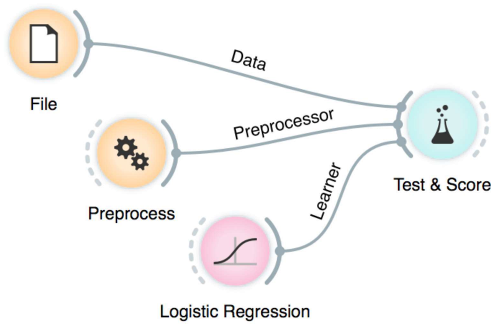
Lesson 14: How to Correctly Perform Test and Score

To put it simply: never, in any way, transform the data prior to cross-validation. Any transformation should happen within cross-validation loop, first on the training set, and then, if required, on a test set. In a relaxed form: it's ok to transform the data, but the transformation should be done independently on the train and the test set and the transformation on the test set should in no way use the information about the class value. Data imputation could be an example of such operation, but again it should be carried out separately for the train and test set and should not consider classes.

The writing on the right looks straightforward. But actually one needs to be extremely careful not to succumb to overfitting when reporting results of cross-validation tests. The literature on systems biology is polluted with reporting on overly optimistic results, and high impact factors provide no guarantee that studies were carried out correctly (in fact, due to a lack of reviewers from the field of machine learning, mistakes likely stay overlooked).

Simon et al. (2003) provides a great read on this topic. He found that many of the early papers in gene expression analysis reported high accuracy simply due to overfitting.

But how do we then correctly apply preprocessing in Orange? The idea of reducing the number of features prior to inferring a predictive model may be still appealing, now that we know we can use it on training data sets (leaving the test set alone). Following are two workflows that do this correctly.

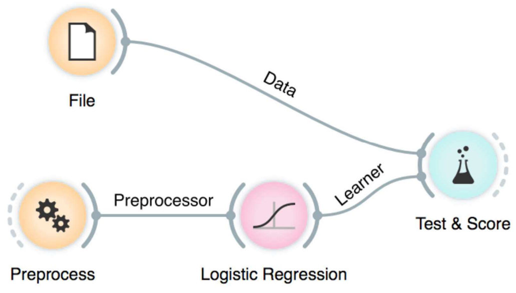


In this first workflow, we gave the Test & Score widget a preprocessor (feature selection was used in this example). The Test & Score widget uses it correctly only on the training sets. This type of workflow is preferred if we would like to test the effect of preprocessing on a number of different learning algorithms.

The Preprocess widget does not necessary require a data set on its input. An alternative use of this widget is to output a method for data preprocessing, which we can then pass to either a learning method or to a widget for cross validation.

This is not the first time we have used a widget that instead of a data passes forward a computation method. All the learners, like Random Forest, do so. A learner could get data on its input and pass a classifier to its output, or simple pass an instance of itself, that is, pass a learning algorithm to whichever widget could use it. For instance, to the Test & Score widget.

Alternatively, we can include a preprocessor in a learning method. The preprocessor is now called on the training data set just before this learner performs inference of the predictive model.



Can you extend this workflow to such an extent that you can test both a learner with preprocessing by feature subset selection and the same learner without this preprocessing? How does the number of selected features affect the cross-validated accuracies? Does the success of this particular combination of machine learning technique depend on the input data set? Does it work better for some machine learning algorithms? Try its performance on k-nearest neighbors learner (warning: use small data sets, this classifier could be very slow).

Somehow, in a shy way, we have also introduced a technique for feature selection, and pointed to its possible utility for classification problems. Feature subset selection, or FSS in short, was and still is, to some extent, an important topic in machine learning. Modern classification algorithms, though, perform it implicitly, and can deal with a large number of features without the help of external procedures for their advanced selection. Random forest is one such technique.