UNIVERSITEIT TWENTE.

# Data Science [201400174]

Course year 2022/2023, Quarter 1B

DATE
November 11, 2022

EXCERPT

## Data Mining [DM]


TEACHERS
Faizan Ahmed
Ellen-Wien Augustijn
Nacir Bouali
Faiza Bukhsh
Rolf de By
Karin Groothuis-Oudshoorn
Maurice van Keulen
Mahdi Khodadadzadeh
Elena Mocanu
Estefania Talavera
Brenda Voorthuis
Shenghui Wang

COURSE COORDINATOR
Karin Groothuis-Oudshoorn (quartile 1A)
Maurice van Keulen (quartile 1B)
Faizan Ahmed (quartile 2A)

PROJECT OWNERS
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Mannes Poel
Michel van Putten
Mohsen Jafari Songhori
Luc Wismans

Topic 2

# Data Mining [DM]

## 2.1 Introduction

Topic teachers: Elena Mocanu and Karin Groothuis-Oudshoorn

Data mining is about discovering patterns or more general information in large data sets using methods from artificial intelligence, machine learning, statistics, and database systems.

The topic gives an introduction in the theory and practice behind three basic machine learning techniques used in data mining; regression, classification and clustering mining. After an introduction to the the theory the techniques are put to practice in the practical assignments using small datasets for educational purposes. Furthermore, you are free to choose between two data mining tools, Python or R. In the project the knowledge and skills should be applied to a ”real life” case.

### 2.1.1 Global description of the practicum and project

In the practical sessions one has to do several pen and paper exercises and some practical exercises. For the ”real life” case one has to select one of the projects described in the project part of this handout. Some projects may be added in due time.

### 2.1.2 Study material and tools

The study material consists of parts of the book Introduction to Statistical Learning with applications in R written by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (second edition). https://www.statlearning.com/ Free download of the PDF. The study material for this topic are:

1. Chapter 1 Introduction
2. Chapter 2 Statistical Learning
3. Chapter 3 Linear Regression
4. Chapter 4 Classification
5. Section 5.1 Cross Validation
6. Section 8.1 The Basics of Decision Trees
7. Section 12.4 Clustering Methods (Sec 10.3 in first edition)

There also some video lectures based on the book. (Link can be found in the digital version.)

In this topic on Data Mining we will use Python or R as our main data mining tools. Please feel free to choose your most preferred tool among the two options.

**Python** ▷ There are many possibilities to install python. We recommend the use of anaconda 3 with jupyter notebook. See also the IENLP topic description paragraph 3.2.2.

**R** ▷ It can be downloaded from the R project website. See also the DPV topic description paragraph 1.4.

### 2.1.3 Deliverables and obligatory items

The practical assignments contain pen and paper assignments and assignments using Python or R. The submission should be one single pdf file containing:

- For the pen and paper assignments the detailed answers (could be a scan of handwritten notes and pictures).
- For the practical assignments the plots, output, interpretation and the commands used (Python or R).

## 2.2 Description of the practical assignments

### 2.2.1 Pen and paper assignments

The pen and paper assignments consists of the exercises.

☞ **2.1** Chapter 2, exercise 7 part a, b and c. □

☞ **2.2** Chapter 3, exercise 3. □

☞ **2.3** Chapter 4, exercise 6. □

☞ **2.4** A retailer wants for marketing purposes distinguish between costumers younger then 35 and customers older then 35. The following table summarizes the data set in the data base of the retailer in an abstract form. The relevant attributes, determined by domain knowledge, are for convenience denoted by $A$ and $B$. The values for $A$ are $a1$, $a2$ and $a3$. The values for $B$ are $b1$ and $b2$. The retailer wants to use Data Mining

| $A$ | $B$ | Number of Instances | |
|---|---|---|---|
| | | $Y$ | $O$ |
| $a1$ | $b1$ | 4 | 10 |
| $a2$ | $b1$ | 6 | 2 |
| $a3$ | $b1$ | 8 | 6 |
| $a1$ | $b2$ | 2 | 8 |
| $a2$ | $b2$ | 6 | 2 |

techniques to classify the costumers in the class "young", denoted by $Y$, and "old", denoted by $O$.

(a) Assume a new costumer enters the web-store and the retailer has no information at all about this customer. How will this new customer be classified based on the above data and explain why.
(b) Now assume a new customer comes in for which the retailer knows the values for attribute $A = a3$ and $B = b2$. Is it possible to apply the standard (non Naive) Bayes to classify this new customer? Explain what the problems is.
(c) Hence the retailer decides to use a naive Bayes classifier for the classification of this new customer.
    - How will this customer now be classified based on the values of $A$ and $B$? Explain your answer.

□

☞ **2.5** Consider the same dataset as in the previous question. Now the retailer (data analyst) wants to use Decision Trees to classify new customers.

(a) What is the classification error rate for attribute $A$?
(b) What is the classification error rate for attribute $B$?
(c) What will be the splitting attribute in the top (root) of the Decision Tree if one uses the classification error rate?
(d) What is the Gini index for attribute $A$?
(e) What is the Gini index for attribute $B$?
(f) What will be the splitting attribute in the top (root) of the Decision Tree if one uses the Gini index?
(g) Construct the full Decision Tree, using the error rate as heuristic, and what is the overall classification error rate on the above dataset?
(h) Is this classification error rate an optimistic or pessimistic estimate of the error rate on unseen new data? Explain your answer.

□

## 2.2.2 Practical assignments

You have a choice for making the practical assignments with Python or with R. For Python, make exercises 2.6 and 2.7 below. For R, make exercises 2.8 and 2.9 below. **NB: You need not make all practical assignments; only Python or R!**

#### Practical assignments using Python

Prerequisites

- install anaconda 3
- install pandas, matplotlib, sklearn, numpy, seaborn, statsmodels libraries
- install from anaconda terminal: conda install -c conda-forge pyreadstat (to read with pandas library the first dataset)

☞ **2.6** Import the data from `voorbeeld7_1.sav` and save the table under the name `chol1`.

(a) Make a scatter plot, with the function `lmplot` (using the seaborn library), from the column cholesterol `chol` (y-axis) and age `leeftijd` (x-axis). Add then a regression line to the graph with `lmplot(..,fit_reg=True)`.
(b) Fit a linear model for chol with leeftijd using the function `ols` (using the statsmodels library). The formula for the model is `chol~leeftijd`. Save the `fitobject` under the name `fit1`. View the result with `fit1.summary()`.
(c) Fit a model `fit2` for `chol` with `leeftijd`, `bmi`, `sekse` and `alcohol`. Which factors are statistically significant?
(d) Add the residuals from the model `fit2` to the table `chol1` and make a histogram from the residuals.

□

☞ **2.7** Import the dataset `births.csv` and call this table `births`. This data set contains data from 49703 childbirths from the year 1995.

(a) Recode the variable `child_birth` into a new variable `home` where `home='at_home'` if the childbirth was a so-called `first line child birth, at home`, if not then `home='not_at_home'`. Use for this the `for` loop and the `if else` statement. So finally the variable `home` should be a factor variable with levels `at_home` and `not_at_home`. Hint: Use a Python list.
(b) Recode the variable `parity` in a new variable `pari` where `pari` has level `primi` if it concerns a first childbirth and `multi` if it is the second or more childbirth. You can do this again with `for` and `if else`.
(c) Recode the variable `etnicity` into a new variable `etni` where `etni` has level `Dutch` if the woman was Dutch and `Not Dutch` if she was not Dutch. Hint: use `unique()` of the pandas library to see which levels are in the variable `etnicity`.

(d) Using the `sklearn` library make a logistic regression model with the function `LogisticRegression` for the probability of childbirth at home with the variables `pari`, `age_cat` (= age categorised), `etni` and `urban` (urbanisation degree). View the outcomes from the model with the `classification_report()` function.
(e) Using the same `sklearn` library make a decision tree for the probability of childbirth at home with the same variables as in the logistic regression model. View the decision tree with the function `tree.plot_tree` from package `tree`.
(f) For assessing which model, the logistic regression model or the decision tree, fits better the data we should fit the models on a training set and calculate accuracy statistics on a test set (or use cross validation). For this we use `sklearn.model_selection` and `cross_val_score`. Which model fits the data better?

□

**Practical assignments using R**

☞ **2.8** Import the data from `voorbeeld7_1.sav` and save the table under the name `chol1`.

(a) Make a scatterplot, with the function `ggplot`, from the column cholesterol `chol` (y-axis) and age `leeftijd` (x-axis). Add then a regression line to the graph with `geom_smooth(method =` lm").
(b) Fit a linear model for chol with leeftijd using the function `lm`. The formula for the model is `chol~leeftijd`. Save the `fitobject` under the name `fit1`. View the result with `summary(fit1)`.
(c) Fit a model for `chol` with `leeftijd`, `bmi`, `sekse` and `alcohol`. Which factors are statistically significant?
(d) Add the residuals from the model `fit2` to the table `chol1` and make a histogram from the residuals.

□

☞ **2.9** Import the dataset `births.csv` and call this table `births`. This data set contains data from 49703 childbirths from the year 1995.

(a) Recode the variable `child_birth` into a new variabele `home` where `home='at_home'` if the childbirth was a so-called `first line child birth, at home`, if not then `home='not_at_home'`. Use for this the function `mutate`, `factor` and `if_else`. So finally the variable `home` should be a factor variable with levels `at_home` and `not_at_home`.
(b) Recode the variable `parity` in a new variable `pari` where `pari` has level `primi` if it concerns a first childbirth and `multi` if it is the second or more childbirth. You can do this again with `mutate` and `if_else`.
(c) Recode the variable `etnicity` into a new variable `etni` where `etni` has level `Dutch` if the woman was Dutch and `Not Dutch` if she was not Dutch. Hint: use `table(etnicity)` to look which levels there are in the variable `etnicity`.
(d) Make a logistic regression model with the function `glm` for the probability of childbirth at home with the variables `pari`, `age_cat` (= age categorised), `etni` and `urban` (urbanisation degree). View the outcomes from the model with the `summary` function.
(e) Using the function `rpart` (with option `method =` class") from the package `rpart` a decision tree model can be estimated on the data. Make a decision tree for the probability of childbirth at home with the same variables as in the logistic regression model. View the decision tree with the function `rpart.plot` from package `rpart.plot`.
(f) For assessing which model, the logistic regression model or the decision tree fits better the data we should fit the models on a training set and calculate accuracy statistics on a test set (or use cross validation). For this we use the package `caret`. The following code will calculate several accuracy measures based on a split-file strategy. Which model fits the data better?

**Note: The following is a sketch of the process, and is not meant to run without additional lines and/or modifications. It should serve as a guideline to show you some important steps you need to do. Or simply: if copied and pasted to R it will result in errors.**

```
library(caret)
set.seed(100)
```

```
mydat <- births
# Step 1: Get row numbers for the training data
trainRowNumbers <- createDataPartition(mydat$home, p=0.8, list=FALSE)
# Step 2: Create the training dataset
trainData <- mydat[trainRowNumbers,c(2,4,8,9,10)]
# Step 3: Create the test dataset
testData <- mydat[-trainRowNumbers,c(2,4,8,9,10)]
#https://www.machinelearningplus.com/machine-learning/caret-package/
fit_logreg <- train(home ~ pari + age_cat + etni + urban, data = trainData,
                            method="glm", family="binomial")
predicted <- predict(fit_logreg, testData)
confusionMatrix(reference = testData$home, data = predicted,
                            mode='everything', positive='at_home')
fit_rpart <- train(home ~ pari + age_cat + etni + urban, data = trainData,
                            method="rpart")
predicted <- predict(fit_rpart, testData)
confusionMatrix(reference = testData$home, data = predicted,
                            mode='everything', positive='at_home')
```

□