Naive_bayes

Park Ju ho

2022 4 17

e1071를 이용한 Naive Bayes

```
library(e1071)

data(iris)
m = naiveBayes(Species~.,data=iris)

m
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##     setosa versicolor  virginica
##  0.3333333  0.3333333  0.3333333
##
## Conditional probabilities:
##             Sepal.Length
## Y             [,1]      [,2]
##   setosa     5.006 0.3524897
##   versicolor 5.936 0.5161711
##   virginica  6.588 0.6358796
##
##             Sepal.Width
## Y             [,1]      [,2]
##   setosa     3.428 0.3790644
##   versicolor 2.770 0.3137983
##   virginica  2.974 0.3224966
##
##             Petal.Length
## Y             [,1]      [,2]
##   setosa     1.462 0.1736640
##   versicolor 4.260 0.4699110
##   virginica  5.552 0.5518947
##
```

```
##             Petal.Width
## Y             [,1]      [,2]
##   setosa     0.246 0.1053856
##   versicolor 1.326 0.1977527
##   virginica  2.026 0.2746501
```

iris data의 정보를 제공해 줌
1. label의 비율은 어떠한지
2. 각 feature별|범주별 mean과std를 보여줌

```
table(predict(m,iris),iris[,5])
```

```
##
##              setosa versicolor virginica
##   setosa         50          0         0
##   versicolor      0         47         3
##   virginica       0          3        47
```

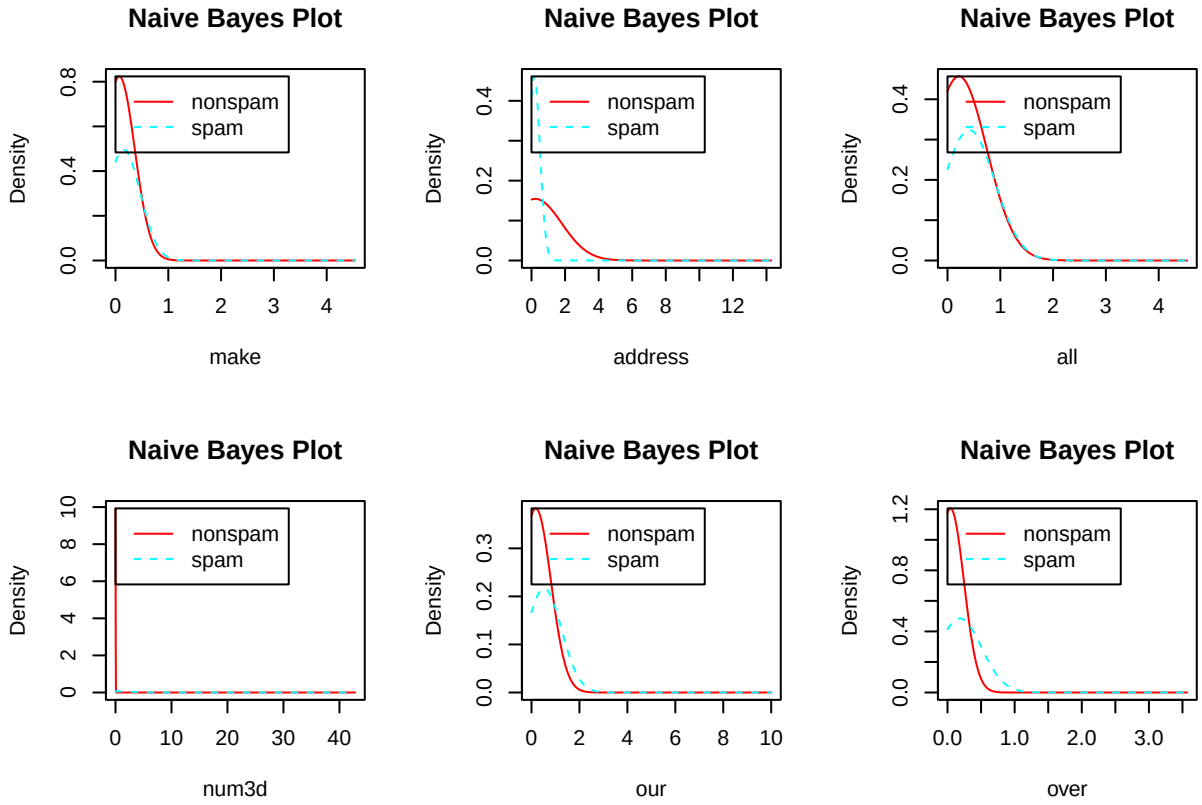
Naive Bayes 분류기를 활용한 분류
총 6개를 잘 못 분류한 것을 확인 할 수 있다

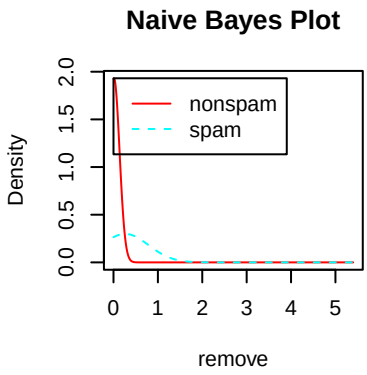
klaR을 이용한 Naive Bayes

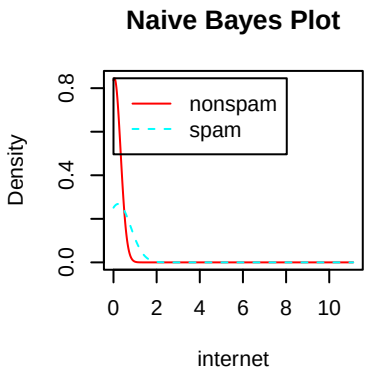
```
library(klaR)
library(kernlab)
data(spam)
colnames(spam)
```

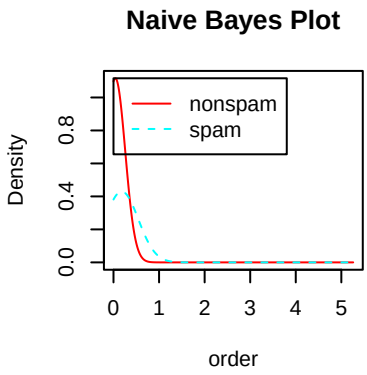
```
##  [1] "make"              "address"           "all"
##  [4] "num3d"             "our"               "over"
##  [7] "remove"            "internet"          "order"
## [10] "mail"              "receive"           "will"
## [13] "people"            "report"            "addresses"
## [16] "free"              "business"          "email"
## [19] "you"               "credit"            "your"
## [22] "font"              "num000"            "money"
## [25] "hp"                "hpl"               "george"
## [28] "num650"            "lab"               "labs"
## [31] "telnet"            "num857"            "data"
## [34] "num415"            "num85"             "technology"
## [37] "num1999"           "parts"             "pm"
## [40] "direct"            "cs"                "meeting"
## [43] "original"          "project"           "re"
## [46] "edu"               "table"             "conference"
## [49] "charSemicolon"     "charRoundbracket"  "charSquarebracket"
## [52] "charExclamation"   "charDollar"        "charHash"
## [55] "capitalAve"        "capitalLong"       "capitalTotal"
## [58] "type"
```

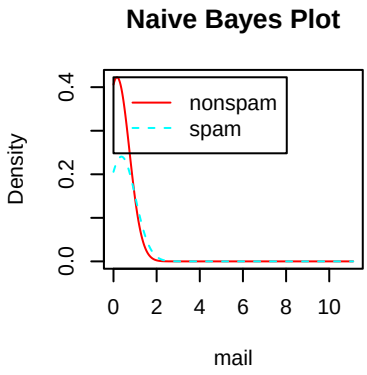
```
train.ind = sample(1:nrow(spam), ceiling(nrow(spam)*2/3), replace=FALSE)
nb.res = NaiveBayes(type ~ ., data=spam[train.ind,])

par(mfrow=c(2,3))
plot(nb.res)
```

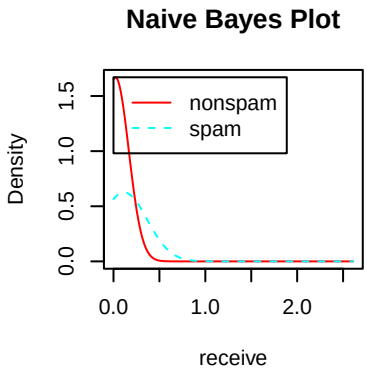


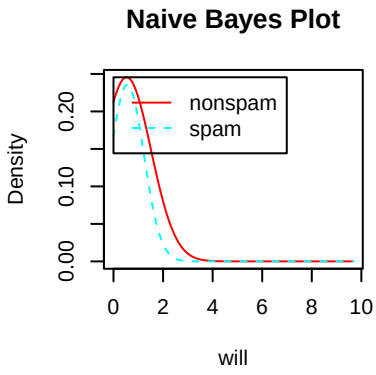

Naive Bayes Plot
nonspam
spam
Density
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
5
remove

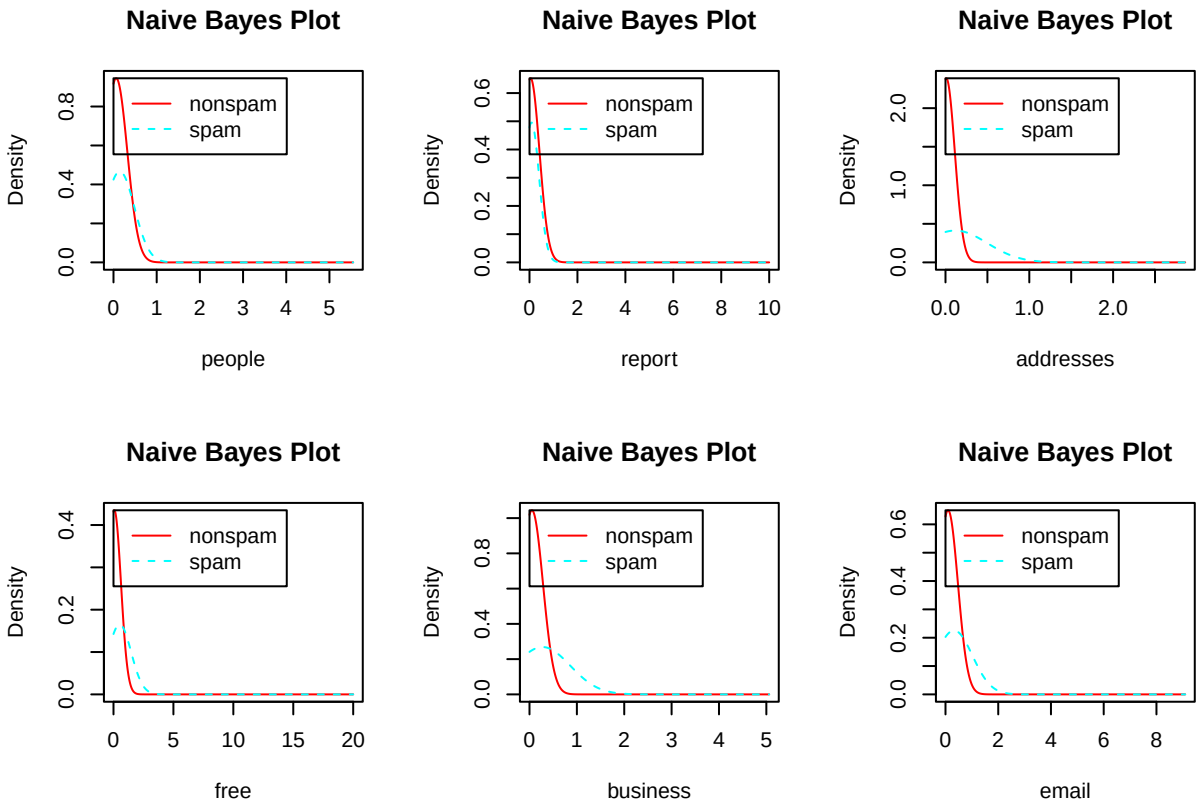

Naive Bayes Plot
nonspam
spam
Density
0.0
0.4
0.8
0
2
4
6
8
10
internet

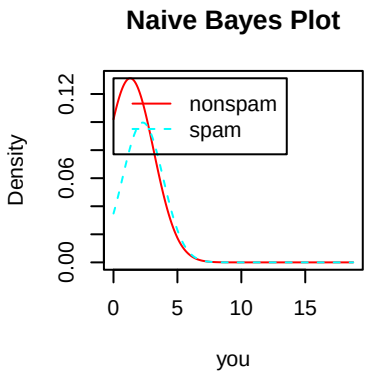

Naive Bayes Plot
nonspam
spam
Density
0.0
0.4
0.8
0
1
2
3
4
5
order

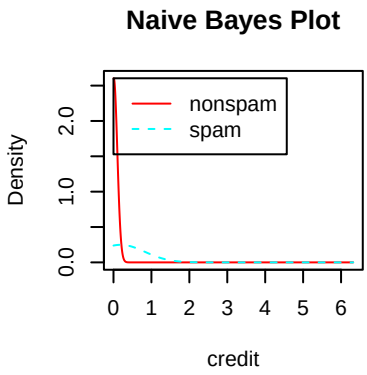

Naive Bayes Plot
nonspam
spam
Density
0.0
0.2
0.4
0
2
4
6
8
10
mail

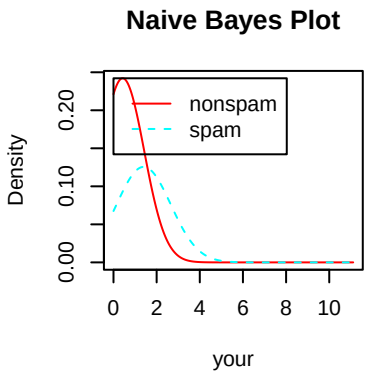

Naive Bayes Plot
nonspam
spam
Density
0.0
0.5
1.0
1.5
0.0
1.0
2.0
receive

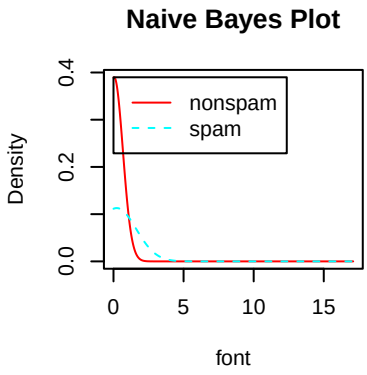

Naive Bayes Plot
nonspam
spam
Density
0.00
0.10
0.20
0
2
4
6
8
10
will

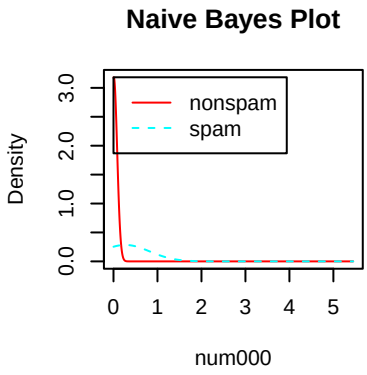

Naive Bayes Plot
Density
nonspam
spam
people
Naive Bayes Plot
Density
nonspam
spam
report
Naive Bayes Plot
Density
nonspam
spam
addresses
Naive Bayes Plot
Density
nonspam
spam
free
Naive Bayes Plot
Density
nonspam
spam
business
Naive Bayes Plot
Density
nonspam
spam
email

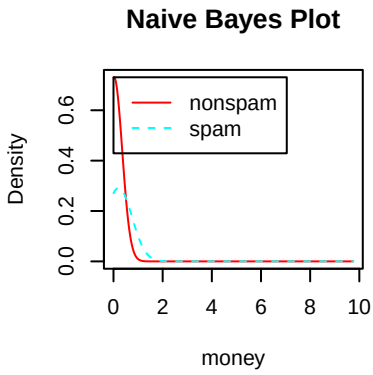

Naive Bayes Plot
Density
0.00
0.06
0.12
nonspam
spam
0
5
10
15
you

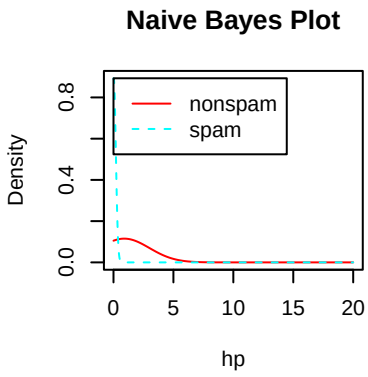

Naive Bayes Plot
Density
0.0
1.0
2.0
nonspam
spam
0
1
2
3
4
5
6
credit

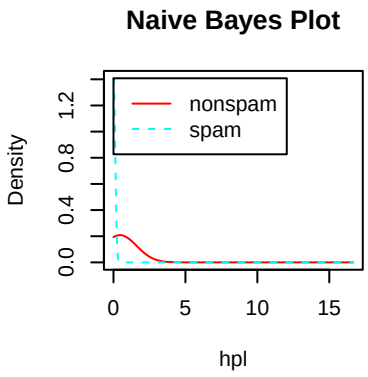

Naive Bayes Plot
Density
0.00
0.10
0.20
nonspam
spam
0
2
4
6
8
10
your

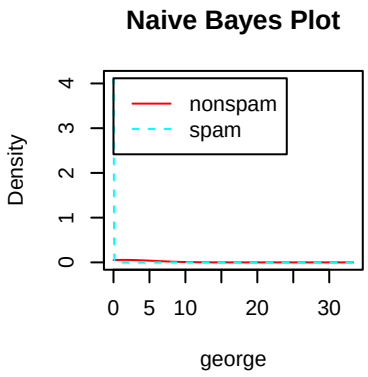

Naive Bayes Plot
Density
0.0
0.2
0.4
nonspam
spam
0
5
10
15
font

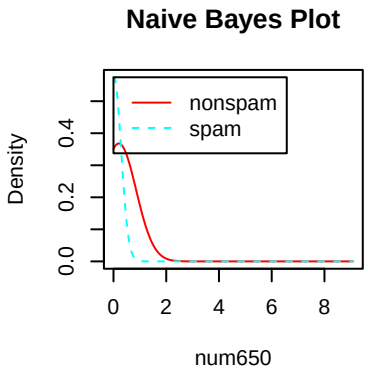

Naive Bayes Plot
Density
0.0
1.0
2.0
3.0
nonspam
spam
0
1
2
3
4
5
num000

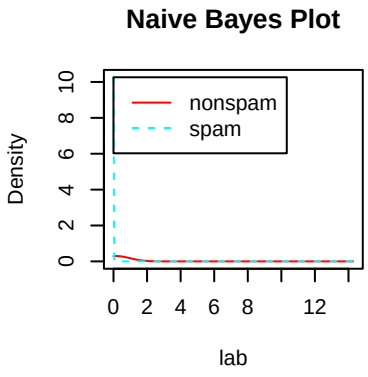

Naive Bayes Plot
Density
0.0
0.2
0.4
0.6
nonspam
spam
0
2
4
6
8
10
money

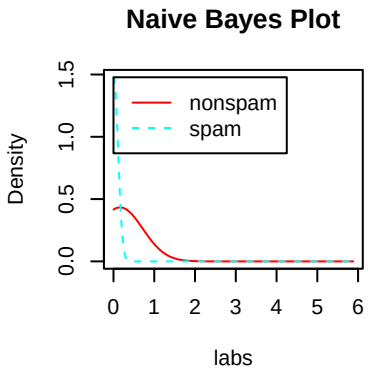

Naive Bayes Plot
nonspam
spam
Density
0.0
0.4
0.8
0
5
10
15
20
hp

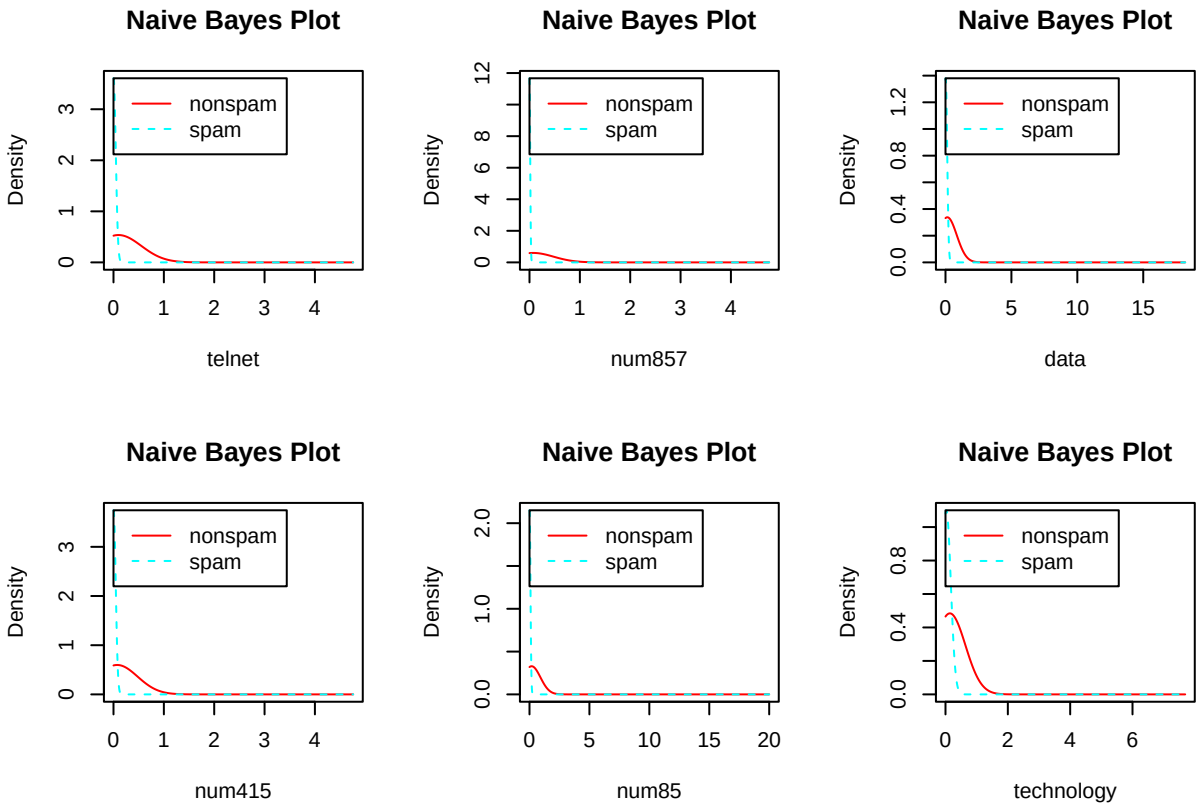

Naive Bayes Plot
nonspam
spam
Density
0.0
0.4
0.8
1.2
0
5
10
15
hpl

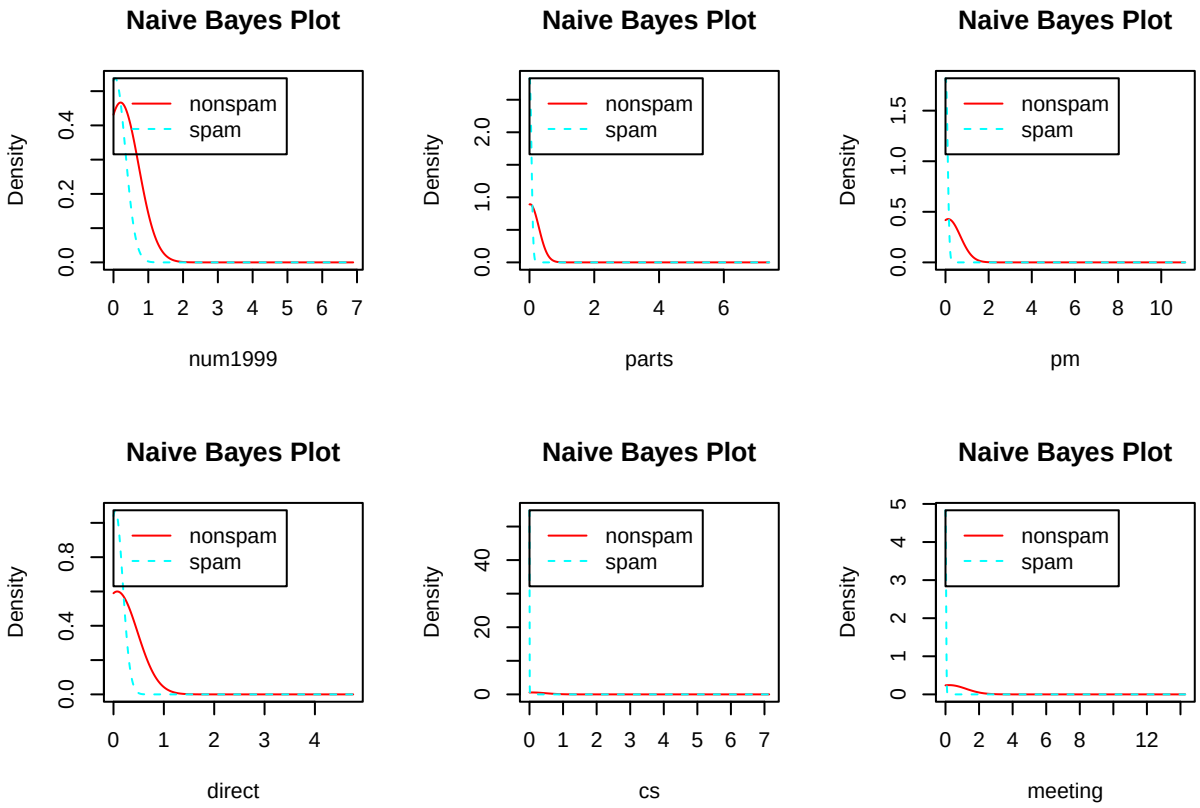

Naive Bayes Plot
nonspam
spam
Density
0
1
2
3
4
0
5
10
20
30
george

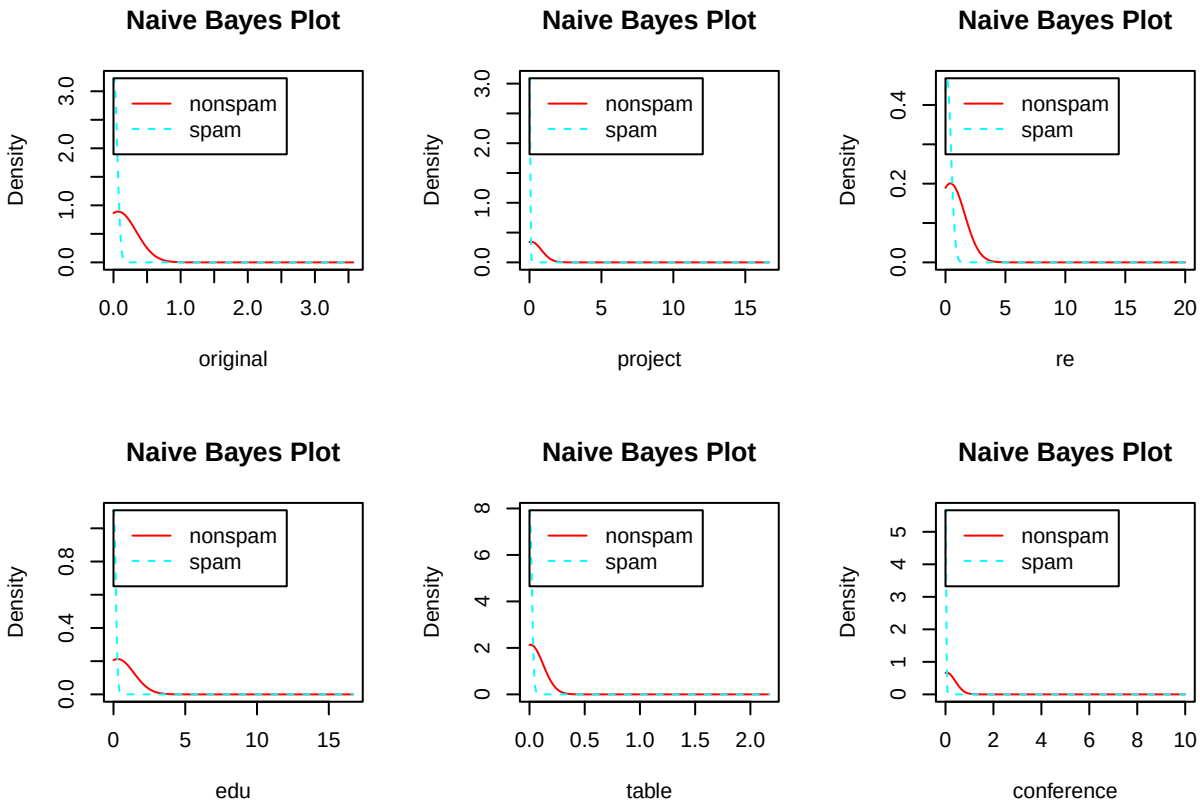

Naive Bayes Plot
nonspam
spam
Density
0.0
0.2
0.4
0
2
4
6
8
num650

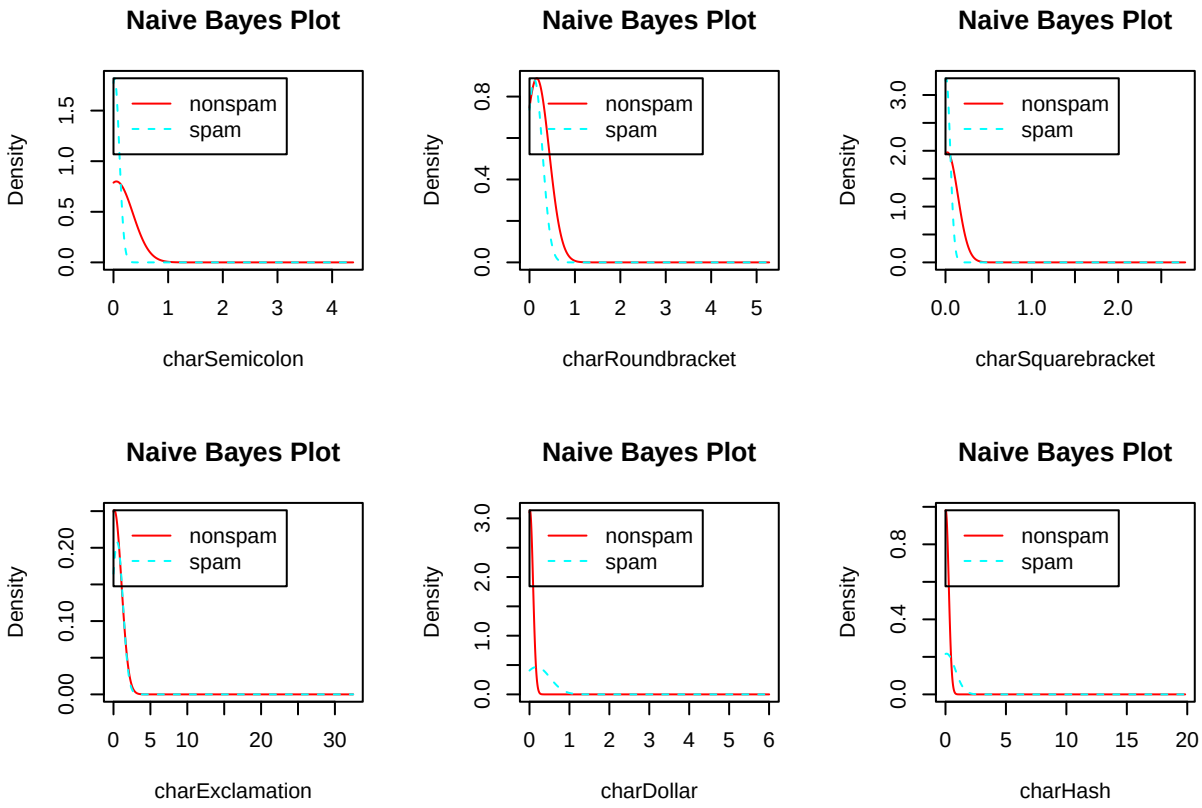

Naive Bayes Plot
nonspam
spam
Density
0
2
4
6
8
10
0
2
4
6
8
12
lab

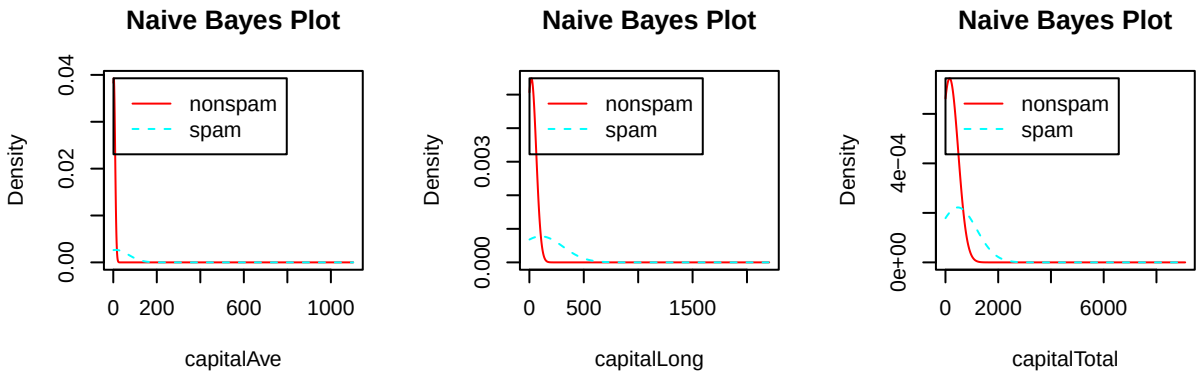

Naive Bayes Plot
nonspam
spam
Density
0.0
0.5
1.0
1.5
0
1
2
3
4
5
6
labs


Naive Bayes Plot
nonspam
spam
Density
telnet
num857
data
num415
num85
technology


Naive Bayes Plot
nonspam
spam
Density
num1999
Naive Bayes Plot
nonspam
spam
Density
parts
Naive Bayes Plot
nonspam
spam
Density
pm
Naive Bayes Plot
nonspam
spam
Density
direct
Naive Bayes Plot
nonspam
spam
Density
cs
Naive Bayes Plot
nonspam
spam
Density
meeting


Naive Bayes Plot
nonspam
spam
Density
original
project
re
edu
table
conference


Naive Bayes Plot
Density
charSemicolon
Naive Bayes Plot
Density
charRoundbracket
Naive Bayes Plot
Density
charSquarebracket
Naive Bayes Plot
Density
charExclamation
Naive Bayes Plot
Density
charDollar
Naive Bayes Plot
Density
charHash
nonspam
spam



변수의 영향력을 알아보는 그래프
선이 겹치지 않을 수록 잘 분류하는 변수로써 중요도가 높다고 할 수 있다

pred

```
nb.pred = predict(nb.res,spam[-train.ind,])
c_mat=table(nb.pred$class,spam[-train.ind,"type"])
c_mat
```

```
##
##           nonspam spam
##   nonspam     522   29
##   spam        410  572
```

```
sum(diag(c_mat))/sum(c_mat)
```

```
## [1] 0.7136334
```

0.7038의 정확도를 보여준다

Na를 포함하고 있는 자료 Naive Bayes

훈련 시: 결측값 포함 시 케이스에서 제외
예측 시: 결측인 속성을 계산 과정에서 생략

```
library(e1071)
library(mlbench)

data("HouseVotes84")
head(HouseVotes84)
```

```
##        Class   V1 V2 V3   V4   V5 V6 V7 V8 V9 V10  V11  V12 V13 V14 V15  V16
## 1 republican    n  y  n    y    y  y  n  n  n   y <NA>    y   y   y   n    y
## 2 republican    n  y  n    y    y  y  n  n  n   n    n    y   y   y   n <NA>
## 3   democrat <NA>  y  y <NA>    y  y  n  n  n   n    y    n   y   y   n    n
## 4   democrat    n  y  y    n <NA>  y  n  n  n   n    y    n   y   n   n    y
## 5   democrat    y  y  y    n    y  y  n  n  n   n    y <NA>   y   y   y    y
## 6   democrat    n  y  y    n    y  y  n  n  n   n    n    n   y   y   y    y
```

```
summary(HouseVotes84)
```

```
##         Class         V1          V2          V3          V4          V5
##  democrat  :267   n   :236   n   :192   n   :171   n   :247   n   :208
##  republican:168   y   :187   y   :195   y   :253   y   :177   y   :212
##                   NA's: 12   NA's: 48   NA's: 11   NA's: 11   NA's: 15
##     V6          V7          V8          V9         V10         V11         V12
##  n   :152   n   :182   n   :178   n   :206   n   :212   n   :264   n   :233
##  y   :272   y   :239   y   :242   y   :207   y   :216   y   :150   y   :171
##  NA's: 11   NA's: 14   NA's: 15   NA's: 22   NA's:  7   NA's: 21   NA's: 31
##    V13         V14         V15         V16
##  n   :201   n   :170   n   :233   n   : 62
##  y   :209   y   :248   y   :174   y   :269
##  NA's: 25   NA's: 17   NA's: 28   NA's:104
```

```
model = naiveBayes(Class~.,data = HouseVotes84)
pred = predict(model,HouseVotes84[,-1])
tab = table(pred,HouseVotes84$Class)

tab
```

```
##
## pred         democrat republican
##   democrat        238         13
##   republican       29        155
```

```
sum(diag(tab))/sum(tab)
```

```
## [1] 0.9034483
```

결측치를 제거하지 않아도 잘 예측하는 것을 확인 할 수 있다