Written Exam
Data Warehousing and Data Mining
course code: 192320201

29 January 2014 (08:45 - 12:15; CR-2K)

Maurice van Keulen & Mannes Poel
& Chintan Amrit & Sjoerd van der Spoel

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.
- One may not consult books, papers or notes, but only one page of A4 size, double-sided printed. The page may contain text (typed or hand-written) and (possibly reduced) images (copied from books, other sources or hand-made). A calculator is allowed (and recommended), but mobile phones are not allowed.
- 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 5 exercises. For each assignment, the number of points is given. In total, there are 70 points.

Assignment 1 (15 pts): Classification

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 X, Y and Z. The values for X are a and b, the values for Y are u and y and the values for Z 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.
The financial company wants to use Data Mining techniques to classify the loan applications.

X	Y	Z	PBT
b	u	g	+
a	u	g	+
a	u	g	+
b	u	g	+
b	u	g	+
b	y	g	+
b	u	g	+
b	u	g	+
a	u	g	-
b	u	g	-
b	y	p	-
b	u	g	-
b	u	p	-
a	y	p	-
b	y	p	-
b	u	g	-

Part a

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.

Part b

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

Part c

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

Part d

Compute the confusion matrix and performance of the decision tree, constructed in part c, on the training set. Is the performance an optimistic or pessimistic estimate of the performance in "real life"? **Explain your answer!**

Assignment 2 (20 pts): Data warehousing case
Case "Holiday cottages"

Imagine a company X that owns tens of thousands of holiday cottages, partly inside holiday parks, partly single. They rent them out to customers through their own website, but also through a number of resellers. Such a reseller has its own website portal offering not only all the cottages of X, but also of others. The management of X wants to make some strategic decisions based on the data they have:

- Should they focus on their own website or on more resellers?
- If the latter, which type of reseller should they expand on?
- For expansion, should they focus on buying more single cottages or in parks, in which country/countries, which size category[1], which luxury category[2], and in city or landscape or beach setting?

The strategy on resellers and expansion are of course based on the profit they (can) earn. Inside the company they have, of course, a database with all data on the cottages including the cost of maintaining them per month. Each cottage is identified with a standardized 'cottage ID'. The sales database contains all bookings including the revenue and uses the same cottage ID. Revenue minus maintenance cost is profit.

They want to set up a data warehouse and a data visualization environment in which they can explore their profit data on the mentioned aspects as trends per month possibly detailed per customer type.[3]

Part a (5 pts)

The fact here is obviously 'profit'. But there is a problem to which month a certain profit 'belongs'. To illustrate, imagine that a certain holiday cottage is rented out to a certain customer for two weeks: from Saturday January 25th until Saturday February 8th. The holiday cottage remained empty for all other days in January and February. Since customers pay when they leave, the sales database has an entry dated February 8th with a revenue of, say, 1400 euro (100 euro per night). Suppose maintenance cost is 200 euro per month. If you do nothing special, then the data visualization will show a profit of -200 for January and 1200 for February. One could also argue

[1] Size categories are in terms of max number of persons, for example, '2', '4', '6', '8+'.

[2] Luxury categories are, for example, 'Basic', 'Comfortable', 'DeLuxe', 'DeLuxe-with sauna', etc.

[3] Customers are categorized into types such as 'business', 'family with kids', 'young couple', 'elderly couple', etc.

that half of the revenue has been gained in January, i.e., the profit is 500 for January and 500 for February.

(i) Which of the two would you choose? Explain your answer.

(ii) Where in the data warehouse architecture would such a redistribution of the money take place: at the data source, in the ETL, in the data warehouse, or in the visualization software? Explain your answer.

Part b (10 pts)

Observe that the case description doesn't mention nor requires to have profits per individual holiday cottage in the data warehouse.

(i) Give a star/snowflake schema for the data warehouse that *does* contain profits per cottage.

(ii) Give a star/snowflake schema for the data warehouse that *does not* contain profits per cottage.

Explain your design by describing the most important design choices and considerations.

(iii) Which one do you prefer? Explain your answer.

Part c (5 pts)

One thing that may misguide the decision makers, is that holiday cottages are not always available. When a cottage, for example, undergoes maintenance, it may be unavailable for days, weeks or even several months. In these periods it doesn't produce any profit, but that doesn't mean that its country, size, luxury type or setting is less attractive for expansion.

What would be your advise on how to solve or deal with this issue such that it doesn't misguide the decision makers anymore. Explain your anwer.

Assignment 3 (10 pts): Visualization

Part a (4 pts)

Suppose you are hired as a consultant to Google for estimating the size of unimportant or waste websites that they have indexed. You do have an idea about some characteristics of a webpage that can influence its usefulness. Characteristics such as:

1. speed of loading
2. time on site
3. bounce rates
4. frequency of access and,
5. number of incoming links.

How will you go about estimating the amount of these unwanted webpages? Describe in some detail whether you would apply a explanatory or a predictive approach. Also, under what specific circumstances would you choose one over the other?

Part b (3 pts)

Suppose you would like to display all the indexed websites in the previous question, briefly describe the kind of visualisation you would use. You could give a small example for 4 websites with differing characteristics. Finally, justify why you would use this visualisation over others that can also deal with multiple dimensions.

Part c (3 pts)

Describe the global structure of a Matrix, Network and a Hierarchy. What kind of data they be used to best represent? In each of the scenarios below explain which of the three visualisations (a Matrix, a Graph or a specific hierarchical Graph) you would use and explain why:

1. An automobile engineer wants to keep track of the engine components and their dependencies
2. Researchers are trying to create more effective drugs to treat a certain form of cancer. To do this, they start with the most promising drug to date, Xantalin, and modify its chemical structure slightly in a variety of ways to create several new drugs. These drugs are then further modified in a variety of ways to create yet more potential cancer drugs, etc. It turns out that if you trace each of the new drugs back to Xantalin, you will find that there is only a single sequence of modifications through

which each can be synthesized. The head of the oncology research division would like a diagram showing how all of the new drugs can be synthesized from Xantalin.

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 ingredi- ents 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 de- scribed at a variety of different levels. The pharmacy professor would like a diagram showing the composition of this medicine.

Assignment 4 (10 pts): Critical reflections

Part a (5 pts)

(i) Describe the difference between explanatory and predictive analysis.
(ii) In which of these categories does data mining fall?

Part b (5 pts)

In table 1 below are the results of two classifiers. The classifiers were used to determine if a document was *waste*(w) or *non-waste*(n). For each document, we have given the actual class of the document and the prediction of both classifiers.

	Prediction	
Actual	**Classifier 1**	**Classifier 2**
w	w	w
w	w	w
n	w	w
w	n	w
n	w	w
w	w	w
w	w	w
w	w	w
w	n	w
n	w	w
n	w	w
n	n	w

Table 1: Classifier results for two classifiers, plus the actual class for each document. Classes are **w**aste and **n**on-waste

(i) Give accuracy for both classifiers, give precision and recall for both classes for both classifiers
(ii) Comment on the performance of the classifiers: which classifier would you consider to be better and why?

Assignment 5 (15 pts): Association rules

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	0	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	1	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 a

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 b

Compute the following frequent itemsets with support ≥ 0.7 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 c

Compute all association rules consisting of at most three items with support ≥ 0.7 and confidence ≥ 0.9.

UNIVERSITY OF TWENTE.

Exam

Module: Business Intelligence and Information Technology
Course: Databases and Business Intelligence

Lecturers: dr. C. Amrit
University of Twente

Name:________________________________

Student nr:______________________________

Open Book Exam
This is an open book exam: All course materials (slides handouts, books, and papers) may be used for reference.

Grading

This exam contains 25 multiple-choice questions (from the lectures, book chapters and from the papers that have been included in the course) and 1 optional question (for 4 points). All the questions have only **1 right answer.** Each of the following questions is for 2 points.

1. In an ER diagram the following are usually multivalued attributes:

A. Date of birth
B. Address
C. Hobbies
D. Age
E. None of the above

2. In the relationship “student takes a course” the primary key is

A. Student id
B. Course id
C. Both Student id and Course id
D. Any of the above depending on the cardinality of the relationship
E. None of the above