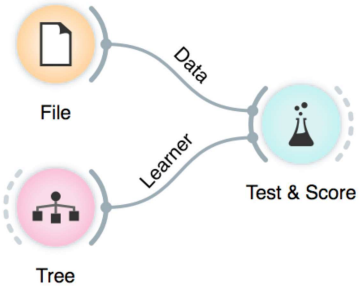
Lesson 10: Cross-Validation

Estimating the accuracy may depend on a particular split of the data set. To increase robustness, we can repeat the measurement several times, each time choosing a different subset of the data for training. One such method is cross-validation. It is available in Orange through the Test & Score widget.

Note that in each iteration, Test & Score will pick part of the data for training, learn the predictive model on this data using some machine learning method, and then test the accuracy of the resulting model on the remaining, test data set. For this, the widget will need on its input a data set from which it will sample data for training and testing, and a learning method which it will use on the training data set to construct a predictive model. In Orange, the learning method is simply called a learner. Hence, Test & Score needs a learner on its input. A typical workflow with this widget is as follows.



For geeks: a learner is an object that, given the data, outputs a classifier. Just what Test & Score needs.

Cross validation splits the data sets into, say, 10 different non-overlapping subsets we call folds. In each iteration, one fold will be used for testing, while the data from all other folds will be used for training. In this way, each data instance will be used for testing exactly once.

This is another way to use the Tree widget. In the workflows from the previous lessons we have used another of its outputs, called Model: its construction required the data. This time, no data is needed for Tree, because all that we need from it a learner.

Here we show Test & Score widget looks like. CA stands for classification accuracy, and this is what we really care for for now. We will talk about other measures, like AUC, later.