MSc Data Science and Advanced Computing
CSMDM21 - Data Analytics and Mining

Practical 4: Classification (Decision Tree)

The classification task of this exercise is to predict the class (Setosa, Versicolour, Virginica) of an iris flower from its four attributes (petal length, petal width, sepal length and sepal width). You can use the following KNIME nodes to generate and to test a predictive model with the iris data set: *Decision Tree Learner, Decision Tree Predictor and Scorer*.

Task1: Resubstitution method (1)

1. Train a decision tree with the iris data set using all records. For this task the configuration of the Decision Tree Learner is shown in figure 1. The minimum number of records per node is 9.
2. Use the generated model to classify the same input data set (all records).
3. Compute the confusion matrix, the accuracy and the classification error (resubstitution error) with the node Scorer.

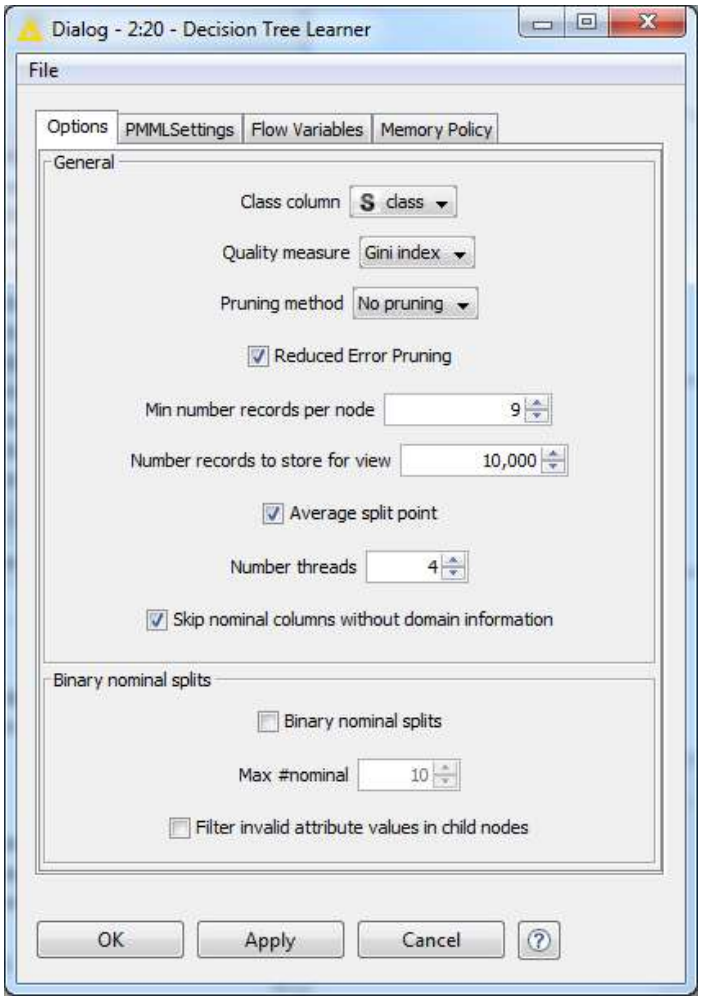

Figure 1: Task 1 configuration dialog of the node Decision Tree Learner

Task2: Resubstitution method (2)

1. Repeat the task 1 with a minimum number of records per node of 1.
2. Report and critically compare the results (accuracy) of task 1 and task 2.

Results and comments:

Task1: ... Task2: ...

Task3: Accuracy estimation with the holdout method

1. Load the wine dataset and add column headers:
   `http://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data`
2. Convert the class attribute from numeric to String. This is necessary because the node *Decision Tree Learner* we are going to use requires string labels as class values.
3. Use the node *Partitioning* to split (e.g. 90%-10%) the rows into training and test sets.
4. Use the node *Decision Tree Learner* to generate a predictive model from the training set.
5. Use the node *Decision Tree Predictor* to apply the predictive model to the test set.
6. Use the node *Scorer* to compute the confusion matrix, the accuracy and the error rate.

Accuracy = Error rate =

Task 4: Accuracy estimation with the cross-validation method

1. Load the wine dataset with column headers.
2. Convert the class attribute from numeric to String. This is necessary because the node *Decision Tree Learner* we are going to use requires string labels as class values.
3. Use the node *X-Partitioner* to split (e.g. 10 folds with random sampling) the rows into training and test sets. This is the starting node of a loop.
4. Use the node *Decision Tree Learner* to generate a predictive model from the training set.
5. Use the node *Decision Tree Predictor* to apply the predictive model to the test set.
6. Use the node X-Aggregator to aggregate the result of each fold. This is the ending node of the loop, which will be repeated 10 time, one for each fold.
7. Use the node *Scorer* to compute the confusion matrix, the accuracy and the error rate.

Accuracy = Error rate =