Exercise Lesson05: Decision trees

Machine Learning using R

Exercise 1: Diabetes in Pima Indian women

This data set includes the test-results of women who are of Pima Indian heritage and were tested for diabetes according to Wold Health Organization criteria.

a) The `Pima.tr` data set is available in the `MASS` package. Load the package to be able to access the data. Get an overview of the data and use the command `?Pima.tr` to see the individual variables' meaning.

b) We want to model the variable `type` (does the woman have diabetes: Yes/No) using a decision tree. Fit a decision tree to the data using the `ctree()` function (**Hint**: `library(party)`)

c) Plot the tree structure. Which variables were used to split the data? What is the meaning of the p-values printed below the splitting variables?

d) Calculate the training error and show the corresponding confusion matrix (**Hint**: `predict(model, newdata=..., method='response')`)

e) Yesterday we looked at a cross-validation function for the KNN classifier. Make the appropriate changes in the function so that it can be used for decision trees (generated with `ctree()`). Evaluate the performance of a `ctree`-generated decision tree (using default options) on the `Pima.tr` data set using cross-validation. What can you say about the predictive performance of the decision tree?