

Homework #4 (due 3/28/11 - 3 weeks)

There will be four more homework assignments (including this one). The other three assignments will cover internet search, multilingual retrieval, and applications of NLP to IR. These assignments should involve less programming work compared to the first three assignments to give you more time to devote to your independent project. Also, you only need to hand in three of the remaining assignments; you can skip the assignment of your choice.

Text Classification (40 points)

(20 pts) Explain the following concepts and what they mean in terms of text classification: (1) *bias-variance tradeoff*; (2) *kernel trick*; (3) and *cross-validation*.

(20 pts) Compute how a naïve Bayes classifier would classify the following 'documents' using the binomial (or Bernoulli) model. The two classes are 'Good' and 'Spam'. Recall in the binomial model that estimates of $P(\text{word}|\text{class})$ are based on the percentage of documents of the class containing the word. Estimates of $P(\text{class})$ and $P(\text{word}|\text{class})$ are given in the tables below.

Document 1: "free drugs willy"

Words	$P(w \text{Good})$	$P(w \text{Spam})$
baker	0.03	0.025
drugs	0.03	0.15
free	0.01	0.25
willy	0.05	0.005

Document 2: "free willy baker"

$P(\text{Good}) = 0.7$
$P(\text{Spam}) = 0.3$

Binary classification using Reuters 21578 dataset (60 points)

(60 pts) Download from the course web site training documents for three Reuters categories (coffee, ship, and wheat) and build a classifier for each. You can use any method (e.g., kNN, naïve Bayes, decision trees, or SVMs). I suggest using the SVMlight tool which is available from <http://svmlight.joachims.org/> (binaries are available for modern operating systems). To use SVMlight you should process the text files, which are similarly formatted to HWs 1-3, and write out files of vectors, one document vector per line. For example:

```
+1 5:1 13:1 78:1 ... 15008:1
+1 5:1 45:1 78:1 15000:1
-1 3:1 13:1 87:1 12000:1
```

“+1” in the leftmost column indicates that the vector is positive for the class and “-1” indicates it is negative. Each *termid:value* element in this example uses binary weights; ‘1’ indicates the presence of a term in the document, and terms not in the document (i.e., the zeros) are not written out. Having created these vectors, train a classifier for each class and then run the classifier on the test sets. Note: it is very important that the termids are consistent in the training and test data (e.g., 'beverage' gets termid=37 for both the coffee.train and coffee.test documents).

The example above is in the format SVMlight expects. To train a model with SVMlight:

```
% svm_learn coffee.train coffee.mod
```

To run a test set against a trained model:

```
% svm_classify coffee.test coffee.mod coffee.out
```

Using the output predictions (+1/-1; above 0 for SVMlight means the prediction is for belonging to the positive class) compute recall, precision, and F_1 scores for each of the three classes (i.e., coffee, ship, and wheat). Show the work in your computation. Recall is the percentage of +1s in the test file that were correctly predicted to belong to the class; precision is the percentage of +1s in the output file that are correct according to the test file labels. $F_1 = 2 * P * R / (P + R)$.

Briefly describe the methods (e.g., do you remove stopwords or perform stemming; do you use binary weights or TF/IDF weights) and which tools you use. Also hand in any source code that you write for the assignment.

Extra credit (4 pts; 2 pts): Compute and report the macro-average for the three classes (i.e., average of the three F_1 scores). The student with the highest average gets +4 points; second place gets +2 points.