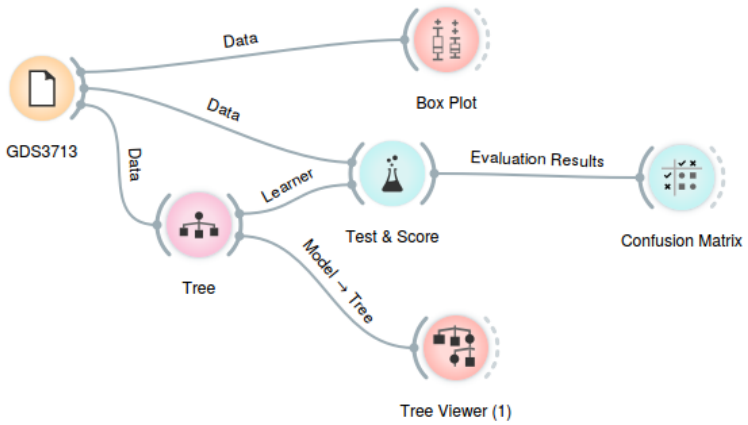
Jana Obšteter
jana.obsteter@kis.si

Homework 3: Classification Accuracy

Workflow shows the procedure used to inspect the data and the prediction of the tree classifier on two data sets including gene expression data and disease status.

	GDS3713	GDS4182
Classification accuracy	0.734	0.802
Area under the curve	0.731	0.600
Precision	0.744	0.364
Recall	0.725	0.250
No cases / controls	40 / 39	16 / 80



A classification tree was used on two datasets GDS4182 and GDS3713, which contain gene expression data and disease state. The classification accuracies (CA) were computed based on 10-fold cross-validation. By comparing the CAs we see that CA of the classification tree on GDS4182 was higher (0.802) that on GDS3713 (0.734). However, in dataset GD3713 there is a large class imbalance since the number of cases is small compared to the number of controls (only 1/6 of the samples are cases). By looking at the confusion matrix we see that this tree performs very poorly when predicting cases, since it correctly predicts only 4 / 11 (36 %) cases and the remaining 7 / 11 (64 %) cases are misclassified as controls. Also, 75 % of actual cases and only 8.8 % of actual controls are misclassified. However, the classification accuracy is still high due to a small number of cases. Even if all the cases were misclassified the CA would decline for only 16 %. The classification tree on GDS4182 has a lower CA (0.734) but performs relatively well for classifying controls as well as cases. The number of misclassified samples is larger that in dataset GDS3713, but the percentage of misclassified is the same among the actual cases and controls (25.6 % and 27.5 %, respectively).

The classification tree on GDS3713 does have a higher CA but it is not useful for classification since it performs well only for classifying controls but does very bad when identifying the cases. This is reflected also in low numbers for other quality parameters, such as precision (a low number of identified positives is true positive), recall (a low number of actual positive is identified) and area under the curve. The latter are substantially higher for the tree on GDS4182. Also in practice – for example for doctors classifying patients - the classifier on GDS4182 would be of more use.