Homework 3 for LING 572

Haotian He

January 25 2014

Collaborators: Haotian He and Jonggun Park

1 Q1

```
The training accuracy is 0.9444444444444444.
The test accuracy is 0.8966666666666666.
```

2 Q3

Table 1: Results of my Bernoulli NB model

cond_prob_delta	Training accuracy	Test accuracy
0.1	0.9603262884686689	0.9158249158249159
0.5	0.9547645532072674	0.898989898989899
1.0	0.9510567296996663	0.8922558922558923
2.0	0.9447534297367445	0.8787878787878788

3 Q4

Table 2: Results of my multinomial NB model

cond_prob_delta	Training accuracy	Test accuracy
0.1	0.9581015943641082	0.9090909090909091
0.5	0.9529106414534668	0.9023569023569024
1.0	0.9480904708935854	0.898989898989899
2.0	0.9406748238783834	0.8922558922558923

4 Q5

Table 3: Results of my multinomial NB model with binary features

cond_prob_delta	Training accuracy	Test accuracy
0.1	0.9595847237671487	0.9124579124579124
0.5	0.9562476826103078	0.9057239057239057
1.0	0.9529106414534668	0.898989898989899
2.0	0.9477196885428254	0.898989898989899

5 Q6

(a):
1. All these three tables show, Bernoulli NB model, multinomial NB model, and multinomial NB model with binary features share the same trend that both the training and test accuracies go down, corresponding to the increase of the conditional probability delta value.
2. The accuracies of Bernoulli NB model go down more than multinomial NB model and multinomial NB model with binary features. 3. When the conditional probability delta is 0.1, the test accuracy of Bernoulli NB model is the highest, whereas when the conditional probability delta is 2.0, that is the lowest.
4. Particularly the performance of the Bernoulli NB model is better than that of the multinomial NB model, and that means those information in the multinomial NB model seems irrelevant and actually reduce its accuracy, while the features in Bernoulli NB model are more relevant.

(b):
In this case, a feature is really important for the Bernoulli NB model, because in the model file, as we could find, the standard deviation of one feature given different class labels is really big, that means the features are more relevant to the class labels.

(c):
In this multinomial NB model, a feature is not important, because in its model file, as we could find, the standard deviation of one feature given different class labels is small, compared with Bernoulli NB model. Thus, the features are less relevant to the class labels.

(d):
The Bernoulli NB model should run faster than the multinomial one, because

their complexities are different. The Bernoulli NB model one is $O(|D| * |V|)$, while the multinomial one is $O(|V| * |C| + |D| * avg(feat/doc))$.