parkinsons

Marat Habibullin

26.12.2014

Load packages:

```
## Loading required package: RColorBrewer
## Loading required package: gplots
##
## Attaching package: 'gplots'
##
## The following object is masked from 'package:stats':
##
##     lowess
##
## Loading required package: ggplot2
##
## Attaching package: 'ggplot2'
##
## The following object is masked from 'package:latticeExtra':
##
##     layer
```

Read and normalize data:

```
pdata <- read.csv(file = "data/parkinsons.csv", comment.char="#")
pdata$status <- factor(pdata$status, labels = c("Healthy", "Sick"))
contrasts(pdata$status)
```

```
##         Sick
## Healthy    0
## Sick       1
```

```
pdata.grouped = pdata
pdata.grouped$name = sapply(pdata.grouped$name,
                            function(x) {x = as.character(x); substr(x, 1, nchar(x) - 2)})
pdata.grouped = aggregate(subset(pdata.grouped, select = c(-name, -status)),
                          list(pdata.grouped$name, pdata.grouped$status), mean)
names(pdata.grouped)[2] = "status"
pdata.grouped <- subset(pdata.grouped, select = -c(Group.1))
pdata.grouped$MDVP.Jitter.Abs. <- pdata.grouped$MDVP.Jitter.Abs. * 1000
```

Lets start with glm:

```
tn.logit <- tune(glm,
                 status ~ .,
                 data = pdata.grouped,
                 family = binomial(link = "logit"),
                 predict.func = simple.predict.glm,
                 tunecontrol = tune.control(sampling = "cross", cross = 10))
```

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

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

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

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

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

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

## Warning: glm.fit: algorithm did not converge

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

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

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

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

## Warning: glm.fit: algorithm did not converge

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

```
tn.logit$performances
```

```
##   dummyparameter     error dispersion
## 1              0 0.4833333  0.2772217
```

We have strange warnings here. Lets look at lda:

```
tn.lda <- tune(lda,
               status ~ .,
               data = pdata.grouped,
               predict.func = simple.predict.da,
               tunecontrol = tune.control(sampling = "cross", cross = 10))
```

```
## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear
```

```
## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear

## Warning in lda.default(x, grouping, ...): variables are collinear
```

```
tn.lda$performances
```

```
##   dummyparameter     error dispersion
## 1              0 0.3416667  0.2305723
```

Warnings again. These warnings mean that predictors may be correlated. Before trying to remove correlated ones, lets look at multinom and naive Bayes:

```
tn.mln <- tune(multinom,
               status ~ .,
               data = pdata.grouped,
               trace = FALSE)
tn.mln$performances
```

```
##   dummyparameter     error dispersion
## 1              0 0.3083333  0.2005009
```

```
tn.nb <- tune(naiveBayes, status ~ ., data = pdata.grouped)
tn.nb$performances
```

```
##   dummyparameter     error dispersion
## 1              0 0.2583333  0.2095336
```

Not very good result, error is pretty high.

Now let's reduce our model with stepAIC:

```
mod.mln <- multinom(status ~ .,
                    data = pdata.grouped,
                    trace = FALSE)
mod.mln.aic <- stepAIC(mod.mln)
```

```
## Start:  AIC=46
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Flo.Hz. + MDVP.Jitter... +
##     MDVP.Jitter.Abs. + MDVP.RAP + MDVP.PPQ + Jitter.DDP + MDVP.Shimmer +
##     MDVP.Shimmer.dB. + Shimmer.APQ3 + Shimmer.APQ5 + MDVP.APQ +
##     Shimmer.DDA + NHR + HNR + RPDE + DFA + spread1 + spread2 +
##     D2 + PPE
```

```
##
##                     Df    AIC
## - PPE                1 44.000
## - MDVP.Shimmer.dB.   1 44.001
## - MDVP.Flo.Hz.       1 44.001
## - spread2            1 44.003
## - MDVP.RAP           1 44.003
## - MDVP.PPQ           1 44.003
## - MDVP.Jitter...     1 44.003
## - Shimmer.APQ5       1 44.003
## - Jitter.DDP         1 44.003
## - Shimmer.APQ3       1 44.003
## - MDVP.APQ           1 44.003
## - MDVP.Shimmer       1 44.003
## - Shimmer.DDA        1 44.003
## - MDVP.Jitter.Abs.   1 44.003
## - NHR                1 44.003
## - DFA                1 44.026
## - RPDE               1 44.096
## - spread1            1 44.597
## <none>                 46.001
## - MDVP.Fo.Hz.        1 46.010
## - HNR                1 46.464
## - MDVP.Fhi.Hz.       1 50.697
## - D2                 1 52.613
##
## Step:  AIC=44
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Flo.Hz. + MDVP.Jitter... +
##     MDVP.Jitter.Abs. + MDVP.RAP + MDVP.PPQ + Jitter.DDP + MDVP.Shimmer +
##     MDVP.Shimmer.dB. + Shimmer.APQ3 + Shimmer.APQ5 + MDVP.APQ +
##     Shimmer.DDA + NHR + HNR + RPDE + DFA + spread1 + spread2 +
##     D2
##
##                     Df    AIC
## - MDVP.Jitter.Abs.   1 42.000
## - NHR                1 42.000
## - Shimmer.DDA        1 42.001
## - MDVP.Jitter...     1 42.001
## - MDVP.PPQ           1 42.001
## - MDVP.RAP           1 42.001
## - MDVP.Shimmer       1 42.001
## - Jitter.DDP         1 42.001
## - Shimmer.APQ3       1 42.001
## - MDVP.APQ           1 42.002
## - spread2            1 42.002
## - Shimmer.APQ5       1 42.002
## - MDVP.Flo.Hz.       1 42.007
## - MDVP.Shimmer.dB.   1 42.018
## - DFA                1 42.359
## - RPDE               1 42.722
## - MDVP.Fo.Hz.        1 42.766
## <none>                 44.000
## - HNR                1 48.010
## - MDVP.Fhi.Hz.       1 49.081
```

```
## - spread1           1 50.343
## - D2                1 50.868
##
## Step:  AIC=42
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Flo.Hz. + MDVP.Jitter... +
##     MDVP.RAP + MDVP.PPQ + Jitter.DDP + MDVP.Shimmer + MDVP.Shimmer.dB. +
##     Shimmer.APQ3 + Shimmer.APQ5 + MDVP.APQ + Shimmer.DDA + NHR +
##     HNR + RPDE + DFA + spread1 + spread2 + D2
##
##                     Df    AIC
## - MDVP.Shimmer       1 40.007
## - Shimmer.DDA        1 40.007
## - NHR                1 40.008
## - spread2            1 40.008
## - MDVP.APQ           1 40.008
## - Shimmer.APQ3       1 40.008
## - Shimmer.APQ5       1 40.008
## - Jitter.DDP         1 40.009
## - MDVP.Jitter...     1 40.009
## - MDVP.PPQ           1 40.009
## - MDVP.RAP           1 40.009
## - MDVP.Fo.Hz.        1 40.032
## - MDVP.Flo.Hz.       1 40.033
## - RPDE               1 40.063
## - MDVP.Shimmer.dB.   1 40.135
## - MDVP.Fhi.Hz.       1 40.565
## <none>                 42.000
## - DFA                1 42.004
## - HNR                1 47.642
## - D2                 1 49.313
## - spread1            1 51.764
##
## Step:  AIC=40.01
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Flo.Hz. + MDVP.Jitter... +
##     MDVP.RAP + MDVP.PPQ + Jitter.DDP + MDVP.Shimmer.dB. + Shimmer.APQ3 +
##     Shimmer.APQ5 + MDVP.APQ + Shimmer.DDA + NHR + HNR + RPDE +
##     DFA + spread1 + spread2 + D2
##
##                     Df    AIC
## - RPDE               1 38.018
## - MDVP.Fo.Hz.        1 38.054
## - MDVP.Flo.Hz.       1 38.145
## - MDVP.RAP           1 38.210
## - NHR                1 38.349
## - spread2            1 38.350
## - MDVP.PPQ           1 38.411
## - MDVP.Jitter...     1 38.457
## - Shimmer.APQ5       1 38.459
## - Jitter.DDP         1 38.465
## - Shimmer.APQ3       1 38.471
## - Shimmer.DDA        1 38.475
## - MDVP.APQ           1 38.522
## - MDVP.Shimmer.dB.   1 38.769
## - DFA                1 39.856
```

```
## <none>                40.007
## - MDVP.Fhi.Hz.      1 44.149
## - HNR               1 46.099
## - D2                1 47.266
## - spread1           1 50.166
##
## Step:  AIC=38.02
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Flo.Hz. + MDVP.Jitter... +
##     MDVP.RAP + MDVP.PPQ + Jitter.DDP + MDVP.Shimmer.dB. + Shimmer.APQ3 +
##     Shimmer.APQ5 + MDVP.APQ + Shimmer.DDA + NHR + HNR + DFA +
##     spread1 + spread2 + D2
##
##                     Df    AIC
## - spread2            1 36.008
## - MDVP.Shimmer.dB.   1 36.057
## - MDVP.APQ           1 36.122
## - DFA                1 36.138
## - MDVP.Jitter...     1 36.140
## - MDVP.PPQ           1 36.140
## - MDVP.RAP           1 36.140
## - Jitter.DDP         1 36.141
## - Shimmer.APQ3       1 36.141
## - Shimmer.APQ5       1 36.149
## - Shimmer.DDA        1 36.155
## - MDVP.Fo.Hz.        1 36.396
## - NHR                1 37.490
## <none>                 38.018
## - MDVP.Flo.Hz.       1 41.585
## - HNR                1 43.198
## - MDVP.Fhi.Hz.       1 44.389
## - D2                 1 45.177
## - spread1            1 48.439
##
## Step:  AIC=36.01
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Flo.Hz. + MDVP.Jitter... +
##     MDVP.RAP + MDVP.PPQ + Jitter.DDP + MDVP.Shimmer.dB. + Shimmer.APQ3 +
##     Shimmer.APQ5 + MDVP.APQ + Shimmer.DDA + NHR + HNR + DFA +
##     spread1 + D2
##
##                     Df    AIC
## - Shimmer.DDA        1 34.045
## - MDVP.APQ           1 34.054
## - Shimmer.APQ5       1 34.057
## - Shimmer.APQ3       1 34.064
## - NHR                1 34.065
## - Jitter.DDP         1 34.069
## - MDVP.RAP           1 34.070
## - MDVP.PPQ           1 34.070
## - MDVP.Jitter...     1 34.070
## - MDVP.Shimmer.dB.   1 34.151
## - MDVP.Flo.Hz.       1 34.681
## - MDVP.Fo.Hz.        1 35.941
## <none>                 36.008
## - DFA                1 40.527
```

```
## - HNR               1 40.933
## - MDVP.Fhi.Hz.      1 42.418
## - D2                1 43.920
## - spread1           1 45.592
##
## Step:  AIC=34.05
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Flo.Hz. + MDVP.Jitter... +
##     MDVP.RAP + MDVP.PPQ + Jitter.DDP + MDVP.Shimmer.dB. + Shimmer.APQ3 +
##     Shimmer.APQ5 + MDVP.APQ + NHR + HNR + DFA + spread1 + D2
##
##                     Df    AIC
## - MDVP.Flo.Hz.       1 32.029
## - MDVP.APQ           1 32.148
## - MDVP.Jitter...     1 32.299
## - MDVP.RAP           1 32.299
## - MDVP.PPQ           1 32.299
## - Jitter.DDP         1 32.299
## - Shimmer.APQ5       1 32.299
## - Shimmer.APQ3       1 32.300
## - NHR                1 32.315
## - MDVP.Shimmer.dB.   1 32.379
## <none>                 34.045
## - MDVP.Fo.Hz.        1 36.164
## - DFA                1 37.050
## - HNR                1 38.142
## - MDVP.Fhi.Hz.       1 40.351
## - D2                 1 41.973
## - spread1            1 43.337
##
## Step:  AIC=32.03
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP +
##     MDVP.PPQ + Jitter.DDP + MDVP.Shimmer.dB. + Shimmer.APQ3 +
##     Shimmer.APQ5 + MDVP.APQ + NHR + HNR + DFA + spread1 + D2
##
##                     Df    AIC
## - Shimmer.APQ3       1 30.340
## - Jitter.DDP         1 30.340
## - MDVP.RAP           1 30.340
## - MDVP.PPQ           1 30.340
## - MDVP.Jitter...     1 30.340
## - Shimmer.APQ5       1 30.340
## - MDVP.APQ           1 30.351
## - NHR                1 30.666
## <none>                 32.029
## - MDVP.Fo.Hz.        1 35.315
## - MDVP.Shimmer.dB.   1 37.296
## - HNR                1 37.429
## - DFA                1 37.622
## - MDVP.Fhi.Hz.       1 39.000
## - D2                 1 40.158
## - spread1            1 42.260
##
## Step:  AIC=30.34
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP +
```

```
##     MDVP.PPQ + Jitter.DDP + MDVP.Shimmer.dB. + Shimmer.APQ5 +
##     MDVP.APQ + NHR + HNR + DFA + spread1 + D2
##
##                     Df    AIC
## - MDVP.APQ           1 28.826
## - Shimmer.APQ5       1 29.242
## - MDVP.RAP           1 29.540
## - MDVP.PPQ           1 29.541
## - MDVP.Jitter...     1 29.566
## - Jitter.DDP         1 29.604
## <none>                 30.340
## - NHR                1 30.912
## - MDVP.Fo.Hz.        1 33.270
## - MDVP.Shimmer.dB.   1 35.487
## - DFA                1 35.694
## - HNR                1 36.995
## - MDVP.Fhi.Hz.       1 37.388
## - D2                 1 38.216
## - spread1            1 40.518
##
## Step:  AIC=28.83
## status ~ MDVP.Fo.Hz. + MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP +
##     MDVP.PPQ + Jitter.DDP + MDVP.Shimmer.dB. + Shimmer.APQ5 +
##     NHR + HNR + DFA + spread1 + D2
##
##                     Df    AIC
## - MDVP.Fo.Hz.        1 28.610
## <none>                 28.826
## - Jitter.DDP         1 31.107
## - MDVP.RAP           1 31.107
## - MDVP.PPQ           1 31.107
## - MDVP.Jitter...     1 31.108
## - Shimmer.APQ5       1 31.110
## - NHR                1 31.771
## - HNR                1 32.282
## - DFA                1 34.027
## - MDVP.Shimmer.dB.   1 34.103
## - MDVP.Fhi.Hz.       1 35.511
## - D2                 1 36.306
## - spread1            1 38.432
##
## Step:  AIC=28.61
## status ~ MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP + MDVP.PPQ +
##     Jitter.DDP + MDVP.Shimmer.dB. + Shimmer.APQ5 + NHR + HNR +
##     DFA + spread1 + D2
##
##                     Df    AIC
## - Jitter.DDP         1 27.757
## - MDVP.Jitter...     1 28.128
## - MDVP.RAP           1 28.140
## - MDVP.PPQ           1 28.140
## <none>                 28.610
## - Shimmer.APQ5       1 28.732
## - HNR                1 31.733
```

```
## - MDVP.Shimmer.dB.  1 32.297
## - DFA               1 32.599
## - MDVP.Fhi.Hz.      1 33.036
## - NHR               1 33.205
## - spread1           1 37.014
## - D2                1 37.729
##
## Step:  AIC=27.76
## status ~ MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP + MDVP.PPQ +
##     MDVP.Shimmer.dB. + Shimmer.APQ5 + NHR + HNR + DFA + spread1 +
##     D2
##
##                     Df    AIC
## <none>                 27.757
## - Shimmer.APQ5       1 28.071
## - MDVP.Jitter...     1 28.316
## - MDVP.PPQ           1 28.331
## - MDVP.RAP           1 28.332
## - HNR                1 29.619
## - MDVP.Shimmer.dB.   1 30.791
## - DFA                1 30.949
## - NHR                1 31.238
## - MDVP.Fhi.Hz.       1 31.365
## - D2                 1 35.770
## - spread1            1 37.627
```

Let's check obtained model:

```
tn.logit <- tune(glm,
                 status ~ MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP + MDVP.PPQ +
                   MDVP.Shimmer.dB. + Shimmer.APQ5 + NHR + HNR + DFA + spread1 + D2,
                 data = pdata.grouped,
                 family = binomial(link = "logit"),
                 predict.func = simple.predict.glm,
                 tunecontrol = tune.control(sampling = "cross", cross = 10))
```

```
## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

```
## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

## Warning: glm.fit: algorithm did not converge

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

```
tn.logit$performances
```

```
##   dummyparameter error dispersion
## 1              0 0.275  0.1573802
```

```
tn.lda <- tune(lda,
               status ~ MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP + MDVP.PPQ +
                 MDVP.Shimmer.dB. + Shimmer.APQ5 + NHR + HNR + DFA + spread1 + D2,
               data = pdata.grouped,
               predict.func = simple.predict.da,
               tunecontrol = tune.control(sampling = "cross", cross = 10))
tn.lda$performances
```

```
##   dummyparameter     error dispersion
## 1              0 0.1916667  0.3558376
```

```
tn.mln <- tune(multinom,
               status ~ MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP + MDVP.PPQ +
                 MDVP.Shimmer.dB. + Shimmer.APQ5 + NHR + HNR + DFA + spread1 + D2,
               data = pdata.grouped,
               trace = FALSE)
tn.mln$performances
```

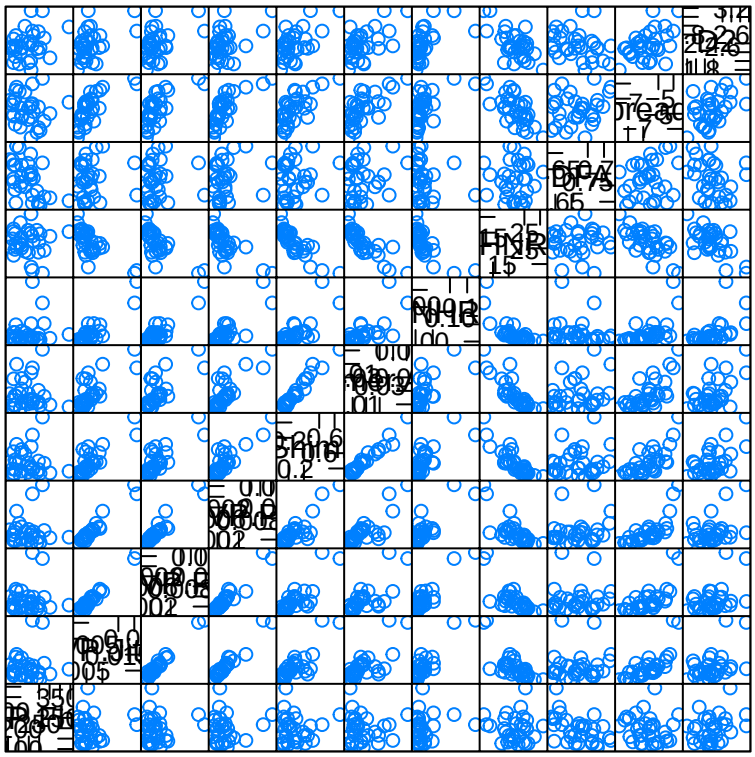
```
##   dummyparameter error dispersion
## 1              0  0.25  0.2078699
```

```
tn.nb <- tune(naiveBayes,
              status ~ MDVP.Fhi.Hz. + MDVP.Jitter... + MDVP.RAP + MDVP.PPQ +
                MDVP.Shimmer.dB. + Shimmer.APQ5 + NHR + HNR + DFA + spread1 + D2,
              data = pdata.grouped)
tn.nb$performances
```

```
##   dummyparameter error dispersion
## 1              0   0.3  0.2810913
```

Now there is no warnings in lda, but still warnings in glm. Let's look at splom:

```
splom(subset(pdata.grouped,
             select = c(MDVP.Fhi.Hz., MDVP.Jitter..., MDVP.RAP, MDVP.PPQ,
                        MDVP.Shimmer.dB., Shimmer.APQ5, NHR, HNR, DFA, spread1, D2)))
```



Scatter Plot Matrix

Let's remove MDVP.RAP, MDVP.PPQ, MDVP.Shimmer.dB., Shimmer.APQ5, NHR:

```
tn.lda <- tune(lda,
               status ~ MDVP.Fhi.Hz. + MDVP.Jitter... + HNR + DFA + spread1 + D2,
               data = pdata.grouped,
               predict.func = simple.predict.da,
               tunecontrol = tune.control(sampling = "cross", cross = 10))
tn.lda$performances
```

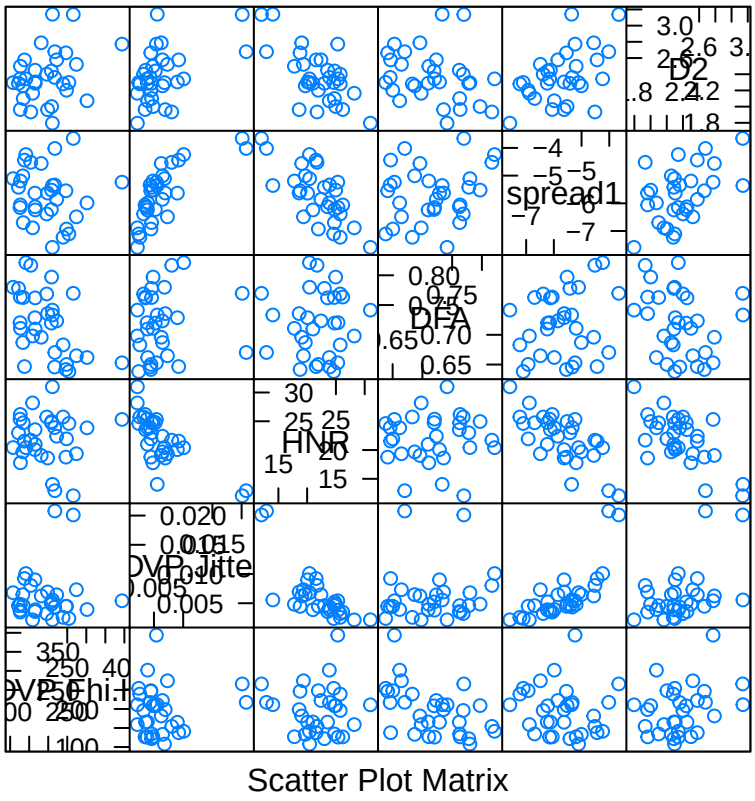
```
##   dummyparameter     error dispersion
## 1              0 0.1583333  0.2305723
```

```
tn.mln <- tune(multinom,
               status ~ MDVP.Fhi.Hz. + MDVP.Jitter... + HNR + DFA + spread1 + D2,
               data = pdata.grouped,
               trace = FALSE)
tn.mln$performances
```

```
##   dummyparameter     error dispersion
## 1              0 0.1916667   0.166898
```

Check correlation again:

```
splom(subset(pdata.grouped,
             select = c(MDVP.Fhi.Hz., MDVP.Jitter..., HNR, DFA, spread1, D2)))
```



Let's remove MDVP.Jitter...:

```
tn.lda <- tune(lda,
               status ~ MDVP.Fhi.Hz. + HNR + DFA + spread1 + D2,
               data = pdata.grouped,
               predict.func = simple.predict.da,
               tunecontrol = tune.control(sampling = "cross", cross = 10))
tn.lda$performances
```

```
##   dummyparameter     error dispersion
## 1              0 0.1333333  0.2490724
```

```
tn.mln <- tune(multinom,
               status ~ MDVP.Fhi.Hz. + HNR + DFA + spread1 + D2,
               data = pdata.grouped,
               trace = FALSE)
tn.mln$performances
```

```
##   dummyparameter error dispersion
## 1              0  0.15  0.1610153
```

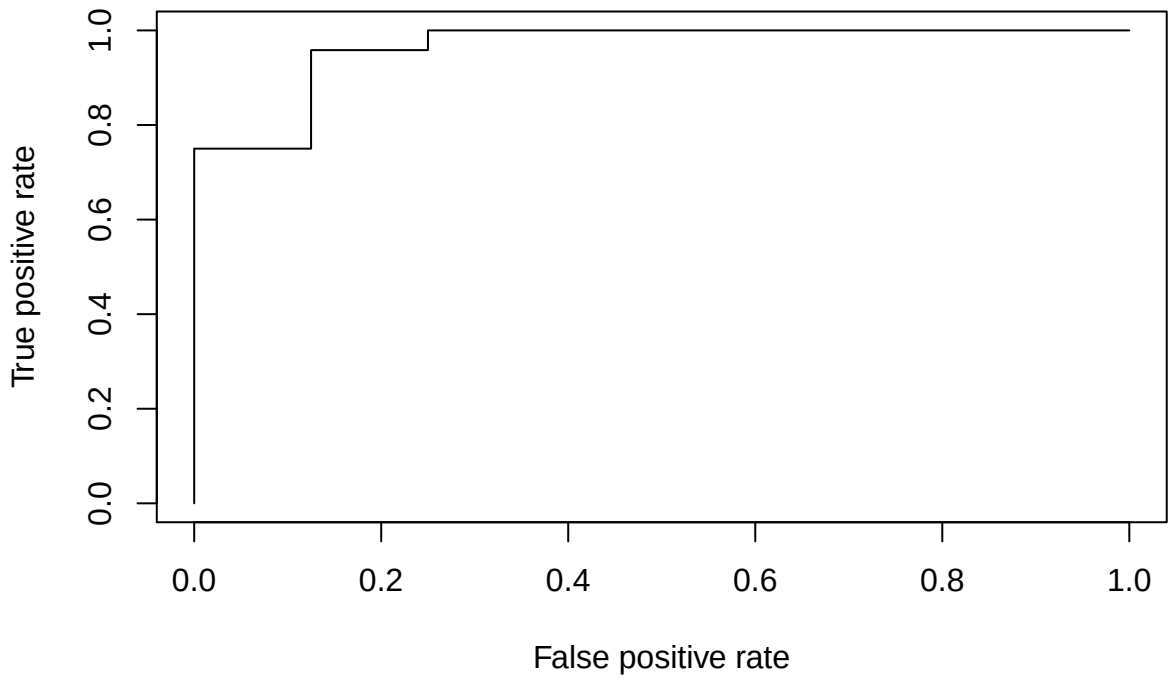
Let it be our final model. Now, let's ROC!:

```
final = subset(pdata.grouped, select = c(status, MDVP.Fhi.Hz., HNR, DFA, spread1, D2))

tbl <- table(predicted = predict(tn.lda$best.model, final)$class, actual = final$status)
tbl
```

```
##          actual
## predicted Healthy Sick
##   Healthy       6    1
##   Sick          2   23
```

```
roc <- ROC(predicted = predict(tn.lda$best.model, final)$x, actual = final$status)
plot(roc)
```




```
AUC(predicted = predict(tn.lda$best.model, final)$x, actual = final$status)
```

```
## [1] 0.9635417
```