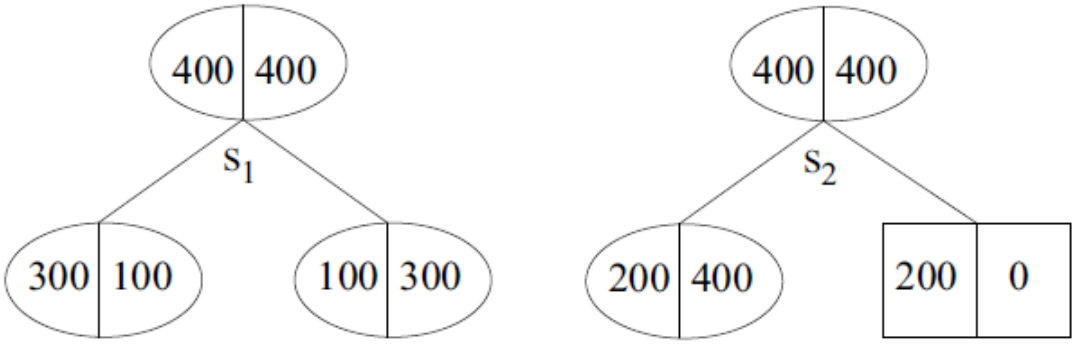
Lecturer: Jan Baumbach Term: Winter 2015/2016

Class: Introduction to Bioinformatics

Exercise sheet – Data Mining

1. Explain/define the "conditional independence" assumption in Naïve Bayes. Describe a case (or cases) when it does not hold?

2. Calculate the re-substitution error and the Gini index for the below example splits (taken from the lecture):



3. Explain / define the term "ensemble methods".

4. Below, you will be given a real gene expression data of 50 clinically relevant genes in breast cancer – the so-called PAM50 data set – for a set of breast cancer patients. The task is to train a Naïve Bayes classifier, which is supposed to assign each of the patients one of the following breast cancer subtype labels: "luminal A", "luminal B", "HER2", "basal" or "normal". You will have 270 such patient samples available for training (including class labels) and 273 patient samples for testing.

(a) Download and unpack the ZIP file from URL1.

(b) Have a look at the supplemental slides (as PDF and PPTX in the ZIP file) and read the paper "Classification of Breast Cancer Subtypes by combining Gene Expression and DNA Methylation Data", which describes the broader scientific background. It can be found here: http://journal.imbio.de/article.php?aid=236

(c) Explain how a "multiclass AUC" is defined. Why can't we use the normal ROC plot / AUC measure here?

(d) In the paper we used a modified version of random forest called "varSelRF". Explain briefly the advantage of recursive feature elimination.

(e) To get started, you first need to install some software:

- Download and install R:
- We recommend you also install Rstudio
- Install the package "e1071"
- Hint: If you have never used R before, you will find useful information and help here. →

  http://www.statmethods.net

(f) From the ZIP file from URL1, load the training (bioinfo1.train.expr.csv) and test data (bioinfo1.test.expr.csv) sets, as well as the class labels (URL3) into R. Find the links to the URLs below. The files come in the so-called CSV format.

(g) Create a Naïve Bayes model based on the training data and the training class labels.

(h) Use your model to predict subtypes for the samples found in the test set.

URL1:
http://www.imada.sdu.dk/~jbaumbac/download/teaching/ws15-16/DM847/exercises/bioinformatics_intro_class_data_mining.zip

Please send your class predictions (exported as CSV) and the R code you have been using via email to your TAs. Also email the names of all group members.