Assignment 10

Machine Learning, Summer term 2013, Ulrike von Luxburg

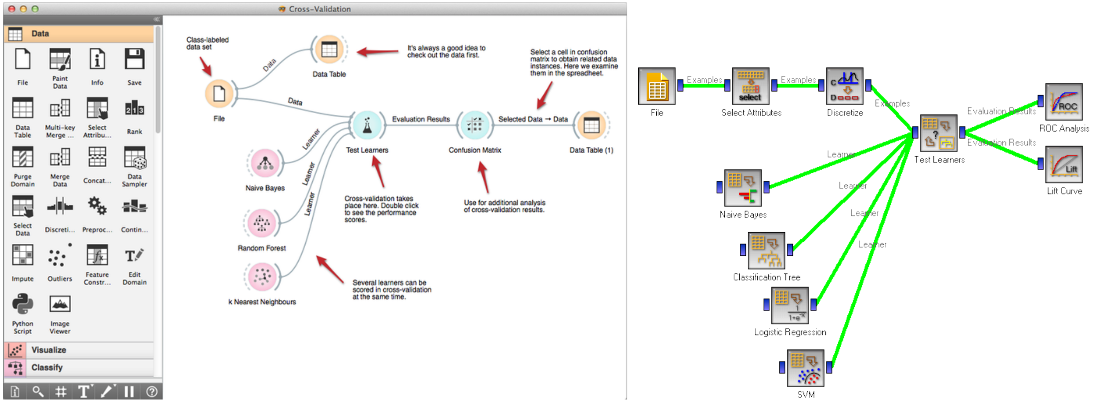
Solutions due by June 24

Exercise 1 (The data processing chain, 16 points) There are several software suites, which provide a visual front-end for doing data analysis, machine learning and visualization. Using these applications, you can run common machine learning tasks by few clicks.
In this exercise, you should choose one of these applications and play with it! You should try different data sets and analyze them: Go through the data processing chain, visualize the data, set training and test data, do classification and parameter selection, do clustering on your data, and compare the results from different classification and clustering algorithms. Each of these tasks can be done by few clicks. For your report, generate a figure or a table from each step.
General guidelines:

- Selecting the data set: For a real world data, use the yeast data set from UCI repository which is also provided in the course web page. The field "label" shows the class labels. Make at least one synthetic data set. In *Orange*, you can generate data sets using *Paint Data*.
- Data pre-processing: Choose 4 classes that have most training points. You can use the visualization tools to see the number of samples in each class.
- Apply different machine learning algorithms. Try at least these units: SVM, k-nearest neighbours classifier, k-means clustering, hierarchical clustering, PCA and a regression method.
- Evaluate your results using different performance scores for classification and the ROC curve.

Here is a list of three free software. Although Rapidminer and KNIME provide more flexibility in using algorithms, we found Orange with a friendlier user interface. You are free to choose your own favorite software.

- Rapidminer: `http://rapid-i.com`
- Orange: `http://orange.biolab.si`
- KNIME: `http://www.knime.org`



Class-labeled data set

It's always a good idea to check out the data first.

Select a cell in confusion matrix to obtain related data instances. Here we examine them in the spreadheet.

Cross-validation takes place here. Double click to see the performance scores.

Use for additional analysis of cross-validation results.

Several learners can be scored in cross-validation at the same time.

Figure 1: Screenshots from Orange