd virus-classifier`), and you will see the following files/directories:

```
NBModel.java  
VirusClassifier.java  
virii/  
non-virii/
```

Now you will need to compile the .java files (`javac *.java`). To start the system, call VirusClassifier as the main class:

```
java VirusClassifier
```

You will be presented with a menu which should be relatively clear.

One extra detail not previously discussed is the “window size”, which is the length of the features being considered. In general, smaller window sizes results in better accuracy because you see more examples of each type.

Task 0

Browse the list of files in the two directories, virii and non-virii:

```
ls -l virii  
ls -l non-virii
```

Jot down the names of a few of these files, remembering whether they are virii or non-virii.

Task 1

First, randomly select 10 random files as your testing set (you can build this set however you like, 5 virus and 5 non-virus, or 1 virus and 9 non-virus, etc).

Then, training the system with 1 virus and 1 non-virus, and see how well it classifies your testing set. Repeat the experiments with 5/5, 10/10, 30/30, and whatever other training sets you consider interesting. Record the accuracy of your experiments, as well as a (rough) estimate of the amount of time taken in training and testing. Also try uneven training sets, like 1/30, 15/30, 22/22, etc.

Remember that the Naïve Bayes algorithm does not “memorize” the training set; that is, even if you trained a model on a given file, there's no guarantee that it will correctly predict the class of that file in testing.

Task 2

Again select 10 random files as your testing set as in Task A.

This time, try varying the window size, and see its effect on classification accuracy. Again keep track of accuracy and performance. Start off with a window size of 1 up to 3.

Optional

The code provided is general enough that it can be used on any dataset. If you find this interesting, speak to me about other experiments you can perform on your own, such as using Naïve Bayes to perform authorship detection or topic categorization.

References

The virus/non-virus dataset was taken from the Columbia CS3134 “Data Structures in Java” course Spring 2007, taught by Professor Shlomo Hershkop.
<http://www.cs.columbia.edu/~sh553/teaching/w3134-s07/>