CS114 (Spring 2015) Homework 3
Naive Bayes Classifier and Evaluation

Due February 27, 2015

This is not a programming assignment. If you write down your solution by hand, scan it and submit on Latte.

1 Naive Bayes Classifier

We will train the sentiment classifier using Naive Bayes Classifier by hand. The data contain a folder for positive movie reviews and anotherfor negative movie review. The files are already tokenized. Each review is in its own file.

For the following questions, we use the notations provided below. We use three features to classify the label. More formally,

- $X_1 = 1$ if the review containing the word 'great'. 0 otherwise.
- $X_2 = 1$ if the review containing the word 'poor'. 0 otherwise.
- $X_3 = 1$ if the review containing the word 'long'. 0 otherwise.
- $Y = 1$ if the review is labeled positive. 0 if labeled negative.

1. Fit the Naive Bayes classifier by hand. The likelihood $P(X_i|Y)$ should be Bernoulli distribution. Show all of the calculation steps and list all of the probability distributions involved in the model.
(Hint: Unix command `grep` is very handy here. We have a brief tutorial on how to use it at the end of this document)
2. If a review contains the word 'great' and the word 'poor', but not the word 'long', what is the probability that the review is positive according to your Naive Bayes Classifier? Show all of the calculation steps.
3. Compute the mutual information $I(X_i, Y)$ for all three features. Which feature is the best feature according to this metric?

2 Evaluation Metrics

You are given a performance report of a classifier in the form of a confusion matrix:

	predicted A	predicted B	predicted C
true A	5	1	1
true B	1	4	2
true C	0	2	4

1. Compute precision, recall, and F_1 for each class.
2. Compute macroaveraged F_1
3. Compute microaveraged F_1
4. Compute accuracy rate

Some useful UNIX commands for CL

We hope you already know how to use cd, mkdir, tar, and ls. If not, please come see one of us. We will show you how to navigate around UNIX.

- `ls -l *.txt`

  List the files end with `.txt` in this folder

- `wc -l *.txt`

  Count the number of lines (`-l`) of all files that end with `.txt` .

- `ls -l *.txt | wc -l`

  Count the number of lines that `ls` command spits out. (this is called piping.) This essentially counts the number of files that end with `.txt`.

- `grep ' great ' *.txt`

  List all of the lines in the files that end with *.txt that has the word great with space surrounding it.

- `grep -l ' great ' *.txt`

  Like above, but only one hit per file.

- `grep -l ' great ' pos/*.txt | wc -l`

  This is useful for counting words in each movie review. Use this.