Assignment 7

Machine Learning, Summer term 2014, Ulrike von Luxburg

To be discussed in exercise groups on June 16-18

You should send the corresponding project files (for example, in *Orange*, the scheme file with extension `.ows`) to your teaching assistant by **email until June 15th** with the subject `Assignment 7`.

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 a 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, and compare the results from different classification algorithms. Each of these tasks can be done by a few clicks.
General guidelines:

- Select the data set:
  - Real world data: Use the `yeast` data set from UCI repository which is also provided on the course web page. The field \label" shows the class labels.
  - Real world data: Use the `phoneme` data set which is provided on the course web page. You can read the data description from here:
    `http://orange.biolab.si/datasets/phoneme.htm`
  - Synthetic data set: For example, in *Orange* you can generate data sets using *Paint Data*.
- Data pre-processing: In the yeast data set, choose 4 classes that have most training points. You can use the visualization tools to see the number of samples in each class. In the phoneme data set, ignore the feature *speaker* and consider phonemes (field g) as class labels.
- Algorithms: Apply different machine learning algorithms. Try at least these units: SVM, k-nearest neighbours classifier, logistic regression and PCA.
- Dimensionality reduction: Reduce the dimensionality of the phoneme data using PCA. Choose the number of components such that you keep at least 85% of the variance of the data. Then apply SVM on the transformed data.
- Choosing parameters: You should choose the parameters of the classification algorithms by cross validation.
- 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`

Marking: For each item, you should produce a confusion matrix and an ROC curve (for each class) as a result.

(a) yeast + SVM : 2 points

(b) yeast + k-nearest neighbor : 2 points

(c) phoneme + SVM : 2 points

(d) phoneme + PCA + SVM : 3 points

(e) phoneme + logistic regression : 2 points

(f) phoneme + PCA + logistic regression : 2 points

(g) phoneme + PCA (2 components) + visualization : 1 points

(h) Synthetic data + two classi cation methods: 2 points

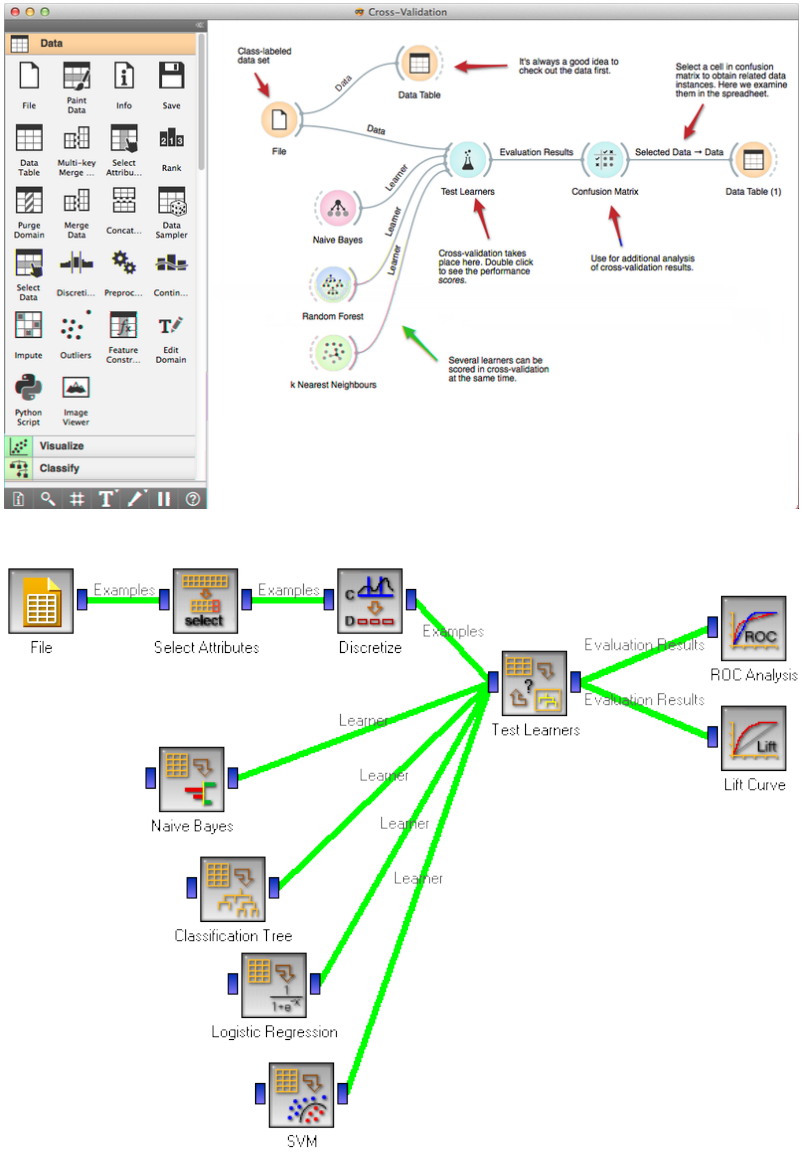

Figure 1: Screenshots from Orange