CS573 - Lab1 Report

Phuong Nguyen

February 2020

1 Learn the Naive Bayes Model

$P_{MLE}(w_k|\omega_j)$ has a lot of zero values because out of all words in the vocabulary, very few appear in one document.

On the other hand, $P_{BE}(w_k|\omega_j)$ has no zero values because we have smoothed it with Laplace estimate. These values are pretty small though.

There are many words only appear on the test data (do not appear on the train data).

2 Evaluate the Performance of the Classifier

We choose the formula:

$$w_{NB} = argmax_{w_j} P(w_j) \prod_{k=1}^{|Vocabulary|} P(w_k|w_j)^{N_k}$$

and then compute using logarithm:

$$w_{NB} = argmax_{w_j} \left[\ln P(w_j) + \sum_{k=1}^{|Vocabulary|} N_k \cdot \ln P(w_k|w_j) \right]$$

3 Performance on Testing Data

The results obtained using the Bayesian estimator on training data are practically the same. But the results obtained using the Bayesian estimator on testing data are much better than the Maximum Likelihood estimator (78.11% vs 9.46%). The reason is because of the if any value fails to occur in the training data, Maximum Likelihood estimate for the corresponding probability will be zero. On the other hand, even with uniform prior, Bayesian estimate for this same probability will be non-zero. Probability estimates of zero can have very bad effects on just about any learning algorithm and we only want zero probability estimates when non-occurrence of an event is justified by prior belief (in such case, the sample size has to be very large).

The Maximum Likelihood estimator performs extremely well on training data (99%) because it literally counts all words it have seen and excludes all words it has not seen before. This will be the problem later on the testing data because there are many words on testing set that are not on the training set.

Therefore, we should prefer Bayesian method, which represents uncertainty about unknown parameters.