Experimental Report

Group 08

In this experiment we want to investigate the performance of different kinds of Naïve Bayes classifiers in spam filtering case.

1 Data

1.1 Initial data

We use csv file "SMSSpamCollection.csv" as initial data which contains 5574 messages with labels "ham" or "spam".

Table 1 Initial SMS data

No.	type	text
1	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2	ham	Ok lar... Joking wif u oni..
3	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
...	...	...
5573	ham	The guy did some bitching but I acted like i'd be interested in buying something else next week and he gave it to us for free
5574	ham	Rofl. Its true to its name

1.2 N-grams vocabulary

In order to apply the Naïve Bayes classifiers, we need to construct the n-grams vocabulary for the messages data. n is the number of words in each element in n-grams sequences and we use n-grams model to convert each text to n-grams words sequence here. For example, we take second message in *Table 1* to build 1-grams sequence and 2-grams sequence.

Table 2 example of n-grams sequence

initial text	Ok lar... Joking wif u oni..
1-grams sequence	**Ok;** lar; **Joking;** wif; **u;** oni;
2-grams sequence	**Ok lar;** lar Joking; **Joking wif;** wif u; **u oni;**

The frequencies of n-grams in sequences for every n-gram in vocabulary are the input value and the types of texts are the output value of Naïve Bayes classifiers.

2 Algorithm

In this experiment we use the algorithms based on different kinds of Naïve Bayes classifiers. And in each algorithm the functions fit(), predict() and evaluate() are implemented.

3 Experimental Setup

3.1 Choose Classifiers

- **Bernoulli Naïve Bayes with 1-grams. (BNB1)**

In this classifier 1-grams model is used and the feature vectors $x = (x_1..., x_N)$ (N is the number of elements in vocabulary.) consist of x_i with value 1 (occurs in 1-grams sequence) or 0 (otherwise).

- **Multinomial Naïve Bayes with 1-grams. (MNB1)**

In this classifier 1-grams model is used and the feature vectors $x = (x_1..., x_N)$ consist of x_i with natural number value which indicate the frequency that the 1-grams occurs in the 1-gram sequence.

- **Bernoulli Naïve Bayes with 2-grams. (BNB2)**

In this classifier 2-grams model is used and the feature vectors $x = (x_1..., x_N)$ consist of x_i with value 1 or 0.

- **Multinomial Naïve Bayes with 2-grams. (MNB2)**

In this classifier 2-grams model is used and the feature vectors $x = (x_1..., x_N)$ consist of x_i with natural number value.

- **Trivial classifier. (TC)**

In this classifier “spam” will always be returned as answer.

- **Random classifier. (RC)**

In this classifier “ham” and “spam” each will have 50% probability to be randomly returned as answer and it is independent with the input value.

3.2 Fit function

For Bernoulli Naïve Bayes classifier, the likelihood of x_i from $x = (x_1..., x_N)$ given label y_k is

$$p(x_i \mid y_k) = p_{ki}^{x_i} \cdot (1 - p_{ki})^{(1-x_i)}.$$

And the likelihood of x given label y_k is

$$p(x \mid y_k) = \prod_{i=1}^{N} p_{ki}^{x_i} \cdot (1 - p_{ki})^{(1-x_i)}.$$

For Multinomial Naïve Bayes classifier, the likelihood of $x = (x_1..., x_N)$ given label y_k is

$$p(x \mid y_k) = \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_i p_{ki}^{x_i}.$$

In order to avoid Zero-Count Problem, we implement Laplace smoothing to estimate likelihood $\hat{p}_{ki}$ instead of $p_{ki} = \frac{m_{ki}}{m_k}$:

$$\hat{p}_{ki} = \frac{m_{ki} + \alpha}{m_k + \alpha N} \quad \alpha \geq 0.$$

3.3 Predict function

For a new input value, we need predict function predict() to predict output value and classify the message.

Based on the Formulas we have

$$P(y \mid x) \propto P(y) \prod_{i=1}^{m} p(x_i \mid y).$$

Then we have the predict function for Bernoulli Naïve Bayes classifier

$$f(x) = \arg\max_k P(y_k) \prod_{i=1}^{m} p(x_i \mid y_k)$$

and the predict function for Multinomial Naïve Bayes classifier

$$f(x) = \arg\max_k P(y_k) \prod_{i=1}^{m} \hat{p}_{ki}^{x_i} .$$

And for Trivial classifier and Ramdom classifier the output value have been raised in 3.1.

3.4 Evaluation function

After step 3.3 we gain the result for test data with 4 possible situations:

- Ham with prediction ham (TP)
- Ham with prediction spam (FN)
- Spam with prediction ham (FP)
- Spam with prediction spam (TN)

Then we evaluate the Accuracy, Precision and Recall for each classifier:

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN},$$

$$precision = \frac{TP}{TP + FP},$$

$$recall = \frac{TP}{TP + FN}.$$

3.5 Train-test split

The number of data is 5574. We choose the first 4500 data to train the classifier and the rest are used to test the classifier.

4 Result and Discussion

4.1 Experiment result

Table 3 and Figure 4 show the accuracy, precision and recall value for each classifier.

Table 3 Performance of different classifiers

	BNB1	MNB1	BNB2	MNB2	TC	RC
Accuracy	**0,9590**	0,9450	0,9580	0,9431	0,8685	0,5299
Precision	0,7163	0,6099	**0,7305**	0,6099	0,0000	0,5035
Recall	**0,9619**	0,9556	0,9364	0,9348	0,0000	0,1406

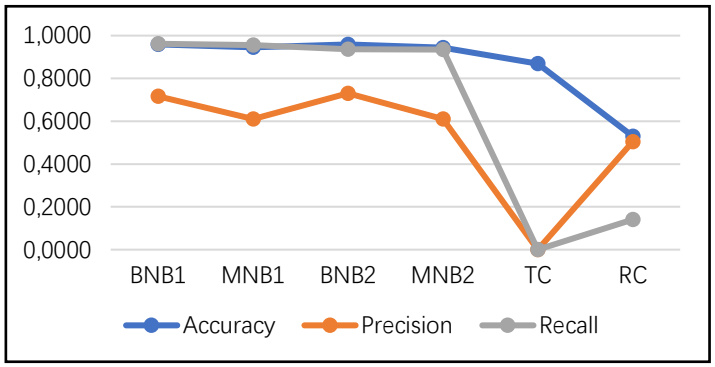

Figure 4 Performance of different classifiers

4.2 Discussion

- The diagrams show that Bernoulli Naïve Bayes with 1-grams has the best performance in accuracy and Bernoulli Naïve Bayes with 2-grams has the best performance in precision. But the accuracies and recalls of the first four classifiers have no great differences.
- Bernoulli Naïve Bayes have better performance than Multinomial Naïve Bayes in precision.
- Comparing to the Trivial classifier and the Random classifier, four Naïve Bayes classifiers have significant effect on all three indexes.
- In spam filtering case, it cost great loss to classify a ham into spam. Therefore, the FN value should be relatively small, simultaneously the recall value will be large according to the formula of recall value. This is to say, that recall value is much more important than the other two indexes value. Thus, the Bernoulli Naïve Bayes with 1-grams proves to be most suitable for spam filtering in our experiment.

5 Conclusion

We show that all Naïve Bayes classifiers have significant effect and the **Bernoulli Naïve Bayes with 1-grams** owns the best performance in spam filtering due to its highest recall value.