# CS838-1 Advanced NLP Homework 5

Due 4/10/2007 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.

Consider the movie review data collected by Bo Pang and Lillian Lee at `http://www.cs.wisc.edu/~cs838-1/dataset/movie/`. There are three files: a readme file, positive reviews, and negative reviews. The corpus consists of the positive and negative reviews. Do not further process the words. Treat punctuations as words. You should get 20612 word types, and 218346 word tokens.

## 1 Mutual information

Each line in positive.txt and negative.txt is a document. All documents in positive.txt have class label $c = 1$, while those in negative.txt have class label $c = -1$. Let $w$ be a binary random variable so that $w = 1$ if the word $w$ appears in a document, and $w = 0$ otherwise.

**Question 1 [10 pt]**. Write down the formula of mutual information $I(w; c)$.

**Question 2 [10 pt]**. To help you compute mutual information, for each word $w$ you want to collect four numbers: the number of documents in which $w$ occurs and $c = 1$; $w$ doesn't occur and $c = 1$; $w$ occurs and $c = -1$; $w$ doesn't occur and $c = -1$. List these four numbers for the words 'the, good, movie' respectively.

**Question 3 [20 pt]**. Compute $I(w; c)$ for all word types. Make sure you use log base 2 to get bits. What is the mutual information for the words 'the, good, movie' respectively?

**Question 4 [10 pt]**. List the top 10 words with the highest mutual information.

**Question 5 [10 pt]**. Browse the list. List 3 words with 'unexpected' mutual information, and explain why they are unexpected.

For example "I thought this word would be a very good one to distinguish positive and negative reviews, but it actually has very low mutual information." Try to explain why they have those mutual information values.

## 2 SVM

For this question we will use SVM-light, available at `http://svmlight.joachims.org/`. Download the code and study the manual. Use all default parameters (linear kernel). Convert positive.txt and negative.txt into appropriate format. Use the first 4000 lines of positive.txt and the first 4000 lines of negative.txt as training data, and remaining lines in the two files as test data.

**Question 6 [20 pt]**. What is the classification accuracy on the test data?

**Question 7 [20 pt]**. The SVM decision boundary is $W^\top x = b$, where in the dual representation $W = \sum_i \alpha_i y_i x_i$. The sum is over the support vectors. In SVM-light's 'model file', the support vectors are listed one per line ($\alpha_i y_i$ in the first column, then the vector $x_i$). Compute $W$ from your model file. List the top 10 words with the largest weights, and bottom 10 words with the smallest (most negative) weights (Please list both the word and its weight).