Written Exam

Data Warehousing and Data Mining

course code: 232020

27 January 2010 (13:45 - 17:15; Sporting Center)

Remarks:

- The exercises are clearly marked as 'DM' for data mining and 'DW' for data warehousing to allow you to start with the topic you feel most confident about.
- To allow parallel correction of the exams, please use three separate sheets: Sheet A for assignments 1 and 2, sheet B for assignments 3 and 4, and sheet C for assignment 5. Do not forget to put your name and student number on every sheet.
- 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. The practicum has 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 40 points.

Assignment 1 (DM; 8 pts; Sheet A): Classification

A retailer wants for marketing purposes distinguish between costumers younger then 35 Y and customers older then 35, O. 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, B and C. The values for A are *a1*, *a2* and *a3*. The values for B are *b1* and *b2*. The values for$C$ are *c1* and *c2*.

A	B	C	Number of Instances	
			Y	O
a1	b1	c1	14	0
a2	b1	c1	0	4
a3	b1	c1	6	2
a1	b2	c1	0	12
a2	b2	c1	6	4
a3	b2	c1	0	6
a1	b1	c2	0	8
a2	b1	c2	8	0
a3	b1	c2	2	0
a1	b2	c2	0	4
a2	b2	c2	2	2
a3	b2	c2	4	0

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 first approach the retailer investigates is Decision Trees. He uses the Gini index as splitting criteria. For each attribute show the contingency table and the corresponding Gini index (pg. 158 of handout of Chapter 4).

Part 1b

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

Part 1c

Also compute, based on the Gini index, the second level of the Decision Tree.

Part 1d

Assume a new customer enters the shop and for this new customer the retailer can determine the attribute values $A = a1$ and $B = b1$. How will this new customer be classified based on the two level decision tree of part c? Explain your answer!

Part 1e

Another approach is to make use of a Naive Bayes Classifier. First explain how a Naive Bayes classifier works and then determine how this new customer of Part d will be classified.

Part 1f

Explain the main differences between a Decision Tree classifier and a (Naive) Bayes classifier.

Assignment 2 (DM; 6 pts; Sheet A): Association Rules

A supermarket stores all the transactions in a large database. These transactions database can be used for "basket analysis". For the sake of simplicity and time we focus only on a small part of the the database and of all the items:

transaction	items
t1	{bread, cheese, milk}
t2	{bread, cheese, jelly, peanut butter}
t3	{cheese, jelly, milk}
t4	{bread, cheese, jelly, milk}
t5	{milk, peanut butter}
t6	{bread, cheese, milk, peanut butter}
t7	{bread, jelly, milk}
t8	{jelly, milk, peanut butter}
t9	{bread, cheese, milk, peanut butter}
t10	{jelly, peanut butter}
t11	{bread, cheese, milk}
t12	{bread, jelly, peanut butter}
t13	{bread, cheese, milk}
t14	{bread, cheese, jelly, milk}
t15	{bread, jelly, milk}

Part of the transaction data base.

Part 2a

Compute, based on the above table, the *support* and *confidence* of the following association rules:

1. {cheese} $\Longrightarrow$ {bread}
2. {bread} $\Longrightarrow$ {cheese}
3. $\emptyset$ $\Longrightarrow$ {peanut butter}with $\emptyset$ the empty set.

Part 2b

Compute the following frequent itemsets with support $\geq$ 50% 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 2c

Compute all the association rules of the form

$$X \implies \{\text{bread}\}$$

with support $s \geq 50\%$ and confidence $\alpha \geq 60\%$.

Part d

Compute the *Lift* of the association rules of Part c.

Assignment 3 (DM; 7 pts; Sheet B): Clustering

Table 1 shows the distance between different cities, in minutes driving time.

	p_1	p_2	p_3	p_4	p_5
p_1	0	263	121	177	110
p_2	263	0	169	63	292
p_3	121	169	0	112	100
p_4	177	63	112	0	175
p_5	110	292	100	175	0

Table 1: Distance Matrix between Cities (Driving Time in Minutes)

Part 3a (3 pts)

Use the distance matrix in Table 1 to perform a *single* link agglomerative hierarchical clustering. Show your results by drawing a dendrogram. The dendrogram should clearly show the order in which the cities are merged. The y-axis should indicate the distance between the clusters.

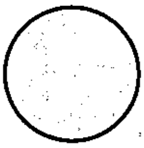
Part 3b (1 pts)

Imagine you created an agglomerative hierarchical clustering of a data set. How do you have to transform the result to make it comparable to the result of a k-means clustering?

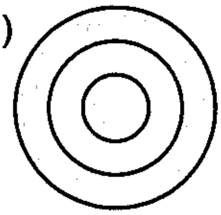
Part 3c (3 pts)

The following sketches show 2-dimensional points inside circles. Assume that the points are uniformly distributed in the circles. Assume you are using k-means clustering with a Euclidean distance function. Assume you are using k=2 clusters for Fig. i) and ii) and k=3 clusters for Fig. iii), and iv). For each of the four figures (i, ii, iii, iv) show how the points would (approximately) be separated into clusters and where (approximately) the centroid of the cluster would be. If there is more than one possible solution, indicate whether each solution is a local or global optimum.
What method would result in a better clustering of the data in Fig. ii? Explain why.

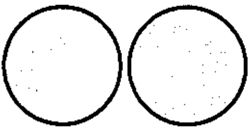
i)



ii)



iii)



iv)

Assignment 4 (DW; 7 pts; Sheet B): Advanced Data Warehousing

The US Department of Education built a data warehouse with application and acceptance data about all universities. Students have to apply to a university and the universities can accept a student or not. The legal officer of the department used the OLAP capabilities of the data warehouse and generated the following cross tables 4, 5, and 3 (each for a different university). These tables use the variables *accepted* (Yes or No), *female* (Yes, No), and department *Medicine* (Yes, No (means all other studies like computer science)). Assume that all differences in the cross tables are statistically significant. The legal officer is very concerned that some universities could be sued for sex discrimination.

The three Figures i), ii), and iii) in Table 2 show different causal graphs, which encode alternative believes about the causal influences between the variables. Assume that each graph shows the complete causal model.

- State for all nine combinations between the three cross tables and the three causal graphs: given the data table, would you reject the causal model (yes, no)? That means, which causal graph is inconsistent with which data table?

- Explain why you think that the causal graph i) is consistent or inconsistent with the Table. 4.

- Explain why you think that the causal graph ii) is consistent or inconsistent with the Table. 3.

- Which of the three causal graphs would represent an illegal discrimination of the applicants based on the gender?

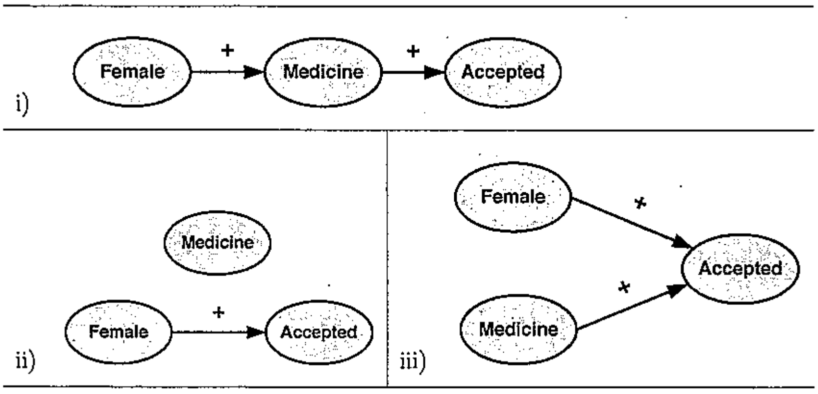


Table 2: Causal Graphs i), ii), and iii). Each graph shows alternative believes about the causal influences between the variables. E.g. Graph i) means that **Females** apply more for **Medicine** and students that apply for **Medicine** get more **Accepted** (because there are more open places for them)

		Medicine						
		Yes			No			Total
Accepted		Yes	No	Total	Yes	No	Total	
Female	Yes	1500	500	2000	50	150	200	2200
	No	150	50	200	500	1500	2000	2200
	Total	1650	550	2200	550	1650	2200	4400

Table 3: **Accepted** (yes, no) depending on sex (**Female** (yes, no)) and department (**Medicine** (yes, no)).

		Accepted		
		Yes	No	Total
Female	Yes	800	200	1000
	No	200	800	1000
	Total	1000	1000	2000

Table 4: **Accepted** (yes, no) depending on Sex (**Female** (yes, no)).

		Accepted		
		Yes	No	Total
Medicine	Yes	1600	400	2000
	No	400	1600	2000
	Total	2000	2000	4000

Table 5: **Accepted** (yes, no) depending on department (**Medicine** (yes, no)).

Assignment 5 (DW; 12 pts; Sheet C): Case 'Recommender System'

Database	Attributes	Description
Sales	Order: OrderID, CustomerID, Date, TotalPrice, Status (ordered, paid, or shipped). OrderLine: OrderID, ProductID, Price.	In one order, a customer can buy several products. Therefore, the sales database contains two tables: Order and Orderline. An integrity constraint ensures that the TotalPrice of the Order is equal to the sum of the prices of the corresponding OrderLines.
Customer	Customer: CustomerID, Name, Address, City, Country, Birthdate, CreditCardNr, Sex.	All customers who bought at least one product during the time the company existed.
Country	Country: Country, Continent.	With this table, one can lookup in which continent the country lies.
Product	Product: ProductId, Name, Description, Image, Type (book, music, movie, or game), Price.	All products that have been on sale during the time the company existed.

Table 6: Available data sources from Nozama

Our case is about the hypothetical company Nozama that has a webshop since 1998 selling books, music, movies, and games. They do well: in 2009 they had about 100,000 registered buyers; They sold on average 1000 products per day for on average 20 euro per product resulting in a revenue of 1000 × 20 × 365 = 7.3 million euro.

Dennis Doubovski, one of our guest speakers, talked about 'recommender systems'. Nozama has a recommender system in their webshop since 2002. It is used to recommend some additional products to the user, that are most likely to be of interest to him or her based on statements like 'people who bought the product you are looking at now, also bought these other products.' In 2007, they hired Dennis to select and install a better recommender system. It is now 2010 and they like to assess the effects of the new recommender system on their sales compared to the old recommender system. They do not have a data warehouse yet, so they plan to realize one. They like the dimensions in rather detailed granularity (e.g., 'per day' instead of 'per month'), because there is enough budget available for the necessary hardware. Available data sources can be found in Table 6.

Nozama management hypothesizes that trends in sales may be different for the different product groups (books, music, movies, and games), different age groups, the different continents buyers come from (europe, north america, south america, africa, asia, australia), and the popularity of the products. The latter is a bit complicated. We can distinguish between low, medium and high popularity based on how often a certain product has been sold. But new products are doomed to be classified as 'low' because they will not have been on sale for so long yet. Therefore, we classify products on popularity in four classes: low, medium, high, and new (less than a year old).

Part 5a (3 pts)

i) Draw a design of a star or snowflake schema for the data warehouse.
ii) Give the corresponding table design (tables and attributes).

Part 5b (2 pts)

Give an estimation of the size (in bytes) of the data warehouse. Also estimate by how much the data warehouse is expected to grow each year if no old data is deleted. Explain your answer.

Part 5c (1 pt)

One of the "do's" that Douwe Huizinga, the other guest speaker, discussed was "Start small; think big".
i) Explain what he meant by this
ii) If you were to follow this advice, suggest two small steps we could start with in this data warehousing project. Explain your answer.

Part 5d (2 pts)

In one of the lectures, we discussed several kinds of pre-processing: filter, standardize, enhance, abstract, join, aggregate. For each, give a concrete example (refer to specific attribute(s) and/or table(s)) where the kind of pre-processing can be used in this case. Explain your answer.

Part 5e (2 pts)

It can happen that a customer who had bought something before, re-registers as a new customer to buy something else, because for example they forgot their username and password. This creates duplicate records in the customer table (two or more distinct tuples belonging to the same customer in the real world). In 2008, Nozama conducted an analysis of their Customer database and they found out that 5% of the tuples were duplicates. Cleaning the databases is rather expensive.
Give arguments for and against cleaning the database.

(Continuation of case)
After the first design of the star schema was shown to Nozama, they added another requirement, namely that they would not only like to see trends in sales, but also in clicks. With clicks we mean how often a user viewing a certain product has clicked on one of the links produced by the recommender system. If a user clicks on such a link, it obviously doesn't mean that he or she also buys the recommended product. Nozama wonders if for this reason the click trends for the different product groups, age groups, continents, and popularity categories differ from the sales trends.
The necessary data can be obtained as follows. The webserver of the webshop maintains a log of all links that have been clicked. The links produced by the recommender system can easily be recognized. The user who clicked on the link, however, cannot always be determined: only if he was logged in; otherwise there is an IP address from which only the country of the user can be determined. A tool could be developed that produces a 'recommender click list': a list with per click: the product the user was viewing, and the user ID (in case it could be determined) or the country (otherwise). The recommender click list can be used as additional data source.

Part 5f (2 pts)

i) For some clicks you have a user ID; for some you only have a country. There are several alternative ways how to incorporate click data in the data warehouse, i.e., in the star or snowflake schema to accommodate the additional requirement. What do you advise?
ii) Draw the new star or snowflake schema design.