# Data Mining: Practical Assignment #4

Due on Thu & Fri, May 18-19 2017, 10:15am-13:15 & 14:15am-17:15

## Task 1

The test of a 3-class classification system on 100 test data points yields the following confusion matrix:

| Predicted \Actual | Class 1 | Class 2 | Class 3 |
|---|---|---|---|
| Class 1 | 20 | 4 | 6 |
| Class 2 | 5 | 16 | 9 |
| Class 3 | 1 | 4 | 35 |

Let us assume that we are interested in results for 'class 1' and want to apply a number of metrics from the lecture. Write down the confusion matrix that results from combining 'class 2' and 'class 3' to 'other' (this should result in four entries). Then compute the following evaluation metric values and give an intuitive description of their meaning:

1. accuracy
2. error
3. true positive rate
4. sensitivity
5. recall
6. false positive rate
7. specificity
8. precision
9. F-score (sometimes simply called F1-score, if $\alpha = 1$ is implied)

## Task 2

In the lecture you learnt about the apriori algorithm. In brief, the algorithm gets data and a minimum support value serving as a threshold. Now, it generates iteratively chunks of data items occurring together.

i.) Open the DAMI *task2/apriori.m* file provided in the data4.zip and go through the code to understand the implementation. Take the data *task2/toy.mat* and run the algorithm. What does the output depict and why? You may need to comment the code (replace the ??? comments with meaningful information) to provide a better understanding of the steps.
*(For further information, refer to ReadMe.txt.)*

ii.) Now let's transfer your insights of the algorithm from the simple toy example to a realistic dataset from grocery shopping. Get the data from the data4.zip as well and investigate: which prominent items and rules you can find in the data?

## Task 3

Decision trees are a prominent tool for categorization large datasets.

| Enough Wind? | Prepared DaMi? | Go Surfing |
|---|---|---|
| Y | Y | Y |
| N | Y | N |
| Y | Y | Y |
| Y | Y | Y |
| Y | N | N |
| N | Y | Y |
| Y | N | Y |
| Y | N | N |
| N | Y | Y |
| N | Y | Y |

i.) Today, we want to decide to whether or not going surfing. The table shows a transcript of two attributes (columns) and the according answers (rows). Your task is to compute the decision based on what you learnt in the lecture concerning the ID3 algorithm.
*Hint: you need to compute the following values:*

1. the entropy $\mathit{Info}(D)$ of the entire data (irrespective the attribute values)
2. the entropy $\mathit{Info}(D_{a_i})$ for each value $i$ of each attribute $a$ (irrespective the other attribute)
3. the average entropy $\mathit{Info}(D_a)$ for each attribute
4. the information gain $\mathit{Info}(D) - \mathit{Info}(D_a)$ for each attribute

Note: the logarithm to base 2 function in Matlab is called `log2`.

ii.) Now let's compute a decision tree on the Auto Miles-per-gallon Data Set with Matlab. A thorough description can be found in the ReadMe.txt file. Before you can call the according tree, you have to prepare the data. From task 2 you should be familiar with the cell array concept.

1. Import the data from the data4.zip into Matlab.
2. Split the data into the classes to decide upon and the features.
3. Inform yourself about the Matlab function for decision trees and perform the computation with the preprocessed data.
4. Visualize the tree.

Clarify how and why the algorithm split the tree.

## @home Task

To prepare the next tutorial, your homework will be to learn about the following topics:

- Unsupervised learning and cluster algorithms, especially learning with Self-Organizing-Maps (SOM, Kohonen Maps), and the K-Means algorithm.
- Genetic algorithms and their most important operators.

Recommended literature:

1. Eiben, A. E., Smith, J. E. Introduction to evolutionary computing. Springer, 2003.
2. Han J. & Kamber, M. Data mining: Concepts and techniques. Elsevier/Morgan Kaufmann, Amsterdam, 2006.
3. Kantardzic, M.: Data mining : concepts, models, methods, and algorithms. Wiley, NY, 2011. (!)
4. Kohonen, T.: Self-organizing maps. Springer, Berlin, 1995.
5. Mitchell Tom M.: Machine learning. McGrawHill, NY, 2010.