Written Exam
Data Warehousing and Data Mining
course code: 192320201

1 February 2012 (13:45 - 17:15; CR-3E)

Maurice van Keulen & Mannes Poel & Chintan Amrit

Remarks:

- The exercises are clearly marked to what topic they belong to allow you to start with the topic you feel most confident about.
- Motivate yours answers. The motivation / argumentation plays an important role in grading the exercise.
- You are allowed to use the study material and notes for the written exam. Both the data mining as well as the data warehousing assignments have to be completed satisfactorily before one is admitted to the written exam. The grade for the written exam is immediately the grade for the course. In case of doubt, the result of the practicum may be taken into account.
- There are ??? exercises. For each assignment, the number of points is given. In total, there are 60 points.

Assignment 1 : Classification (10 pts.)

A financial company want to automatize the applications for a loan. In order to build a classification model the company uses the part of the database concerning loan applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data. The relevant attributes, determined by domain experts, are for convenience denoted by A, B and C. The values for A are a and b, the values for B are u and y and the values for C are g and p. PBT stands for "Payed Back in Time" and a $+$ means that the customer did pay his or her loan in time, a $-$ means that he or she did not pay back the loan in time.

A	B	C	PBT
b	u	g	+
a	u	g	+
a	u	g	+
b	u	g	+
b	u	g	+
b	u	g	+
b	u	g	+
b	u	g	-
a	u	g	-
b	u	g	-
b	y	p	-
b	u	g	-
b	u	g	-
a	y	p	-
b	y	p	-
b	u	g	-

The financial company wants to use Data Mining techniques to classify the loan applications.

Part 1a

As a first approach the financial company investigates Decision Trees. They use the Gini index as splitting criteria. For each attribute show the contingency table and the corresponding Gini index (pg. 158 of handout Chapter 4).

Part 1b

Based on the Gini index, which attribute will be the top most attribute in the Decision Tree?

Part 1c

Compute the first two levels of the decision tree for which the Gini index is used as heuristic for selecting the best attribute.

Part 1d

The financial company tested three different approaches. For each approach you can find the confusion matrix below.

Classified as	Real class	
	+	-
+	1206	6
-	6	38

Table 1: Approach 1: Naive Bayes classifier

Classified as	Real class	
	+	-
+	1212	12
-	0	32

Table 2: Approach 3: Bayes classifier

Classified as	Real class	
	+	-
+	1200	0
-	12	44

Table 3: Approach 3: Decision tree classifier

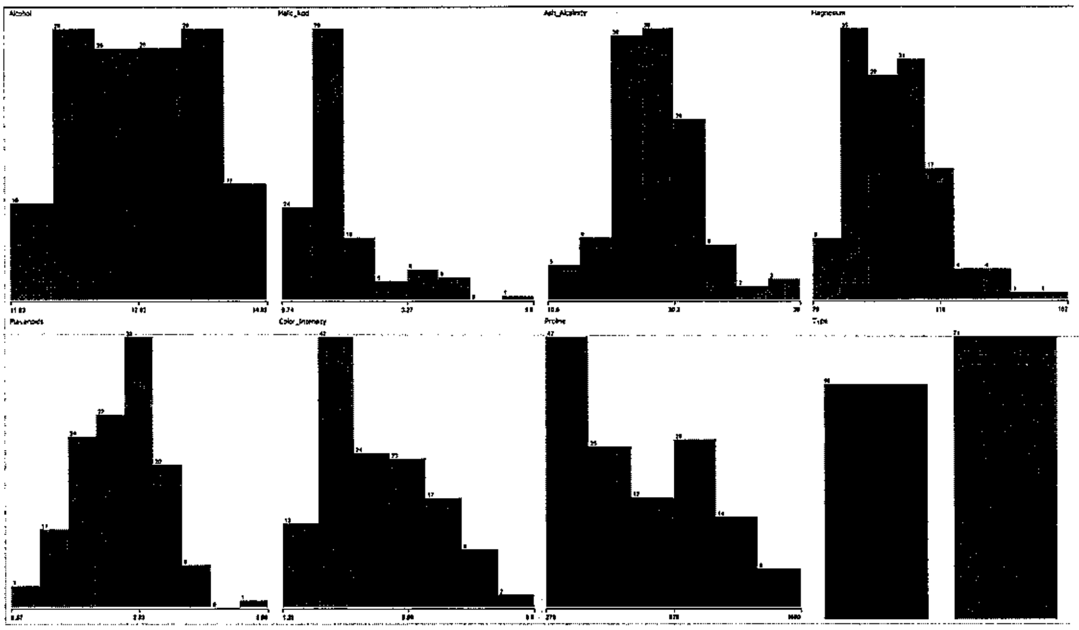
Which Data Mining approach (1, 2, or 3) would you use for classifying new customers? What are the influencing factors for the decision? **Explain your answer(s)!**

Assignment 2 : Visualization (8 pts.)

Part 2a (2 pts)

Briefly describe what you understand by Ontological Analysis and Semiotic Clarity. What are the four mapping problems that can occur?

Part 2b (2 pts)



The figures show the different attributes for two types of wine - shown in light and dark grey. By analyzing the visualized attributes, which one do you think is the most predictive for the type of wine? Why?

Part 2c (4 pts)

Explain when and for what kind of data would you use a Matrix, a Graph or a specific hierarchical Graph to visualize the data.
In each of the scenarios below describe which of the three vizualisations you would use and explain why:

1. In the psychiatric ward of a certain hospital, each patient sees only one doctor, who is responsible for diagnosis and treatment. A researcher is interested in determining whether patients would receive different diagnoses from different doctors. Therefore, she selected a group of

newly-admitted patients and asked all of the staff psychiatrists to submit a diagnosis for each patient in the group. The department chair would like a diagram showing each doctor's diagnosis for each patient.

2. Sisters of Mercy Hospital serves a close-knit religious community that observes strict laws forbidding the exchange of blood with people who are not members of their sect. These complicated laws also govern the exchange of blood among community members (e.g., unmarried women over the age of 20 cannot donate blood to married women). Although the constraints on who can donate blood to whom are well-specified by law, they do not follow any coherent pattern. The hospital would like a diagram showing who in the community may donate blood to whom.
3. In her next lecture, a pharmacy professor plans to discuss the composition of a new over-the-counter flu medicine. The medicine contains several active ingredients and several inactive ingredients. Each of these ingredients can be broken down into its basic chemical structure. Thus the composition of the flu medicine can be described at a variety of different levels. The pharmacy professor would like a diagram showing the composition of this medicine.

Assignment 3 : Data warehousing case (20 pts) Case "Consultation Bureau — *continued*"

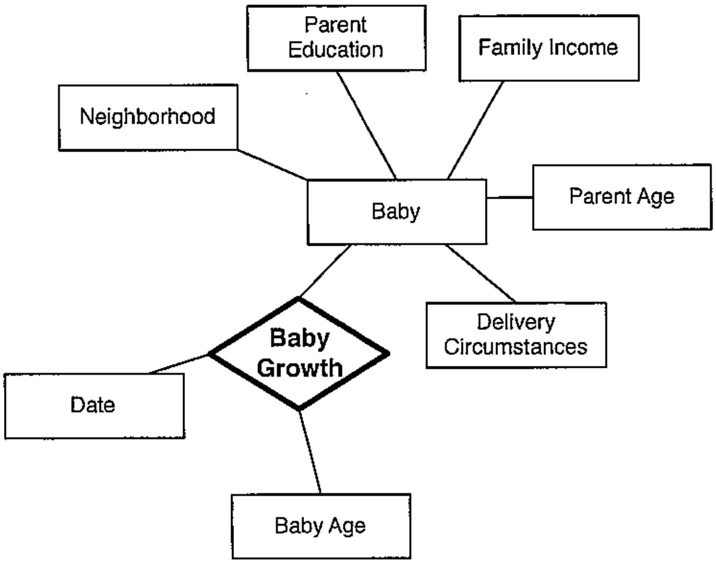


Figure 1: Star schema of Consultation Bureau case

The exam questions below are based on the same case as the Data Warehousing assignment. If you didn't bring the case description and the commentary with you, then ask the teacher for a copy, because the questions below refer to these documents. The first few questions are about the case directly. Then, an extension/continuation of the case is given to which the subsequent questions apply.

Part 3a (4 pts)

Figure 1 contains one of the possible the star schemas suggested in the commentary. A next step in the design is to design a table structure. Give a table structure for this star schema in terms of a list of tables with for each table a list of attributes with their types.

Part 3b (2 pts)

In the commentary it is suggested to *exclude* all data for babies who died before or just after child birth. At the end, it is also suggested to *exclude* data about babies who have only one parent, or who have parents of the same sex, or whose parents are seperated.

1. Explain why it is *correct and safe* to simply exclude data in this way.
2. Explain why it may be *good* to exclude data for such circumstances.

Part 3c (4 pts)

The case description roughly describes the possible delivery circumstances as "delivery at home / hospital / initial delivery at home but rushed to hospital during delivery, various complications such as Cesarean section, etc., possibly death of baby and/or mother". As indicated in the commentary, some dimensions, including this one, may need some thinking and refinement.

1. Give a list of concrete values for the dimension 'Delivery Circumstances'
2. Explain whether or not the dimension is *complete.*
3. Explain whether or not the dimension is *disjoint.*

(Case continuation)
Assume that the Consultation Bureau data warehouse has been realized and used for a number of years. They now like to extend it to be able to answer questions about *the well-being of teen mothers.* Let us define a teen mother as a mother who was less than 18 of age at child birth. Obviously data about the babies of teen mothers is already contained in the data warehouse and is continued to be supplemented over time.
The additional source data is composed of two things: in addition to the regular questions during the intake, a teen mother is asked whether she is working, studying or unemployed (we call this a teen mother's *status*). Moreover, a teen mother is asked to fill in a questionnaire at each visit to the CB. The questions are:

(i) How *tired* were you generally since the last visit?
(5-point scale from "very tired" to "very energetic")

(ii) Did you have moments of feeling *depressed*?
(5-point scale from "many" to "none")

(iii) What is your current working/studying/unemployed status?

The goal for the extension is to determine the influence of income, education, neighborhood, and status on the well-being of the teen mothers (well-being measured in terms of tiredness and depression on the given scales).

Part 3d (3 pts)

The Kimball model advises to use a 'Data Mart Bus Architecture Linked by Conformed Dimensions'. Explain what "conformed dimensions" are and how they can be used to link data marts.

Part 3e (7 pts)

Design a star schema (no attributes) for the data warehouse extension preferrably according to Kimball, i.e., linked to the existing data warehouse by conformed dimensions.

1. Give this star schema
2. Explain any complications with the status and how you advise to solve them.
3. Estimate the number of rows the fact table and each of the dimension tables of the data warehouse extension. Provide the full calculation including any assumptions you make.
4. Give an example of a business question that you can answer with the extended data warehouse that you cannot answer with either the original or the extension seperately.

Assignment 4 : Association rules (10 pts.)

Consider the following market transaction database, in a binary 0/1 representation.

TID	Bread	Milk	Diapers	Beer	Eggs	Cola
1	1	1	1	1	1	1
2	0	1	1	1	0	1
3	1	0	1	1	1	1
4	0	1	0	1	0	0
5	1	1	1	0	1	0
6	1	1	1	0	0	0
7	0	0	1	1	0	1
8	1	0	1	1	1	1
9	1	0	0	1	0	1
10	1	1	1	1	0	0
11	1	1	1	1	1	1
12	1	1	1	1	0	1
13	1	1	1	1	0	1
14	1	0	1	1	0	1
15	1	1	0	1	0	0

Part 4a

Compute the support, confidence and lift of the following association rules:

1. {Diapers} $\Longrightarrow$ {Milk}
2. {Milk} $\Longrightarrow$ {Diapers}
3. {Beer} $\Longrightarrow$ {Diapers, Milk}

Part 4b

Compute the following frequent itemsets with support ≥ 0.6 obtained by the *candidate generation procedure* using the $F_{k-1} \times F_1$ merging strategy:

1. Frequent 1-itemsets
2. Frequent 2-itemsets; list first the candidate 2-itemsets.
3. Frequent 3-itemsets; list first the candidate 3-itemsets.

Part 4c

Compute all association rules consisting of exactly three items with support ≥ 0.6 and confidence ≥ 0.9.

Assignment 5 :
Advanced Topics Visualization (12 pts.)

Part 5a (6 pts)

A local oil company offers medical benefits to its entire staff. Their database contains logs of all the medical claims ever made by their staff. Each medical claim record in the database has the following structure:

StaffID	ClaimDate	ClaimAmount
41006703	21-04-2007	750.00

Table 4: Available data sources

In the above database, StaffID is the foreign key referencing the Staff table which has the following structure:

StaffID	fName	lName	Position	Salary
41006703	Peter	Estibeiro	Manager	85000

Table 5: Available data sources

You have been asked by the company to create a visualization of their medical claims data.

1. With the help of simple sketches, describe your design for visualizing medical claims data.
2. If you are allowed to use the data from the Staff table, describe your new visualization scheme.

Part 5b (6 pts)

Consider the data set of Table 6 showing values of two variables time and depth measured by a scuba diver in the "pre dive computer era" using a wrist watch and a depth meter during his ascent

Time	Depth
6.00	38.91
10.00	102.34
3.00	10.21
7.00	43.89
8.00	70.12
4.00	13.56
5.00	30.80
2.00	4.45
9.00	27.65

Table 6: Available data sources

Using a rough sketch, design a visualization to communicate (visually) the contents of this data set clearly stating your design steps and any assumptions you make.

1. What are the "messages" your visualization communicates to the end user (diver)?

2. Discuss the merits and demerits of different techniques used for evaluating information visualizations?