ETC3250 Lab 8

Di Cook

15 September 2015

Supervised classification, and data visualisation

Assignment

Similarly examine, build a classifier and assess the model fit for the spam data from the ggobi web site. Here is some code to get you started.

The error rate from the logistic regression model is quite small. The most important error is real mail getting misclassified as spam, and this is about the same as the overall error rate. (The actual numbers will change depending on your split of training and test data.)

The most important variables according to the model are weekends, nbox, nlocal and digits. Mail on the weekends is less likely to be spam. real mail is more likely to come from someone in the user's address book, and also from the same domain. Spam is more likely to come from someone with a lot of numbers in their email address.

```
spam <- read.csv("http://www.ggobi.org/book/data/spam.csv",
                 stringsAsFactors = FALSE)
spam$nbox <- ifelse(spam$box == "yes", 1, 0)
spam$nlocal <- ifelse(spam$local == "yes", 1, 0)
spam$cspam <- ifelse(spam$spam == "yes", 1, 0)
spam$day.of.week <- factor(spam$day.of.week,
                           levels=c("Mon", "Tue", "Wed", "Thu", "Fri", "Sat", "Sun"))
spam.sub <- select(spam, day.of.week, time.of.day, size.kb, nbox,
                   nlocal, digits, cappct, cspam)

qplot(nbox, data=spam.sub, geom="bar") + facet_wrap(~cspam, ncol=1)
qplot(nlocal, data=spam.sub, geom="bar") + facet_wrap(~cspam, ncol=1)
qplot(day.of.week, data=spam.sub, geom="bar") + facet_wrap(~cspam, ncol=1)
qplot(time.of.day, data=spam.sub, geom="histogram", binwidth=1) +
  facet_wrap(~cspam, ncol=1)
qplot(size.kb, data=spam.sub, geom="histogram", binwidth=1) +
  facet_wrap(~cspam, ncol=1) + scale_x_log10()
qplot(digits, data=spam.sub, geom="histogram", binwidth=1) +
  facet_wrap(~cspam, ncol=1) + scale_x_log10()
qplot(cappct, data=spam.sub, geom="histogram", binwidth=0.1) +
  facet_wrap(~cspam, ncol=1)

spam.sub <- arrange(spam.sub, cspam)
indx <- sort(c(sample(1:1461, 974), sample(1462:2171, 473)))
spam.sub.tr <- spam.sub[indx,]
spam.sub.ts <- spam.sub[-indx,]
glm.fit <- glm(cspam~., data=spam.sub.tr,
               family=binomial)
summary(glm.fit)
```

```
##
```

```
## Call:
## glm(formula = cspam ~ ., family = binomial, data = spam.sub.tr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6388  -0.2985  -0.0487   0.4855   3.6323
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.005450   0.324187   3.101  0.00193 **
## day.of.weekTue  0.271503   0.335291   0.810  0.41808
## day.of.weekWed  0.215732   0.327388   0.659  0.50993
## day.of.weekThu -0.274318   0.321628  -0.853  0.39371
## day.of.weekFri  0.548295   0.368998   1.486  0.13730
## day.of.weekSat  1.575517   0.498100   3.163  0.00156 **
## day.of.weekSun  1.085567   0.406135   2.673  0.00752 **
## time.of.day    -0.021294   0.014749  -1.444  0.14882
## size.kb        -0.001796   0.001838  -0.977  0.32853
## nbox           -4.029279   0.296889 -13.572  < 2e-16 ***
## nlocal         -3.712739   0.347253 -10.692  < 2e-16 ***
## digits          0.319805   0.095863   3.336  0.00085 ***
## cappct          0.497668   0.522451   0.953  0.34081
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1828.86  on 1446  degrees of freedom
## Residual deviance:  724.32  on 1434  degrees of freedom
## AIC: 750.32
##
## Number of Fisher Scoring iterations: 7
```

```
spam.ts.p <- predict(glm.fit, spam.sub.ts, type="response")
spam.sub.ts$pcspam <- 0
spam.sub.ts$pcspam[spam.ts.p > 0.5] <- 1
table(spam.sub.ts$cspam, spam.sub.ts$pcspam)
```

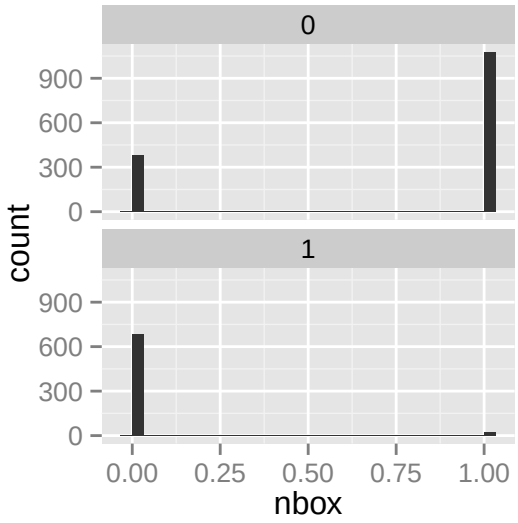
```
##
##       0   1
##   0 423  64
##   1  12 225
```

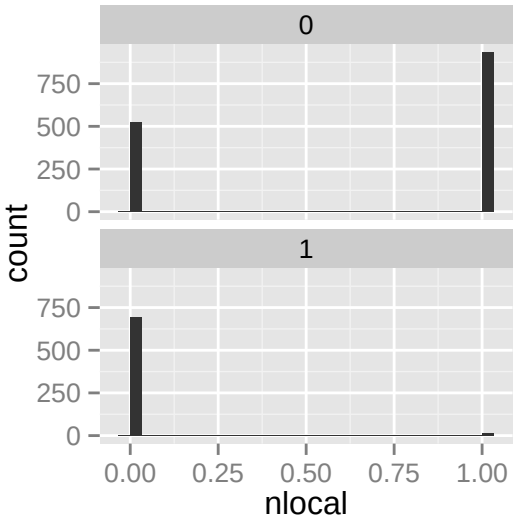
```
x <- table(spam.sub.ts$cspam, spam.sub.ts$pcspam)
1-sum(diag(x))/sum(x)
```

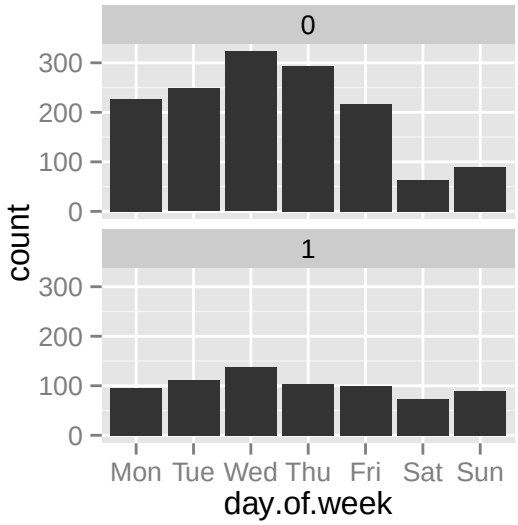
```
## [1] 0.1049724
```

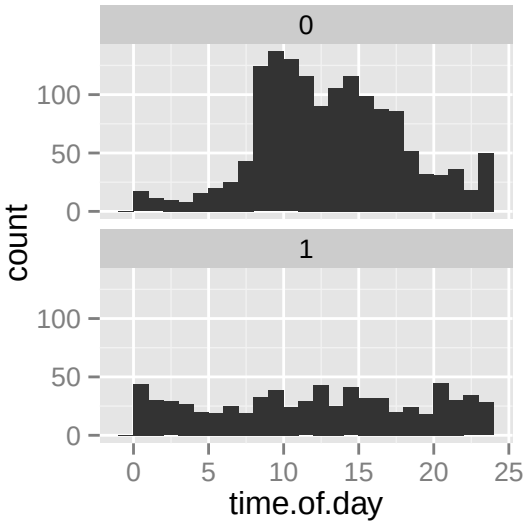
```
x[2,1]/(x[2,1]+x[2,2])
```

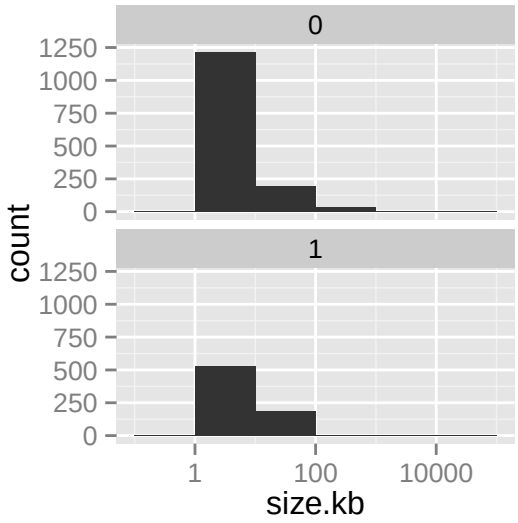
```
## [1] 0.05063291
```

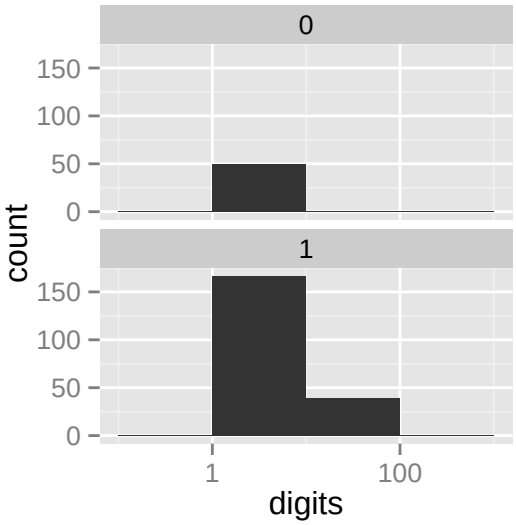

0
1
count
900
600
300
0
0.00
0.25
0.50
0.75
1.00
nbox

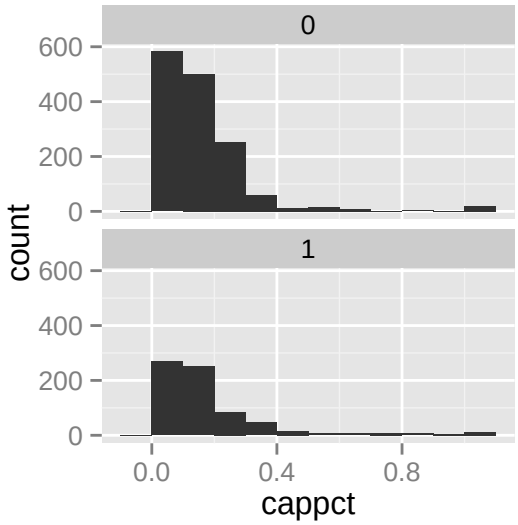

0
1
count
750
500
250
0
0.00
0.25
0.50
0.75
1.00
nlocal


0
1
count
300
200
100
0
Mon
Tue
Wed
Thu
Fri
Sat
Sun
day.of.week


0
1
count
100
50
0
0
5
10
15
20
25
time.of.day


0
1
count
size.kb
1
100
10000


0
1
count
digits
1
100


0
1
count
cappct
0.0
0.4
0.8