Laboratory 2

Santiago Álvarez Sepúlveda, *saalvarezse@unal.edu.co, Cód: 25481031*
Universidad Nacional de Colombia, Sede Bogotá

***Abstract*—The following laboratory report contains the development of the first two problems of the second laboratory of the course Machine Learning at National University of Colombia. The first problem is to train two types of Naive Bayes classifier, the first is applying a Multinomial probabilistic model, and the other one applying the Multivariate Bernoulli model, to classify documents that belong or not to the class China. Although both methods are different in essence, both gave the same result, the message to classify doesn't belong to China. On the second problem there is an optimization of the complexity of the algorithm ApplyMultinomialNB, by estimating the score of the document for each class using the histogram of the document, avoiding us to execute repeated calculations, but instead abbreviating with the coefficients of the histogram.**

***Palabras Claves*—Naive Bayes Classifier, Multinomial Model, Multivariate Bernoulli Model, Algorithmic Complexity.**

I. Problem 1

Based on the data in Table 13.10 from [1] (i) estimate a multinomial Naive Bayes classifier, (ii) apply the classifier to the test document, (iii) estimate a Bernoulli NB classifier, (iv) apply the classifier to the test document.

A. Multinomial Naive Bayes Classifier

The table 13.10 from [1] gives us the following information:

	docID	words in document	in $c =$ China
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

It shows a set of documents that have been previously classified as related to China or not related to China. The purpose of the multinomial Naive Bayes classifier is to find out, from the previous knowledge of the words contained in each class of document, whether a new document belongs to class China or not. What we expect to get at the output of the classifier is the conditional probability of a given document d belongs to the class c, or in mathematical language $P(c|d)$.

Applying the Bayes theorem to this problem, in terms of the literal content of the document, we must know the probability of each word in the document (t_k) to belong also to a document of a given class, in other words, the probability that a document that contains certain words t_k belongs to the class c. This can be computed with the following equation:

$$P(c|d) = P(c) \prod_{n=1} P(t_k|c)$$

That way, the class with the maximum score of conditional probability will be the class to which the document belongs.

And in order to calculate the values of $P(c)$ and $P(t_k|c)$, we apply the *maximum likelyhood estimate*, or estimate the relative frequencies of each class with respect to all of the possible classes, and from the words of the document with respect to each of the words in the documents of class c. In mathematical language:

$$P(c) = \frac{N_c}{N} \tag{2}$$

Where N is the number of documents to train the classifier and N_C is the number of documents that are known to belong to the class c.

Class	$P(c)$
c	$\frac{2}{4} = \frac{1}{2}$
$\bar{c}$	$\frac{2}{4} = \frac{1}{2}$

$$P(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B'} \tag{3}$$

Where T_{ct} is the number of times the word t appears in the whole documents of class c and the denominator is the total number of words included in class c. A Laplacian filter is applied to avoid singularities that generate errors in the total probability, so we add 1 to all the numerators, in case the word is not in any document of a given class (it would cancel the whole probability), and in the denominator we add the term $B' = |V|$ which is the number of words in the vocabulary in the case there is a class without any number of words.

Vocabulary	T_{ct}	$T_{\bar{c}t}$	$P(t_k\|c)$	$P(t_k\|\bar{c})$
Taipei	1	0	$\frac{1+1}{5+7} = \frac{1}{6}$	$\frac{0+1}{5+7} = \frac{1}{12}$
Taiwan	2	1	$\frac{2+1}{5+7} = \frac{1}{4}$	$\frac{1+1}{5+7} = \frac{1}{6}$
Macao	1	0	$\frac{1+1}{5+7} = \frac{1}{6}$	$\frac{0+1}{5+7} = \frac{1}{12}$
Shanghai	1	0	$\frac{1+1}{5+7} = \frac{1}{6}$	$\frac{0+1}{5+7} = \frac{1}{12}$
Japan	0	1	$\frac{0+1}{5+7} = \frac{1}{12}$	$\frac{1+1}{5+7} = \frac{1}{6}$
Sapporo	0	2	$\frac{0+1}{5+7} = \frac{1}{12}$	$\frac{2+1}{5+7} = \frac{1}{4}$
Osaka	0	1	$\frac{0+1}{5+7} = \frac{1}{12}$	$\frac{1+1}{5+7} = \frac{1}{6}$
Totals	5	5		

This resulting matrix contains all the information needed to train the classifier for each word in a document, now we can

use this information to estimate the probability of a document to belong to a class given its content.

B. Apply Multinomial Naive Bayes Classifier

From equation [1] and table [2] we can estimate the conditional probability of a document to belong to a given class. In this case we have the following content for the document:

$$P(c|d) = \frac{1}{2}\frac{1}{4}\frac{1}{4}\frac{1}{12} = \frac{1}{384} \tag{4}$$

$$P(\bar{c}|d) = \frac{1}{2}\frac{1}{6}\frac{1}{6}\frac{1}{4} = \frac{1}{288} \tag{5}$$

After applying the classifier, we find out that the document should not belong to the class China.

C. Multivariate Bernoulli Naive Bayes Classifier

The multivariate Bernoulli Naive Bayes classifier has a different way to predict if a given document belongs to a certain class. The Bernoulli model generates a binary (Boolean) indicator (vector) for each term of the vocabulary, each component x_k of the vector $\vec{x_d} = (x_1, ..., x_M)$ indicates the presence ("1") or absence ("0") of the word x_k of the vocabulary V in the document d.

Vocabulary	$\vec{x_1}$	$\vec{x_2}$	$\vec{x_3}$	$\vec{x_4}$
Taipei	1	0	0	0
Taiwan	1	1	0	1
Macao	0	1	0	0
Shanghai	0	1	0	0
Japan	0	0	1	0
Sapporo	0	0	1	1
Osaka	0	0	0	1

With each of this vectors, the Bernoulli model to estimate the conditional probability of a document to belong to a certain class is the following:

$$P(t_k|c) = \frac{N_{ct} + 1}{N_c + 2} \tag{6}$$

Where N_{ct} is the number of documents in the class c that contain the word t and N_c the number of documents of class c, for this case $N_c = 2$ and $N_{\bar{c}} = 2$.

From the vector form of each document, and the function to estimate the conditional probability by the model of Bernoulli, we get the following table:

Vocabulary	N_{ct}	$N_{\bar{c}t}$	$P(t_k\|c)$	$P(t_c\|\bar{c})$
Taipei	1	0	$\frac{1+1}{2+2} = \frac{1}{2}$	$\frac{0+1}{2+2} = \frac{1}{4}$
Taiwan	2	1	$\frac{2+1}{2+2} = \frac{3}{4}$	$\frac{1+1}{2+2} = \frac{1}{2}$
Macao	1	0	$\frac{1+1}{2+2} = \frac{1}{2}$	$\frac{0+1}{2+2} = \frac{1}{4}$
Shanghai	1	0	$\frac{1+1}{2+2} = \frac{1}{2}$	$\frac{0+1}{2+2} = \frac{1}{4}$
Japan	0	1	$\frac{0+1}{2+2} = \frac{1}{4}$	$\frac{1+1}{2+2} = \frac{1}{2}$
Sapporo	0	2	$\frac{0+1}{2+2} = \frac{1}{4}$	$\frac{2+1}{2+2} = \frac{3}{4}$
Osaka	0	1	$\frac{0+1}{2+2} = \frac{1}{4}$	$\frac{1+1}{2+2} = \frac{1}{2}$

D. Apply Multivariate Bernoulli Naive Bayes Classifier

From equation [1] and the previous table we can estimate the conditional probability of a document to belong to a given class:

$$P(c|d) = \frac{1}{2}\frac{3}{4}\frac{3}{4}\frac{1}{4} = \frac{9}{128} \tag{7}$$

$$P(\bar{c}|d) = \frac{1}{2}\frac{1}{2}\frac{1}{2}\frac{3}{4} = \frac{3}{32} \tag{8}$$

After applying the classifier, we find out again that the document should not belong to the class China.

II. Problem 2

Exercise 13.6 from [1]

The function **ApplyMultinomialNB** in the following figure has time complexity $\Theta(L_a + |\mathbb{C}|L_a)$. How would you modify the function so that its time complexity is $\Theta(L_a + |\mathbb{C}|M_a)$?

Algorithm 1 Apply Multinomial NB Algorithm

```
1: procedure APPLYMULTINOMIALNB(C, V, P, cP, d)
2:     W ← ExtractTokensFromDoc(V, d)
3:     for c ∈ C do
4:         score[c] ← log(P[c])
5:         for t ∈ W do
6:             score[c] += log(cP[t][c])
7:     return argmax_{c∈C} score[c]
```

In order to change the complexity of this algorithm, we see that we are adding multiple times the repeated words to the total score of each class, this means that the algorithm is adding a term to the score for every word in the document. But if instead, we first get the histogram of the document (W', ω) and for every class c we estimate the score of each word t in the subset W' and multiply it by the number of times it appears in the document ω, we get the complexity of the algorithm to depend on the size of the vocabulary included in the document, more than on the numbers of words this document has. This is implemented in the following figure:

Algorithm 2 Apply Multinomial NB Algorithm Using Histogram

```
1: procedure NEWAPPLYMULTINOMIALNB(C, V, P, cP, d)
2:     W', ω ← ExtractHistogramOfTokensFromDoc(V, d)
3:     for c ∈ C do
4:         score[c] ← log(P[c])
5:         for t ∈ W' do
6:             score[c] += ω[t] log(cP[t][c])
7:     return argmax_{c∈C} score[c]
```

This improves the complexity of the algorithm because it makes it be a function of $\Theta(L_a + |\mathbb{C}|M_a)$ where M_a is the length of the vocabulary, L_a the length of the document and

$|\mathbb{C}|$ the number of classes. This is better when a very long document has to be classified, because many words will be repeated and the length of the document will be bigger than the length of the vocabulary, and in that case it will be better to use the **newApplyNaiveBayesClassifier**.

REFERENCIAS

[1] Naive Bayes Text Classification – C.D. Manning, P. Raghavan and H. Schuetze (2008). Introduction to Information Retrieval. Cambridge University Press, pp. 234-265 – nlp.stanford.edu