CS838-1 Advanced NLP Homework 4

Due 3/27/2007 in class

Instructor: Jerry Zhu, jerryzhu@cs.wisc.edu

Type your answers and hand in a printed version to the instructor in class on the due date. The homework is worth 50% if it is no later than 48 hours (you may email me a pdf file), and worth nothing after that. I will not accept homeworks to the TA or in the physical mailbox.

Note: This homework lets you explore the Naive Bayes classifier, and cross validation. It is very hands-on: start early.

Question 1 [20 pt]. Prove equations (14, 15) in the naive Bayes lecture notes `http://www.cs.wisc.edu/~jerryzhu/cs838/NB.pdf`, starting from equations (11–13).

Question 2 [10 pt each]. Download the dataset at `http://www.cs.wisc.edu/~cs838-1/dataset/tinySRAA/tinySRAA.tgz`. This dataset contains postings from four discussion groups: real automobile, real aviation, simulated automobile, simulated aviation. Your task is to classify automobile vs. aviation postings, and measure its accuracy with 5-fold cross validation on the dataset.

1. Describe how you set up the class labels y.
2. Describe how you convert the postings into bag-of-word representations, in particular any text processing to the postings, and how you create the vocabulary.
3. Describe how you train a Naive Bayes classifier from a training set, including whether you smooth the parameters.
4. Describe how you perform 5-fold cross validation.
5. List the cross validation accuracy you obtain. Also break it down into accuracies on each fold.
6. Collect the posterior probabilities $p(y = \text{automobile}|x)$, for all postings x. Plot a histogram of the posterior probability. For example, you may show how many falls into each bin of $[0, 0.1], [0.1, 0.2], \ldots, [0.9, 1]$. You may use a different number of bins. Discuss what you observe.
7. Discuss the effect of k in k-fold cross validation, in particular the advantages and disadvantages in having a small or large k, in estimating the future performance of the classifier.

8. Now we want to instead classify real vs. simulation on the same dataset. Describe what you do differently. List the 5-fold cross validation accuracy.