Project 4 - STAT 3022

Henrique Lispector *ID: 4839516* *lispe001@umn.edu*

April 29, 2016

Project Description

Install the library TH.data. Use the data named GlaucomaM in this library. The GlaucomaM data has 196 observations in two classes. 62 variables are derived from a confocal laser scanning image of the optic nerve head, describing its morphology. Observations are from normal and glaucomatous eyes, respectively. Use the help file to know more about the dataset. Your goal is to predict whether a person will have glaucoma based on the 62 variables. Identify the predictors and the response variable in the dataset. Randomly select 70% of the data as training data and the remaining 30% as test data. Install the package glmnet and use elastic net method on the training data to determine an appropriate model. Then use this model to do predictions on the test dataset. Report which covariates were selected in the model. You do not need to interpret any coefficient estimate. Prediction and variable selection are the main focus of your analysis.

Loading the data

```
library(TH.data)
```

```
## Loading required package: survival

## Loading required package: MASS

##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```
GlaucomaM_data <- GlaucomaM
head(GlaucomaM_data)
```

```
##       ag    at    as    an    ai   eag   eat   eas   ean   eai  abrg  abrt
## 2  2.220 0.354 0.580 0.686 0.601 1.267 0.336 0.346 0.255 0.331 0.479 0.260
## 43 2.681 0.475 0.672 0.868 0.667 2.053 0.440 0.520 0.639 0.454 1.090 0.377
## 25 1.979 0.343 0.508 0.624 0.504 1.200 0.299 0.396 0.259 0.246 0.465 0.209
## 65 1.747 0.269 0.476 0.525 0.476 0.612 0.147 0.017 0.044 0.405 0.170 0.062
## 70 2.990 0.599 0.686 1.039 0.667 2.513 0.543 0.607 0.871 0.492 1.800 0.431
## 16 2.917 0.483 0.763 0.901 0.770 2.200 0.462 0.637 0.504 0.597 1.311 0.394
##     abrs  abrn  abri    hic   mhcg  mhct   mhcs   mhcn   mhci   phcg
## 2  0.107 0.014 0.098  0.214  0.111 0.412  0.036  0.105 -0.022 -0.139
## 43 0.257 0.212 0.245  0.382  0.140 0.338  0.104  0.080  0.109 -0.015
## 25 0.112 0.041 0.103  0.195  0.062 0.356  0.045 -0.009 -0.048 -0.149
## 65 0.000 0.000 0.108 -0.030 -0.015 0.074 -0.084 -0.050  0.035 -0.182
```

```
## 70 0.494 0.601 0.274  0.383  0.089 0.233  0.145  0.023  0.007 -0.131
## 16 0.365 0.251 0.301  0.442  0.128 0.375  0.049  0.111  0.052 -0.088
##      phct   phcs   phcn   phci   hvc  vbsg  vbst  vbss  vbsn  vbsi  vasg
## 2   0.242 -0.053  0.010 -0.139 0.613 0.303 0.103 0.088 0.022 0.090 0.062
## 43  0.296 -0.015 -0.015  0.036 0.382 0.676 0.181 0.186 0.141 0.169 0.029
## 25  0.206 -0.092 -0.081 -0.149 0.557 0.300 0.084 0.088 0.046 0.082 0.036
## 65 -0.097 -0.125 -0.138 -0.182 0.373 0.048 0.011 0.000 0.000 0.036 0.070
## 70  0.163  0.055 -0.131 -0.115 0.405 0.889 0.151 0.253 0.330 0.155 0.020
## 16  0.281 -0.067 -0.062 -0.088 0.507 0.972 0.213 0.316 0.197 0.246 0.043
##     vast  vass  vasn  vasi  vbrg  vbrt  vbrs  vbrn  vbri  varg  vart  vars
## 2  0.000 0.011 0.032 0.018 0.075 0.039 0.021 0.002 0.014 0.756 0.009 0.209
## 43 0.001 0.007 0.011 0.010 0.370 0.127 0.099 0.050 0.093 0.410 0.006 0.105
## 25 0.002 0.004 0.016 0.013 0.081 0.034 0.019 0.007 0.021 0.565 0.014 0.132
## 65 0.005 0.030 0.033 0.002 0.005 0.001 0.000 0.000 0.004 0.380 0.032 0.147
## 70 0.001 0.004 0.008 0.007 0.532 0.103 0.173 0.181 0.075 0.228 0.011 0.026
## 16 0.001 0.005 0.028 0.009 0.467 0.136 0.148 0.078 0.104 0.540 0.008 0.133
##     varn  vari   mdg   mdt   mds   mdn   mdi    tmg    tmt    tms    tmn
## 2  0.298 0.240 0.705 0.637 0.738 0.596 0.691 -0.236 -0.018 -0.230 -0.510
## 43 0.181 0.117 0.898 0.850 0.907 0.771 0.940 -0.211 -0.014 -0.165 -0.317
## 25 0.243 0.177 0.687 0.643 0.689 0.684 0.700 -0.185 -0.097 -0.235 -0.337
## 65 0.151 0.050 0.207 0.171 0.022 0.046 0.221 -0.148 -0.035 -0.449 -0.217
## 70 0.105 0.087 0.721 0.638 0.730 0.730 0.640 -0.052 -0.105  0.084 -0.012
## 16 0.232 0.167 0.927 0.842 0.953 0.906 0.898 -0.040  0.087  0.018 -0.094
##       tmi    mr   rnf  mdic   emd    mv  Class
## 2  -0.158 0.841 0.410 0.137 0.239 0.035 normal
## 43 -0.192 0.924 0.256 0.252 0.329 0.022 normal
## 25 -0.020 0.795 0.378 0.152 0.250 0.029 normal
## 65 -0.091 0.746 0.200 0.027 0.078 0.023 normal
## 70 -0.054 0.977 0.193 0.297 0.354 0.034 normal
## 16 -0.051 0.965 0.339 0.333 0.442 0.028 normal
```

Data Pre-Processing

```r
summary(GlaucomaM_data)
```

```
##        ag              at               as               an        
##  Min.   :1.312   Min.   :0.2010   Min.   :0.3450   Min.   :0.3970  
##  1st Qu.:2.139   1st Qu.:0.3708   1st Qu.:0.5385   1st Qu.:0.6810  
##  Median :2.533   Median :0.4445   Median :0.6305   Median :0.8085  
##  Mean   :2.607   Mean   :0.4590   Mean   :0.6518   Mean   :0.8359  
##  3rd Qu.:2.943   3rd Qu.:0.5280   3rd Qu.:0.7382   3rd Qu.:0.9520  
##  Max.   :5.444   Max.   :0.9670   Max.   :1.3400   Max.   :1.7650  
##        ai              eag              eat              eas        
##  Min.   :0.3690   Min.   :0.415   Min.   :0.1370   Min.   :0.0170  
##  1st Qu.:0.5505   1st Qu.:1.309   1st Qu.:0.3157   1st Qu.:0.3807  
##  Median :0.6320   Median :1.843   Median :0.4025   Median :0.4685  
##  Mean   :0.6600   Mean   :1.874   Mean   :0.4064   Mean   :0.4864  
##  3rd Qu.:0.7498   3rd Qu.:2.317   3rd Qu.:0.4833   3rd Qu.:0.6055  
##  Max.   :1.3730   Max.   :4.125   Max.   :0.8480   Max.   :1.2250  
##       ean              eai              abrg              abrt        
##  Min.   :0.0080   Min.   :0.0980   Min.   :0.0030   Min.   :0.0030
```

```
##  1st Qu.:0.2805   1st Qu.:0.3725   1st Qu.:0.6817   1st Qu.:0.2450  
##  Median :0.5035   Median :0.4840   Median :1.3120   Median :0.3225  
##  Mean   :0.5012   Mean   :0.4801   Mean   :1.2919   Mean   :0.3248  
##  3rd Qu.:0.6895   3rd Qu.:0.5948   3rd Qu.:1.7352   3rd Qu.:0.4295  
##  Max.   :1.5680   Max.   :0.9610   Max.   :4.9800   Max.   :0.8270  
##       abrs             abrn             abri             hic         
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :-0.1890  
##  1st Qu.:0.1928   1st Qu.:0.0885   1st Qu.:0.1693   1st Qu.: 0.1958  
##  Median :0.3250   Median :0.2520   Median :0.3255   Median : 0.3240  
##  Mean   :0.3295   Mean   :0.3125   Mean   :0.3251   Mean   : 0.3050  
##  3rd Qu.:0.4512   3rd Qu.:0.4520   3rd Qu.:0.4595   3rd Qu.: 0.4190  
##  Max.   :1.3400   Max.   :1.7650   Max.   :1.2090   Max.   : 0.8870  
##       mhcg              mhct              mhcs         
##  Min.   :-0.14700   Min.   :-0.0470   Min.   :-0.17200  
##  1st Qu.: 0.04675   1st Qu.: 0.1610   1st Qu.: 0.00175  
##  Median : 0.09450   Median : 0.2110   Median : 0.07050  
##  Mean   : 0.09415   Mean   : 0.2142   Mean   : 0.06123  
##  3rd Qu.: 0.13825   3rd Qu.: 0.2742   3rd Qu.: 0.11825  
##  Max.   : 0.32200   Max.   : 0.4770   Max.   : 0.29300  
##       mhcn               mhci              phcg               phct        
##  Min.   :-0.21200   Min.   :-0.1610   Min.   :-0.28600   Min.   :-0.1210  
##  1st Qu.: 0.01975   1st Qu.:-0.0035   1st Qu.:-0.13300   1st Qu.: 0.0950  
##  Median : 0.07950   Median : 0.0640   Median :-0.08800   Median : 0.1540  
##  Mean   : 0.07380   Mean   : 0.0647   Mean   :-0.07853   Mean   : 0.1477  
##  3rd Qu.: 0.12250   3rd Qu.: 0.1300   3rd Qu.:-0.01650   3rd Qu.: 0.2052  
##  Max.   : 0.66000   Max.   : 0.4540   Max.   : 0.14500   Max.   : 0.4300  
##       phcs               phcn               phci               hvc        
##  Min.   :-0.24700   Min.   :-0.28500   Min.   :-0.28600   Min.   :0.1100  
##  1st Qu.:-0.08925   1st Qu.:-0.08900   1st Qu.:-0.11200   1st Qu.:0.2860  
##  Median :-0.02850   Median :-0.03350   Median :-0.04700   Median :0.3470  
##  Mean   :-0.03105   Mean   :-0.03238   Mean   :-0.04238   Mean   :0.3604  
##  3rd Qu.: 0.02600   3rd Qu.: 0.02400   3rd Qu.: 0.02700   3rd Qu.:0.4283  
##  Max.   : 0.16000   Max.   : 0.39800   Max.   : 0.37100   Max.   :0.9690  
##       vbsg             vbst              vbss              vbsn       
##  Min.   :0.0200   Min.   :0.00700   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.3315   1st Qu.:0.07575   1st Qu.:0.09275   1st Qu.:0.0525  
##  Median :0.5960   Median :0.12200   Median :0.16800   Median :0.1185  
##  Mean   :0.6334   Mean   :0.13399   Mean   :0.18581   Mean   :0.1494  
##  3rd Qu.:0.8632   3rd Qu.:0.17400   3rd Qu.:0.26225   3rd Qu.:0.2157  
##  Max.   :2.1260   Max.   :0.44600   Max.   :0.81700   Max.   :0.6960  
##       vbsi              vasg              vast               vass       
##  Min.   :0.00600   Min.   :0.00800   Min.   :0.000000   Min.   :0.0010  
##  1st Qu.:0.08275   1st Qu.:0.02200   1st Qu.:0.001000   1st Qu.:0.0040  
##  Median :0.15600   Median :0.03600   Median :0.001000   Median :0.0070  
##  Mean   :0.16420   Mean   :0.04967   Mean   :0.002077   Mean   :0.0101  
##  3rd Qu.:0.22375   3rd Qu.:0.06425   3rd Qu.:0.002000   3rd Qu.:0.0110  
##  Max.   :0.49000   Max.   :0.75100   Max.   :0.026000   Max.   :0.2390  
##       vasn              vasi              vbrg             vbrt        
##  Min.   :0.00100   Min.   :0.00100   Min.   :0.0000   Min.   :0.00000  
##  1st Qu.:0.00900   1st Qu.:0.00400   1st Qu.:0.1338   1st Qu.:0.04075  
##  Median :0.01750   Median :0.00800   Median :0.3540   Median :0.08100  
##  Mean   :0.02561   Mean   :0.01186   Mean   :0.4256   Mean   :0.09719  
##  3rd Qu.:0.03125   3rd Qu.:0.01425   3rd Qu.:0.5540   3rd Qu.:0.13300  
##  Max.   :0.39700   Max.   :0.10500   Max.   :3.7000   Max.   :0.39900  
```

```
##       vbrs              vbrn              vbri              varg
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0160
##  1st Qu.:0.03875   1st Qu.:0.01275   1st Qu.:0.03275   1st Qu.:0.1450
##  Median :0.10150   Median :0.05650   Median :0.08750   Median :0.2780
##  Mean   :0.12374   Mean   :0.09906   Mean   :0.10552   Mean   :0.2962
##  3rd Qu.:0.16475   3rd Qu.:0.14025   3rd Qu.:0.14925   3rd Qu.:0.3900
##  Max.   :1.09900   Max.   :1.62000   Max.   :0.58900   Max.   :1.3250
##       vart              vars              varn             vari
##  Min.   :0.00100   Min.   :0.00000   Min.   :0.0000   Min.   :0.00100
##  1st Qu.:0.00400   1st Qu.:0.03375   1st Qu.:0.0630   1st Qu.:0.03400
##  Median :0.00700   Median :0.06950   Median :0.1170   Median :0.06750
##  Mean   :0.01050   Mean   :0.07595   Mean   :0.1298   Mean   :0.07991
##  3rd Qu.:0.01225   3rd Qu.:0.10100   3rd Qu.:0.1780   3rd Qu.:0.11050
##  Max.   :0.06500   Max.   :0.39700   Max.   :0.5970   Max.   :0.26600
##       mdg              mdt              mds              mdn
##  Min.   :0.1210   Min.   :0.1170   Min.   :0.0220   Min.   :0.0230
##  1st Qu.:0.5773   1st Qu.:0.4910   1st Qu.:0.5760   1st Qu.:0.4585
##  Median :0.6825   Median :0.6015   Median :0.6915   Median :0.6320
##  Mean   :0.6853   Mean   :0.6095   Mean   :0.6951   Mean   :0.6115
##  3rd Qu.:0.8125   3rd Qu.:0.7183   3rd Qu.:0.8110   3rd Qu.:0.7768
##  Max.   :1.2980   Max.   :1.2150   Max.   :1.3510   Max.   :1.2600
##       mdi              tmg                tmt
##  Min.   :0.1160   Min.   :-0.35300   Min.   :-0.291000
##  1st Qu.:0.5298   1st Qu.:-0.16150   1st Qu.:-0.101000
##  Median :0.6370   Median :-0.08100   Median :-0.018500
##  Mean   :0.6365   Mean   :-0.09298   Mean   :-0.004658
##  3rd Qu.:0.7455   3rd Qu.:-0.02525   3rd Qu.: 0.087750
##  Max.   :1.2470   Max.   : 0.19200   Max.   : 0.366000
##       tms                tmn                tmi                 mr
##  Min.   :-0.44900   Min.   :-0.51000   Min.   :-0.40500   Min.   :0.6470
##  1st Qu.:-0.13525   1st Qu.:-0.23100   1st Qu.:-0.12750   1st Qu.:0.8260
##  Median :-0.03150   Median :-0.14650   Median :-0.03600   Median :0.8995
##  Mean   :-0.03981   Mean   :-0.14720   Mean   :-0.03651   Mean   :0.9050
##  3rd Qu.: 0.06800   3rd Qu.:-0.05625   3rd Qu.: 0.04950   3rd Qu.:0.9685
##  Max.   : 0.35800   Max.   : 0.24500   Max.   : 0.41800   Max.   :1.3170
##       rnf               mdic             emd               mv
##  Min.   :-0.2970   Min.   :0.0120   Min.   :0.0470   Min.   :0.00000
##  1st Qu.: 0.1197   1st Qu.:0.1440   1st Qu.:0.2305   1st Qu.:0.02100
##  Median : 0.1820   Median :0.2270   Median :0.2980   Median :0.02800
##  Mean   : 0.1824   Mean   :0.2313   Mean   :0.3089   Mean   :0.03354
##  3rd Qu.: 0.2370   3rd Qu.:0.2993   3rd Qu.:0.3792   3rd Qu.:0.03825
##  Max.   : 0.4510   Max.   :0.6630   Max.   :0.7430   Max.   :0.18300
##       Class
##  glaucoma:98
##  normal  :98
##
##
##
##
```

No NA's are present in the summary of the data, so we do not need to worry about NA's.

Now let's check if the variables are coded with correct data types:

```r
lapply(GlaucomaM_data, class)
```

```
## $ag
## [1] "numeric"
##
## $at
## [1] "numeric"
##
## $as
## [1] "numeric"
##
## $an
## [1] "numeric"
##
## $ai
## [1] "numeric"
##
## $eag
## [1] "numeric"
##
## $eat
## [1] "numeric"
##
## $eas
## [1] "numeric"
##
## $ean
## [1] "numeric"
##
## $eai
## [1] "numeric"
##
## $abrg
## [1] "numeric"
##
## $abrt
## [1] "numeric"
##
## $abrs
## [1] "numeric"
##
## $abrn
## [1] "numeric"
##
## $abri
## [1] "numeric"
##
## $hic
## [1] "numeric"
##
## $mhcg
## [1] "numeric"
##
```

```
## $mhct
## [1] "numeric"
##
## $mhcs
## [1] "numeric"
##
## $mhcn
## [1] "numeric"
##
## $mhci
## [1] "numeric"
##
## $phcg
## [1] "numeric"
##
## $phct
## [1] "numeric"
##
## $phcs
## [1] "numeric"
##
## $phcn
## [1] "numeric"
##
## $phci
## [1] "numeric"
##
## $hvc
## [1] "numeric"
##
## $vbsg
## [1] "numeric"
##
## $vbst
## [1] "numeric"
##
## $vbss
## [1] "numeric"
##
## $vbsn
## [1] "numeric"
##
## $vbsi
## [1] "numeric"
##
## $vasg
## [1] "numeric"
##
## $vast
## [1] "numeric"
##
## $vass
## [1] "numeric"
##
```

```
## $vasn
## [1] "numeric"
##
## $vasi
## [1] "numeric"
##
## $vbrg
## [1] "numeric"
##
## $vbrt
## [1] "numeric"
##
## $vbrs
## [1] "numeric"
##
## $vbrn
## [1] "numeric"
##
## $vbri
## [1] "numeric"
##
## $varg
## [1] "numeric"
##
## $vart
## [1] "numeric"
##
## $vars
## [1] "numeric"
##
## $varn
## [1] "numeric"
##
## $vari
## [1] "numeric"
##
## $mdg
## [1] "numeric"
##
## $mdt
## [1] "numeric"
##
## $mds
## [1] "numeric"
##
## $mdn
## [1] "numeric"
##
## $mdi
## [1] "numeric"
##
## $tmg
## [1] "numeric"
##
```

```
## $tmt
## [1] "numeric"
##
## $tms
## [1] "numeric"
##
## $tmn
## [1] "numeric"
##
## $tmi
## [1] "numeric"
##
## $mr
## [1] "numeric"
##
## $rnf
## [1] "numeric"
##
## $mdic
## [1] "numeric"
##
## $emd
## [1] "numeric"
##
## $mv
## [1] "numeric"
##
## $Class
## [1] "factor"
```

All variables are classified correctly.

Let us now create our training and test datasets so we can move on to model fitting:

```
set.seed(5) #random sample remains fixed in every run in R.
index_training <- sample(1:nrow(GlaucomaM_data), round(0.7*nrow(GlaucomaM_data)))
training_data <- GlaucomaM_data[index_training,]
test_data <- GlaucomaM_data[-index_training,]
```

Model Fitting: Elastic Net

Our response variable is "Class", while the predictors are all the other variables.

```
X <- as.matrix(training_data[,-63])
Y <- training_data[,63]

library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 3.2.4
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-5
```

```r
fit1 <- cv.glmnet(X,Y, family='binomial')
```

The next command tells which covariates are selected. In addition, it gives the estimated coefficients. In this project, we care more about whether we have a good predictor for glaucoma or not, and not so much how individual variables affect the outcome. So we do not care much about the values of the estimated coefficients.

```r
coef(fit1, s = "lambda.min")
```

```
## 63 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept) -1.6431367
## ag           .
## at           .
## as           .
## an           .
## ai           .
## eag          .
## eat          .
## eas          .
## ean          .
## eai          .
## abrg         .
## abrt         .
## abrs        -1.5224756
## abrn         .
## abri         .
## hic          .
## mhcg         .
## mhct         .
## mhcs         .
## mhcn        -0.5013093
## mhci        -6.9850653
## phcg         .
## phct         .
## phcs         .
## phcn        -1.7603040
## phci        -3.2963381
## hvc          .
## vbsg         .
## vbst         .
## vbss         .
## vbsn         .
## vbsi         .
## vasg         .
## vast         .
## vass         .
## vasn         .
## vasi         .
## vbrg         .
```

```
## vbrt          .
## vbrs          .
## vbrn          .
## vbri          .
## varg          2.6557843
## vart          .
## vars         16.1340404
## varn          .
## vari          .
## mdg           .
## mdt           .
## mds           .
## mdn           .
## mdi           .
## tmg           .
## tmt           .
## tms          -0.1314503
## tmn           .
## tmi          -1.5718128
## mr            .
## rnf           1.4466328
## mdic          .
## emd           .
## mv            .
```

The selected covariates in this model were "abrs", "mhcn", "mhci", "phcn", "phci", "varg", "vars", "tms", "tmi", and "rnf".

Let us see the performance in the training data first:

```
#The function "show" calculates missclassification error, i.e. how many people cases of "normal" or "gl
show <- function(tt){
  print(tt)
  cat(paste("Misclassification rate =", round(1-sum(diag(tt))/sum(tt),2),"\n"))
  invisible()
}

nx <- as.matrix(training_data[,-63])
nrow(training_data)
```

```
## [1] 137
```

```
nrow(test_data)
```

```
## [1] 59
```

```
show(with(training_data, table(actual=Y,
                               predicted=predict(fit1, newx = nx, s="lambda.min", type="class"))))
```

```
##           predicted
## actual     glaucoma normal
##   glaucoma       65      8
##   normal          9     55
## Misclassification rate = 0.12
```

Now let us see the performance on the test data:

```
nx <- as.matrix(test_data[,-63])
show(with(test_data, table(actual=test_data[,63],
                           predicted=predict(fit1, newx = nx, s="lambda.min", type="class"))))
```

```
##           predicted
## actual     glaucoma normal
##   glaucoma       22      3
##   normal          9     25
## Misclassification rate = 0.2
```

Conclusion

The misclassification rate overall in the training data is okay, but in the test data the misclassification rate is high. This could suggest:

- The model is overfit.
- The predictors are not good enough, i.e., we need better predictors.
- The model is not good enough.
- There is too much "randomness" in the data.

However, if we consider that the most "dangerous" prediction to be made is predict that a person would not have glaucoma, when actually the person had glaucoma, the results in the training and test data do not differ much. Only 3 people actually had glaucoma when the prediction did not say so in the test data, resulting in a misclassification rate of $0.12 = 3/(22+3)$, while in the training data this rate was around $0.109 = 8/(65+8)$.