732A95 INTRODUCTION MACHINE LEARNING

LAB 2A BLOCK 2: ENSEMBLE METHODS

JOSE M. PEÑA
IDA, LINKÖPING UNIVERSITY, SWEDEN

INSTRUCTIONS

Each student must submit a report with his/her solutions to the lab. Submission is done via LISAM and before the deadline. The submission file should be named Name_LastName.pdf. The report must be concise but complete. It should include (i) the code implemented or the calls made to existing functions, (ii) the results of such code or calls, and (iii) explanations for (i) and (ii).

PhD students pass the lab if their individual report is of sufficient quality. TDDE01 and 732A95 students pass the lab as follows. The students must discuss their lab solutions in a group. Each group must compile a collaborative report that will be used for presentation at the seminar. The report should clearly state the names of the students that participated in its compilation and a short description of how each student contributed to the report. This report should be submitted via LISAM and before the deadline. The file should be named Group_X.pdf where X is the group number. The collaborative reports are corrected and graded. The individual reports are also checked, but feedback on them will not be given. The students pass the lab if their group report passes the seminar and their individual reports have reasonable quality, otherwise the students must complete their individual reports by correcting the mistakes in them.

Please use `set.seed(1234567890)` at the beginning of each assignment to ensure reproducibility.

RESOURCES

The lab is designed to be solved with the R packages `tree`, `mboost` and `randomForest`.

ASSIGNMENT 1

Prove that the squared error of the bagging regression is $1/B$ of the average error of the individual regressions, under the assumption that the error terms $\epsilon^b(x)$ have zero mean and are uncorrelated. In other words, prove that

$$E_X[(f_{bag}(x) - h(x))^2] = \frac{1}{B}\Big[\frac{1}{B}\sum_b E_X[(f^b(x) - h(x))^2]\Big]$$

For simplicity, you can take $B = 3$. Moreover, recall that $E[\alpha U + \beta V] = \alpha E[U] + \beta E[V]$. Recall also that uncorrelated means zero covariance, and that the latter is defined as $E[(U - E[U])(V - E[V])]$.

ASSIGNMENT 2

The file `bodyfatregression.xls` contains records of waist measure (cm), weight (kg) and body fat (%) for 110 persons. Your task is to carry out a regression tree analysis using the function `tree()` of the R package `tree`. Let the body fat be the response variable and the waist measure and weight the predictor variables.

2.1. Estimate an upper bound of the squared error of the bagging regression tree. To do that, compute the average error of a set of individual regression trees. To estimate these individual errors, use 2/3 of the data for training and 1/3 as hold-out test data.

2.2. Repeat the previous question using 3-fold cross-validation instead of hold-out test data.

2.3. For the two previous scenarios, compute the bagging regression tree that you would return to the user.

ASSIGNMENT 3

The file `bodyfatregression.xls` contains records of waist measure (cm), weight (kg) and body fat (%) for 110 persons. Your task is to carry out a boosting regression tree analysis using the function `blackboost()` of the R package `mboost`. Let the body fat be the response variable and the waist measure and weight the predictor variables.

3.1. Interpret the plot resulting from the code below.

```
bf=read.csv2("bodyfatregression.csv")
m=blackboost(Bodyfat_percent~Waist_cm+Weight_kg, data=bf)
mstop(m)
cvf=cv(model.weights(m), type = "kfold")
cvm=cvrisk(m, folds =cvf, grid = 1:100)
plot(cvm)
```

3.2. Estimate the squared error of the boosting regression tree. Use 2/3 of the data for training and 1/3 as hold-out test data. Let the boosting procedure choose the appropriate number of trees by adding the parameter `control=boost_control(mstop=mstop(cvm))` to the function `blackboost()`.

ASSIGNMENT 4

The file `spambase.xls` contains information about the frequency of various words, characters, etc. for a total of 4601 e-mails. Furthermore, these e-mails have been classified as spams (spam = 1) or regular e-mails (spam = 0). You can find more information about these data at `https://archive.ics.uci.edu/ml/datasets/Spambase`

Your task is to evaluate the performance of Adaboost classification trees and random forests on the spam data. Specifically, provide a plot showing the error rates when the number of trees considered are $10, 20, \ldots, 100$. To estimate the error rates, use 2/3 of the data for training and 1/3 as hold-out test data.

To learn Adaboost classification trees, use the function `blackboost()` of the R package `mboost`. Specify the loss function corresponding to Adaboost with the parameter `family`. To learn random forests, use the function `randomForest` of the R package `randomForest`.