Naïve Bayes Assignment

Create a KNIME workflow that utilizes the Naive Bayes method to train a model on the Adult training data set. This data set can be found at the UCI machine learning repository.

http://archive.ics.uci.edu/ml/datasets/Adult

This data set was developed by Barry Becker by extracting from the 1994 Census database. Prediction task is to determine whether a person makes over 50K a year.

Use the file reader node to read in the training data file from:

http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

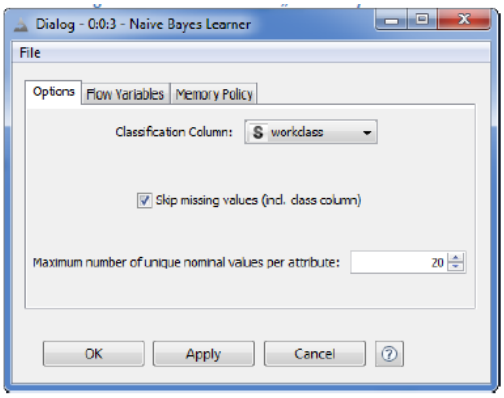
The Naive Bayes method is available in the Node repository panel under "Mining"-> "Bayes" category. There are two different nodes: "Naïve Bayes Learner" and "Naïve Bayes Predictor". Naïve Bayes learner node creates a Bayesian model from the input training data. The Naïve Bayes Predictor node applies an existing Bayesian modes to the input data table. Additionally, the scorer node should be added at the end of the workflow in order to measure classifiers' performance.

Naïve Bayes Learner

The "Naïve Bayes Learner" node creates a Bayesian model from the input training data. It calculates the distributions and probabilities to define the Bayesian model's rules from the training data.

In the configuration window you need to specify:

- The class column (= the column containing the row classes)
- How to deal with missing values (skip vs. keep)
- The maximum number of unique nominal values allowed per column. If a column contains more than this maximum number of unique nominal values, it will be excluded from the training process.

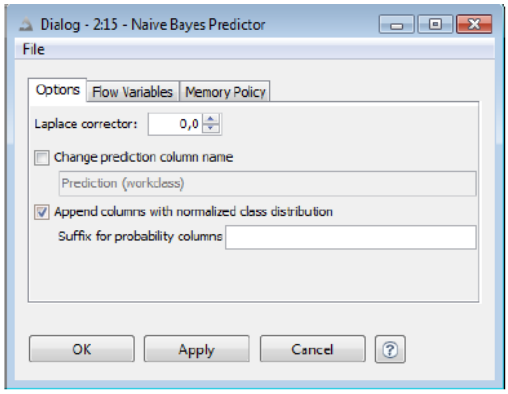


Naïve Bayes Predictor

The "Naïve Bayes Predictor" node applies an existing Bayesian model to the input data table.

In the configuration window you can:

- Append the normalized class distribution values for all classes to the input data table
- Use a Laplace corrector as the initial count for nominal columns with 0 count. A 0 value indicates no correction.
- Customize the column name for the predicted class

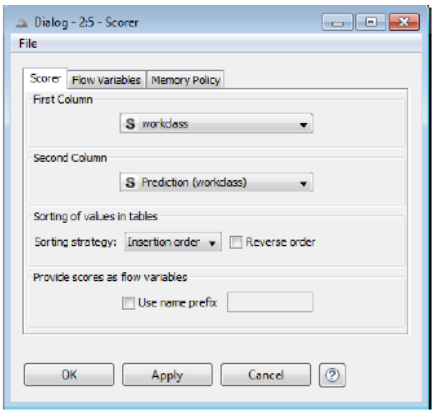


Scorer

The "Scorer" node is located in the "Node Repository" panel in the "Mining" -> "Scoring" category. It compares the values of two columns (reference column and prediction column) in the data table and shows the confusion matrix and some accuracy measures.

The "Scorer" node has a View option, where the confusion matrix is displayed and the "Hilite" functionality is available. Thus, the user can hilite the cells of the confusion matrix to isolate the underlying rows and to see them, for example, by means of an "Interactive Table" node.

The configuration window requires the selection of the two columns to compare. It also provides a flag to enable the storing of the score values as flow variables (flow variables though are not explained in this beginner's book).



Task #1.

By utilizing the simple Naïve Bayes learner, predictor and scorer nodes – how well was the Model able to learn to predict income (<=50K or >50K). Hint: take a look at the confusion matrix within the scorer node. Since in this exercise we are using the entire training data set to evaluate the model – we are looking at what is called the re-substitution error.

Task #2.

In order to properly evaluate the trained model – lets' split the data into training and testing by utilizing the "Partitioning" node. Split into 66%-33% subset for training and testing. What is the evaluation of the model performance estimated now? Looking at the accuracy statistics – which class has a better Precision and F-Measure?

Task #3

Upload the Iris data set. Train the Naïve Bayes model on the partitioned data set. Which one of the 3 Iris class has the highest number of true positives?