UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data and Web Science Group
Prof. Dr. Heiko Paulheim
B6 – B0.22
68131 Mannheim

Data Mining – Exercise 3

3.1. Should we play golf?

The *Golf 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 naïve bayes model from the Golf data set (*GaussianNB* classifier in scikit learn). Use this model to classify the examples in the Golf test set. 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 calling *confusion_matrix and accuracy_score*. Examine the confusion matrix. What is the accuracy of your classifier?
3. Does a k-NN classifier work better for this task? 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 Naïve Bayes algorithm first. Create a train/test split (with function *train_test_split*) with 30% test size and stratified sampling. Evaluate the accuracy of the learned model.
2. Try a k-NN classifier on the problem. Does it perform better?

3.3 More Classification

In the lecture, you learned about the Nearest Centroid Classifier (NearestCentroid classifier in scikit learn).

1. Compare kNN and Nearest Centroid Classification using the "Weighting" dataset.