UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Data and Web Science Group
B6 – B1.15
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 decision tree model from the Golf data set (*DecisionTreeClassifier* 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 Decision Tree 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

Learn and evaluate different classifiers on the "Weighting" dataset.

- Compare kNN and Decision Tree Classification using the "Weighting" dataset.

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

ChatGPT Bonus Exercises

Reminder: Do not take the answers of ChatGPT at face value! Always cross-check with lecture slides, literature and/or the teaching staff!

C.1. Discuss Overfitting with ChatGPT

Discuss Overfitting with ChatGPT.

- What is the problem with a machine learning model overfitting to the training data? What are the symptoms of overfitting for decision trees and KNN models? What can you do to prevent overfitting (in the case of decision trees, random forests and KNN models)?

C.2. Coding

Experiment with different methods to prevent overfitting of the Decision Tree on the Iris dataset.

- Ask ChatGPT how your code from 3.2 needs to be changed to implement the methods proposed in C.1. (Hint: You can also start by asking ChatGPT to propose a solution for 3.2)

Test the generated code snippets and ask for clarification if it does not work.

C.3. Self-Assessment

Test your knowledge about tree-based classification with ChatGPT

- Ask ChatGPT to create a pen-and-paper exercise for you that lets you practice the algorithm to create a Decision Tree, which applies methods to prevent overfitting, using a small example dataset. After completing the exercise ask ChatGPT for the solution and critically evaluate it. You can also ask ChatGPT for code to run the exercise using scikit-learn.
- Ask ChatGPT to generate three multiple-choice questions on overfitting tree-based classification models. (Hint: If the questions are too simple, add that you are a grad student.)