APA-L5

September 6, 2018

1 APA Laboratori 5 - LDA/QDA/NBayes/RegLog

```
In [1]: # uncomment to install missing libraries
        #install.packages('klaR')
        #install.packages('mlbench')
        # install.packages('e1071')
        # install.packages('class')
        # install.packages('kernlab')
```

```
In [2]: options(repr.plot.width=6, repr.plot.height=6)
```

```
In [3]: library(MASS)
```

1.1 Example 1: Visualizing and classifying wines with LDA and QDA

We have the results of an analysis on wines grown in a region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 chemical constituents found in each of the three types of wines. The goal is to separate the three types of wines:

```
In [4]: wine <- read.table("wine.data", sep=",", dec=".", header=FALSE)

        dim(wine)

        colnames(wine) <- c('Wine.type','Alcohol','Malic.acid','Ash',
                            'Alcalinity.of.ash','Magnesium','Total.phenols',
                            'Flavanoids','Nonflavanoid.phenols','Proanthocyanins',
                            'Color.intensity','Hue','OD280/OD315','Proline')
```

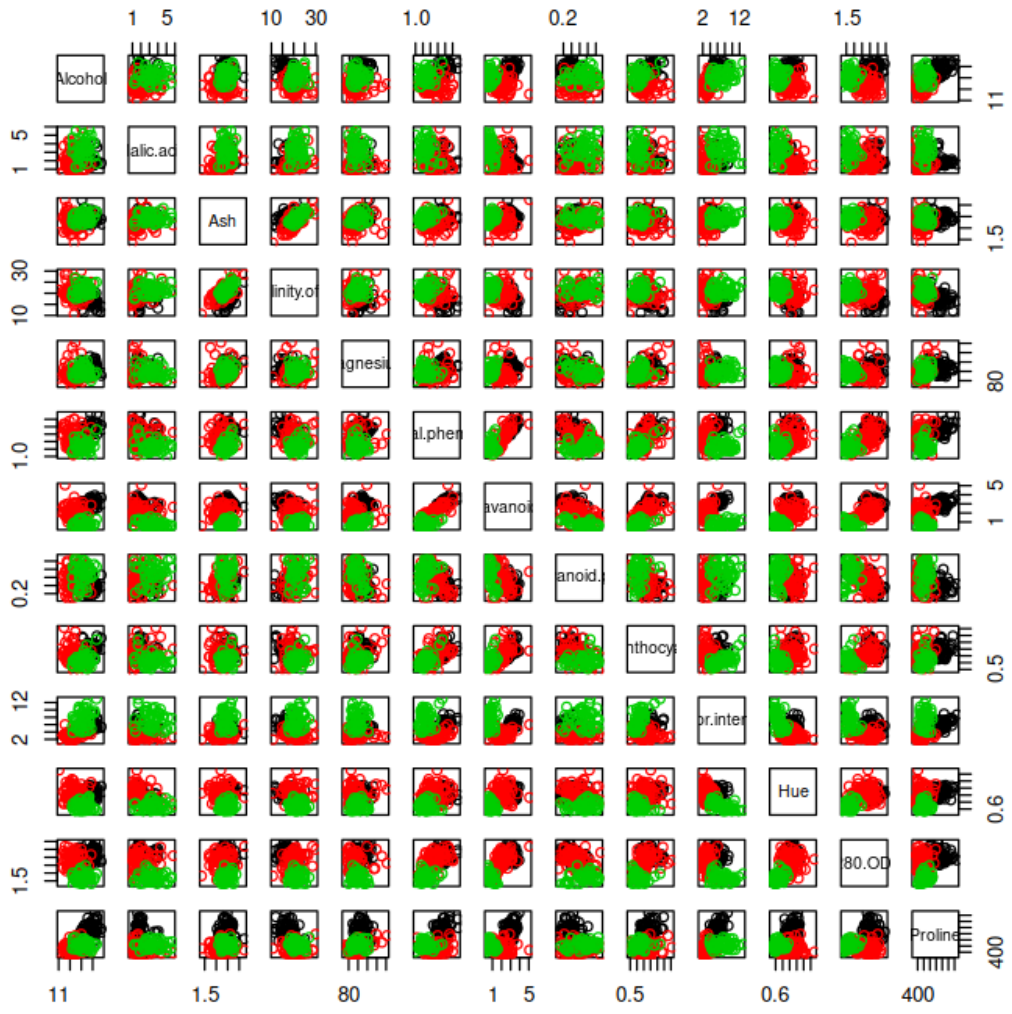
1. 178 2. 14

Clean up column names

```
In [5]: colnames(wine) <- make.names(colnames(wine))

        wine$Wine.type <- as.factor(wine$Wine.type)

        summary(wine)
```

```
 Wine.type    Alcohol         Malic.acid          Ash         Alcalinity.of.ash
 1:59       Min.   :11.03   Min.   :0.740   Min.   :1.360   Min.   :10.60
 2:71       1st Qu.:12.36   1st Qu.:1.603   1st Qu.:2.210   1st Qu.:17.20
 3:48       Median :13.05   Median :1.865   Median :2.360   Median :19.50
            Mean   :13.00   Mean   :2.336   Mean   :2.367   Mean   :19.49
            3rd Qu.:13.68   3rd Qu.:3.083   3rd Qu.:2.558   3rd Qu.:21.50
            Max.   :14.83   Max.   :5.800   Max.   :3.230   Max.   :30.00
   Magnesium      Total.phenols     Flavanoids    Nonflavanoid.phenols
 Min.   : 70.00   Min.   :0.980   Min.   :0.340   Min.   :0.1300
 1st Qu.: 88.00   1st Qu.:1.742   1st Qu.:1.205   1st Qu.:0.2700
 Median : 98.00   Median :2.355   Median :2.135   Median :0.3400
 Mean   : 99.74   Mean   :2.295   Mean   :2.029   Mean   :0.3619
 3rd Qu.:107.00   3rd Qu.:2.800   3rd Qu.:2.875   3rd Qu.:0.4375
 Max.   :162.00   Max.   :3.880   Max.   :5.080   Max.   :0.6600
 Proanthocyanins Color.intensity       Hue           OD280.OD315
 Min.   :0.410   Min.   : 1.280   Min.   :0.4800   Min.   :1.270
 1st Qu.:1.250   1st Qu.: 3.220   1st Qu.:0.7825   1st Qu.:1.938
 Median :1.555   Median : 4.690   Median :0.9650   Median :2.780
 Mean   :1.591   Mean   : 5.058   Mean   :0.9574   Mean   :2.612
 3rd Qu.:1.950   3rd Qu.: 6.200   3rd Qu.:1.1200   3rd Qu.:3.170
 Max.   :3.580   Max.   :13.000   Max.   :1.7100   Max.   :4.000
    Proline
 Min.   : 278.0
 1st Qu.: 500.5
 Median : 673.5
 Mean   : 746.9
 3rd Qu.: 985.0
 Max.   :1680.0
```

In [6]:
```
plot(subset(wine,select=-Wine.type),col=unclass(wine$Wine.type))
```



For this example let's practice a different call mode to lda(), using a formula; this is most useful when our data is in a dataframe format:

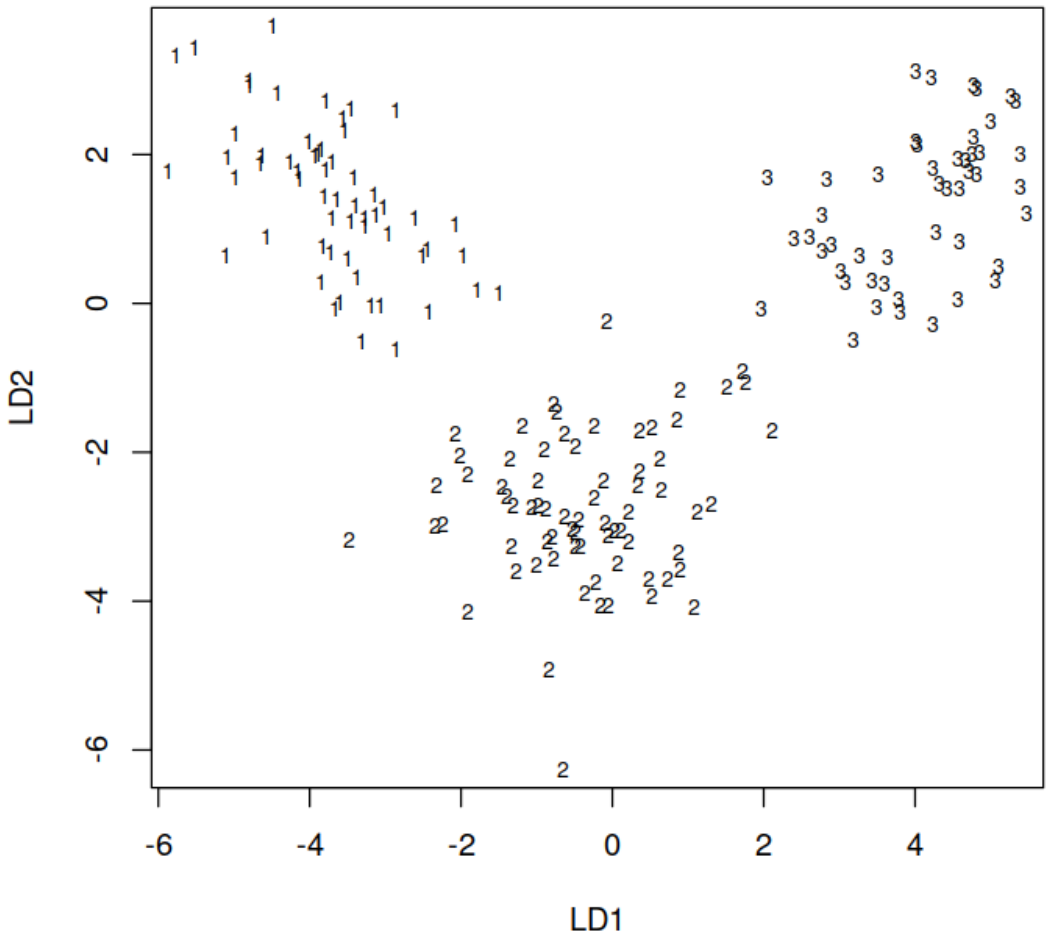
```
In [7]: lda.model <- lda (Wine.type ~ ., data = wine)

        lda.model

Call:
lda(Wine.type ~ ., data = wine)

Prior probabilities of groups:
        1         2         3
0.3314607 0.3988764 0.2696629
```

```
Group means:
   Alcohol Malic.acid      Ash Alcalinity.of.ash Magnesium Total.phenols
1 13.74475   2.010678 2.455593          17.03729  106.3390      2.840169
2 12.27873   1.932676 2.244789          20.23803   94.5493      2.258873
3 13.15375   3.333750 2.437083          21.41667   99.3125      1.678750
  Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue
1  2.9823729             0.290000        1.899322        5.528305 1.0620339
2  2.0808451             0.363662        1.630282        3.086620 1.0562817
3  0.7814583             0.447500        1.153542        7.396250 0.6827083
  OD280.OD315   Proline
1    3.157797 1115.7119
2    2.785352  519.5070
3    1.683542  629.8958

Coefficients of linear discriminants:
                               LD1           LD2
Alcohol              -0.403399781  0.8717930699
Malic.acid            0.165254596  0.3053797325
Ash                  -0.369075256  2.3458497486
Alcalinity.of.ash     0.154797889 -0.1463807654
Magnesium            -0.002163496 -0.0004627565
Total.phenols         0.618052068 -0.0322128171
Flavanoids           -1.661191235 -0.4919980543
Nonflavanoid.phenols -1.495818440 -1.6309537953
Proanthocyanins       0.134092628 -0.3070875776
Color.intensity       0.355055710  0.2532306865
Hue                  -0.818036073 -1.5156344987
OD280.OD315          -1.157559376  0.0511839665
Proline              -0.002691206  0.0028529846

Proportion of trace:
   LD1    LD2
0.6875 0.3125
```

We can see that neither Magnesium or Proline seem useful to separate the wines; while Flavanoids and Nonflavanoid.phenols do. Ash is mainly used in the LD2.

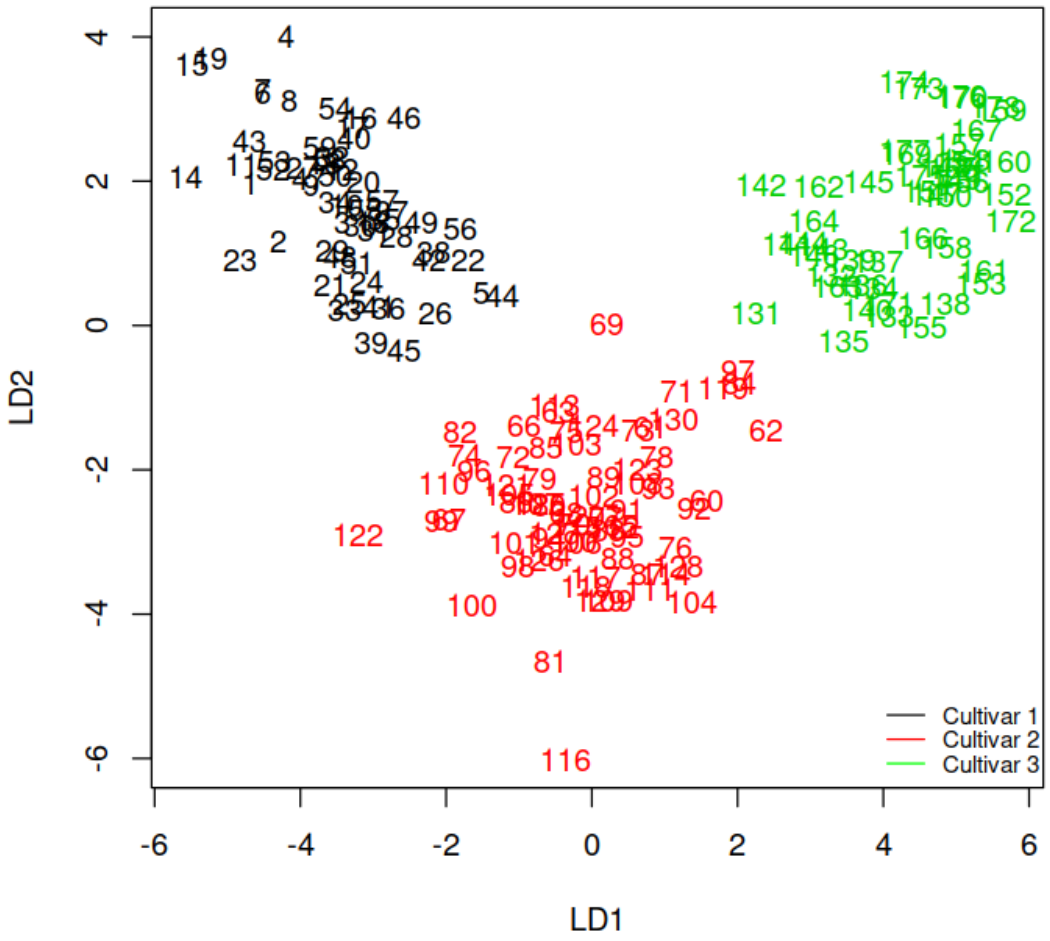
Plot the projected data in the first two LDs We can see that the discrimination is very good

```
In [8]: options(repr.plot.width=6, repr.plot.height=6)
        plot(lda.model)
```



alternatively, we can do it ourselves, with more control on color and text (wine number)

```
In [9]: wine.pred <- predict(lda.model)
        plot(wine.pred$x,type="n")
        text(wine.pred$x,labels=as.character(rownames(wine.pred$x)),col=as.integer(wine$Wine.typ
        legend('bottomright', c("Cultivar 1","Cultivar 2","Cultivar 3"), lty=1, col=c('black', '
```

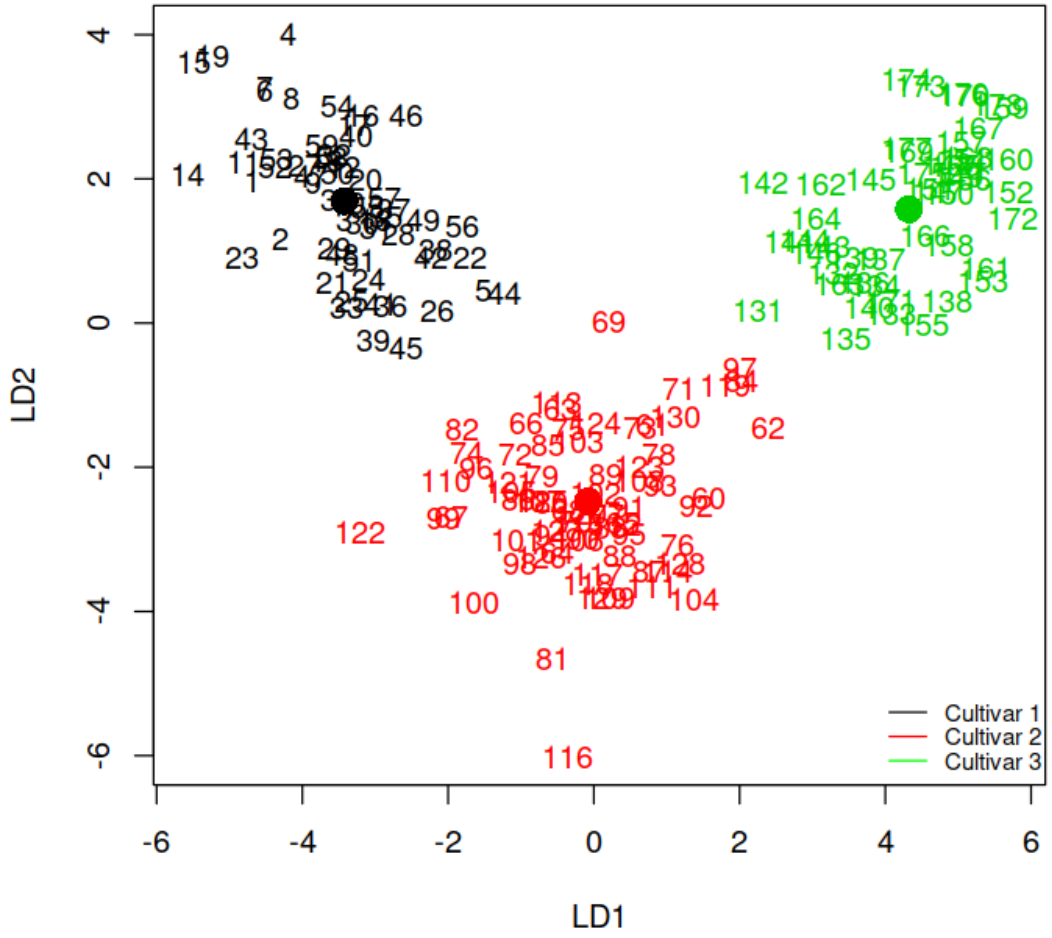


If need be, we can add the (projected) means to the plot

```
In [10]: plot(wine.pred$x,type="n")
         text(wine.pred$x,labels=as.character(rownames(wine.pred$x)),
              col=as.integer(wine$Wine.type))
         legend('bottomright', c("Cultivar 1","Cultivar 2","Cultivar 3"),
                lty=1, col=c('black', 'red', 'green'), bty='n', cex=.75)
         plot.mean <- function (class)
         {
           m1 <- mean(subset(wine.pred$x[,1],wine$Wine.type==class))
           m2 <- mean(subset(wine.pred$x[,2],wine$Wine.type==class))
           print(c(m1,m2))
           points(m1,m2,pch=16,cex=2,col=as.integer(class))
         }
```

```
plot.mean ('1')
plot.mean ('2')
plot.mean ('3')
```

```
[1] -3.422489  1.691674
[1] -0.07972623 -2.47265573
[1] 4.324737 1.578120
```



indeed classification is perfect

```
In [11]: table(wine$Wine.type, wine.pred$class)
```

```
    1  2  3
1  59  0  0
2   0 71  0
3   0  0 48
```

Let us switch to leave-one-out cross-validation

```
In [12]: wine.predcv <- update(lda.model,CV=TRUE)
         head(wine.predcv$posterior)
         print(table(wine$Wine.type,wine.predcv$class))
```

	1	2	3
1	1.0000000	2.797215e-09	2.071649e-18
2	0.9999996	4.380414e-07	7.695679e-17
3	0.9999970	3.024498e-06	1.071531e-13
4	1.0000000	1.251896e-12	1.095841e-16
5	0.8667524	1.332472e-01	3.998307e-07
6	1.0000000	2.236339e-11	1.017938e-17

```
    1  2  3
1  59  0  0
2   1 69  1
3   0  0 48
```

2 mistakes (on 178 observations): 1.12% error

Quadratic Discriminant Analysis is the same, replacing 'lda' by 'qda'

problems may arise if for some class there are less (or equal) observations than dimensions (is not the case for the wine data)

```
In [13]: qda.model <- qda (Wine.type ~ ., data = wine)

         qda.model

Call:
qda(Wine.type ~ ., data = wine)

Prior probabilities of groups:
        1         2         3
0.3314607 0.3988764 0.2696629

Group means:
   Alcohol Malic.acid      Ash Alcalinity.of.ash Magnesium Total.phenols
1 13.74475   2.010678 2.455593          17.03729  106.3390      2.840169
2 12.27873   1.932676 2.244789          20.23803   94.5493      2.258873
3 13.15375   3.333750 2.437083          21.41667   99.3125      1.678750
  Flavanoids Nonflavanoid.phenols Proanthocyanins Color.intensity       Hue
1  2.9823729             0.290000        1.899322        5.528305 1.0620339
```

```
2  2.0808451             0.363662        1.630282        3.086620 1.0562817
3  0.7814583             0.447500        1.153542        7.396250 0.6827083
  OD280.OD315   Proline
1    3.157797 1115.7119
2    2.785352  519.5070
3    1.683542  629.8958
```

There is no projection this time (because projection is a linear operator and the QDA boundaries are quadratic ones)

but let's have a look at classification:

```
In [14]: wine.pred <- predict(qda.model)
         table(wine$Wine.type, wine.pred$class)
```

```
    1  2  3
1 59  0  0
2  1 70  0
3  0  0 48
```

Let us switch to leave-one-out cross-validation

```
In [15]: wine.predcv <- update(qda.model,CV=TRUE)
         head(wine.predcv$posterior)

         print(table(wine$Wine.type,wine.predcv$class))
```

	1	2	3
1	1.0000000	8.920821e-12	4.507563e-103
2	1.0000000	1.769818e-09	4.454565e-92
3	0.9999999	8.114981e-08	6.097512e-110
4	1.0000000	3.912182e-19	1.829719e-137
5	0.9995416	4.583964e-04	4.452314e-53
6	1.0000000	4.786107e-18	1.777360e-134

```
    1  2  3
1 59  0  0
2  1 70  0
3  0  0 48
```

1 mistake (on 178 observations): 0.56% error

it would be nice to ascertain which wine is the "stubborn" one: it is a wine of type '2' classified as class '1'. Maybe there is something special with this wine ...

In the event of numerical errors (insufficient number of observations per class), we can use 'rda'

```
In [16]: library(klaR)
         (rda.model <- rda (Wine.type ~ ., data = wine))

Call:
rda(formula = Wine.type ~ ., data = wine)

Regularization parameters:
      gamma      lambda
1.000000000 0.001690939

Prior probabilities of groups:
        1         2         3
0.3314607 0.3988764 0.2696629

Misclassification rate:
       apparent: 27.528 %
cross-validated: 26.95 %
```

Look at the gamma and lambda coefficients, and note gamma=0, lambda=1 corresponds to LDA

1.2 Example 2: The Naïve Bayes classifier

```
In [17]: library (e1071)
```

Naive Bayes Classifier for Discrete Predictors: we use the 1984 United States Congressional Voting Records;

This data set includes votes for each of the U.S. House of Representatives Congressmen on 16 key votes In origin they were nine different types of votes:

- voted for, paired for, and announced for (these three simplified to yea or 'y'),
- voted against, paired against, and announced against (these three simplified to nay or 'n'),
- voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an 'unknown' disposition)

The goal is to classify Congressmen as Republican or Democrat as a function of their voting profiles, which is not immediate because in the US Congressmen have a large freedom of vote (obviously linked to their party but also to their own feelings, interests and compromises with voters)

```
In [18]: data (HouseVotes84, package="mlbench")
```

add meaningful names to the votes

```
In [19]: colnames(HouseVotes84) <- c("Class","handicapped.infants","water.project.sharing",
                                     "budget.resolution","physician.fee.freeze",
```

```
                                    "el.salvador.aid","religious.groups.in.schools",
                                    "anti.satellite.ban","aid.to.nicaraguan.contras",
                                    "mx.missile","immigration","synfuels.cutback",
                                    "education.spending","superfund","crime","duty.free.exports
                                    "export.South.Africa")

        summary(HouseVotes84)
```

```
       Class     handicapped.infants water.project.sharing budget.resolution
democrat  :267   n   :236            n   :192              n   :171
republican:168   y   :187            y   :195              y   :253
                 NA's: 12            NA's: 48              NA's: 11
physician.fee.freeze el.salvador.aid religious.groups.in.schools
n   :247             n   :208        n   :152
y   :177             y   :212        y   :272
NA's: 11             NA's: 15        NA's: 11
anti.satellite.ban aid.to.nicaraguan.contras mx.missile immigration
n   :182           n   :178                  n   :206   n   :212
y   :239           y   :242                  y   :207   y   :216
NA's: 14           NA's: 15                 NA's: 22   NA's:  7
synfuels.cutback education.spending superfund   crime     duty.free.exports
n   :264         n   :233           n   :201   n   :170   n   :233
y   :150         y   :171           y   :209   y   :248   y   :174
NA's: 21         NA's: 31          NA's: 25   NA's: 17   NA's: 28
export.South.Africa
n   : 62
y   :269
NA's:104
```

1 = democrat, 0 = republican Note "unknown dispositions" have been treated as missing values!

```
In [20]: set.seed(1111)

         N <- nrow(HouseVotes84)
```

We first split the available data into learning and test sets, selecting randomly 2/3 and 1/3 of the data.

We do this for a honest estimation of prediction performance

```
In [21]: learn <- sample(1:N, round(2*N/3))

         nlearn <- length(learn)
         ntest <- N - nlearn
```

First we build a model using the learn data

```
In [22]: model <- naiveBayes(Class ~ ., data = HouseVotes84[learn,])
```

we get all the probabilities

```
In [23]: model
```

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
  democrat republican
 0.6344828  0.3655172

Conditional probabilities:
            handicapped.infants
Y                    n         y
  democrat   0.4034091 0.5965909
  republican 0.8076923 0.1923077

            water.project.sharing
Y                    n         y
  democrat   0.5000000 0.5000000
  republican 0.4468085 0.5531915

            budget.resolution
Y                    n         y
  democrat   0.1073446 0.8926554
  republican 0.8653846 0.1346154

            physician.fee.freeze
Y                     n          y
  democrat   0.94350282 0.05649718
  republican 0.01923077 0.98076923

            el.salvador.aid
Y                     n          y
  democrat   0.78160920 0.21839080
  republican 0.04854369 0.95145631

            religious.groups.in.schools
Y                     n          y
  democrat   0.50282486 0.49717514
  republican 0.06730769 0.93269231

            anti.satellite.ban
Y                    n         y
```

```
  democrat   0.2290503 0.7709497
  republican 0.7623762 0.2376238

            aid.to.nicaraguan.contras
Y                    n         y
  democrat   0.1722222 0.8277778
  republican 0.8484848 0.1515152

            mx.missile
Y                     n          y
  democrat   0.25581395 0.74418605
  republican 0.92307692 0.07692308

            immigration
Y                    n         y
  democrat   0.5054945 0.4945055
  republican 0.4038462 0.5961538

            synfuels.cutback
Y                    n         y
  democrat   0.4514286 0.5485714
  republican 0.8787879 0.1212121

            education.spending
Y                    n         y
  democrat   0.8546512 0.1453488
  republican 0.1250000 0.8750000

            superfund
Y                    n         y
  democrat   0.7109827 0.2890173
  republican 0.1313131 0.8686869

            crime
Y                     n          y
  democrat   0.66101695 0.33898305
  republican 0.01010101 0.98989899

            duty.free.exports
Y                    n         y
  democrat   0.3885714 0.6114286
  republican 0.8888889 0.1111111

            export.South.Africa
Y                     n          y
  democrat   0.07518797 0.92481203
  republican 0.41111111 0.58888889
```

predict the outcome of the first 20 Congressmen

```
In [24]: predict(model, HouseVotes84[1:20,-1])
```

1. republican 2. republican 3. republican 4. democrat 5. democrat 6. democrat 7. republican 8. republican 9. republican 10. democrat 11. republican 12. republican 13. democrat 14. democrat 15. republican 16. republican 17. democrat 18. democrat 19. republican 20. democrat

Levels: 1. 'democrat' 2. 'republican'

same but displaying posterior probabilities

```
In [25]: predict(model, HouseVotes84[1:20,-1], type = "raw")
```

democrat	republican
1.247826e-07	9.999999e-01
6.159235e-08	9.999999e-01
7.532654e-03	9.924673e-01
9.992485e-01	7.515351e-04
9.480874e-01	5.191264e-02
6.952961e-01	3.047039e-01
1.506125e-04	9.998494e-01
9.107411e-06	9.999909e-01
9.672681e-08	9.999999e-01
1.000000e+00	1.965474e-11
1.850068e-06	9.999981e-01
7.472722e-06	9.999925e-01
1.000000e+00	1.986725e-09
1.000000e+00	5.671126e-10
5.406035e-07	9.999995e-01
1.191933e-07	9.999999e-01
9.999987e-01	1.292354e-06
1.000000e+00	3.345366e-11
6.781291e-08	9.999999e-01
1.000000e+00	2.581527e-13

compute now the apparent error

```
In [26]: pred <- predict(model, HouseVotes84[learn,-1])
```

form and display confusion matrix & overall error

```
In [27]: tab <- table(pred, HouseVotes84[learn,]$Class)
         tab
         1 - sum(tab[row(tab)==col(tab)])/sum(tab)
```

```
pred         democrat republican
  democrat        164         10
  republican       20         96
```

0.103448275862069
compute the test (prediction) error

```
In [28]: pred <- predict(model, newdata=HouseVotes84[-learn,-1])
```

form and display confusion matrix & overall error

```
In [29]: tab <- table(pred, HouseVotes84[-learn,]$Class)
         tab
         1 - sum(tab[row(tab)==col(tab)])/sum(tab)
```

```
pred         democrat republican
  democrat         74          3
  republican        9         59
```

0.0827586206896552
note how most errors (9/12) correspond to democrats wrongly predicted as republicans
in the event of **empty empirical probabilities**, this is how we would setup Laplace correction (aka smoothing):

```
In [30]: model <- naiveBayes(Class ~ ., data = HouseVotes84[learn,], laplace = 1)
```

1.3 Example 3: The kNN classifier

We are going to use the famous (Fisher's or Anderson's) Iris data set, which gives the measurements in centimeters of the sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of Iris. The species are Iris setosa, versicolor, and virginica.

```
In [31]: library (class)
         data(iris3)
```

first we split a separate test set of relative size 30%

```
In [32]: learn.inputs   <- rbind(iris3[1:35,,1], iris3[1:35,,2], iris3[1:35,,3])
         learn.classes <- factor(c(rep("s",35), rep("c",35), rep("v",35)))

         test.inputs  <- rbind(iris3[36:50,,1], iris3[36:50,,2], iris3[36:50,,3])
         test.classes <- factor(c(rep("s",15), rep("c",15), rep("v",15)))
```

setup a kNN model with 3 neighbours Notice there is no "learning" ... the data is the model (just test!)

```
In [33]: myknn <- knn (learn.inputs, test.inputs, learn.classes, k = 3, prob=TRUE)

         tab <- table(myknn, test.classes)
         1 - sum(tab[row(tab)==col(tab)])/sum(tab)
         tab
```

0

```
     test.classes
myknn  c  s  v
    c 15  0  0
    s  0 15  0
    v  0  0 15
```

rows are predictions, columns are true test targets
one can use the function 'knn1()' when k=1 (just one neighbour)
How do we optimize k? One way is by using LOOCV

In [34]:
```
myknn.cv <- knn.cv (learn.inputs, learn.classes, k = 3)

tab <- table(myknn.cv, learn.classes)
1 - sum(tab[row(tab)==col(tab)])/sum(tab)
```

0.0571428571428572
aha! now you see that previous training error (0%) was a little bit optimistic
Let's loop over k

In [35]:
```
set.seed (23)

neighbours <- c(1:sqrt(nrow(learn.inputs)))
errors <- matrix (nrow=length(neighbours), ncol=2)
colnames(errors) <- c("k","LOOCV error")

for (k in neighbours)
{
  myknn.cv <- knn.cv (learn.inputs, learn.classes, k = neighbours[k])

  # fill in no. of neighbours and LOO validation error
  errors[k, "k"] <- neighbours[k]

  tab <- table(myknn.cv, learn.classes)
  errors[k, "LOOCV error"] <- 1 - sum(tab[row(tab)==col(tab)])/sum(tab)
}

errors
```

k	LOOCV error
1	0.05714286
2	0.09523810
3	0.05714286
4	0.05714286
5	0.06666667
6	0.05714286
7	0.05714286
8	0.04761905
9	0.06666667
10	0.06666667

It seems that k=8 is the best value.

Now we *refit* with k=8 and predict the test set

```
In [36]: myknn <- knn (learn.inputs, test.inputs, learn.classes, k = 8, prob=TRUE)
```

```
tab <- table(myknn, test.classes)
1 - sum(tab[row(tab)==col(tab)])/sum(tab)
tab
```

0.0222222222222223

```
     test.classes
myknn  c  s  v
    c 15  0  1
    s  0 15  0
    v  0  0 14
```

so our error is 2.2%

1.4 Example 4: Logistic Regression using artificial data

The goal of this example is to get acquainted with the call to glm() glm() is used to fit generalized linear models (of which both linear and logistic regression are particular cases)

You may need to recall at this point the logistic regression model ...

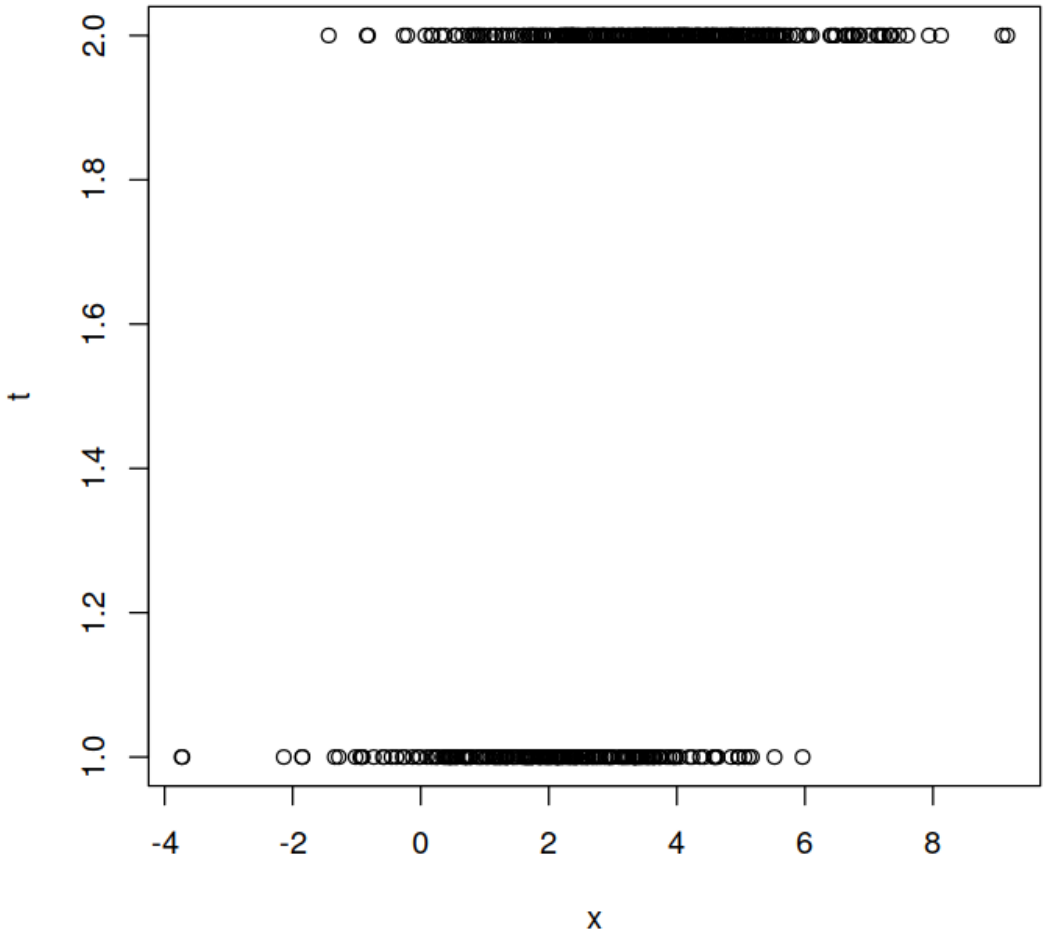
Let x represent a single continuous predictor

Let y represent a class ('0' or '1'), with a probability of being 1 that is related linearly to the predictor via the logit funtion, that is $logit(p) = a * x + b$ (or $beta_1 * x + beta_0$ if you prefer)

```
In [37]: set.seed (1968)
```

```
N <- 500
x <- rnorm(n=N, mean=3, sd=2)      # generate the x_n (note x is a vector)
a <- 0.6 ; b <- -1.5               # this is the ground truth, which is unknown
```

```
p <- 1/(1+exp( -(a*x + b) ))      # generate the p_n (note p is a vector)
t <- rbinom(n=N,size=1,prob=p)
t <- as.factor(t)                 # generate the targets according to p

plot(x,t)
```




```
In [38]: glm.res <- glm (t~x, family = binomial)
```

look at the coefficients!
'Intercept' is b, 'x' is a

```
In [39]: summary(glm.res)
```

```
Call:
glm(formula = t ~ x, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1174  -1.0121   0.5127   0.9089   2.1940

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.44926    0.21178  -6.843 7.75e-12 ***
x            0.60011    0.06687   8.974  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 683.87  on 499  degrees of freedom
Residual deviance: 571.30  on 498  degrees of freedom
AIC: 575.3

Number of Fisher Scoring iterations: 4
```

Obviously x is very significant (and the Intercept is always significant)

Therefore, our estimated model is $logit(p_n) = 0.60011 * x_n - 1.44926$ quite close to the ground truth

In general you get this as:

- glm.res$coefficients["x"]
- glm.res$coefficients["(Intercept)"]

Interpretation of the coefficients:

- For a 1 unit increase in x, there is an increase in the odds for t by a factor of ...

```
In [40]: exp(glm.res$coefficients["x"])
```

x: 1.82231020552275

that is almost doubling the odds (82.2% more)

1.5 Example 5: Logistic regression for classifying spam mail

This example will also illustrate how to change the 'cut point' for prediction, when there is an interest in minimizing a particular source of errors

```
In [41]: library(kernlab)

         data(spam)
```

Type `help(spam)` for some basic information about the dataset
We do some basic pre-processing

```
In [42]: spam[,55:57] <- as.matrix(log10(spam[,55:57]+1))

         spam2 <- spam[spam$george==0,]
         spam2 <- spam2[spam2$num650==0,]
         spam2 <- spam2[spam2$hp==0,]
         spam2 <- spam2[spam2$hpl==0,]

         george.vars <- 25:28
         spam2 <- spam2[,-george.vars]

         moneys.vars <- c(16,17,20,24)
         spam3 <- data.frame( spam2[,-moneys.vars], spam2[,16]+spam2[,17]+spam2[,20]+spam2[,24])

         colnames(spam3)[51] <- "about.money"

         dim(spam3)
```

1. 2999 2. 51

```
In [43]: set.seed (4321)
         N <- nrow(spam3)
         learn <- sample(1:N, round(0.67*N))
         nlearn <- length(learn)
         ntest <- N - nlearn
```

Fit a GLM in the learning data

```
In [44]: spamM1 <- glm (type ~ ., data=spam3[learn,], family=binomial)
```

```
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Simplify it using the AIC (this may take a while, since there are many variables)

```
In [45]: suppressWarnings(spamM1.AIC <- step (spamM1))
```

```
Start:  AIC=940.25
type ~ make + address + all + num3d + our + over + remove + internet +
    order + mail + receive + will + people + report + addresses +
    email + you + your + font + num000 + lab + labs + telnet +
    num857 + data + num415 + num85 + technology + num1999 + parts +
    pm + direct + cs + meeting + original + project + re + edu +
```

```
    table + conference + charSemicolon + charRoundbracket + charSquarebracket +
    charExclamation + charDollar + charHash + capitalAve + capitalLong +
    capitalTotal + about.money

                    Df Deviance     AIC
- font               1   838.27  938.27
- num1999            1   838.31  938.31
- order              1   838.31  938.31
- telnet             1   838.32  938.32
- charSquarebracket  1   838.33  938.33
- report             1   838.39  938.39
- all                1   838.41  938.41
- num415             1   838.46  938.46
- direct             1   838.58  938.58
- num857             1   838.62  938.62
- charHash           1   838.71  938.71
- people             1   838.78  938.78
- parts              1   838.88  938.88
- addresses          1   838.98  938.98
- table              1   839.09  939.09
- capitalLong        1   839.17  939.17
- num3d              1   839.32  939.32
- labs               1   839.48  939.48
- original           1   839.61  939.61
- will               1   839.83  939.83
- email              1   840.03  940.03
- address            1   840.18  940.18
<none>                   838.25  940.25
- make               1   840.55  940.55
- over               1   840.75  940.75
- num85              1   840.92  940.92
- receive            1   841.74  941.74
- mail               1   842.45  942.45
- charRoundbracket   1   842.90  942.90
- your               1   843.30  943.30
- technology         1   844.88  944.88
- internet           1   844.89  944.89
- pm                 1   845.36  945.36
- num000             1   845.42  945.42
- charExclamation    1   846.96  946.96
- data               1   847.27  947.27
- lab                1   847.71  947.71
- project            1   848.33  948.33
- capitalAve         1   848.69  948.69
- you                1   848.95  948.95
- meeting            1   855.17  955.17
- charSemicolon      1   855.24  955.24
- re                 1   856.76  956.76
```

```
- cs                1   856.98  956.98
- conference        1   859.08  959.08
- charDollar        1   866.10  966.10
- our               1   867.79  967.79
- remove            1   875.27  975.27
- about.money       1   886.48  986.48
- capitalTotal      1   907.88 1007.88
- edu               1   937.89 1037.89

Step:  AIC=938.27
type ~ make + address + all + num3d + our + over + remove + internet +
    order + mail + receive + will + people + report + addresses +
    email + you + your + num000 + lab + labs + telnet + num857 +
    data + num415 + num85 + technology + num1999 + parts + pm +
    direct + cs + meeting + original + project + re + edu + table +
    conference + charSemicolon + charRoundbracket + charSquarebracket +
    charExclamation + charDollar + charHash + capitalAve + capitalLong +
    capitalTotal + about.money

                   Df Deviance     AIC
- num1999           1   838.33  936.33
- order             1   838.34  936.34
- telnet            1   838.35  936.35
- charSquarebracket 1   838.35  936.35
- report            1   838.41  936.41
- all               1   838.44  936.44
- num415            1   838.49  936.49
- direct            1   838.59  936.59
- num857            1   838.65  936.65
- charHash          1   838.73  936.73
- people            1   838.80  936.80
- parts             1   838.91  936.91
- addresses         1   839.02  937.02
- table             1   839.11  937.11
- capitalLong       1   839.18  937.18
- num3d             1   839.35  937.35
- labs              1   839.52  937.52
- original          1   839.64  937.64
- will              1   839.84  937.84
- email             1   840.05  938.05
- address           1   840.22  938.22
<none>                  838.27  938.27
- make              1   840.57  938.57
- over              1   840.77  938.77
- num85             1   840.94  938.94
- receive           1   841.75  939.75
- mail              1   842.50  940.50
- charRoundbracket  1   842.90  940.90
```

```
- your               1   843.33  941.33
- internet           1   844.92  942.92
- technology         1   844.93  942.93
- pm                 1   845.38  943.38
- num000             1   845.45  943.45
- charExclamation    1   846.99  944.99
- data               1   847.29  945.29
- lab                1   847.75  945.75
- project            1   848.36  946.36
- capitalAve         1   848.73  946.73
- you                1   849.03  947.03
- meeting            1   855.20  953.20
- re                 1   856.82  954.82
- cs                 1   856.98  954.98
- conference         1   859.08  957.08
- charDollar         1   866.40  964.40
- our                1   867.84  965.84
- charSemicolon      1   871.99  969.99
- remove             1   875.27  973.27
- about.money        1   886.70  984.70
- capitalTotal       1   907.97 1005.97
- edu                1   937.89 1035.89

Step:  AIC=936.33
type ~ make + address + all + num3d + our + over + remove + internet +
    order + mail + receive + will + people + report + addresses +
    email + you + your + num000 + lab + labs + telnet + num857 +
    data + num415 + num85 + technology + parts + pm + direct +
    cs + meeting + original + project + re + edu + table + conference +
    charSemicolon + charRoundbracket + charSquarebracket + charExclamation +
    charDollar + charHash + capitalAve + capitalLong + capitalTotal +
    about.money

                   Df Deviance     AIC
- order              1   838.39  934.39
- telnet             1   838.40  934.40
- charSquarebracket  1   838.40  934.40
- report             1   838.47  934.47
- all                1   838.48  934.48
- num415             1   838.54  934.54
- direct             1   838.64  934.64
- num857             1   838.70  934.70
- charHash           1   838.81  934.81
- people             1   838.85  934.85
- parts              1   838.96  934.96
- addresses          1   839.07  935.07
- table              1   839.17  935.17
- capitalLong        1   839.21  935.21
```

```
- num3d              1   839.41  935.41
- labs               1   839.58  935.58
- original           1   839.83  935.83
- will               1   839.91  935.91
- email              1   840.12  936.12
- address            1   840.26  936.26
<none>                   838.33  936.33
- make               1   840.63  936.63
- over               1   840.82  936.82
- num85              1   840.99  936.99
- receive            1   841.79  937.79
- mail               1   842.64  938.64
- charRoundbracket   1   843.27  939.27
- your               1   843.45  939.45
- internet           1   845.01  941.01
- technology         1   845.05  941.05
- pm                 1   845.41  941.41
- num000             1   845.55  941.55
- charExclamation    1   847.03  943.03
- data               1   847.41  943.41
- lab                1   847.77  943.77
- project            1   848.43  944.43
- capitalAve         1   848.76  944.76
- you                1   849.09  945.09
- meeting            1   855.21  951.21
- re                 1   856.83  952.83
- cs                 1   857.02  953.02
- conference         1   859.73  955.73
- charDollar         1   866.66  962.66
- our                1   867.95  963.95
- charSemicolon      1   872.00  968.00
- remove             1   875.50  971.50
- about.money        1   887.36  983.36
- capitalTotal       1   907.97 1003.97
- edu                1   939.85 1035.85

Step:  AIC=934.39
type ~ make + address + all + num3d + our + over + remove + internet +
    mail + receive + will + people + report + addresses + email +
    you + your + num000 + lab + labs + telnet + num857 + data +
    num415 + num85 + technology + parts + pm + direct + cs +
    meeting + original + project + re + edu + table + conference +
    charSemicolon + charRoundbracket + charSquarebracket + charExclamation +
    charDollar + charHash + capitalAve + capitalLong + capitalTotal +
    about.money

                    Df Deviance     AIC
- telnet             1   838.47  932.47
```

```
- charSquarebracket  1   838.47  932.47
- report             1   838.53  932.53
- all                1   838.54  932.54
- num415             1   838.60  932.60
- direct             1   838.71  932.71
- num857             1   838.76  932.76
- charHash           1   838.88  932.88
- people             1   838.94  932.94
- parts              1   839.02  933.02
- addresses          1   839.14  933.14
- capitalLong        1   839.24  933.24
- table              1   839.25  933.25
- num3d              1   839.46  933.46
- labs               1   839.64  933.64
- original           1   839.89  933.89
- will               1   840.00  934.00
- email              1   840.16  934.16
- address            1   840.32  934.32
<none>                   838.39  934.39
- make               1   840.70  934.70
- over               1   840.86  934.86
- num85              1   841.04  935.04
- receive            1   841.89  935.89
- mail               1   842.76  936.76
- charRoundbracket   1   843.33  937.33
- your               1   843.89  937.89
- internet           1   845.08  939.08
- technology         1   845.10  939.10
- pm                 1   845.46  939.46
- num000             1   845.59  939.59
- charExclamation    1   847.14  941.14
- data               1   847.54  941.54
- lab                1   847.84  941.84
- project            1   848.51  942.51
- capitalAve         1   848.76  942.76
- you                1   849.10  943.10
- meeting            1   855.31  949.31
- cs                 1   857.12  951.12
- re                 1   857.13  951.13
- conference         1   859.92  953.92
- charDollar         1   866.95  960.95
- our                1   867.95  961.95
- charSemicolon      1   872.11  966.11
- remove             1   875.52  969.52
- about.money        1   887.55  981.55
- capitalTotal       1   907.99 1001.99
- edu                1   940.64 1034.64
```

```
Step:  AIC=932.47
type ~ make + address + all + num3d + our + over + remove + internet +
    mail + receive + will + people + report + addresses + email +
    you + your + num000 + lab + labs + num857 + data + num415 +
    num85 + technology + parts + pm + direct + cs + meeting +
    original + project + re + edu + table + conference + charSemicolon +
    charRoundbracket + charSquarebracket + charExclamation +
    charDollar + charHash + capitalAve + capitalLong + capitalTotal +
    about.money

                    Df Deviance     AIC
- charSquarebracket  1   838.54  930.54
- report             1   838.61  930.61
- all                1   838.62  930.62
- num415             1   838.67  930.67
- direct             1   838.79  930.79
- num857             1   838.84  930.84
- charHash           1   838.96  930.96
- people             1   839.01  931.01
- parts              1   839.09  931.09
- addresses          1   839.21  931.21
- capitalLong        1   839.31  931.31
- table              1   839.32  931.32
- num3d              1   839.54  931.54
- labs               1   839.71  931.71
- original           1   839.97  931.97
- will               1   840.07  932.07
- email              1   840.23  932.23
- address            1   840.39  932.39
<none>                   838.47  932.47
- make               1   840.77  932.77
- over               1   840.94  932.94
- num85              1   841.12  933.12
- receive            1   841.96  933.96
- mail               1   842.84  934.84
- charRoundbracket   1   843.39  935.39
- your               1   843.97  935.97
- internet           1   845.16  937.16
- technology         1   845.18  937.18
- pm                 1   845.53  937.53
- num000             1   845.66  937.66
- charExclamation    1   847.22  939.22
- data               1   847.61  939.61
- lab                1   847.92  939.92
- capitalAve         1   848.81  940.81
- project            1   849.09  941.09
- you                1   849.20  941.20
- meeting            1   855.37  947.37
```

```
- re                1   857.19  949.19
- cs                1   857.20  949.20
- conference        1   860.00  952.00
- charDollar        1   867.02  959.02
- our               1   868.04  960.04
- charSemicolon     1   872.18  964.18
- remove            1   875.62  967.62
- about.money       1   887.64  979.64
- capitalTotal      1   908.26 1000.26
- edu               1   940.69 1032.69

Step:  AIC=930.54
type ~ make + address + all + num3d + our + over + remove + internet +
    mail + receive + will + people + report + addresses + email +
    you + your + num000 + lab + labs + num857 + data + num415 +
    num85 + technology + parts + pm + direct + cs + meeting +
    original + project + re + edu + table + conference + charSemicolon +
    charRoundbracket + charExclamation + charDollar + charHash +
    capitalAve + capitalLong + capitalTotal + about.money

                  Df Deviance     AIC
- report           1   838.69  928.69
- all              1   838.69  928.69
- num415           1   838.75  928.75
- direct           1   838.86  928.86
- num857           1   838.91  928.91
- charHash         1   839.03  929.03
- people           1   839.08  929.08
- parts            1   839.17  929.17
- addresses        1   839.29  929.29
- capitalLong      1   839.36  929.36
- table            1   839.40  929.40
- num3d            1   839.62  929.62
- labs             1   839.78  929.78
- original         1   840.04  930.04
- will             1   840.12  930.12
- email            1   840.30  930.30
- address          1   840.45  930.45
<none>                 838.54  930.54
- make             1   840.83  930.83
- over             1   841.00  931.00
- num85            1   841.18  931.18
- receive          1   842.03  932.03
- mail             1   842.92  932.92
- charRoundbracket 1   843.51  933.51
- your             1   844.05  934.05
- internet         1   845.26  935.26
- technology       1   845.30  935.30
```

```
- pm                1   845.61  935.61
- num000            1   845.69  935.69
- charExclamation   1   847.35  937.35
- data              1   847.74  937.74
- lab               1   848.01  938.01
- capitalAve        1   848.82  938.82
- project           1   849.18  939.18
- you               1   849.34  939.34
- meeting           1   855.64  945.64
- cs                1   857.26  947.26
- re                1   857.63  947.63
- conference        1   860.10  950.10
- charDollar        1   867.42  957.42
- our               1   868.21  958.21
- charSemicolon     1   872.18  962.18
- remove            1   875.74  965.74
- about.money       1   887.88  977.88
- capitalTotal      1   908.28  998.28
- edu               1   940.74 1030.74

Step:  AIC=928.69
type ~ make + address + all + num3d + our + over + remove + internet +
    mail + receive + will + people + addresses + email + you +
    your + num000 + lab + labs + num857 + data + num415 + num85 +
    technology + parts + pm + direct + cs + meeting + original +
    project + re + edu + table + conference + charSemicolon +
    charRoundbracket + charExclamation + charDollar + charHash +
    capitalAve + capitalLong + capitalTotal + about.money

                  Df Deviance     AIC
- all              1   838.84  926.84
- num415           1   838.89  926.89
- direct           1   839.01  927.01
- num857           1   839.06  927.06
- charHash         1   839.17  927.17
- people           1   839.19  927.19
- parts            1   839.32  927.32
- addresses        1   839.44  927.44
- capitalLong      1   839.49  927.49
- table            1   839.53  927.53
- num3d            1   839.76  927.76
- labs             1   839.92  927.92
- original         1   840.19  928.19
- will             1   840.24  928.24
- email            1   840.41  928.41
- address          1   840.62  928.62
<none>                 838.69  928.69
- make             1   841.00  929.00
```

```
- over              1   841.12  929.12
- num85             1   841.34  929.34
- receive           1   842.17  930.17
- mail              1   843.07  931.07
- charRoundbracket  1   843.64  931.64
- your              1   844.15  932.15
- internet          1   845.36  933.36
- num000            1   845.85  933.85
- pm                1   845.85  933.85
- technology        1   846.22  934.22
- charExclamation   1   847.44  935.44
- data              1   847.95  935.95
- lab               1   848.17  936.17
- capitalAve        1   848.96  936.96
- project           1   849.39  937.39
- you               1   849.41  937.41
- meeting           1   855.81  943.81
- cs                1   857.41  945.41
- re                1   857.76  945.76
- conference        1   860.33  948.33
- charDollar        1   867.61  955.61
- our               1   868.37  956.37
- charSemicolon     1   872.39  960.39
- remove            1   875.77  963.77
- about.money       1   888.29  976.29
- capitalTotal      1   908.30  996.30
- edu               1   941.56 1029.56

Step:  AIC=926.84
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + people + addresses + email + you +
    your + num000 + lab + labs + num857 + data + num415 + num85 +
    technology + parts + pm + direct + cs + meeting + original +
    project + re + edu + table + conference + charSemicolon +
    charRoundbracket + charExclamation + charDollar + charHash +
    capitalAve + capitalLong + capitalTotal + about.money

                  Df Deviance     AIC
- num415           1   839.04  925.04
- direct           1   839.17  925.17
- num857           1   839.20  925.20
- charHash         1   839.34  925.34
- people           1   839.38  925.38
- parts            1   839.46  925.46
- addresses        1   839.58  925.58
- capitalLong      1   839.63  925.63
- table            1   839.79  925.79
- num3d            1   839.93  925.93
```

```
- labs              1   840.07  926.07
- original          1   840.31  926.31
- will              1   840.45  926.45
- email             1   840.60  926.60
- address           1   840.78  926.78
<none>                  838.84  926.84
- make              1   841.21  927.21
- over              1   841.34  927.34
- num85             1   841.45  927.45
- receive           1   842.24  928.24
- mail              1   843.21  929.21
- charRoundbracket  1   843.88  929.88
- your              1   844.23  930.23
- internet          1   845.60  931.60
- num000            1   845.98  931.98
- pm                1   846.22  932.22
- technology        1   846.47  932.47
- charExclamation   1   847.55  933.55
- data              1   848.00  934.00
- lab               1   848.23  934.23
- capitalAve        1   849.22  935.22
- you               1   849.43  935.43
- project           1   849.43  935.43
- meeting           1   855.96  941.96
- cs                1   857.54  943.54
- re                1   857.79  943.79
- conference        1   860.38  946.38
- charDollar        1   867.80  953.80
- our               1   868.59  954.59
- charSemicolon     1   872.39  958.39
- remove            1   876.10  962.10
- about.money       1   888.31  974.31
- capitalTotal      1   908.70  994.70
- edu               1   941.58 1027.58

Step:  AIC=925.04
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + people + addresses + email + you +
    your + num000 + lab + labs + num857 + data + num85 + technology +
    parts + pm + direct + cs + meeting + original + project +
    re + edu + table + conference + charSemicolon + charRoundbracket +
    charExclamation + charDollar + charHash + capitalAve + capitalLong +
    capitalTotal + about.money

                  Df Deviance     AIC
- direct           1   839.38  923.38
- num857           1   839.40  923.40
- charHash         1   839.54  923.54
```

```
- people            1   839.59  923.59
- parts             1   839.66  923.66
- addresses         1   839.78  923.78
- capitalLong       1   839.82  923.82
- table             1   839.99  923.99
- num3d             1   840.13  924.13
- labs              1   840.27  924.27
- original          1   840.52  924.52
- will              1   840.68  924.68
- email             1   840.79  924.79
- address           1   840.98  924.98
<none>                  839.04  925.04
- make              1   841.34  925.34
- over              1   841.53  925.53
- num85             1   841.66  925.66
- receive           1   842.51  926.51
- mail              1   843.46  927.46
- charRoundbracket  1   844.12  928.12
- your              1   844.43  928.43
- internet          1   845.80  929.80
- num000            1   846.15  930.15
- pm                1   846.42  930.42
- technology        1   846.66  930.66
- charExclamation   1   847.94  931.94
- data              1   848.24  932.24
- lab               1   848.44  932.44
- capitalAve        1   849.40  933.40
- you               1   849.60  933.60
- project           1   849.65  933.65
- meeting           1   856.17  940.17
- cs                1   857.78  941.78
- re                1   858.15  942.15
- conference        1   860.63  944.63
- charDollar        1   868.03  952.03
- our               1   868.74  952.74
- charSemicolon     1   872.69  956.69
- remove            1   876.23  960.23
- about.money       1   888.40  972.40
- capitalTotal      1   909.01  993.01
- edu               1   941.93 1025.93

Step:  AIC=923.38
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + people + addresses + email + you +
    your + num000 + lab + labs + num857 + data + num85 + technology +
    parts + pm + cs + meeting + original + project + re + edu +
    table + conference + charSemicolon + charRoundbracket + charExclamation +
    charDollar + charHash + capitalAve + capitalLong + capitalTotal +
```

```
    about.money

                   Df Deviance     AIC
- charHash          1   839.71  921.71
- num857            1   839.73  921.73
- people            1   839.93  921.93
- parts             1   839.98  921.98
- addresses         1   840.10  922.10
- capitalLong       1   840.17  922.17
- table             1   840.37  922.37
- num3d             1   840.48  922.48
- labs              1   840.61  922.61
- original          1   840.85  922.85
- will              1   840.96  922.96
- email             1   841.10  923.10
- address           1   841.29  923.29
<none>                  839.38  923.38
- make              1   841.64  923.64
- over              1   841.90  923.90
- num85             1   842.04  924.04
- receive           1   842.95  924.95
- mail              1   843.79  925.79
- charRoundbracket  1   844.34  926.34
- your              1   844.88  926.88
- internet          1   846.19  928.19
- num000            1   846.42  928.42
- pm                1   846.75  928.75
- technology        1   847.03  929.03
- charExclamation   1   848.34  930.34
- data              1   848.50  930.50
- lab               1   848.79  930.79
- capitalAve        1   849.76  931.76
- project           1   849.88  931.88
- you               1   850.20  932.20
- meeting           1   856.48  938.48
- cs                1   858.15  940.15
- re                1   858.44  940.44
- conference        1   860.98  942.98
- charDollar        1   868.39  950.39
- our               1   869.15  951.15
- charSemicolon     1   872.73  954.73
- remove            1   876.26  958.26
- about.money       1   888.82  970.82
- capitalTotal      1   909.68  991.68
- edu               1   941.94 1023.94

Step:  AIC=921.71
type ~ make + address + num3d + our + over + remove + internet +
```

```
    mail + receive + will + people + addresses + email + you +
    your + num000 + lab + labs + num857 + data + num85 + technology +
    parts + pm + cs + meeting + original + project + re + edu +
    table + conference + charSemicolon + charRoundbracket + charExclamation +
    charDollar + capitalAve + capitalLong + capitalTotal + about.money

                   Df Deviance     AIC
- num857            1   840.06  920.06
- people            1   840.29  920.29
- parts             1   840.31  920.31
- addresses         1   840.43  920.43
- capitalLong       1   840.54  920.54
- table             1   840.71  920.71
- num3d             1   840.82  920.82
- labs              1   840.94  920.94
- original          1   841.19  921.19
- will              1   841.26  921.26
- email             1   841.42  921.42
- address           1   841.62  921.62
<none>                  839.71  921.71
- make              1   842.00  922.00
- over              1   842.32  922.32
- num85             1   842.41  922.41
- receive           1   843.30  923.30
- mail              1   844.07  924.07
- charRoundbracket  1   844.74  924.74
- your              1   845.15  925.15
- internet          1   846.52  926.52
- num000            1   846.82  926.82
- pm                1   847.12  927.12
- technology        1   847.53  927.53
- charExclamation   1   848.70  928.70
- data              1   848.86  928.86
- lab               1   849.19  929.19
- project           1   850.11  930.11
- you               1   850.76  930.76
- capitalAve        1   850.86  930.86
- meeting           1   856.91  936.91
- cs                1   858.65  938.65
- re                1   858.99  938.99
- conference        1   861.57  941.57
- charDollar        1   868.89  948.89
- our               1   869.43  949.43
- charSemicolon     1   872.84  952.84
- remove            1   876.90  956.90
- about.money       1   889.13  969.13
- capitalTotal      1   911.70  991.70
- edu               1   942.16 1022.16
```

```
Step:  AIC=920.06
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + people + addresses + email + you +
    your + num000 + lab + labs + data + num85 + technology +
    parts + pm + cs + meeting + original + project + re + edu +
    table + conference + charSemicolon + charRoundbracket + charExclamation +
    charDollar + capitalAve + capitalLong + capitalTotal + about.money

                  Df Deviance     AIC
- people           1   840.65  918.65
- parts            1   840.66  918.66
- addresses        1   840.78  918.78
- capitalLong      1   840.92  918.92
- table            1   841.06  919.06
- num3d            1   841.17  919.17
- labs             1   841.29  919.29
- original         1   841.54  919.54
- will             1   841.62  919.62
- email            1   841.75  919.75
- address          1   841.93  919.93
<none>                 840.06  920.06
- make             1   842.37  920.37
- over             1   842.66  920.66
- num85            1   842.76  920.76
- receive          1   843.68  921.68
- mail             1   844.40  922.40
- charRoundbracket 1   845.08  923.08
- your             1   845.57  923.57
- internet         1   846.84  924.84
- num000           1   847.15  925.15
- pm               1   847.47  925.47
- technology       1   847.86  925.86
- charExclamation  1   849.05  927.05
- data             1   849.24  927.24
- lab              1   849.57  927.57
- project          1   850.48  928.48
- you              1   851.07  929.07
- capitalAve       1   851.28  929.28
- meeting          1   857.26  935.26
- cs               1   859.05  937.05
- re               1   859.35  937.35
- conference       1   861.97  939.97
- charDollar       1   869.13  947.13
- our              1   869.71  947.71
- charSemicolon    1   873.32  951.32
- remove           1   877.18  955.18
- about.money      1   889.53  967.53
```

```
- capitalTotal      1   912.62  990.62
- edu               1   942.65 1020.65

Step:  AIC=918.65
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + addresses + email + you + your +
    num000 + lab + labs + data + num85 + technology + parts +
    pm + cs + meeting + original + project + re + edu + table +
    conference + charSemicolon + charRoundbracket + charExclamation +
    charDollar + capitalAve + capitalLong + capitalTotal + about.money

                  Df Deviance     AIC
- parts            1   841.23  917.23
- addresses        1   841.28  917.28
- capitalLong      1   841.55  917.55
- table            1   841.62  917.62
- num3d            1   841.75  917.75
- labs             1   841.93  917.93
- original         1   842.10  918.10
- email            1   842.20  918.20
- will             1   842.25  918.25
- address          1   842.48  918.48
<none>                 840.65  918.65
- make             1   843.00  919.00
- over             1   843.17  919.17
- num85            1   843.34  919.34
- receive          1   844.23  920.23
- mail             1   845.10  921.10
- charRoundbracket 1   845.54  921.54
- your             1   846.24  922.24
- internet         1   847.63  923.63
- num000           1   847.67  923.67
- pm               1   848.25  924.25
- technology       1   848.68  924.68
- data             1   849.57  925.57
- charExclamation  1   849.65  925.65
- lab              1   850.01  926.01
- project          1   851.19  927.19
- you              1   851.68  927.68
- capitalAve       1   852.49  928.49
- meeting          1   857.82  933.82
- cs               1   859.58  935.58
- re               1   859.81  935.81
- conference       1   862.27  938.27
- our              1   870.25  946.25
- charDollar       1   870.82  946.82
- charSemicolon    1   873.72  949.72
- remove           1   877.46  953.46
```

```
- about.money       1   890.30  966.30
- capitalTotal      1   912.63  988.63
- edu               1   944.39 1020.39

Step:  AIC=917.23
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + addresses + email + you + your +
    num000 + lab + labs + data + num85 + technology + pm + cs +
    meeting + original + project + re + edu + table + conference +
    charSemicolon + charRoundbracket + charExclamation + charDollar +
    capitalAve + capitalLong + capitalTotal + about.money

                  Df Deviance     AIC
- addresses         1   841.86  915.86
- capitalLong       1   842.14  916.14
- table             1   842.21  916.21
- num3d             1   842.33  916.33
- labs              1   842.53  916.53
- original          1   842.68  916.68
- email             1   842.81  916.81
- will              1   842.81  916.81
- address           1   843.04  917.04
<none>                  841.23  917.23
- make              1   843.58  917.58
- over              1   843.78  917.78
- num85             1   843.91  917.91
- receive           1   844.85  918.85
- mail              1   845.71  919.71
- charRoundbracket  1   846.03  920.03
- your              1   846.91  920.91
- num000            1   848.27  922.27
- internet          1   848.34  922.34
- pm                1   848.81  922.81
- technology        1   849.31  923.31
- data              1   850.15  924.15
- charExclamation   1   850.29  924.29
- lab               1   850.58  924.58
- project           1   851.78  925.78
- you               1   852.62  926.62
- capitalAve        1   853.20  927.20
- cs                1   860.17  934.17
- re                1   860.40  934.40
- meeting           1   860.79  934.79
- conference        1   862.84  936.84
- our               1   871.11  945.11
- charDollar        1   871.41  945.41
- charSemicolon     1   874.36  948.36
- remove            1   878.31  952.31
```

```
- about.money       1   890.34  964.34
- capitalTotal      1   913.63  987.63
- edu               1   944.88 1018.88

Step:  AIC=915.86
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + email + you + your + num000 + lab +
    labs + data + num85 + technology + pm + cs + meeting + original +
    project + re + edu + table + conference + charSemicolon +
    charRoundbracket + charExclamation + charDollar + capitalAve +
    capitalLong + capitalTotal + about.money

                  Df Deviance     AIC
- capitalLong       1   842.86  914.86
- table             1   842.89  914.89
- num3d             1   842.95  914.95
- labs              1   843.15  915.15
- original          1   843.31  915.31
- will              1   843.44  915.44
- address           1   843.60  915.60
- email             1   843.86  915.86
<none>                  841.86  915.86
- make              1   844.18  916.18
- over              1   844.47  916.47
- num85             1   844.56  916.56
- receive           1   845.53  917.53
- mail              1   846.39  918.39
- charRoundbracket  1   846.64  918.64
- your              1   847.60  919.60
- internet          1   848.98  920.98
- num000            1   849.12  921.12
- pm                1   849.47  921.47
- technology        1   849.93  921.93
- data              1   850.78  922.78
- charExclamation   1   850.96  922.96
- lab               1   851.31  923.31
- project           1   852.49  924.49
- you               1   853.43  925.43
- capitalAve        1   853.83  925.83
- cs                1   860.96  932.96
- re                1   861.24  933.24
- meeting           1   861.54  933.54
- conference        1   863.49  935.49
- our               1   871.92  943.92
- charDollar        1   872.37  944.37
- charSemicolon     1   874.96  946.96
- remove            1   879.58  951.58
- about.money       1   890.72  962.72
```

```
- capitalTotal      1   916.04  988.04
- edu               1   945.96 1017.96

Step:  AIC=914.86
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + email + you + your + num000 + lab +
    labs + data + num85 + technology + pm + cs + meeting + original +
    project + re + edu + table + conference + charSemicolon +
    charRoundbracket + charExclamation + charDollar + capitalAve +
    capitalTotal + about.money

                   Df Deviance     AIC
- table             1   843.87  913.87
- num3d             1   843.91  913.91
- labs              1   844.14  914.14
- original          1   844.35  914.35
- will              1   844.51  914.51
- address           1   844.79  914.79
- email             1   844.80  914.80
<none>                  842.86  914.86
- make              1   845.15  915.15
- over              1   845.45  915.45
- num85             1   845.55  915.55
- receive           1   846.51  916.51
- charRoundbracket  1   847.37  917.37
- mail              1   847.56  917.56
- your              1   848.49  918.49
- internet          1   850.20  920.20
- num000            1   850.38  920.38
- pm                1   850.52  920.52
- technology        1   851.21  921.21
- charExclamation   1   852.01  922.01
- lab               1   852.02  922.02
- data              1   852.59  922.59
- project           1   853.27  923.27
- you               1   854.09  924.09
- capitalAve        1   857.85  927.85
- cs                1   861.45  931.45
- meeting           1   862.04  932.04
- re                1   862.62  932.62
- conference        1   864.88  934.88
- our               1   872.82  942.82
- charDollar        1   873.96  943.96
- charSemicolon     1   875.13  945.13
- remove            1   880.33  950.33
- about.money       1   891.09  961.09
- edu               1   948.78 1018.78
- capitalTotal      1   992.89 1062.89
```

```
Step:  AIC=913.87
type ~ make + address + num3d + our + over + remove + internet +
    mail + receive + will + email + you + your + num000 + lab +
    labs + data + num85 + technology + pm + cs + meeting + original +
    project + re + edu + conference + charSemicolon + charRoundbracket +
    charExclamation + charDollar + capitalAve + capitalTotal +
    about.money

                  Df Deviance     AIC
- num3d            1   844.92  912.92
- labs             1   845.14  913.14
- original         1   845.36  913.36
- will             1   845.42  913.42
- email            1   845.51  913.51
- address          1   845.74  913.74
<none>                 843.87  913.87
- make             1   846.12  914.12
- over             1   846.52  914.52
- num85            1   846.58  914.58
- receive          1   847.33  915.33
- charRoundbracket 1   848.55  916.55
- mail             1   848.57  916.57
- your             1   849.01  917.01
- internet         1   851.21  919.21
- num000           1   851.33  919.33
- pm               1   851.56  919.56
- technology       1   852.19  920.19
- lab              1   853.03  921.03
- charExclamation  1   853.08  921.08
- data             1   853.68  921.68
- project          1   854.30  922.30
- you              1   854.73  922.73
- capitalAve       1   858.97  926.97
- cs               1   862.37  930.37
- meeting          1   863.20  931.20
- re               1   863.36  931.36
- conference       1   866.01  934.01
- our              1   874.35  942.35
- charDollar       1   875.21  943.21
- charSemicolon    1   876.42  944.42
- remove           1   881.48  949.48
- about.money      1   892.83  960.83
- edu              1   950.04 1018.04
- capitalTotal     1   995.29 1063.29

Step:  AIC=912.92
type ~ make + address + our + over + remove + internet + mail +
```

```
    receive + will + email + you + your + num000 + lab + labs +
    data + num85 + technology + pm + cs + meeting + original +
    project + re + edu + conference + charSemicolon + charRoundbracket +
    charExclamation + charDollar + capitalAve + capitalTotal +
    about.money

                   Df Deviance     AIC
- labs              1   846.17  912.17
- original          1   846.41  912.41
- will              1   846.52  912.52
- email             1   846.53  912.53
- address           1   846.81  912.81
<none>                  844.92  912.92
- make              1   847.16  913.16
- over              1   847.51  913.51
- num85             1   847.60  913.60
- receive           1   848.46  914.46
- mail              1   849.59  915.59
- charRoundbracket  1   849.71  915.71
- your              1   849.98  915.98
- internet          1   852.25  918.25
- num000            1   852.31  918.31
- pm                1   852.48  918.48
- technology        1   853.56  919.56
- lab               1   854.13  920.13
- charExclamation   1   854.13  920.13
- data              1   854.86  920.86
- project           1   855.42  921.42
- you               1   855.99  921.99
- capitalAve        1   860.58  926.58
- cs                1   863.55  929.55
- meeting           1   864.27  930.27
- re                1   864.61  930.61
- conference        1   867.29  933.29
- our               1   875.37  941.37
- charDollar        1   876.03  942.03
- charSemicolon     1   877.77  943.77
- remove            1   882.35  948.35
- about.money       1   894.06  960.06
- edu               1   952.02 1018.02
- capitalTotal      1   998.81 1064.81

Step:  AIC=912.17
type ~ make + address + our + over + remove + internet + mail +
    receive + will + email + you + your + num000 + lab + data +
    num85 + technology + pm + cs + meeting + original + project +
    re + edu + conference + charSemicolon + charRoundbracket +
    charExclamation + charDollar + capitalAve + capitalTotal +
```

```
    about.money

                   Df Deviance     AIC
- original          1   847.68  911.68
- email             1   847.73  911.73
- will              1   847.79  911.79
- address           1   848.07  912.07
<none>                  846.17  912.17
- make              1   848.47  912.47
- num85             1   848.88  912.88
- over              1   848.93  912.93
- receive           1   849.56  913.56
- mail              1   850.78  914.78
- charRoundbracket  1   850.94  914.94
- your              1   851.02  915.02
- internet          1   853.35  917.35
- num000            1   853.43  917.43
- pm                1   853.54  917.54
- technology        1   854.68  918.68
- lab               1   855.22  919.22
- charExclamation   1   855.29  919.29
- data              1   856.25  920.25
- project           1   856.54  920.54
- you               1   857.09  921.09
- capitalAve        1   861.47  925.47
- cs                1   863.89  927.89
- meeting           1   865.70  929.70
- re                1   865.97  929.97
- conference        1   868.75  932.75
- our               1   876.42  940.42
- charDollar        1   877.19  941.19
- charSemicolon     1   879.06  943.06
- remove            1   883.24  947.24
- about.money       1   895.40  959.40
- edu               1   954.33 1018.33
- capitalTotal      1  1001.69 1065.69

Step:  AIC=911.68
type ~ make + address + our + over + remove + internet + mail +
    receive + will + email + you + your + num000 + lab + data +
    num85 + technology + pm + cs + meeting + project + re + edu +
    conference + charSemicolon + charRoundbracket + charExclamation +
    charDollar + capitalAve + capitalTotal + about.money

                   Df Deviance     AIC
- email             1   849.24  911.24
- will              1   849.36  911.36
- address           1   849.53  911.53
```

```
<none>                  847.68  911.68
- make              1   849.93  911.93
- num85             1   850.35  912.35
- over              1   850.45  912.45
- receive           1   851.04  913.04
- mail              1   852.32  914.32
- charRoundbracket  1   852.47  914.47
- your              1   852.55  914.55
- internet          1   854.77  916.77
- num000            1   854.96  916.96
- pm                1   855.27  917.27
- technology        1   856.35  918.35
- lab               1   856.62  918.62
- charExclamation   1   856.95  918.95
- data              1   857.69  919.69
- project           1   857.96  919.96
- you               1   858.66  920.66
- capitalAve        1   863.13  925.13
- cs                1   865.62  927.62
- meeting           1   867.04  929.04
- re                1   868.00  930.00
- conference        1   870.11  932.11
- our               1   877.91  939.91
- charDollar        1   879.12  941.12
- charSemicolon     1   880.63  942.63
- remove            1   884.94  946.94
- about.money       1   897.26  959.26
- edu               1   956.38 1018.38
- capitalTotal      1  1002.65 1064.65

Step:  AIC=911.24
type ~ make + address + our + over + remove + internet + mail +
    receive + will + you + your + num000 + lab + data + num85 +
    technology + pm + cs + meeting + project + re + edu + conference +
    charSemicolon + charRoundbracket + charExclamation + charDollar +
    capitalAve + capitalTotal + about.money

                   Df Deviance     AIC
- address           1   850.64  910.64
- will              1   850.83  910.83
<none>                  849.24  911.24
- make              1   851.67  911.67
- over              1   851.97  911.97
- num85             1   852.00  912.00
- receive           1   852.62  912.62
- mail              1   853.67  913.67
- charRoundbracket  1   854.05  914.05
- your              1   854.09  914.09
```

```
- num000            1   856.20  916.20
- internet          1   856.43  916.43
- pm                1   857.11  917.11
- technology        1   858.03  918.03
- lab               1   858.22  918.22
- charExclamation   1   858.61  918.61
- data              1   859.41  919.41
- project           1   859.63  919.63
- you               1   861.68  921.68
- capitalAve        1   864.57  924.57
- cs                1   866.77  926.77
- meeting           1   868.96  928.96
- re                1   870.93  930.93
- conference        1   871.90  931.90
- our               1   880.82  940.82
- charDollar        1   881.87  941.87
- charSemicolon     1   883.04  943.04
- remove            1   888.98  948.98
- about.money       1   903.03  963.03
- edu               1   958.40 1018.40
- capitalTotal      1  1006.90 1066.90

Step:  AIC=910.64
type ~ make + our + over + remove + internet + mail + receive +
    will + you + your + num000 + lab + data + num85 + technology +
    pm + cs + meeting + project + re + edu + conference + charSemicolon +
    charRoundbracket + charExclamation + charDollar + capitalAve +
    capitalTotal + about.money

                  Df Deviance     AIC
- will             1   852.15  910.15
<none>                 850.64  910.64
- make             1   852.98  910.98
- num85            1   853.33  911.33
- over             1   853.48  911.48
- receive          1   854.00  912.00
- mail             1   854.81  912.81
- charRoundbracket 1   855.16  913.16
- your             1   855.53  913.53
- num000           1   857.65  915.65
- internet         1   858.02  916.02
- pm               1   858.53  916.53
- technology       1   859.67  917.67
- lab              1   859.68  917.68
- charExclamation  1   860.27  918.27
- data             1   860.63  918.63
- project          1   860.81  918.81
- you              1   863.83  921.83
```

```
- capitalAve        1   865.47  923.47
- cs                1   868.25  926.25
- meeting           1   870.01  928.01
- re                1   872.11  930.11
- conference        1   873.22  931.22
- our               1   882.54  940.54
- charDollar        1   883.49  941.49
- charSemicolon     1   884.47  942.47
- remove            1   890.23  948.23
- about.money       1   904.16  962.16
- edu               1   959.24 1017.24
- capitalTotal      1  1014.93 1072.93

Step:  AIC=910.15
type ~ make + our + over + remove + internet + mail + receive +
    you + your + num000 + lab + data + num85 + technology + pm +
    cs + meeting + project + re + edu + conference + charSemicolon +
    charRoundbracket + charExclamation + charDollar + capitalAve +
    capitalTotal + about.money

                  Df Deviance     AIC
<none>                 852.15  910.15
- make             1   854.93  910.93
- num85            1   855.09  911.09
- over             1   855.20  911.20
- mail             1   855.82  911.82
- receive          1   856.11  912.11
- charRoundbracket 1   856.36  912.36
- your             1   856.80  912.80
- num000           1   859.25  915.25
- internet         1   859.86  915.86
- technology       1   861.28  917.28
- lab              1   861.53  917.53
- pm               1   861.70  917.70
- data             1   862.07  918.07
- charExclamation  1   862.16  918.16
- project          1   862.36  918.36
- you              1   864.73  920.73
- capitalAve       1   867.23  923.23
- cs               1   869.66  925.66
- meeting          1   871.86  927.86
- re               1   872.87  928.87
- conference       1   875.54  931.54
- our              1   884.59  940.59
- charSemicolon    1   885.45  941.45
- charDollar       1   885.62  941.62
- remove           1   892.29  948.29
- about.money      1   907.91  963.91
```

```
- edu               1   959.35 1015.35
- capitalTotal      1  1016.54 1072.54
```

do not worry about these warnings: they are fitted probabilities numerically very close to 0 or 1

We define now a convenience function:

'P' is a parameter; whenever our filter assigns spam with probability at least P then we predict spam

```
In [46]: spam.accs <- function (P=0.5)
         {
           ## Compute accuracy in learning data

           spamM1.AICpred <- NULL
           spamM1.AICpred[spamM1.AIC$fitted.values<P] <- 0
           spamM1.AICpred[spamM1.AIC$fitted.values>=P] <- 1

           spamM1.AICpred <- factor(spamM1.AICpred, labels=c("nonspam","spam"))

           print(M1.TRtable <- table(Truth=spam3[learn,]$type,Pred=spamM1.AICpred))

           print(100*(1-sum(diag(M1.TRtable))/nlearn))

           ## Compute accuracy in test data

           gl1t <- predict(spamM1.AIC, newdata=spam3[-learn,],type="response")
           gl1predt <- NULL
           gl1predt[gl1t<P] <- 0
           gl1predt[gl1t>=P] <- 1

           gl1predt <- factor(gl1predt, labels=c("nonspam","spam"))

           print(M1.TEtable <- table(Truth=spam3[-learn,]$type,Pred=gl1predt))

           print(100*(1-sum(diag(M1.TEtable))/ntest))
         }

         spam.accs()

         Pred
Truth     nonspam spam
  nonspam     792   81
  spam         64 1072
[1] 7.217521
         Pred
Truth     nonspam spam
  nonspam     357   42
```

```
  spam         28  563
[1] 7.070707
```

gives 7.21% TRAINING ERROR and 7.07% TESTING ERROR

Although the errors are quite low still one could argue that we should try to lower the probability of predicting spam when it is not We can do this (at the expense of increasing the converse probability) by:

```
In [47]: spam.accs(0.7)
```

```
         Pred
Truth     nonspam spam
  nonspam     821   52
  spam        142  994
[1] 9.656546
         Pred
Truth     nonspam spam
  nonspam     372   27
  spam         75  516
[1] 10.30303
```

gives 9.66% TRAINING ERROR and 10.3% TESTING ERROR

So we get a much better spam filter; notice that the filter has a very low probability of predicting spam when it is not (which is the delicate case), of about 6.77%