Homework #2: Classification Accuracy

Consider two datasets from Gene Expression Omnibus (GEO, data sets with accession numbers GDS3713, GDS4182). Please download the data from http://file.biolab.si/files/hm2-data.zip.

Both data sets contain a binary (two-valued) class, and the task is to predict a class value based on gene expression profile. In brief, the data comprises tissue samples (in rows) profiled through their gene expressions. Each tissue is from one of two classes (say, disease or healthy). The question is if one can classify the tissue based on its gene expression profile. Questions like these are essential for systems biology and clinical decision making, as profiling tissues through gene expression can improve diagnostic and prognostic tools.

For the homework, construct a workflow where you use cross-validation to estimate the classification accuracy (abbreviated with CA in Orange's Test & Score widget) of the classification trees. Report on this accuracy for each of the two data sets. Tell us for which of the two datasets the machine learning method of classification trees performed better. Please explain how you have reached this conclusion.

Submit the homework as a short report in PDF. The report should include a title of the homework, your name and email, and the following sections:

> **Introduction**: your brief description of the what is homework is about (up to two sentences),
>
> **Material**: a paragraph with up to three sentences on the datasets used, including the report on the size of the data (number of samples and features), and on the type of the features included;
>
> **Methods**: include a screenshot of the workflow you have used and a brief comment (up to two sentences) on what this workflow does;
>
> **Results and Discussion**: one paragraph (up to three sentences) on results. Make sure you end with a sentence with your conclusion that states for which dataset the classification trees were a better predictor.

The report should not exceed two pages. The limit on page length and the limits in the number of paragraphs and sentences are strict. Use 11 pt Calibre or Arial or similar sans-serif font, and 1.2 spacing between lines. Use 6 pt separation between paragraphs.

Submit your report as a PDF document (not Word). Name this document as lastname-firstname-2.pdf (like smith-mary-2.pdf; notice there are no spaces in the name) where lastname is your last name and firstname your first name. Email the report to bzupan@gmail.com with subject DM-HW2 (copy the subject title and then paste it into the email title field; notice there are no spaces in the subject title).

The submission deadline is this Friday, Feb 12, at 9:00 am.

A hint: Gadi claims he's a psychic. Blaz introduced him to a student he met in front of Baylor, and Gadi says he can predict, with over 95% accuracy, whether she studies a medicine or mechanical engineering. Blaz doesn't seem impressed. Why?

PS The two data sets are coming from Gene Expression Omnibus. For the homework, we have reduced the number of features (genes) in the data. You may look up the two data sets on GEO (say, for GDS3713, go to http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS3713) to obtain more information about the datasets. The description of the data sets is however of no relevance to the answer to this homework.