Heart dataset logistic regression

Aigerim Madakimova

1) Given a2+b2+c**2 = 0, find the roots (values of a,b,c that satisfies the above constraint). Explain. What optimization method will you use? And why?

Answer: By analogy with conic sections, there are also degenerate surfaces of the second order. So, the second-order equation a^2 = 0 describes a pair of coincident planes, equation a^2 = 1 describes a pair of parallel planes, equation a^2 − b^2 = 0 describes a pair of intersecting planes. The equation a^2 + b^2 + c^2 = 0 describes a point with coordinates (0; 0; 0). In general this equation has no solutions. In order to solve the equation we could use minimization since we have to get rid of the power while finding the minimum.

2) consider the following R code

——————variableImportance——————

```
if(!require(tfdatasets)) install.packages(c('tfdatasets'))
```

```
## Loading required package: tfdatasets
```

```
library(tfdatasets)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data(hearts)
hearts$thal <- as.numeric(factor(hearts$thal))
if(!require(rpart)) install.packages(c('rpart'))
```

```
## Loading required package: rpart
```

```
library(rpart)
require(rpart)
tree.heart<-rpart(target~.,data=hearts)
tree.heart$variable.importance
```

```
##    oldpeak         cp    thalach      slope      exang       thal         ca
## 14.6813936 12.1272338  9.5276438  6.4108006  5.5175952  5.0662849  3.2482943
##   trestbps        age       chol        fbs        sex    restecg
##  2.2255524  1.9139321  0.9934904  0.4120926  0.2846873  0.1454545
```

```
head(hearts)
```

```
## # A tibble: 6 x 14
##     age   sex    cp trestbps  chol   fbs restecg thalach exang oldpeak slope
##   <int> <int> <int>    <int> <int> <int>   <int>   <int> <int>   <dbl> <int>
## 1    63     1     1      145   233     1       2     150     0     2.3     3
## 2    67     1     4      160   286     0       2     108     1     1.5     2
## 3    67     1     4      120   229     0       2     129     1     2.6     2
## 4    37     1     3      130   250     0       0     187     0     3.5     3
## 5    41     0     2      130   204     0       2     172     0     1.4     1
## 6    56     1     2      120   236     0       0     178     0     0.8     1
## # ... with 3 more variables: ca <int>, thal <dbl>, target <int>
```

————————variableImportance———————————— ### (A) set.seed(your_favorite_seed) Split your dataset 70/30 into training and test sets. Perform the below tasks using the same training and tests:

```
set.seed(42)

## 70% of the sample size
smp_size <- floor(0.7 * nrow(hearts))

## set the seed to make your partition reproducible
train_ind <- sample(seq_len(nrow(hearts)), size = smp_size)

train <- hearts[train_ind, ]
test <- hearts[-train_ind, ]
```

(A.1) Run Logistic Regression with the top 5 variables as indicated by variable.importance for the heart dataset

```
model_1 = glm(target ~ oldpeak + cp+ thalach+slope+exang, data = train, family = binomial)
summary(model_1)
```

```
##
## Call:
## glm(formula = target ~ oldpeak + cp + thalach + slope + exang,
##     family = binomial, data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.93880  -0.52808  -0.29028  -0.03845   2.30516
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.395810   2.162845  -2.032 0.042111 *
```

```
## oldpeak      0.807799   0.220758   3.659 0.000253 ***
## cp           0.880077   0.290988   3.024 0.002491 **
## thalach     -0.009002   0.009638  -0.934 0.350306
## slope        0.068332   0.420037   0.163 0.870770
## exang        1.275491   0.436028   2.925 0.003442 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 238.43  on 211  degrees of freedom
## Residual deviance: 153.29  on 206  degrees of freedom
## AIC: 165.29
##
## Number of Fisher Scoring iterations: 5
```

(A.2) repeat with the top 3 variables

```r
model_2 = glm(target ~ oldpeak + cp+ thalach, data = train, family = binomial)
summary(model_2)
```

```
##
## Call:
## glm(formula = target ~ oldpeak + cp + thalach, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6567  -0.5904  -0.3191  -0.0216   2.5460
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.459287   1.959866  -1.765 0.077553 .
## oldpeak      0.859166   0.193561   4.439 9.05e-06 ***
## cp           1.088757   0.297099   3.665 0.000248 ***
## thalach     -0.015995   0.008928  -1.791 0.073227 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 238.43  on 211  degrees of freedom
## Residual deviance: 162.00  on 208  degrees of freedom
## AIC: 170
##
## Number of Fisher Scoring iterations: 6
```

(B) Determine the better performing model!

```r
# Anova of two best models selected
anova(model_1, model_2, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ oldpeak + cp + thalach + slope + exang
## Model 2: target ~ oldpeak + cp + thalach
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       206     153.29
## 2       208     162.00 -2  -8.7099  0.01284 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(C) Explain how you determined the better model?

Answer: According to the likelihood ratio test results of anova, model_2 values are statistically more significant than model_1 values, therefore we picked model_2 to construct classification of heart diseases.

(D) Assume you had no variableImportance, can you think of any other method to be selective. Can you think of any other "objective" method to identify features that can better fit?

Answer:Since we have multiple independent variables, we run chi square test to understand the relationship between predictor and each of the independent variables. we can try to fit the all values and pic the ones that are more statistically significant, so more stars near the column - the more statistically significant it is. Also we can try to use varImp(model) this is almost the same thing.

```r
summary(glm(target ~., data = train, family = binomial))
```

```
##
## Call:
## glm(formula = target ~ ., family = binomial, data = train)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.71314  -0.38587  -0.18173  -0.01881   2.16107
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.082e+01  3.994e+00  -2.710 0.006738 **
## age          4.815e-03  3.283e-02   0.147 0.883377
## sex          1.073e+00  7.284e-01   1.473 0.140724
## cp           7.465e-01  3.165e-01   2.358 0.018349 *
## trestbps     8.157e-03  1.447e-02   0.564 0.572849
## chol         9.076e-04  5.563e-03   0.163 0.870406
## fbs          1.822e-01  5.913e-01   0.308 0.758033
## restecg      1.301e-01  2.434e-01   0.535 0.592850
## thalach     -5.052e-03  1.206e-02  -0.419 0.675136
## exang        1.216e+00  5.332e-01   2.280 0.022623 *
## oldpeak      6.811e-01  2.501e-01   2.724 0.006457 **
## slope        3.464e-01  4.721e-01   0.734 0.463135
## ca           1.005e+00  2.735e-01   3.675 0.000238 ***
```

```
## thal         5.759e-01  3.807e-01   1.513 0.130334
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 238.43  on 211  degrees of freedom
## Residual deviance: 121.06  on 198  degrees of freedom
## AIC: 149.06
##
## Number of Fisher Scoring iterations: 6
```

D.1) Using an objective method, identify top 5 features and top 3 features as you did in (B) without using variableImportance, run two different models and compare the model with the corresponding better models you identified in (B). So in total you have run 4 different models. Compare and contrast the results.

```r
model_3 = glm(target ~ ca + oldpeak+ exang+sex+thal, data = train, family = binomial)
summary(model_3)
```

```
##
## Call:
## glm(formula = target ~ ca + oldpeak + exang + sex + thal, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6630  -0.4099  -0.2134  -0.0474   2.5668
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -7.9348     1.6974  -4.675 2.95e-06 ***
## ca            1.0773     0.2297   4.690 2.73e-06 ***
## oldpeak       0.7347     0.2007   3.661 0.000252 ***
## exang         1.8301     0.4615   3.965 7.33e-05 ***
## sex           0.6253     0.5836   1.071 0.283995
## thal          0.8296     0.3619   2.293 0.021872 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 238.43  on 211  degrees of freedom
## Residual deviance: 133.00  on 206  degrees of freedom
## AIC: 145
##
## Number of Fisher Scoring iterations: 6
```

```r
model_4 = glm(target ~ ca + oldpeak+ exang, data = train, family = binomial)
summary(model_4)
```

```
##
```

```
## Call:
## glm(formula = target ~ ca + oldpeak + exang, family = binomial,
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5040  -0.4826  -0.2366  -0.1023   2.5366
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.9467     0.5062  -7.796 6.39e-15 ***
## ca            1.1039     0.2244   4.919 8.70e-07 ***
## oldpeak       0.7704     0.1930   3.993 6.53e-05 ***
## exang         2.0306     0.4460   4.553 5.30e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 238.43  on 211  degrees of freedom
## Residual deviance: 140.92  on 208  degrees of freedom
## AIC: 148.92
##
## Number of Fisher Scoring iterations: 5
```

```r
# Anova of two best models selected
anova(model_3, model_4, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: target ~ ca + oldpeak + exang + sex + thal
## Model 2: target ~ ca + oldpeak + exang
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       206     133.00
## 2       208     140.92 -2  -7.9179  0.01908 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: According to the likelihood ratio test results of anova, model_4 values are statistically more significant than model_3 values, therefore we picked model_4 to construct classification of heart diseases

(E) Write a summary explaining your findings.

Answer: In this work, we have examined the presense of a heart disease from a set of variables including chollesterol, age, sex and others provided in the dataset. We found logistic regression model_4 to be a better model for prediction. Using anova and likelihood ratio test.