Machine Learning Report 2-Classification: Spam Filtering

NCKU Statistics
Grade:107
ID:H24034019
Name:陳育婷

1. Introduction

We focus on the problem with text classification that are closely related to our life. Email is an important tool for communication in our daily life. However; we often receive a lot of spams, which is really annoying and dangerous; since many of them may contain virus. Therefore; it's very important to find a way to filter and classify spams. There are several classification techniques which can help us dealing with this problem. All of these widely used classifiers are based on solid statistical theory.

2. Data Description and Transformation

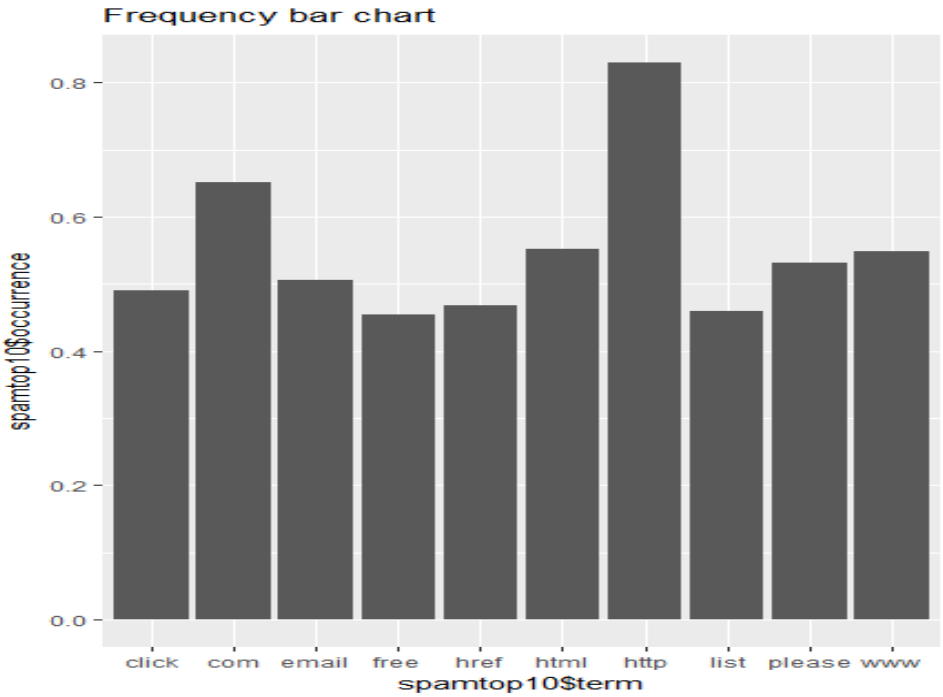
This raw data set comes from the SpamAssassin public corpus. The dataset we used is only a portion of the raw data, which will further be analyzed in this report.

The data contains 2500 easy hams and 500 spams which are used as training set. The test data contains 248 hard hams. Hard hams are still hams, but they have features similar to spams. The test sets also contain 1400 spams and 1400 easy hams.

First, we need to transform our raw text data into a set of features that describe qualitative concepts in a quantitative way. We will need a strategy for turning the text in our email to numbers. When working with text, historically the most important type of feature that has been used is word count. If we think that the text of "html" and "table" are strong indicators of whether an email is spam, we can pick terms like "html" and "table" and how often they occur in one type of document versus the others.

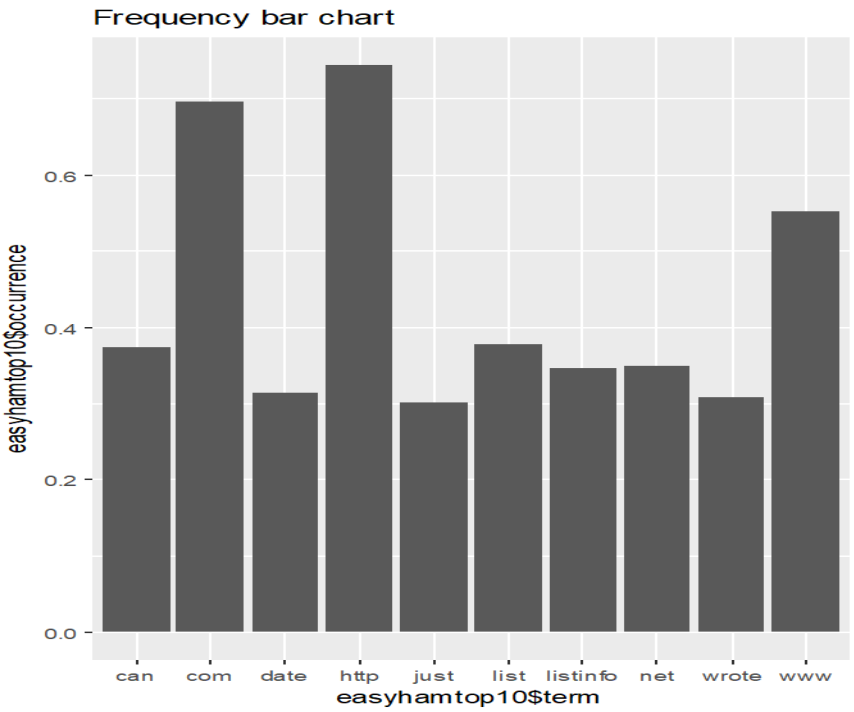
We first transformed our data into the number of times of each term appears in our mails. Then we find the most frequently appeared terms of both easy ham and spam. We can visualize the occurrence of most common terms by graphing bar plot of easy ham data and spam data.

Frequency Bar Chart of Spams:



Above is the bar plot of ten terms which has the highest occurrence-the conditional probability of a message being spam based on how many messages contain the term. We can see that the term with highest occurrence is “click”, which is not surprising at all for email users like you and me, since many spams contain links to ads and encourage you to open them. Next, we plot the bar plot of ten highest occurrence terms in easy hams:

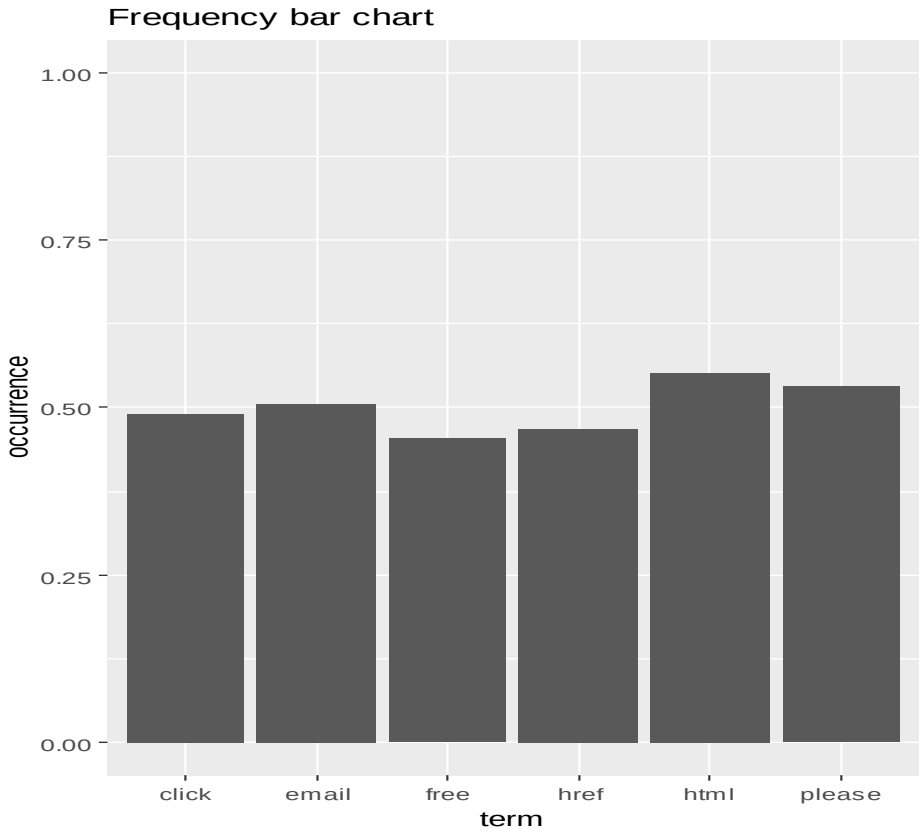
Frequency Bar Chart of Easy Hams:



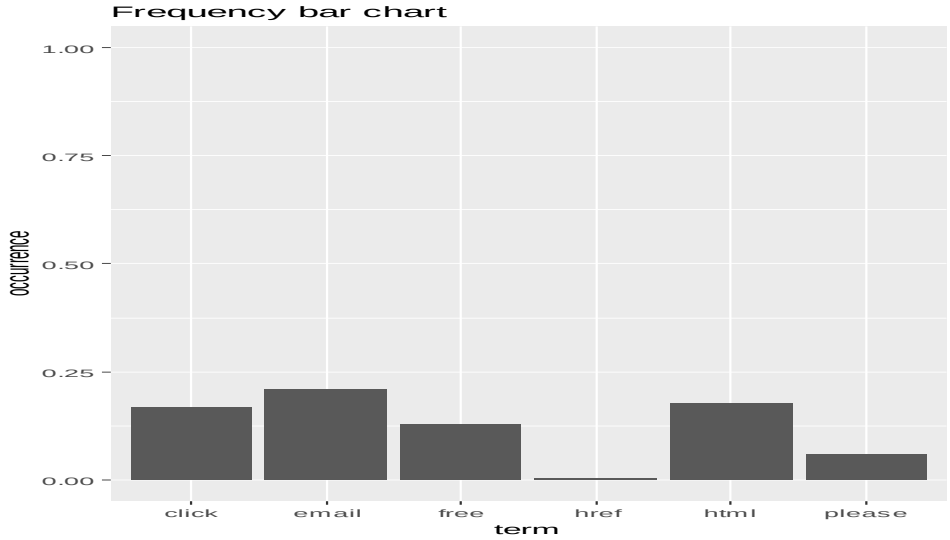
From this plot, we can see that the term with highest occurrence is “can”. The term “com” appears a lot in the easy hams, too, so this term is a weak classification index, we will not use this term.

To choose which term should be used as our predictors, I utilize the terms shown in the previous plots. First, I deleted the intersect terms, which are “com”, “http”, “list”, “www”. Now I have six most frequently appeared terms both in easy hams and spams. I chose to use the remaining six highest occurrence features in spams, since the goal for us is detecting spams rather than hams. To make sure they are good classification features, I plotted the bar plot of these six features’ occurrence both in spams and in easy hams training data set:

Frequency Bar Chart of Spams:



Frequency Bar Chart of Easy Hams:



We can see that the occurrence of these six terms in easy hams are relatively low (under 25%,"href" 's occurrence is even near zero), comparing to the occurrence in spams data set. The result implies that it's adequate to use these six terms as our variables. We expect that the frequency of these terms is low in easy hams, moderate in hard hams, and high in spams. We can confirm that by checking a small part of our test set concluding 1400 spams, 1400 easy hams, and 248 hard hams.

Frequency of First Six Messages in Spams:

	html	please	email	click	href	free
1	0	2	4	1	0	3
2	2	4	3	0	4	2
3	2	4	3	1	5	4
4	2	4	3	1	5	4
5	2	6	2	2	2	0
6	3	0	1	0	0	0

By observing the first six rows of spams, we can find out that spams have high probability of observing these six terms several times in each mail.

Frequency of First Six Messages in Easy Hams:

	html	please	email	click	href	free
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0

By observing the first six rows of easy hams, we can make inference that few of them contain these six terms.

	html	please	email	click	href	free
1	0	0	2	0	0	2
2	0	1	3	2	0	4
3	0	1	0	0	0	0
4	1	2	0	0	0	2
5	0	3	2	0	0	4
6	3	0	0	0	0	0

By observing the first six rows of hard hams, it is likely that some of the mails contain these six terms, the number of counts is less than spams but higher than easy hams.

3. Performing Logistic, LDA, QDA, KNN

When we construct our classifier, we will assume that each message has an equal probability of being ham or spam. As such, it is good practice to ensure that our training data reflects our assumptions. We only have 500 spam messages, so we will limit on ham training set to 500 messages as well. We use 1000 messages in our training set, half of them are spams; the other half are hams.

i. Logistic Regression

We first model the relationship between the type of email and the counts of six selected terms using logistic regression. We use the number of counts on “html”,”please”,”email”,”click”,”href”,”free” as predictors, the response is binary, either equals spam(1) or ham(0) .

Result is obtained as follows:

	Estimate	Std. Error	z value	p value
Intercept	-1.441	0.112	-12.76	<0.0001
html	0.212	0.078	2.71	0.006
please	2.193	0.227	9.67	<0.0001
email	-0.024	0.080	-0.30	0.76
click	0.825	0.158	5.23	<0.0001
href	0.267	0.099	2.69	0.007
free	0.375	0.099	3.79	0.0002

Null Deviance: 1386.3 df:999	Residual Deviance:880.4 df:993

Almost all the variables are significant, and the coefficients are all positive except for the email term, it is also the only feature whose coefficient is not significant. It makes sense, recalling that the bar chart of that term has relative low occurrence in spams but relative high occurrence in easy hams. However; overall we can conclude that the possibility of being classified as spams becomes higher as the number of counts on these six terms increase. The difference in deviance is 505.9 with 6 degree of freedom, which indicates that coefficients other than intercept aren't zeroes.

ii. Linear Discriminant Analysis

The linear discriminant analysis result is obtained as follows:

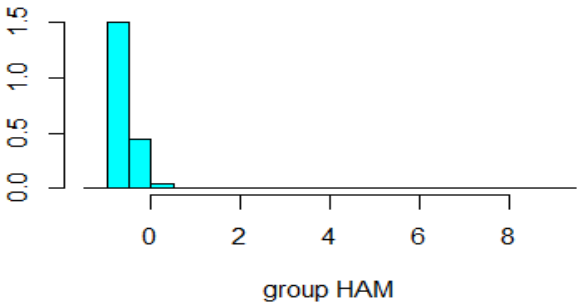
Prior Probabilities of the Groups	
ham	spam
0.5	0.5

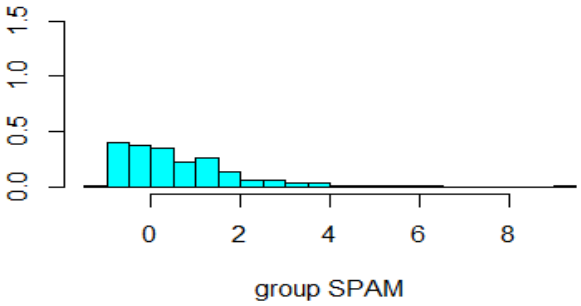
Group Means						
	html	please	email	click	href	free
ham	0.354	0.052	0.308	0.194	0.124	0.246
spam	1.792	0.972	1.238	0.862	1.836	1.474

Coefficients of linear discriminants						
Variables	html	please	email	click	href	free
coefficients	0.10301	0.78324	-0.03864	0.39241	0.01368	0.1045

We first notice groups means of each term within ham are much smaller in group means in spam, the difference is roughly around 1.The coefficients are all positive except term "email". Overall, we can infer that LDA will classify the message to spam if the linear discriminant is large; to the opposite if it's small.

We then plot the linear discriminants by computing 0.10301*html+0.78324*please+-0.03864*email+0.39241*click+0.01368*href +0.1045*free. The plot is obtained as follows:





From the plot, we found that the distributions of two discriminants are all right skewed. Also, the values of ham tend to be negative while much more of the discriminants 'values of spams are positive. It is the difference that makes us capable of classifying.

Since QDA and KNN do not give us details of modeling outputs, we will just present their prediction results. We standardized the variable when using KNN, and then all variables are given a mean of zero and a standard deviation of 1.

4. Prediction

i. Prediction on hard ham testing set (248 obs.)

Logistic	Ham
Ham	50
Spam	198
Accuracy rate: 0.202	

QDA	Ham
Ham	17
Spam	231
Accuracy rate:0.0685	

LDA	Ham
Ham	32
Spam	216
Accuracy rate:0.129	

KNN	Ham
Ham	150
Spam	98
Accuracy rate:0.605	

For predicting the hard hams testing set, KNN performs way better than other method with accuracy rate 60.5% ,comparing to other methods with accuracy rate around 5%-20%.The false positive rate for Logistic, LDA, QDA is very high, meaning that it's easy to classify hard hams as spams. Overall, it's difficult to not classify hard hams to spams, because they contain terms like spams, our highest accuracy rate is only 60%, slightly higher than random guessing. The KNN method is best for detecting hard hams. QDA performs the worst, indicating that the decision boundary in this case may be highly nonlinear and complex.

ii. Prediction on easy ham testing set (1400obs.)

Logistic	Ham
Ham	1376
Spam	24
Accuracy rate: 0.983	

QDA	Ham
Ham	1169
Spam	231
Accuracy rate:0.835	

LDA	Ham
Ham	1346
Spam	54
Accuracy rate: 0.961	

KNN	Ham
Ham	1149
Spam	251
Accuracy rate:0.815	

For predicting the easy hams testing set, Logistic Regression and LDA perform better than QDA and KNN with accuracy rate higher than 95%, comparing to other methods with accuracy rate around 80%.The false positive rate for Logistic, LDA, QDA is very low, meaning that it's easy to classify easy hams correctly. Overall, all classifiers perform better than classifying hard hams. Our highest accuracy rate is 98.3% using Logistic method. KNN method has the worst result in this case, but it still can detect over 80% of easy hams. The decision boundary in this case may be linear since LDA and Logistic Regression has best performance.

iii. Prediction on spam testing set (1400 obs.)

Logistic	Spam
Ham	553
Spam	884
Accuracy rate: 0.604	

QDA	Spam
Ham	316
Spam	1081
Accuracy rate:0.774	

LDA	Spam
Ham	493
Spam	904
Accuracy rate:0.647	

KNN	Spam
Ham	464
Spam	933
Accuracy rate:0.668	

For predicting the spams testing set, QDA and KNN performs slightly better than Logistic Regression and LDA. The accuracy rates (true positive rates) for each method do not differ a lot, all are between 60% ~ 77%.The results are not satisfying enough but are ok. Though there are still about 23%~40% of spams will be seen by users since we decide they are hams and didn't filter out those messages. The decision boundary in this case may be nonlinear since QDA gives better results.

5. Conclusion

Using number of counts of highest occurrence in spams training set as our predictors clearly is not adequate enough. However; it still can give us a guide on classifying our messages, especially for identifying hams. The first shortage of using number of counts of highest occurrence as our predictors is that -it cannot detect spams very well; the true positive rate is about 60%~70%,which is not bad, but there are still 30% of spams will be classified wrongly. The second con of our selection of variables is that it makes us very difficult to identify hard hams as hams; the false positive rate of hard hams is very high. It's not surprising, because hard hams contain terms similar to hams, and our predictors are based on counts of terms. In spam filtering, it is more important to detect spams than detecting hams, but our method works better on detecting hams. KNN works best on detecting hard hams with almost 50% higher accuracy rate than other methods. KNN also works better on detecting spams than LDA and Logistic Regression. As a result, I will recommend picking KNN-the nonparametric approach as our final analysis tool.