Dimensionality reduction and feature selection

Project 4

Exercises

Exercise 4.1. [25 p.]

This exercise will focus on `Arcene` https://archive.ics.uci.edu/ml/datasets/Arcene. The task here falls into binary classification category (with continuous features), i.e., we want to distinguish cancer versus normal patterns from mass-spectrometric data. However, data consists of 10000 attributes with only 900 observations. Moreover, 3000 variables are completely useless - they were added to dataset as distractors. As we can read on data webpage:

> ARCENE was obtained by merging three mass-spectrometry datasets to obtain enough training and test data for a benchmark. The original features indicate the abundance of proteins in human sera having a given mass value. Based on those features one must separate cancer patients from healthy patients. We added a number of distractor feature called 'probes' having no predictive power. The order of the features and patterns were randomized.

This dataset is one of 5 datasets of the NIPS 2003 feature selection challenge.

The goal of this investigation is to obtain the **best predictions and to select the smallest possible subset of relevant input variables (features)**.

You can look up the winning solutions or try something different - anything goes!

Results

As usual, present your results and conclusions in `Jupyter Notebook` or `knitr` short report.