Data Analysis in IC4.0 2022/2023

Second Assignment

This assignment involves the `Boston` dataset which is part of the `MASS` package.

1. (0.5 points) Create a binary variable which replaces `crim` in `Boston` dataset that contains a 1 with probability 0.7 if `crim` contains a value above its median, and a 0 with probability 0.7 if `crim` contains a value below its median.

2. (0.5 points) Split the dataset into two parts randomly such that one part contains (approximately) 80% of the observations, and the other the remaining 20%. From now on, first subset is the training data and second one the test data unless a different thing is said.

3. (1 point) Use the training data to perform a logistic regression with the new `crim` variable obtained in exercise 1. as the response and the remaining 13 variables as predictors. Compute the confusion matrix and the overall fraction of correct predictions (aka accuracy), sensitivity and specificity for the held out data (that is, the test set of observations) and a cut off value of 0.6.

4. (0.5 points) Do any of the predictors appear to be statistically significant at 1% significance level? If so, which ones? Repeat exercise 3. using this subset of predictors and compare the accuracy obtained with the previous one.

5. (1 point) Use the training data to perform a linear discriminant analysis with the new `crim` variable obtained in exercise 1. as the response and the remaining 13 variables as predictors. Compute the confusion matrix and the overall fraction of correct predictions (aka accuracy), sensitivity and specificity for the held out data (that is, the test set of observations).

6. (1.5 points) Use the training data to perform a K-Nearest Neighbors classification with the new `crim` variable obtained in exercise 1. as the response and the remaining 13 variables as predictors. Compute the confusion matrix and the overall fraction of correct predictions (aka accuracy), sensitivity and specificity for the held out data (that is, the test set of observations) for values of K from 1 to 20. For which value(s) of K do you obtain the largest accuracy? And sensitivity? And specificity?

7. (0.5 points) Make plots representing accuracy, sensitivity and specificity, respectively, for the different values of K using `ggplot`.

8. (1 point) Use the training data to perform a Random Forest classification with the new `crim` variable obtained in exercise 1. as the response and the remaining 13 variables as predictors, 5 variables sampled as candidates at each split and 100 trees. To do so, first convert the variable `crim` into a factor using `as.factor()`. Compute the confusion matrix and the overall fraction of correct predictions (aka accuracy), sensitivity and specificity for the held out data (that is, the test set of observations).

9. (2.5 points) Using the relevance of predictors provided by Random Forest in 8., perform a model selection (i.e. identify the best subset of predictors) using a forward approach such that the order in which the variables are added in each step is the one given by the Mean Decrease Accuracy plot (from top to bottom). Do it for logistic regression, linear discriminant analysis, K-nearest neighbors (using the best K obtained in 6) and a classification tree, training the models with the training set of observations, and make the decision based on the accuracy in the test set.

10. (0.5 points) Make plots representing (test) accuracy for the models in 8. for each number of variables using `ggplot`.

11. (0.5 points) Which of the obtained classifiers in 3, 5, 6 and 8 would you say is the best for predicting `crim`? Why? And what subset of variables would you use for the prediction according to the analysis in 9.?