# Computer lab ML 5-6: Tree methods, Support Vector Machines and clustering.

## Learning objectives

The main objective of this computer lab is to make the student familiar with regression trees, The Random Forests, the Support Vector Machine and clustering techniques.

After completing the lab the student shall be able to:

1. Perform and compare analyses with regression trees and the Random Forests methods.
2. Implement Support Vector Machine with different kernels.
3. Apply model based clustering and evaluate the usefulness.

## Recommended reading

Chapter 8 - 10 in James et al. (2013). An Introduction to Statistical Learning.

## Assignment 1: Comparison of prediction properties of the Random Forests and Partial Least Squares methods

The file **tecator.csv** contains the results from 215 samples of finely chopped meat aimed to investigate whether a near infrared absorbance spectrum can be used to predict the fat content. For each meat sample the data consists of a 100 channel spectrum of absorbance records and the levels of moisture (water), fat and protein. The absorbance is -log10 of the transmittance measured by the spectrometer. The moisture, fat and protein are determined by analytic chemistry.

Read the data into R and create one training data of rows 1:150 and test data of rows 151:215. Make sure that the first column, and the moisture and protein columns are excluded since we are only interested the NIR and fat data.

**1.a.** Use the tree library to construct and plot a regression tree with fat as response variable and all the 100 channels as predictors.

**1.b.** Perform a Random Forests analysis with the randomForest package on the same variables as in 1.a. Sample 1000 trees and make sure to set the variable importance option to true. Provide a variable importance plot and discuss the result compared to 1.a. Use the test data to predict new test values and calculate residual mean squared prediction error.

**1.c.** The next task is to compare the prediction error from 1.b with the prediction error from a Partial Least Squares analysis. Use the plsr() function on the training data with CV to find the optimal number of PLS components and make test predictions based on this model. Calculate residual mean squared prediction error and compare with these from the RF analysis. Discuss your result. Hint: calculation of correlations between the channel variables will give you important information.

***Assignment 2: Create a spam filter using Support Vector Machines with different kernels***

The data file **spambase.csv** contains information about the frequency of various words, characters etc for a total of 4601 e-mails. Furthermore, these e-mails have been classified as spams (spam = 1) or regular e-mails (spam = 0). Your task is to use SVM to develop a model that can be used as a spam filter. Start by randomly dividing your data into 70% training and 30% test as you did in Ass. 3 of Lab 1-2.

**2.a.** Use the svm() function in the e1071() library to tune the cost parameter over the values c(1,5,10,20,50) for the linear kernel. Predict test observations and calculate PER.

**2.b.** Repeat the analysis in **2.a.** with the radial basis function kernel and calculate new PER. Which kernel would you prefer? Discuss your result.

***Assignment 3: Model based clustering***

The seeds data set <http://archive.ics.uci.edu/ml/machine-learning-databases/00236/seeds_dataset.txt> contains measurements of seven geometric parameters of wheat kernels from three different varieties. Your task is to investigate if model based clustering using the mclust() library supports the classification into three distinct varieties. Provide evidence for the number of clusters and which measurements that cluster together, as well as how the clustering corresponds to the real classification. Note that the last variable contains the variety information and should not be included in the clustering, and make sure that you standardize the data.

## To hand in

A written report (preferably a Word or .pdf document) where you summarize your main findings in the assignments. Submit your report to Patrik.Waldmann@slu.se.