Heidelberg University
Database Systems Research Group
Dr. Erich Schubert
Sebastian Lackner, Christian Straßberger

23.10.2017
Knowledge Discovery in Databases
WT 2017/2018

Assignment 1: Introduction to Data Mining

Due: Thursday, November 2nd, 9am

Organizational: For the assignments please form **teams of up to three students**. Only *one* person per group should upload the solution, but **the names of all team members must be given both on the PDF and in the source code**. Teams cannot be changed during the semester. You *must not* share the answers to a larger group.

Please follow the upload instructions at the end of the sheet closely.

Problem 1-1 Data mining tasks — 5 points

Which data mining tasks (association rule mining, clustering, outlier detection, classification, etc.) are hiding in the following use cases? Are the tasks supervised or unsupervised?

(a) **Optical character recognition/OCR**:
When crossing the alps using the Brenner Autobahn, there is the option to pay electronically in advance. When approaching the toll station, the barrier automatically opens when the number plate was recognized. The reconition happens fully automatically by a digital camera system.

(b) **Computer Aided Diagnosis:**
Patients that suffer from blood cancer can be characterized in two categories (ALL and AML). The therapies for these two types partially differ, and the therapy for AML can sometimes be detrimental to patients suffering from ALL and the other way around. To avoid these complications, special gene expression data is used to differentiate between these two types by comparing them to the data from patients where the cancer type is already known.

(c) **Cheat Detection:**
The operator of a multi player online game wants to protect his system against various violations of the terms of service. Particular problems are the use of game bot programs, the manipulation of timestamps in the communication protocol and attempts to predict random numbers used. To prevent this misuse, data mining is used on the available user data.

(d) **Online Shopping:**
An online shopping portal wants to recommend products to registered customers upon login. The available data includes products previously bought by the customer to predict his interests. For example a user that bought the book "Lord of the rings" might be offered the DVDs of the movie triology. A related task might be suggesting additional products for already chosen products as a bundled offer.

(e) **News Aggregation:**
A news aggregator web site collects current news from various sites to keep the visitor informed. However, news reports about the same subject are common and should be grouped by subject. This happens at multiple levels: there are broad categories like politics and sports, and subcategories such as soccer. But even on a single soccer game, there will be different news sites reporting. Some articles will be identical to the report of a major agency, some will only be slightly modified, others will be original works.

Problem 1-2 Probability Theory **3 points**

(a) Given two non-empty disjoint events $A \cap B = \emptyset$, with $P(A) \neq 0$, $P(B) \neq 0$.
Prove or refute: A and B are independent.

(b) Derive Bayes' Rule using only Conditional Probability (and the Law of Total Probability).

Problem 1-3 Sample Space and Events **8 points**

Consider the experiment of tossing two (distinguishable) fair coins and rolling a 12-sided fair dice at once.

(a) Define the sample space of possible outcomes and specify the size of the sample space.

(b) Define the event A_1 that both coins turn up heads and that the dice shows a number greater than 5. Compute the probability $P(A_1)$ and show how you arrived at your solution.

(c) Define the event A_2 that at least one coin turns up head and that the dice shows a number greater than 7. Compute the probability $P(A_2)$ and show how you arrived at your solution.

(d) Define the event A_3: 'number of heads/number on the dice ≥ 0.5'. (/ is the usual division)
Compute the probability $P(A_3)$ and show how you arrived at your solution.

Problem 1-4 Bayesian Reasoning **7 Points**

Lets denote "The answer to the ultimate question of life, the universe, and everything" as A_{42}.
Suppose the probability to find A_{42}, given that you studied Computer Science, is 0.01, and the probability to find A_{42}, given that you studied something else instead, is $2 \cdot 10^{-5}$. Let the probability that you decide to study Computer Science be 0.25.

(a) What is the probability that you find A_{42}?

(b) What is the probability that you studied Computer Science, (1) given that you found A_{42}, (2) given that you have not found A_{42}? Compare this to the prior probability.

Problem 1-5 Descriptive Statistics in Python *or* R **7 Points**

Load the well-known "iris flower petal" data set (which is included in both sklearn and R). For the tasks below, only consider the numeric attributes "petal length", "petal width", "sepal length", and "sepal width."

To load the iris data in Python, use:

```
import sklearn.datasets
iris = sklearn.datasets.load_iris()
iris = iris.data # Use numeric attributes only (for this exercise)
```

To instead load the iris data in R, use:

```
> data(iris)
> D <- iris[,1:4] # Numeric attributes only
```

This data set has $p = 4$ numeric attributes and $n = 150$ rows.

For this exercise, choose *either Python or R*. You *may* use libraries such as `numpy` and `matplotlib`.

(a) Print the empirical mean and the empirical standard deviation of the attributes.

(b) Create a boxplot containing the above attributes (4 boxes).

(c) Calculate the pairwise (Pearson) correlation coefficients between the attributes and show the results in a $p \times p$ matrix.

(d) Create a scatter plot for the two attributes with highest positive correlation coefficients (or between any two attributes in case you have problems with (c)). A scatter plot simply plots the sample values of two given attributes on the x- and y-axis respectively (e.g., see slides 2:18, 2:55).

The written solutions to (1-1 to 1-4) need to be **uploaded as a single PDF in Moodle**.

The source code is uploaded *separately* as a `.zip` file only (do not include your main PDF file, do not include multiple solutions – there should be a single file, ending in `.py` or `.r`).

Make sure to put the **names of all group members** into the source code *and* on the PDF.

Make sure that the source can be executed from scratch, and that all necessary libraries / packages are loaded at the start of the program (i.e., make sure that your program re-runs reproducable, not just interactively).

Only `.pdf` and `.zip` files can be uploaded (Moodle does not know `.py` or `.r` files).

Do *not* include the PDF with the *other* solutions in the zip file, but upload A) the `.pdf`, B) the source `.zip` because they will be graded separately.