Asteroid 4

Annika Lin

2023-04-04

```
df <- read.csv("~/Documents/Georgetown/Spring23/Statistical Learning & Data Science/Project/NASA-asteroid-Classification-master/nasa_4_4_23.csv")
df <- df[ , !(names(df) %in% c("X"))]
```

#Lasso

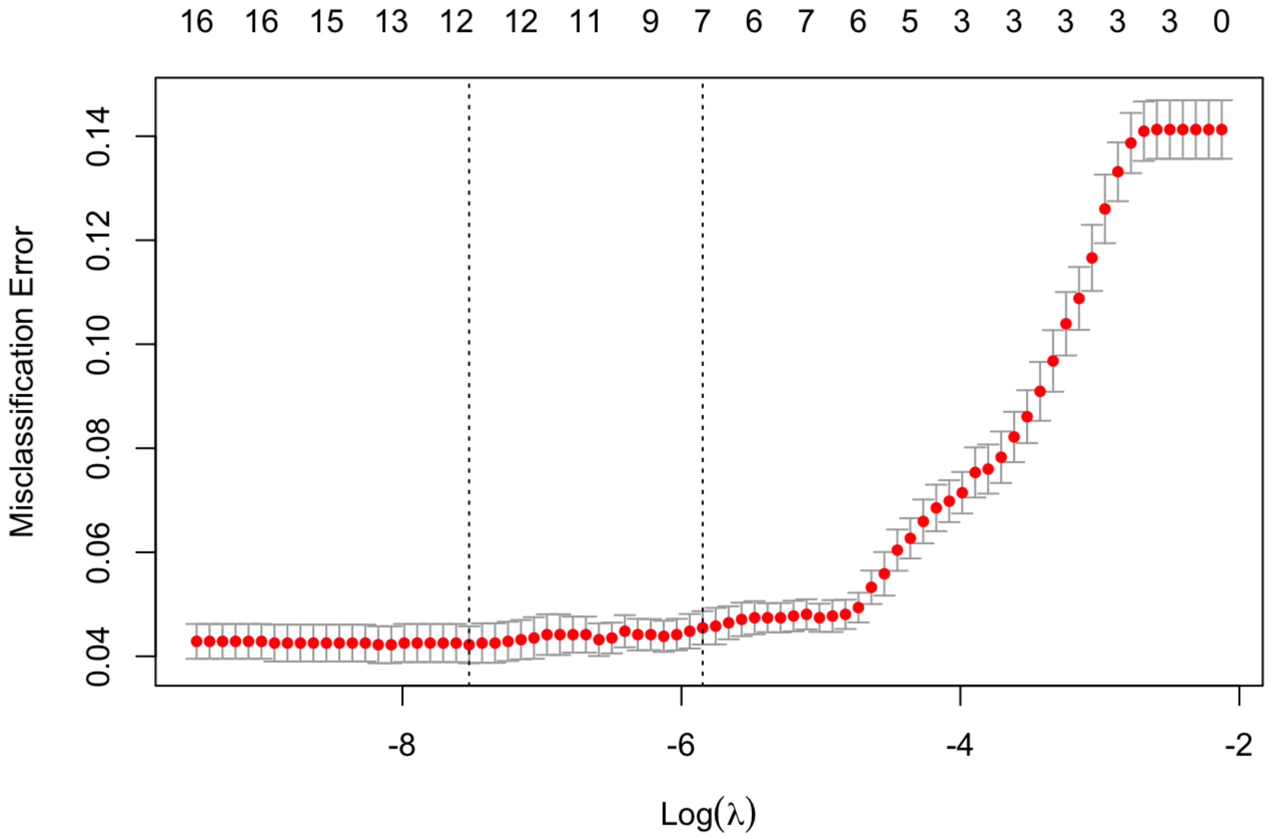
```
library(glmnet)
```

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

```
## Loaded glmnet 4.1-7
```

```
# we use the function model.matrix to create the design matrix
X = model.matrix(Hazardous ~ ., data=df)
Y = as.numeric(df$Hazardous=="True")

# cv.glmnet is the main function to do cross-validation.
# Here we use "class", the misclassification error, as criterion.
# Other options include "deviance" (the default) and "auc""
set.seed(1)
cvfit = cv.glmnet(x=X[,-1], y=Y, family="binomial", type.measure="class")
plot(cvfit)
```



```
coef(cvfit, s=cvfit$lambda.1se)
```

```
## 21 x 1 sparse Matrix of class "dgCMatrix"
##                                          s1
## (Intercept)                   3.850936e+01
## Absolute.Magnitude           -1.647259e+00
## Est.Dia.in.KM.min.           -4.964057e+00
## Est.Dia.in.KM.max.            .           
## Close.Approach.Date           .           
## Relative.Velocity.km.per.sec  .           
## Miss.Dist..kilometers.        .           
## Orbit.Uncertainity           -1.155934e-01
## Minimum.Orbit.Intersection   -7.502262e+01
## Jupiter.Tisserand.Invariant   .           
## Eccentricity                  .           
## Semi.Major.Axis               .           
## Inclination                   4.250769e-04
## Asc.Node.Longitude            .           
## Orbital.Period                .           
## Perihelion.Distance           .           
## Perihelion.Arg                .           
## Aphelion.Dist                 .           
## Mean.Anomaly                  .           
## Mean.Motion                  -2.583585e-01
## Range.Dia.in.KM              -3.477517e-01
```

```
sel.vars <- which(coef(cvfit, s=cvfit$lambda.1se)!=0)[-1]-1
sel.names <- colnames(df)[sel.vars]
sel.names
```

```
## [1] "Absolute.Magnitude"         "Est.Dia.in.KM.min."        
## [3] "Orbit.Uncertainity"         "Minimum.Orbit.Intersection"
## [5] "Inclination"                "Mean.Motion"               
## [7] "Range.Dia.in.KM"
```

Logistic Regression

```
df$Hazardous <- df$Hazardous=="True"

fit.lasso <- glm(df$Hazardous ~ Absolute.Magnitude+Est.Dia.in.KM.min.+Orbit.Uncertainity
+Minimum.Orbit.Intersection+Inclination+Mean.Motion+Range.Dia.in.KM,
                 family="binomial", data=df)
```

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

```
summary(fit.lasso)
```

```
##
## Call:
## glm(formula = df$Hazardous ~ Absolute.Magnitude + Est.Dia.in.KM.min. +
##     Orbit.Uncertainity + Minimum.Orbit.Intersection + Inclination +
##     Mean.Motion + Range.Dia.in.KM, family = "binomial", data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3682  -0.0356  -0.0026   0.0000   6.6252
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 7.028e+01  5.265e+00  13.349  < 2e-16 ***
## Absolute.Magnitude         -2.990e+00  2.292e-01 -13.047  < 2e-16 ***
## Est.Dia.in.KM.min.         -5.990e+09  1.828e+09  -3.277  0.00105 **
## Orbit.Uncertainity         -1.453e-01  4.845e-02  -2.999  0.00271 **
## Minimum.Orbit.Intersection -1.238e+02  8.372e+00 -14.786  < 2e-16 ***
## Inclination                 1.343e-02  1.185e-02   1.133  0.25739
## Mean.Motion                -5.795e-01  3.283e-01  -1.765  0.07757 .
## Range.Dia.in.KM             4.846e+09  1.479e+09   3.277  0.00105 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2508.03  on 3078  degrees of freedom
## Residual deviance:  544.57  on 3071  degrees of freedom
## AIC: 560.57
##
## Number of Fisher Scoring iterations: 9
```

Signficiant variables (0.001 level)

Absolute.Magnitude+Est.Dia.in.KM.min.+Orbit.Uncertainity+Minimum.Orbit.Intersection++Range.Dia.in.KM

#PCR

```
library(pls)
```

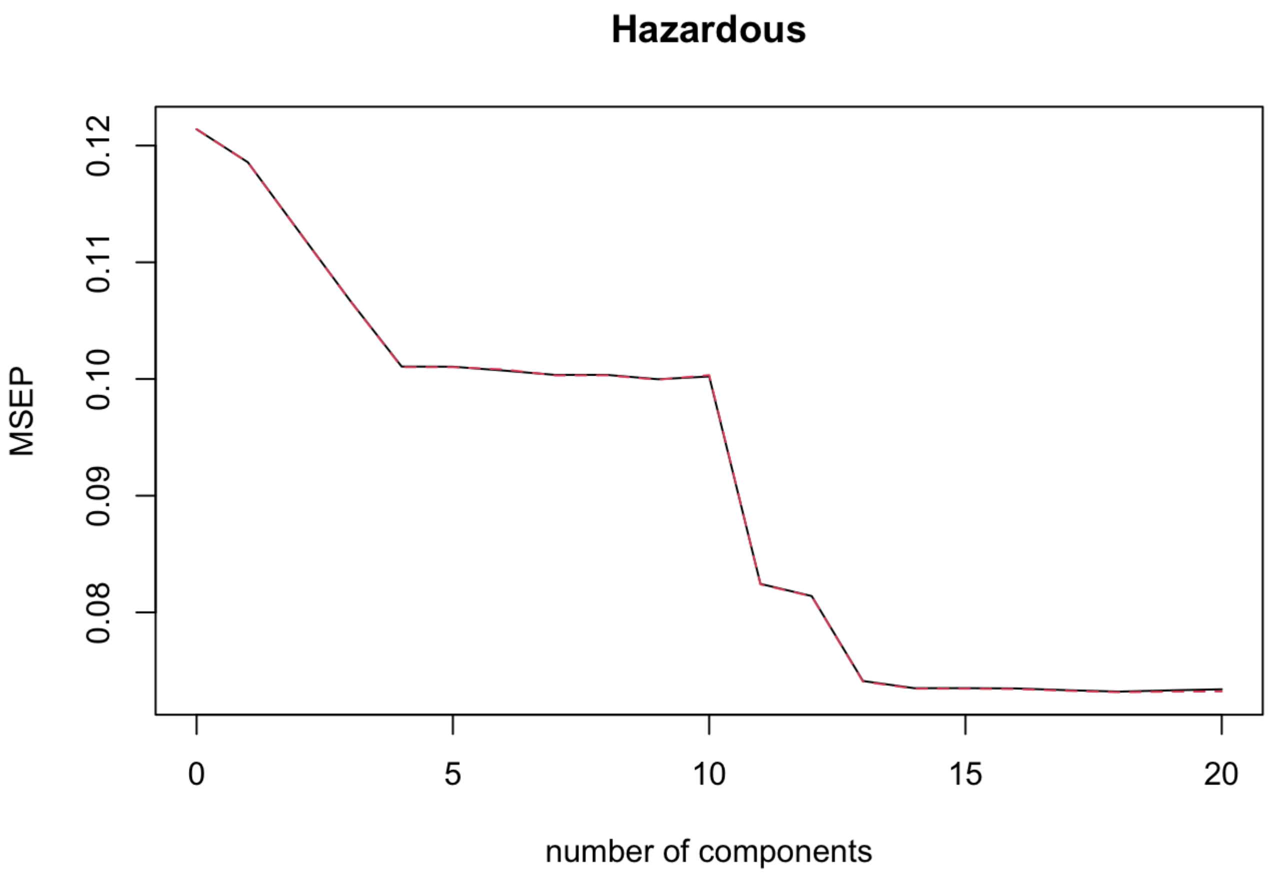
```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
 set.seed(1)
pcr.fit <- pcr(Hazardous ~ ., scale=T, validation="CV", segments=10,
               data=df)
summary(pcr.fit)
```

```
## Data:    X dimension: 3079 20 
##  Y dimension: 3079 1
## Fit method: svdpc
## Number of components considered: 20
## 
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV          0.3484   0.3443   0.3356   0.3266   0.3179   0.3179   0.3174
## adjCV       0.3484   0.3443   0.3356   0.3266   0.3178   0.3178   0.3175
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV      0.3168   0.3168   0.3162    0.3166    0.2871    0.2853    0.2723
## adjCV   0.3167   0.3167   0.3161    0.3168    0.2871    0.2853    0.2722
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
## CV       0.2711    0.2711    0.2711    0.2708    0.2706    0.2708    0.2709
## adjCV    0.2710    0.2710    0.2710    0.2707    0.2705    0.2706    0.2706
## 
## TRAINING: % variance explained
##            1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X           30.594   52.940    62.23    68.57    73.85    78.72    83.59
## Hazardous    2.432    7.535    12.47    17.29    17.31    17.45    17.93
##            8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
## X            88.35    91.86     94.57     97.10      98.2     99.15     99.84
## Hazardous    17.98    18.32     18.32     32.73      33.7     39.60     40.10
##            15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
## X             99.99    100.00    100.00    100.00    100.00    100.00
## Hazardous     40.18     40.24     40.38     40.51     40.51     40.51
```

```
validationplot(pcr.fit, val.type="MSEP")
```



```
pca.nasa <- prcomp(df[,1:20], scale=T)
summary(pca.nasa)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.4736 2.1140 1.36288 1.12629 1.02750 0.98729 0.98678
## Proportion of Variance 0.3059 0.2235 0.09287 0.06343 0.05279 0.04874 0.04869
## Cumulative Proportion  0.3059 0.5294 0.62227 0.68570 0.73849 0.78722 0.83591
##                            PC8     PC9   PC10   PC11    PC12    PC13    PC14
## Standard deviation     0.97536 0.83817 0.7362 0.7113 0.46971 0.43429 0.37257
## Proportion of Variance 0.04757 0.03513 0.0271 0.0253 0.01103 0.00943 0.00694
## Cumulative Proportion  0.88348 0.91860 0.9457 0.9710 0.98203 0.99146 0.99840
##                           PC15    PC16    PC17      PC18      PC19      PC20
## Standard deviation     0.17540 0.03026 0.01629 1.736e-10 4.266e-15 3.846e-15
## Proportion of Variance 0.00154 0.00005 0.00001 0.000e+00 0.000e+00 0.000e+00
## Cumulative Proportion  0.99994 0.99999 1.00000 1.000e+00 1.000e+00 1.000e+00
```

```
pc.dat <- data.frame(pca.nasa$x[,1:4], df$Hazardous)
pc.logit <- glm(df.Hazardous ~ PC1+PC2+PC3+PC4,
                family="binomial", data=pc.dat)
summary(pc.logit)
```

```
##
## Call:
## glm(formula = df.Hazardous ~ PC1 + PC2 + PC3 + PC4, family = "binomial",
##     data = pc.dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8422  -0.5101  -0.3614  -0.2413   2.6489
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.21853    0.06904 -32.133  < 2e-16 ***
## PC1         -0.13113    0.02188  -5.994 2.05e-09 ***
## PC2          0.26396    0.02565  10.293  < 2e-16 ***
## PC3         -0.54492    0.04393 -12.404  < 2e-16 ***
## PC4          0.60242    0.05586  10.785  < 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: 2508.0  on 3078  degrees of freedom
## Residual deviance: 2021.1  on 3074  degrees of freedom
## AIC: 2031.1
##
## Number of Fisher Scoring iterations: 5
```

PCs not significant since p value is large.