UNIVERSITY OF MANNHEIM

Data and Web Science Group
Prof. Dr. Christian Bizer
B6 – B1.15
68159 Mannheim

Data Mining – FSS 2020

Exercise 3: Classification

3.1. Should we play golf?

The *Golf data* set is one of the examples that are delivered together with RapidMiner. The data set models different aspects of the weather (outlook, temperature, humidity, forecast) that are relevant for deciding whether one should play golf or not.

1. Learn a decision tree from the Golf data set (Operator: *Decision Tree*). Use this tree to classify the examples in the Golf-Testset, which is also delivered together with RapidMiner (Operator: *Apply Model*). Think about ways how you can evaluate the performance of your model. What measures can be calculated from the resulting dataset?
2. Evaluate the performance of your model by adding a *Performance (Classification)* operator to your process. Examine the confusion matrix. What is the accuracy of your classifier?
3. Does a k-nearest-neighbor classifier work better for this task? Replace the *Decision Tree* operator with a *K-NN* operator and check how the accuracy of your classifier changes to find out. Do different values of k improve the performance?

3.2. Learning a classifier for the Iris Data Set

You want to learn and evaluate a classifier for recognizing different types of Iris flowers.

1. Let's try the ID3 tree building algorithm first. Build a process that (1) discretizes all attributes of the Iris data set by frequency into three bins. (2) Afterwards, the process should use the *Split Validation* operator (split ratio=0.7, stratified sampling) to generate a training and test data set. (3) As inner operator of the split validation, the process should use the *ID3* operator to learn a decision tree and the *Performance (Classification)* operator to evaluate the accuracy of the learned model.
2. Remove the discretization operator and change the ID3 operator into RapidMiner's standard *DecisionTree* building operator. Run the process again. Does the accuracy change? Compare the complexity of the two models. Which model should be preferred according to Occam's razor?
3. Try a k-nearest-neighbor classifier on the problem. Does it perform better?

3.3 More Classification

In the lecture, you learned about the Nearest Centroid Classifier. For this classifier, RapidMiner does not provide you with an operator. So you have to install the “Mannheim RapidMiner Toolbox” in order to use this classifier.

1. Install the”Mannheim RapidMinerToolbox” from the Marketplace (“Help” -> “Marketplace”)
2. Compare kNN and Nearest Centroid Classification using the “Weighting” dataset from the RapidMiner Samples Repository.