NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Computer and Information Science

Examination paper for TDT4300 Data warehousing and data mining

Academic contact during examination: Kjetil Nørvåg

Phone: 735 96755

Examination date: May 29th

Examination time (from-to): 09.00-13.00

Permitted examination support material: D: No tools allowed except approved simple calculator.

Other information:

Language: English

Number of pages: 4

Number of pages enclosed: 0

Checked by:

_____

Date Signature

Problem 1 – Various – 15% (all having same weight)

a) Assume two bit vectors *p* and *q*:

p = 1 0 0 0 0 0 0 0 0 1
q = 0 0 0 0 0 0 1 0 0 1

Calculate the Jaccard coefficient for the bit vectors p and q. What is the advantage of Jaccard compared to ”simple matching”?

b) Explain 3 techniques that can be employed for pre-processing of numerical data.

c) Explain the principles behind bitmap indexes. For what type of data is this indexing methods suitable, for what data is it not suitable?

Problem 2 – Modeling – 20% (15% on a, 5% on b)

Design a data warehouse for a regional weather bureau. The weather bureau has about 1000 probes, which are scattered throughout various land and ocean locations in the region to collect basic weather data, including air pressure, temperature, and precipitation at each hour. All data are sent to the central station, which has collected such data for over 10 years. Your design should facilitate efficient querying and on-line analytical processing, and derive general weather patterns.

The description is somewhat imprecise formulated and it is part of the task to select what should be included. We are primarily looking for you to show modeling principles for data warehousing. Explain any assumptions you find it necessary to do.

a) Make a star or snowflake schema for the described case.

b) Make two different concept hierarchies (you can chose dimensions).

Problem 3 – Clustering – 20 % (5% on a, 15% on b)

a) Explain possible disadvantages of hierarchical clustering.

b) 1) Explain the DBSCAN algorithm.
2) Assume a two-dimensional data set as shown in the table to the right. Cluster this data set using DBSCAN, given MinPts=3 and Eps=3.

X	Y
2	3
4	5
6	4
6	5
7	5
7	12
8	2
8	10
8	14
9	12
9	13
10	12
11	16
13	16
13	18
16	16
16	19

Problem 4 – Association analysis – 20 %

Assume the market basket to the right, Use the apriori-algorithm to find all frequent itemsets with minimum support of 50 % (e.g., *minimum support count* is 4). Chose one of the frequent 3-itemsets and find all association rules based on that set, given confidence of 75 %.

TransactionID	Element
T1	A,B,C,D
T2	A,G
T3	A,C,E,F
T4	B,C,G
T5	A,C,E,F
T6	C,D
T7	A,B,C,E,F
T8	A,B,C,E,F,G

Problem 5 – Classification– 25 % (5% on a and b, 15% on c)

a) Explain the purpose with classification. Give three examples of typical problems that we can solve using classification.

b) Explain briefly the principles behind *nearest neighbor classification.*

c) Magnus Carlsen and Vishy Anand will play the World Championship final against each other later this year. It has been decided that the final will be in India. Imagine that these two have previously faced each other several times, and that we have obtained data and information from their games. We also know that Carlsen's results have previously been dependent on how much effort he has put to prepare for his games. In the following table we have a dataset that shows the value 1 if Carlsen used full strength, while the value 0 means his efforts have not been top notch. We also take into account the type of matches, match time and place for the matches.

Time	**Match Type**	**Place**	**Effort**	**Outcome**
Morning	Master	Indoor	1	C
Afternoon	Grand Tour	Indoor Crowded	1	C
Night	Show	Mall	0	C
Afternoon	Show	Mixed	0	A
Afternoon	Master	Indoor Crowded	1	A
Afternoon	Grand Tour	Indoor	1	C
Afternoon	Grand Tour	Mall	1	C
Afternoon	Grand Tour	Mall	1	C
Morning	Master	Indoor	1	C
Afternoon	Grand Tour	Indoor Crowded	1	A
Night	Show	Mall	0	C
Night	Master	Mixed	1	A
Afternoon	Master	Indoor Crowded	1	A
Afternoon	Master	Indoor	1	C
Afternoon	Grand Tour	Mall	1	C
Afternoon	Grand Tour	Indoor Crowded	1	C

Games that ended in a draw (remis) are not included in this dataset.

Assume that we will use *decision tree* as a classification method. We will use the above dataset as our training data. To decide the best split we need to use **Entropy** for a node *t*, given by $Entropy(t) = -\sum_j p(j \mid t) \log p(j \mid t)$, where *p(j|t)* is the probability for class *j* given node *t (i.e.*, the portion of class *j* in node *t)* . For each split, the "information gain" defined by $GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$, where n_i is the number of elements in node *i* and *n* the total number of elements in the parent node *p*.

Task: The goal of the classification is to be able to predict the outcome of future matches between Carlsen and Anand. Compute the GAIN for splitting by attribute (1) "**Time**" and (2) "**Match Type**". Which of these splits would you chose to start building your decision tree? Justify your answer.