Exercises for the Applied Machine Learning Lecture Summer 2016

Problem Set 1

Bernhard Renard - renardB@rki.de

April 19, 2016

This week's exercise focuses on reviewing basic classification analysis on a pathogenicity prediction data set. The underlying research question is whether it is possible to predict whether a bacteria is human pathogenic based on sequence and structure features that can be determined e.g. via a NGS run. This could then be used for diagnoistics and risk assessment.

Step 1
Load the Paprbag R package with pathogenicity data from `https://github.com/crarlus/paprbag`. If you are up for a challenge, load the big data objects from `https://github.com/crarlus/data4paprbag` instead. Prepare the data for analysis as stated on the Paprbag package website in the section training data. The first column gives the label, the other columns are the features.

Step 2
Split the data set into an equal sized training and test data set and train a classifier of your choice (probably the easiest choice is to run a random forest) on the training data and apply it to the test data. Obtain a confusion matrix and comment on sensitivity and specificity.

Step 3
Apply a leave-one-out crossvalidation on the data and obtain a 95% bootstrap confidence interval for sensitivity and specificity of your classifier.

Step 4 (optional)
Analyze the impact of data balancing on the big Paprbag data.

Please email your solution to the problem set to RenardB@rki.de until Wednesday, **April 27th, 2016, 9am**. You cannot get credit for solutions turned in late and there are no exceptions from this rule. Feel free to work in groups of two and hand in a single joint solution with both names clearly stated. Feel free to discuss with other groups, but each group has to hand in their own solution in their own words (and please do not plagiarize or copy from the internet). Please note that we cannot give an active participation passing grade to those cheating.