

## ComS 573 Machine Learning

### Lab Assignment 1

# Naive Bayes Classifier for Text Classification

Naive Bayes classifiers are among the most successful known algorithms for learning to classify text documents. In this lab assignment, you will implement in Java a Naive Bayes algorithm to solve the “20 Newsgroups” classification problem.

## 1 Data Set

The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. It was originally collected by Ken Lang, probably for his *Newsweeder: Learning to filter netnews*[1] paper, though he does not explicitly mention this collection. The 20 newsgroups collection has become a popular data set for experiments in text applications of machine learning techniques, such as text classification and text clustering.

The data is organized into 20 different newsgroups, each corresponding to a different topic. Here is a list of the 20 newsgroups:

<code>alt.atheism</code>	<code>sci.crypt</code>
<code>comp.graphics</code>	<code>sci.electronics</code>
<code>comp.os.ms-windows.misc</code>	<code>sci.med</code>
<code>comp.sys.ibm.pc.hardware</code>	<code>sci.space</code>
<code>comp.sys.mac.hardware</code>	<code>soc.religion.christian</code>
<code>comp.windows.x</code>	<code>talk.politics.guns</code>
<code>misc.forsale</code>	<code>talk.politics.mideast</code>
<code>rec.autos</code>	<code>talk.politics.misc</code>
<code>rec.motorcycles</code>	<code>talk.religion.misc</code>
<code>rec.sport.baseball</code>	
<code>rec.sport.hockey</code>	

The original data set can be downloaded from <http://qwone.com/~jason/20Newsgroups/>. In this lab, you won't need to process the original data set. Instead, a processed version of the data set is provided (see `20newsgroups.zip`). This processed version represents 18824 documents which have been divided to two subsets: training (11269 documents) and testing (7505 documents). After unzipping the file, you will find six files: `map.csv`, `train_label.csv`, `train_data.csv`, `test_label.csv`, `test_data.csv`, `vocabulary.txt`. The `vocabulary.txt` contains all distinct words and other tokens in the 18824 documents. `train_data.csv` and `test_data.csv` are formatted "`docIdx`, `wordIdx`, `count`", where `docIdx` is the document id, `wordIdx` represents the word id (in correspondence to `vocabulary.txt`) and `count` is the frequency of the word in the document. `train_label.csv` and `test_label.csv` are simply a list of label id's indicating which newsgroup each document belongs to. The `map.csv` maps from label id's to label names.

## 2 What You Will Do

In general, you will implement a Java program that takes the six files as input, builds a Naive Bayes classifier and outputs relevant statistics. You should call your code as follows:

```
java NaiveBayes vocabulary.txt map.csv training_label.csv training_data.csv
testing_label.csv testing_data.csv
```

(Note: you may use a different name for the Java class.)

You will learn your Naive Bayes classifier from the training data (`train_label.csv`, `train_data.csv`), then evaluate its performance on the testing data (`test_label.csv`, `test_data.csv`). Specifically, your program will accomplish the following two tasks.

### 2.1 Learn Naive Bayes Model

You will implement the *multinomial model* (“a bag of words” model) discussed in the lecture. In the learning phase, you will estimate the required probability terms using the training data.

For each target value  $\omega_j$  (each newsgroup)

- Calculate class prior  $P(\omega_i)$
- Calculate  $n$ : total number of words in all documents in class  $\omega_j$
- For each word  $w_k$  in Vocabulary
  - Calculate  $n_k$ : number of times word  $w_k$  occurs in all documents in class  $\omega_j$ .
  - Calculate Maximum Likelihood estimator  $P_{MLE}(w_k|\omega_j) = \frac{n_k}{n}$  and Bayesian estimator  $P_{BE}(w_k|\omega_j) = \frac{n_k+1}{n+|Vocabulary|}$  (this is Laplace estimate).

Your program should output the class priors, for example

```
P(Omega = 1) = 0.0426
P(Omega = 2) = 0.0516
P(Omega = 3) = 0.0508
...
```

You need not print out all  $P_{MLE}(w_k|\omega_j)$  or  $P_{BE}(w_k|\omega_j)$ , just take a look at the values of these two types of estimators. What do you observe? Discuss what you have observed in your report.

### 2.2 Evaluate the Performance of the Classifier

In this task, you will evaluate your Naive Bayes classifier on both the training and the testing data. You will use your Naive Bayes classifier to make classification decision on these data set then calculate relevant statistics such as overall accuracy, class accuracy, confusion matrix, etc. When make classification, considering all word positions in a document that contain tokens found in Vocabulary, return  $\omega_{NB}$  where

$$\omega_{NB} = \underset{\omega_j}{\operatorname{argmax}} P(\omega_j) \prod_{i \text{ in positions}} P(x_i|\omega_j). \quad (1)$$

You may find it more convenient to compute

$$\omega_{NB} = \underset{\omega_j}{\operatorname{argmax}} [\ln P(\omega_j) + \sum_{i \text{ in positions}} \ln P(x_i|\omega_j)]. \quad (2)$$

### 2.2.1 Performance on Training Data

First you will evaluate your Naive Bayes classifier on the training data. You will only make classification using your Bayesian estimators  $P_{BE}(w_k|\omega_j)$ . After making prediction for each document in the training data set, you will calculate and output the *overall accuracy*, i.e, the percentage of correctly classified documents

$$\text{Overall Accuracy} = \frac{\text{number of correctly classified documents}}{\text{total number of documents}}, \quad (3)$$

as well as the *class accuracy* for each newsgroup. A sample output is given as follows.

```
Overall Accuracy = 0.9211
Class Accuracy:
Group 1:  0.9667
Group 2:  0.9191
...
```

Further, you will construct and output the confusion matrix ([http://en.wikipedia.org/wiki/Confusion\\_matrix](http://en.wikipedia.org/wiki/Confusion_matrix)). An example is showed as follows:

```
464    0    0    0    ...
1   534    6   15    ...
1    10   503   23    ...
0    10    4   546    ...
...
```

A cell  $(i, j)$  in the matrix represents the number of documents in group  $i$  that are predicted to be in group  $j$ .

### 2.2.2 Performance on Testing Data

Now you will evaluate your classifiers on the testing data set. First, repeat the experiments described in Section 2.2.1 on testing dataset. Compare the results obtained with the results you have obtained in Section 2.2.1. What do you observe? Discuss.

Next repeat the experiments (on testing data) using your Maximum Likelihood estimator  $P_{MLE}(w_k|\omega_j)$  instead of the Bayesian estimators. Compare your results to the results obtained using your Bayesian estimators. Can you observe the difference? Which one is better?

## 2.3 Summary

In summary, your program will output the following relevant statistics:

- Class priors
- Performance on training data (using Bayesian estimators): overall accuracy, class accuracy, confusion matrix.
- Performance on testing data (using both MLE and BE estimators): overall accuracy, class accuracy, confusion matrix for MLE, overall accuracy, class accuracy, confusion matrix for BE.

### 3 Where to Compile

The TA will grade your work on `pyrite`, which has the `Java SDK 1.8.0` installed. You should be able to use standard `javac` and `java` comments to compile and execute your program. Make sure that your program compiles well on `pyrite` before you submit it.

### 4 What to Turn In

Turn in via Blackboard a compressed file (`.zip` `.rar` or `.tar.gz`) containing the following:

- All of your commented source code (`.java` files). Make sure the debugging flag is set to false. **Do not turn in the results of running the code with `debugging = true`.**
- A `README` file explaining how to compile and run the program.
- A short lab report that includes answers to questions listed above and the experimental results obtained.

### References

- [1] Ken Lang, Newsweeder: Learning to filter netnews, Proceedings of the Twelfth International Conference on Machine Learning, 331-339 (1995).