Exercise 8

Advanced Methods for Regression and Classification

January 12, 2023

Take the bank data set (see Exercise 5) for classification. The goal is to predict if the client will subscribe a term deposit or not. This information is represented by the binary variable y (last one). Select randomly a training set of a reasonable size, compute the classifer, and evaluate the classifier based on the test set.

1. *Classification trees*: function `rpart()` from the R package `rpart`
   (a) Compute an initial tree T_0 (see `help(rpart)` or lecture notes).
   (b) Visualize the tree with the function `plot()` and `text()`, and interpret the results.
   (c) Predict the class variable for the test set (see `help(predict.rpart)` or lecture notes). Report the corresponding misclassification rate.
   (d) Show and interpret results of cross-validation obtained by using `printcp()` und `plotcp()`. What is the optimal tree complexity?
   (e) Prune the tree T_0 to the optimal complexity using `prune()`. Visualize und interpret the results.
   (f) Predict the class variable for the test set and report the corresponding misclassification rate. Do we observe any improvement?
   (g) Just for your thoughts: Is there any possibility to reduce the misclassification error for the "yes" clients?
2. *Random forests*: function `randomForest()` from the R package `randomForest`
   (a) Use Random Forests to classify the training data and predict the class variable for the test data. Report the resulting misclassification error?
   (b) Use the option `importance=TRUE` in the function `randomForest()`, and plot the result object with `plot()` and `varImpPlot()`. How can you interpret these plots?
   (c) Try to improve the misclassification error of the "yes" clients (by keeping the overall misclassificatione error still small) with different strategies.
      – Modify the parameter `sampsize` in the `randomForest()` function. What is it doing?
      – Modify the parameter `classwt` in the `randomForest()` function. What is it doing?
      – Modify the parameter `cutoff` in the `randomForest()` function. What is it doing?
      – Modify the parameter `strata` in the `randomForest()` function. What is it doing?

      Which approach leads to the overall best solution?