Introduction to Machine Learning in R
NPFL 054

Easy homework assigned on April 30, 2021

Contact teacher: Martin Holub
holub@ufal.mff.cuni.cz

In all tasks specified below you will work with dataset 'Auto', which is a part of ISLR package available in R. You will do a regression task and build a model to predict the value of mpg attribute using 7 features:

mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin.

> library(ISLR)
> str(Auto)
'data.frame': 392 obs. of 9 variables:
$ mpg : num 18 15 18 16 17 15 14 14 14 15 ...
$ cylinders : num 8 8 8 8 8 8 8 8 8 8 ...
$ displacement: num 307 350 318 304 302 429 454 440 455 390 ...
$ horsepower : num 130 165 150 150 140 198 220 215 225 190 ...
$ weight : num 3504 3693 3436 3433 3449 ...
$ acceleration: num 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
$ year : num 70 70 70 70 70 70 70 70 70 70 ...
$ origin : Factor w/ 3 levels "USA","Europe",..: 1 1 1 1 1 1 1 1 1 1 ...
$ name : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 2 ...

Here is an example how to build a Boosting Trees (BT) regression model using gbm package:

> library(gbm)
> model.boosting = gbm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin,
data = Auto, distribution = "gaussian",
n.trees = 500, shrinkage = 0.01, interaction.depth = 2)

You can use the example code related to boosting trees and posted on the course web page.

Task 1 – Boosting Trees and the number of splits d
Build a BT model and experiment with parameter d (interaction.depth). Take into consideratin also d = 1 (stumps). For different values of d make plots to show how the model performance depends on the number of trees. To estimate generalization error use 8-fold or 4-fold cross-validation.

Task 2 – Boosting Trees and overfitting
Can a BT model overfit? Experiment with parameters λ (shrinkage) and B (n.trees) and find some settings in which the model obviously underfits or overfits. Also, try to find an 'optimal' settings. To estimate generalization error use 8-fold or 4-fold cross-validation. Make a plot to illustrate how the model performance depends on the model complexity.
Task 3 – Boosting trees and Random Forest
Compare Boosting Trees and Random Forest models with 1000 trees. Try to tune other parameters to get best performance. To estimate generalization error use 8-fold or 4-fold cross-validation. Which of the two methods is better for the given data?

Task 4 – Boosting trees and Linear Regression
Compare Boosting Trees and Linear Regression models. To estimate generalization error use 8-fold or 4-fold cross-validation. When you compare the two methods, also look at the distribution of residuals in the union of cross-validation test sets.

Illustration – output of the example code posted on the course web page.