Written Exam
Data Warehousing and Data Mining

course code: 192320201

30 January 2013 (08:45 - 12:15; Hal-B 2C)

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 retailer wants for marketing purposes to distinguish between costumers younger then 35 Y and customers older then 35, O. The following table summarizes the data set in the database 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* and *a2*, the values for B are *b1* and *b2* and the values for C are *c1* and *c2*.

A	B	C	Number of Instances	
			Y	O
a1	b1	c1	16	2
a1	b1	c2	0	4
a1	b2	c1	6	4
a1	b2	c2	0	12
a2	b1	c1	8	4
a2	b1	c2	0	6
a2	b2	c1	6	0
a2	b2	c2	4	8
			40	40

The retailer wants to use Data Mining techniques to classify the costumers in the class “young”, denoted by Y, and “old”, denoted by O.

Part 1a

The approach the retailer investigates is Decision Trees. He uses the Gini index as splitting criteria. Compute for the attribute C the contingency table and the corresponding Gini index.

Part 1b

Assume that the attribute C is the top level decision node of the Decision Tree. What will be the next attribute to split on for the case $C = c1$, if one uses the Gini index as criterium for attribute selection.

Part 1c

The retailer computes the performance of the Decision Tree on the training set. Is this performance an optimistic or pessimistic estimate of the performance of the Decision Tree in real life setting. **Explain your answer!**
What would be a better approach to assess the performance of a Decision Tree on this classification problem?

Assignment 2 (15 pts): Data warehousing case Case "Mobile app beta tester service"

The company *Pear* sells smartphones. For downloading and updating mobile applications they have the "Pear app store". Pear management has a new business idea: a beta tester service for game developers. The purpose of beta testing is mainly for getting feedback on game play: do you become bored too quickly, is it challenging enough, etc. The purpose is not so much finding bugs. Good gamers can offer themselves as beta tester and developers can hire them. Developers pay Pear for a subscription plus a fee for hiring a beta tester of which a small percentage goes to Pear.

A subscription includes support for developers in finding the "best" beta testers for their particular application. Pear needs to realize a data warehouse for this purpose with all gaming data of the beta testers. They have in their databases timestamps of start of game app, switch to other app, and relevant events (turn, next level, completion of game, etc.) as well as scores of completed games, and data on the apps: name, solitary/multiplayer, name of developer, a fine-grained category, etc.

(you may assume more data to be available if it is reasonable a company as Pear would have it; make these assumptions explicit)

A beta tester receives a small amount of money for joining in exchange for allowing Pear to disclosing their data to developers. A second purpose of the data warehouse is that during a beta test, Pear provides the developer with information on how much time the beta tester has devoted to beta testing the game.

Part 2a (4 pts)

For being able to determine the right fact(s) for the star schema of this data warehouse, an important question is "What is *good*", i.e., "What makes a particular gamer a good beta tester for a particular kind of game?" More concretely, how can you determine one or more scores that quantify this 'goodness' or 'suitability', which obviously needs to be calculated from the available data.

Propose one or more scores with for each a definition, a sketch of how they could be derived, and arguments why they are an indicator for a good beta tester.

NB: for the remainder of this assignment, use 'goodness score' as fact.

Part 2b (2 pts)

What is the business question, or what are the business questions in this case? Formulate them as accurately as possible.

Part 2c (6 pts)

Give a star schema for the data warehouse. Explain your design by describing the most important design choices and considerations.

Part 2d (3 pts)

Bertheussen IT, the developer of the popular word puzzle game 'WordFeud', subscribed to this service. Describe a report that Bertheussen IT could use to determine the best candidates for beta testing WordFeud, which can be derived with a sequence of OLAP operations SLICE, DICE, ROLL-UP, DRILL-DOWN on the in Part 2c designed cube. Give this sequence. And explain how they can use the report for this purpose.

Assignment 3 (10 pts): Visualization

Part 3a (2 pts)

Suppose you would like to display all the files in a hard disk, briefly describe the kind of visualization you would use. You could give a small example for 4 files of differing sizes and file types.

Part 3b (2 pts)

Describe the Anscombe's quartet, which statistical technique is useful in discovering the differences among the data? Based on this, can you think of a relevant best practice in data analysis?

Part 3c (3 pts)

In each of the scenarios below explain which of the 3 visualizations (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. The same automobile engineer wants to keep track of who is working on which component
3. An automobile engineer wants to keep track of the components of the entire car and all the dependencies
4. 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.

Part 3d (3 pts)

Let's say a professor has records of every class he has ever taught, including names of classes, year and semester of classes, names of all students, their final grades, and their class standing (1st/2nd/.../final year). List two specific tasks (possibly related) that might require an information visualisation (that is not trivial with excel). Briefly describe a visualization and how to use it to accomplish those tasks.

Assignment 4 (15 pts): Critical reflections

Part 4a (3 pts)

Describe the characteristics of a dataset, for which you would rather use datamining and not a standard statistical technique to analyse.

Case "Simpson Pharmaceuticals"

Simpson Pharmaceuticals has been developing a new treatment for a terminal disease. They have investigated the effect of the treatment on people suffering from the disease. The group, existing of an equal amount of men and women, was split into two groups: a group that received the treatment and a group that did not receive the treatment (the so-called control group). For each patient, the researchers recorded whether the patient *recovered* or *died*.

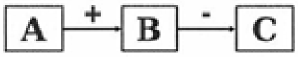

Figure 1: An example causal graph

Figure 1 shows an example causal graph. A causal graph shows the influence variables have on each other. Here, A positively influences B. B negatively influences C.

Part 4b (2 pts)

	Recovery	Death	Total	Recovery rate
Treatment	40	40	80	50 %
No treatment	48	32	80	60 %
Total	**88**	**72**	**160**	**55 %**

Table 1: Results of the Simpson Pharmaceuticals experiment

The results of the experiment at Simpson Pharmaceuticals are in table 1. Does the treatment work? Explain your answer by making a causal graph.

Part 4c (2 pts)

Upon more careful investigation (by drilling down on the previous table), the researchers find the results in table 2.

	Recovery	Death	Total	Recovery rate
		Male		
Treatment	24	36	60	**40 %**
No treatment	6	14	20	**30 %**
Total	**30**	**50**	**80**	**37.5 %**
		Female		
Treatment	16	4	20	**80 %**
No treatment	42	18	60	**70 %**
Total	**58**	**22**	**80**	**72.5 %**

Table 2: Simpson Pharmaceuticals experiment: results of drilling-down

Does the treatment work? Explain your answer by making a causal graph.

Part 4d (5 pts)

Table 1 shows recovery rate without treatment to be higher than the recovery rate *with* treatment. Table 2 shows the opposite. How do you explain this difference?

Part 4e (3 pts)

The researchers at Simpson Pharmaceuticals try to train a classifier on the data from their experiment. Its goal is to predict whether a patient will or will not recover from the treatment: a two class problem with classes "Recovery" and "Death". The classifier training results in the confusion matrix in table 3. Calculate the *accuracy* as well as *precision* and *recall* for **both classes**.

		Actual class	
		Recovery	Death
Predicted class	Recovery	20	6
	Death	8	6

Table 3: Confusion matrix

Assignment 5 (15 pts): Association rules

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

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

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

1. {Diapers} $\Longrightarrow$ {Milk}
2. {Milk} $\Longrightarrow$ {Diapers}

Part 5b

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 5c

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