CSE250B Homework 2

Qiao Zhang

January 21, 2016

1 Text classification using multinomial Naive Bayes

1.1 Model

1.1.1 baseline

The multinomial Naive Bayes model described in instructions is regarded as the baseline.
The classifier is

$h(x) = argmax_j \; log\pi_j + \sum_{i=1}^{|V|} x_i logp_{ji}$

$p_{jw} = \frac{number\ of\ words\ w\ in\ j\ class\ +\ 1*\alpha}{number\ of\ all\ words\ in\ j\ class\ +\ |V|*\alpha}$ (with Laplacian smoothing $\alpha = 1$)

The test error rate of this simple model is approximately 0.2189.

1.1.2 removing stopwords

In order to improve the earlier model, we remove stopwords from vocabulary thus reducing its size. The other parts of the baseline model still apply.

1.1.3 logrithm frequency

Another approach for improvement is to replace the frequency f of a word in a document by log(1+f). Here f refers to the counts of a word in a document.

1.2 Performance

In order to decide between options, we randomly split the training data into a smaller training set and a validation set with proportion of 8:2.
We choose error rate as the measurement of performance.

Error rate = $\frac{wrongly\ classified}{correctly\ classified + wrongly\ classified}$

The result of performance is reported in the following table:

error rate	baseline	remove stopwords	log frequency
validation error	0.1589	0.1447	0.1793
test error	0.2404	0.2195	0.2630

Table 1: error rate

We can see from the table that the removing stopwords model has the best performance on validation set. Since we usually do not have access to test data, should be our final model. Its error rate on the test set is 0.2195, which is also the best among all models provided above. The performance of these models can be promoted by tuning parameter α. As a result of lack of time, α is set to one in all the experiments.

2 Classification with an abstain option

The classifier should be

$$h(x) = \begin{cases} 0 \quad if\ \eta(x) < \theta \\ 1 \quad if\ \eta(x) \geq 1-\theta \\ abstain \quad if\ \theta \leq \eta(x) < 1-\theta \end{cases}$$