Biostatistics: Exercise 11

Beate Sick, Lisa Herzog

24.11.2020

Exercise 1: Random Forest for classification

The goal in this exercise is to use a Random Forest for classification. The data set summarizes the chemical concentration of 9 different elements (e.g. Na and Mg) and each observation corresponds to one out of 6 classes corresponding to different types of glass fragments. You can download the training `train.fgl.RData` and the test data set `test.fgl.RData` from the Website.

- Load the training and the test data into R using the function `load()` and become acquainted with the data. Perform a descriptive analysis. Assess how the target variable `type` is distributed in the training and the test data and evaluate the pair-wise relationships between the explanatory variables and the target `type`. Comment on your analysis results.
- Use the training data to fit a classification RF. Set the arguments to `importance=TRUE` for a later assessment of the importance of the different explanatory variables on the target and `ntree=1000`.
  - How large is the out-of-bag error over all classes?
  - Which class(es) are especially hard to classify correctly?
  - Which class(es) are most easy to classify correctly?
- Use the trained RF to predict the classes in the test data. Determine the test confusion matrix, the accuracy and the misclassification rate. Comment on your results.
- Which explanatory variables are most important for the classification?

Exercise 2: Random Forest versus lm for a regression model with continuous outcome

- The data set Boston is available in the package `MASS`. Load it and explore the help page to grab a minimal understanding of the data.
- Randomly split the data into two subsets, a training and a test data set, using the proportion of 70% - 30%.
- Fit a regression model (once with lm and once with randomForest) with `medv` as target variable and all other variables as predictors. Fit the models using the training set.
- Get the predictions for the test data using the fitted models (lm and rf) and plot the observed `medv` values in the test set versus the predicted values. Based on the plot – how do both models compare?
- Calculate the mean squared error (MSE) of these predictions on the test set. Is the MSE better for lm or for the random forest? (Hint: MSE $= \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$

f) Assess the influence of the predictors `rm` and `lstat` in the linear model and the random forest. What do you oberseve. (R-Hint: `varImpPlot()`, `partialPlot()`)