STA 141 Assignment 4

Junxiao Bu

November 13, 2014

In this project, I generated 20 potential variables to classify if one email is SPAM or HAM. The code for generate those variables are in the appendix. In the main part of the report, there are only some codes for analyzing each variable's characteristic.

The 20 variables I chose are the followings:

isSpam , numlinesInBody , numAttachments , numRecipients , isInReplyTo , bodyCharacterCount , multipartText , subjectSpamWords , percentSubjectBlanks , averageWordLength , messageIdHasNoHostname , fromnumericEnd , hourSent , isYelling , isDear , numDollarSigns, subjectExclamationCount , subjectQuestCount , percentCapitals , isRe.

Obviously, not every one of these variables is useful to classify SPAM and HAM. The following variables are useful for classifying HAM and SPAM:

For each variable, I will do some exploratory data analysis in order to check if they are useful to predict the SPAM and HAM.

1. isSpam.

For this variable, I don't need to do any analysis. This variable is mainly used to combine with other variables. If isSpam == TRUE, it represents that the email is a spam email. I can see the number of SPAM and HAM in the training data set.

1. isRe

The mosaic plot is omitted here. But from the result, it seems that the emails are less likely to contain "Re" no matter they are HAM or SPAM. So this variable may not be useful to classify HAM and SPAM.

1. numlinesInBody

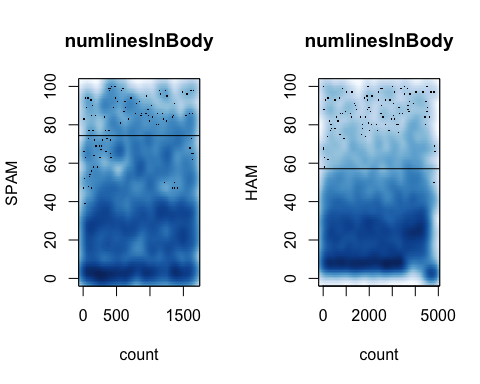
summary(numlinesInBody[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 16.00 38.00 74.37 89.00 1694.00

summary(numlinesInBody[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 14.00 27.00 57.13 44.00 6319.00

par(mfrow=c(1,2))  
smoothScatter(numlinesInBody[isSpam],ylim=c(0,100),xlab="count",ylab="SPAM",main="numlinesInBody")  
abline(h = mean(numlinesInBody[isSpam]))  
  
smoothScatter(numlinesInBody[!isSpam],ylim=c(0,100),xlab="count",ylab="HAM",main="numlinesInBody")  
abline(h = mean(numlinesInBody[!isSpam]))



For this variable, I can find that the average number of lines in SPAM is 74.37, which is much more than the average number of lines in HAM. It seems that there are more lines in the SPAM than HAM. So this variable is relevant to clarify the SPAM and HAM.

4.numAttachments

summary(numAttachments[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.2922 0.0000 6.0000

table(numAttachments[isSpam])

##   
## 0 1 2 3 5 6   
## 1383 112 173 7 1 1

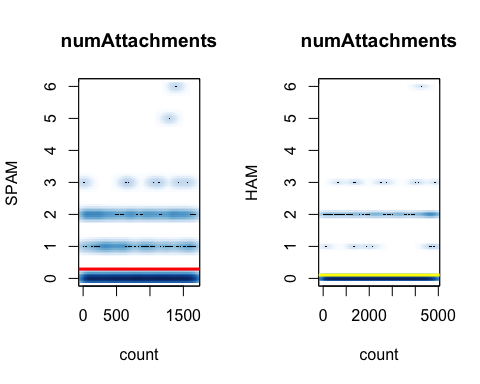
summary(numAttachments[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 0.104 0.000 6.000

table(numAttachments[!isSpam])

##   
## 0 1 2 3 6   
## 4612 10 233 8 1

par(mfrow = c(1,2))  
smoothScatter(numAttachments[isSpam],xlab="count",ylab="SPAM",main="numAttachments")  
abline(h = mean(numAttachments[isSpam]),lwd=3,col=rainbow(1))  
smoothScatter(numAttachments[!isSpam],xlab="count",ylab="HAM",main="numAttachments")  
abline(h = mean(numAttachments[!isSpam]),col="yellow",lwd=3)



It seems that SPAM emails contain more attachments than HAM emails in average(from summary and smoothscatter plot). From the scatter plot of combining numAttachment and numlinesInBody, it seems that there are more attachment for isSpam==1(SPAM) rather than isSpam==0(HAM). They have a pretty similar distribution for number of lines in body in both groups.

So this variable is relevant to clarify the SPAM and HAM.

5&6. subjectExclamationCount & subjectQuestCount

Since these two variables are both counting punctuation marks in the subject of emails. I can analyze them together.

summary(subjectExclamationCount[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.4114 1.0000 42.0000

table(subjectExclamationCount[isSpam])

##   
## 0 1 2 3 4 5 8 42   
## 1250 295 80 31 9 8 3 1

summary(subjectExclamationCount[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.00000 0.03495 0.00000 3.00000

table(subjectExclamationCount[!isSpam])

##   
## 0 1 2 3   
## 4718 131 6 9

summary(subjectQuestCount[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.1747 0.0000 4.0000

table(subjectQuestCount[isSpam])

##   
## 0 1 2 3 4   
## 1520 110 2 1 44

summary(subjectQuestCount[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.1258 0.0000 8.0000

table(subjectQuestCount[!isSpam])

##   
## 0 1 2 3 4 8   
## 4315 509 26 9 4 1

It seems that SPAM emails contain far more exclamation marks than the HAM emails in average. But for question marks, SPAM emails are slightly more than HAM emails in average. So maybe subjectExclamationCount is the more useful variable than subjectQuestCount variable for classifying SPAM and HAM. Because there are no significant differences in counting the numbers of question marks. This subjectQuestCount may not be a good variable to predict if an email is SPAM or HAM.

We can also find the same patterns from frequency tables. For subjectExclamationCount,a relative high percentage of SPAM emails have one exclamation, while the percentage is lower for HAM email. But for subjectQuestCount, this pattern doesn't hold.

1. numRecipients

summary(numRecipients[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.000 2.000 3.775 2.000 108.000

table(numRecipients[isSpam])

##   
## 0 2 4 6 8 10 12 14 16 18 20 22 24 26 28   
## 1 1521 12 5 22 27 11 8 4 8 18 7 4 1 2   
## 32 38 42 44 46 48 50 60 66 68 74 96 108   
## 2 1 2 2 1 4 2 1 2 1 2 4 2

summary(numRecipients[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 2.000 2.000 2.158 2.000 22.000

table(numRecipients[!isSpam])

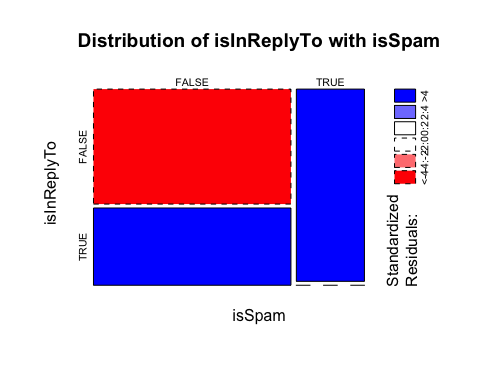
##   
## 2 4 6 8 10 12 14 22   
## 4553 275 21 6 2 4 2 1

It seems that SPAM emails contain more number of recipients than HAM emails. From the frequency table, I can find that most of HAM emails only have recipients less or equal to 4. But for SPAM emails, there are more recipients. From the scatter plot, I find that SPAM emails have more extreme value than HAM emails, which strengthens the conclusion.

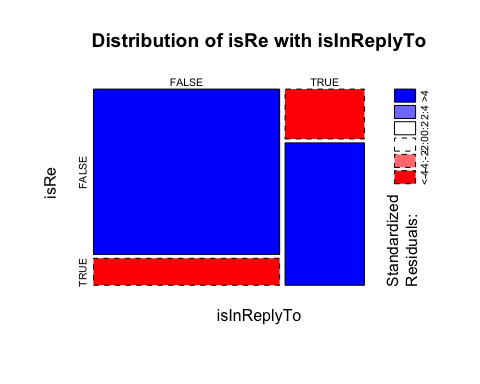
So numRecipients is a useful variable to classify the SPAM and HAM.

1. isInReplyTo

par(mfrow = c(1,1))  
mosaicplot(table(isSpam, isInReplyTo),shade = TRUE,main="Distribution of isInReplyTo with isSpam",col=rainbow(2))



mosaicplot(table(isInReplyTo, isRe),shade = TRUE,main="Distribution of isRe with isInReplyTo",col=rainbow(2))



From the first mosaic plot, I can find that most of SPAM emails don't have In-Reply-To field. Only a very small part of SPAM emails in the training dataset have In-Reply-to field in the header.

I can also compare isRe variable and isInReplyTo variable. In the second mosaic plot,I find that most of emails both have "Re" in the subject and In-Reply-To in the header or neither. If an email has both "Re" and In-Reply-To, this email has higher probability to belong to HAM emails.

So isInReplyTo is useful to classify SPAM and HAM.

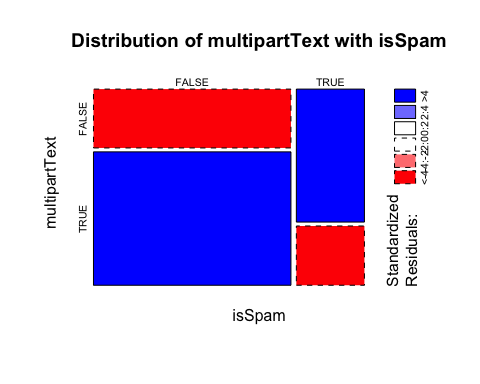
1. bodyCharacterCount

Although the average numbers of body part's character in SPAM emails are bigger than the average numbers in HAM emails, the difference here is not significant. If giving more data, this difference may be more significant. Generally, I don't think this variable is very useful for prediction. (I don’t include graphs and summary here since I don’t think this variable is useful).

1. multipartText

I assume the TRUE represents that there are only texts in this email. The FALSE represents that there are attachments in this email.

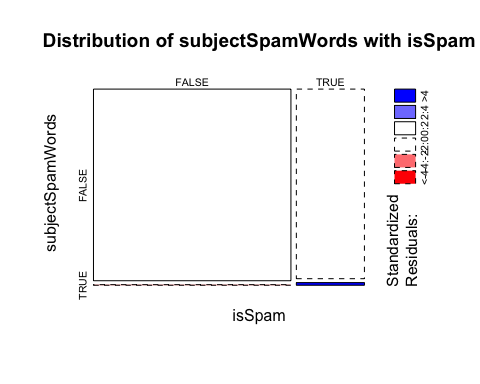
par(mfrow = c(1,1))  
mosaicplot(table(isSpam, multipartText),shade = TRUE,main="Distribution of multipartText with isSpam",col=rainbow(2))



In the mosaic plot, it shows a pattern that if the message is without attachment, the email is HAM. If the email only has attachments, it is more likely to be a SPAM email. In my opinion, this variable is useful to decide if one email is HAM but hard to use for deciding if one email is SPAM. Because there are a lot of HAM emails with attachments.

11.subjectSpamWords

par(mfrow = c(1,1))  
mosaicplot(table(isSpam, subjectSpamWords),shade = TRUE,main="Distribution of subjectSpamWords with isSpam",col=rainbow(2))



length(subjectSpamWords[isSpam])

## [1] 1677

length(subjectSpamWords[!isSpam])

## [1] 4864

This variable is useful since those spam words are very common. It seems that if one email has those spam words, this email belongs to the SPAM. So this variable is useful to classify SPAM and HAM.

12.percentCapitals:

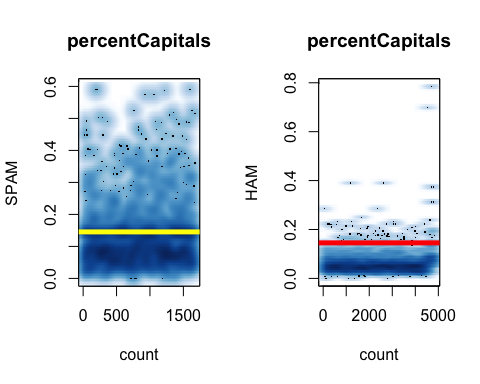
summary(percentCapitals[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00000 0.06869 0.11450 0.14540 0.16790 1.00000 97

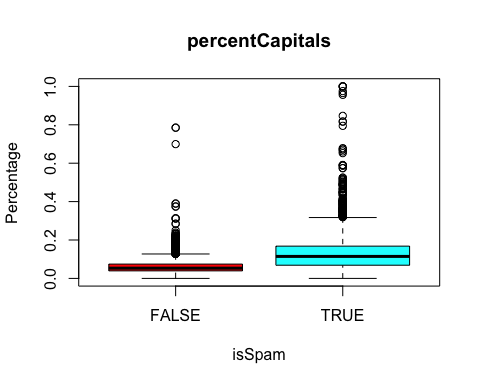
summary(percentCapitals[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00000 0.03922 0.05333 0.06176 0.07454 0.78520 53

par(mfrow = c(1,2))  
smoothScatter(percentCapitals[isSpam],ylim=c(0,0.6),xlab="count",ylab="SPAM",main="percentCapitals")  
abline(h = mean(percentCapitals[isSpam],na.rm=TRUE),lwd=5,col="yellow")  
smoothScatter(percentCapitals[!isSpam],xlab="count",ylab="HAM",main="percentCapitals")  
abline(h = mean(percentCapitals[isSpam],na.rm=TRUE),lwd=5,col=rainbow(1))



par(mfrow = c(1,1))  
boxplot(percentCapitals ~ isSpam,main="percentCapitals",xlab="isSpam",ylab="Percentage",col=rainbow(2))



I can find that SPAM emails contains more capital letters in the body of the message. From the box plot, I can also find the mean of SPAM is higher than HAM. And 75 percentile of SPAM is much more higher than HAM. It seems that it has higher probability in SPAM to have extreme value for this variable. We see that this is the case for the percent capitals with spam having a larger mean and greater spread than ham. So percentCapitals should be a useful variable to classify SPAM and HAM.

1. averageWordLength

summary(averageWordLength[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 65.0 183.0 520.8 515.0 14490.0

summary(averageWordLength[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 58.75 141.00 361.90 278.20 42510.00

par(mfrow = c(1,2))  
smoothScatter(averageWordLength[isSpam],xlab="count",ylab="SPAM",main="averageWordLength")  
abline(h = mean(averageWordLength[isSpam],na.rm=TRUE),lwd=5,col="yellow")  
smoothScatter(averageWordLength[!isSpam],xlab="count",ylab="HAM",main="averageWordLength")

abline(h = mean(averageWordLength[isSpam],na.rm=TRUE),lwd=5,col=rainbow(1))

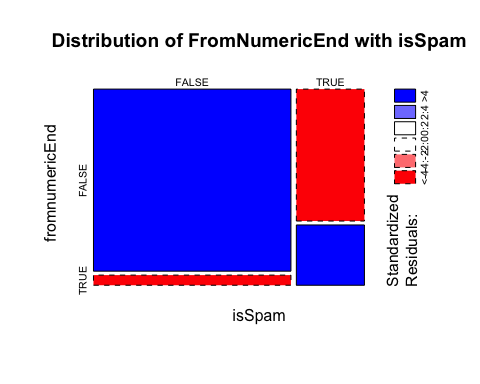
I can find that SPAM emails' average word lengths are longer than HAM emails'. From scatter plot, I can also find the same pattern. So this variable is useful to classify the SPAM and HAM.

1. messageIdHasNoHostname

If messageIdHasNoHostname == 1, it means this email has host name. This variable may not be very useful. Although it shows the pattern that if one email has host name, it is more likely to be HAM. But it also shows the pattern that if one email doesn’t have host name, it is more likely to be HAM rather than SPAM relatively. So this variable cannot distinguish HAM and SPAM clearly.

1. fromnumericEnd:

mosaicplot(table(isSpam, fromnumericEnd),shade = TRUE,main="Distribution of FromNumericEnd with isSpam",col=rainbow(2))



It seems that if one email doesn’t have numeric end, the HAM emails are more likely to have an end without any numbers in the address. So from the training dataset, I can know that if one email has a normal end in the From: field, this email is very possible from HAM emails.

This variable is useful to classify the SPAM and HAM.

1. hourSent

summary(hourSent[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.00 7.00 13.00 12.51 19.00 85.00 1

summary(hourSent[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 8.0 12.0 12.3 16.0 93.0

table(hourSent[isSpam])

##   
## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 31   
## 44 74 52 71 55 59 57 51 64 70 84 69 64 74 68 64 78 70 84 96 92 73 81 76 1   
## 33 40 41 80 85   
## 1 1 1 1 1

table(hourSent[!isSpam])

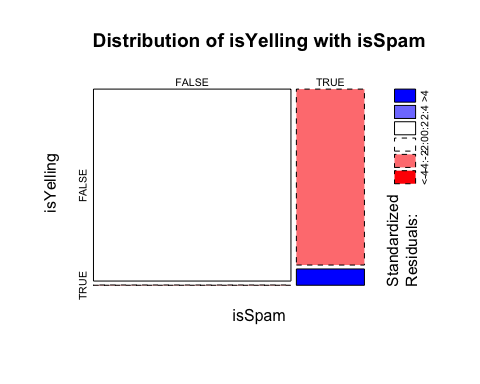
##   
## 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17   
## 113 90 139 68 41 20 75 61 883 219 301 266 290 260 243 391 266 203   
## 18 19 20 21 22 23 93   
## 167 167 127 142 155 175 2

For this variable, I don't think mean and median can provide useful information to classify SPAM and HAM. When observing the frequency tables of hour\_Sent for HAM and SPAM. The distribution of HAM emails is not uniform. There are more emails during the work hours(8:00-18:00). On the contrary, the distribution of SPAM email is pretty uniform distributed.

So if we have some new email data, we can calculate the frequency table for hour\_Sent. If the distribution is roughly uniform, those emails may come from SPAM emails.

17.isYelling

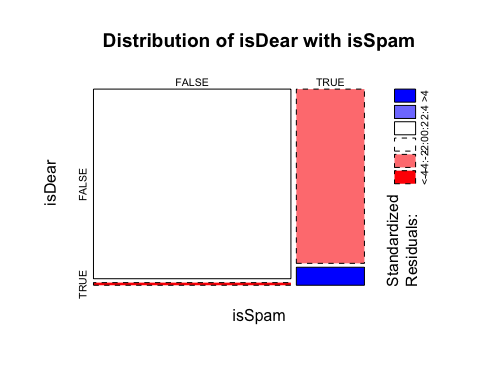
mosaicplot(table(isSpam, isYelling),shade = TRUE,main="Distribution of isYelling with isSpam",col=rainbow(2))



It seems that if one email has the subject, which is not all the capital letters, it is more likely to be HAM emails. So if an email's subject is not all capital letters. This email is highly possible to belong to HAM email. If an email's subject is all capital letters. This email is highly possible to belong to SPAM email.

1. isDear

mosaicplot(table(isSpam, isDear),shade = TRUE,main="Distribution of isDear with isSpam",col=rainbow(2))



It seems that most of HAM emails don't have "Dear" in their body messages. If one email's body message has "Dear", this email is more likely to come from SPAM emails.

So this variable is useful to classify SPAM and HAM.

1. numDollarSigns

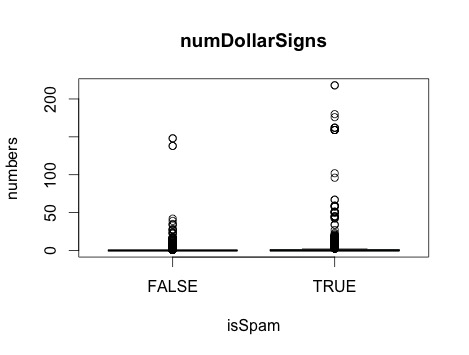
summary(numDollarSigns[isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 3.225 1.000 218.000

summary(numDollarSigns[!isSpam])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.5302 0.0000 148.0000

boxplot(numDollarSigns ~ isSpam,main="numDollarSigns",xlab="isSpam",ylab="numbers",col=rainbow(2))



It is obvious that the average number of dollar sign in SPAM emails is much higher than the average number in HAM emails. Besides, this variable with SPAM has a greater spread than HAM. So this variable is useful for classifying the SPAM and HAM.

1. percentSubjectBlanks

It seems that the percent of subject spaces doesn't have big difference between SPAM emails and HAM emails. So this variable may not be very useful for classifying SPAM and HAM.