BUAN 6356.003 Spring 2017 (Johnston)

Project 3: Cross Validation and Ensemble Estimation

Due by: 26 Mar 2017

Description:

This project is set up in 3 segments. Deliverables for each segment are cumulative. Use alpha = 0.05 wherever applicable. Do NOT save any information to disk from your submitted R code. Use a random number seed value of 385766359. Calculate any needed residual values as original value – fitted value.

Code for each segment is to be submitted through eLearning. Multiple code submissions are allowed. Each instance of submitted code will be run and the deliverables compared to a reference result.

Segment 1.

Download the file "Boston.csv" from the BUAN_6356>projects area. Use this data to generate a "training" sample (90%) and a "testing" sample (10%). Define and extract the "testing" sample first. Keep all variables originally in each sample. Use the "training" set to build a final regression model explaining the median home value ("medv") using all numeric variables in the data set.

Calculate fitted values for the "testing" sample.

Update both "training" and "testing" data frames to include fitted values (named "olsFit") and residuals (named "olsResid").

Segment 1 Deliverables:

1. Updated "training" data frame (name: housingTrain)
2. Updated "testing" data frame (name: housingTest)
3. Final regression model results (name: modelTrain)

Segment 2.

After successfully performing the actions specified in Segment 1, create regression trees to explain the median home value ("medv") from the original variables in the "training" sample using both rpart() and tree().

Calculate fitted values for the "testing" sample.

Update both "training" and "testing" data frames to include fitted values for each regression tree procedure (named "rpartFit" and "treeFit") as well as residuals (named "rpartResid" and "treeResid").

Segment 2 Deliverables:

1. Updated "training" data frame (name: housingTrain)
2. Updated "testing" data frame (name: housingTest)
3. Final rpart() model results (name: rpartTrain)
4. Final tree() model results (name: treeTrain

Segment 3.

After successfully performing the actions specified in Segment 2, use the "training" set to build a final regression model explaining the median home value ("medv") using only the fitted data values now present in data frame "housingTrain".

Calculate fitted values from this new model for the "testing" sample.

Update both "training" and "testing" data frames to include fitted values for the final regression procedure (named "eFit") as well as residuals (named "eResid").

Segment 3 Deliverables:

1. Updated "training" data frame (name: housingTrain)
2. Updated "testing" data frame (name: housingTest)
3. Final regression model results (name: eTrain)