

Homework #4: Classification and Regression Trees, Random Forests and Boosting

Biostat 216: Machine Learning in R for the Biomedical Sciences

Classification and Regression Trees

1. Draw an example (of your own invention) of a partition of two-dimensional feature space that could occur from recursive binary splitting. Your partition should contain at least 5 regions. Draw a decision tree that corresponds to this partition. Label the figures showing the regions $R_1, R_2, \dots, R_d$, the cutpoints $t_1, t_2, \dots, t_{m-1}$ in the partition, and the cutpoints $t_1, t_2, \dots, t_{m-1}$ at the branches of the tree and corresponding regional means $m_1, m_2, \dots, m_d$ at the terminal nodes. (4 points)
2. Draw an example (of your own invention) of a partition of two-dimensional feature space that could not occur from recursive binary splitting. Your partition should contain at least 5 regions. Briefly explain why the partition cannot be represented by a decision tree. (2 points)
3. Read in the breast cancer imaging data “ispy1doctored.csv” into a data frame called **dat**
4. Load the **rpart** package
5. Generate a histogram of the **MRI_LD_Tfinal** variable that will be our outcome to predict (1 point)
6. Set the random seed to 10
7. Split the dataset into a training set of size 70 and a testset consisting of the remaining data (1 point)
8. Fit a regression tree to the training data with **MRI_LD_Tfinal** as outcome and all other variables as candidate predictors. Make sure that you specify the correct method for regression (2 points)
9. Plot the fitted tree and add text labels (2 points)
10. How many partitions does the fitted tree have? (1 point)
11. Determine how many nodes are in the tree and confirm your answer by examining the output of the **summary** command applied to the tree (2 points)
12. Determine the mean squared error (MSE) of the fitted tree based on the test data (2 points)

Bagging

12. Load the **randomForest** package and set the random seed to 4
13. Perform bagging with **MRI_LD_Tfinal** as outcome and all other variables as candidate predictors (2 points)
14. Determine the mean squared error (MSE) of the bagging based on the test data (1 point)

Random Forests

15. Set the seed to 4 again and perform random forests with tree sizes of 6, **MRI_LD_Tfinal** as outcome and all other variables as candidate predictors (2 points)
16. Determine the mean squared error (MSE) of the random forest based on the test data (1 point)

Boosting

17. Load the `gbm` package and set the random seed to 7
18. Fit a gradient boosted model with `MRI_LD_Tfinal` as outcome and all other variables as candidate predictors. Specify the Gaussian distribution since this is a regression. Use 5000 trees and set the `interaction.depth` to 1 (2 points)
19. Determine the mean squared error (MSE) of this gradient boosted model based on the test data with `n.trees=5000` (1 point)
20. Refit the gradient boosted model but this time set `interaction.depth` to 2 (1 point)
21. Determine the mean squared error (MSE) of this second gradient boosted model based on the test data with `n.trees=5000` (1 point)