Statistical Learning, Homework #2

Annalisa Xamin

Introduction

In the following analysis, we will focus on cancer data to investigate the correlation between the level of prostate-specific antigen (`lpsa`, in ng/ml and log scaled) and a number of clinical measures, measured in 97 men who were about to receive a radical prostatectomy. In particular, the 9 explanatory variables are:

- `lcavol`: log(cancer volume in cm^3)
- `lweight`: log(prostate weight in g)
- `age` in years
- `lbph`: log(amount of benign prostatic hyperplasia in cm^2)
- `svi`: seminal vesicle invasion (1 = yes, 0 = no)
- `lcp`: log(capsular penetration in cm)
- `gleason`: Gleason score for prostate cancer (6,7,8,9)
- `pgg45`: percentage of Gleason

During the analysis we will use three different methods (cost-complexity decision trees, random forests and boosting) to later on compare their performances.

To start, the data is loaded and a summary is printed.

```r
df <- read.csv("./prostate.csv")
summary(df)
```

```
##      lcavol           lweight           age             lbph        
##  Min.   :-1.3471   Min.   :2.375   Min.   :41.00   Min.   :-1.3863  
##  1st Qu.: 0.5128   1st Qu.:3.376   1st Qu.:60.00   1st Qu.:-1.3863  
##  Median : 1.4469   Median :3.623   Median :65.00   Median : 0.3001  
##  Mean   : 1.3500   Mean   :3.629   Mean   :63.87   Mean   : 0.1004  
##  3rd Qu.: 2.1270   3rd Qu.:3.876   3rd Qu.:68.00   3rd Qu.: 1.5581  
##  Max.   : 3.8210   Max.   :4.780   Max.   :79.00   Max.   : 2.3263  
##       svi              lcp             gleason          pgg45       
##  Min.   :0.0000   Min.   :-1.3863   Min.   :6.000   Min.   :  0.00  
##  1st Qu.:0.0000   1st Qu.:-1.3863   1st Qu.:6.000   1st Qu.:  0.00  
##  Median :0.0000   Median :-0.7985   Median :7.000   Median : 15.00  
##  Mean   :0.2165   Mean   :-0.1794   Mean   :6.753   Mean   : 24.38  
##  3rd Qu.:0.0000   3rd Qu.: 1.1787   3rd Qu.:7.000   3rd Qu.: 40.00  
##  Max.   :1.0000   Max.   : 2.9042   Max.   :9.000   Max.   :100.00  
##       lpsa
##  Min.   :-0.4308
##  1st Qu.: 1.7317
##  Median : 2.5915
##  Mean   : 2.4784
##  3rd Qu.: 3.0564
##  Max.   : 5.5829
```

As we can see from the summary, in the data there are no NAs.

Decision Tree

Fit a decision tree on the whole data and plot the results. Choose the tree complexity by cross-validation and decide whether you should prune the tree based on the results. Prune the tree if applicable and interpret the fitted model.

```
set.seed(1)
```

Random Forest

Consider now a random forest and let m be the number of variables to consider at each split. Set the range for m from 1 to the number of explanatory variables, say nvar, and define a k-fold cross-validation schema for the selection of this tuning parameter, with k of your choice. Prepare a matrix with nvar rows and 2 columns and fill the first column with the average cross-validation error corresponding to each choice of m and the second column with the OOB error (from the full dataset). Are the CV and OOB error different? Do they reach the minimum at the same value of m? Interpret the optimal model (either using the CV or the OOB error).

Boosted regression trees

Fit boosted regression trees making a selection of the number of boosting iterations (n.trees) by CV. Interpret your selected optimal model.

Comparison

Compare the performance of the three methods (cost-complexity decision trees, random forests and boosting) using cross-validation. Make sure that the model complexity is re-optimized at each choice of the training set (either using another CV or using the OOB error).

Conclusion

Draw some general conclusions about the analysis and the different methods that you have considered.