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: 41440433

Examination date: June 5th 2015

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 (front page excluded): 3

Number of pages enclosed: 0

Checked by:

_____

Date Signature

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

a) Explain *asymmetrical attribute.* Give an example of such an attribute.

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

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

Calculate the Jaccard coefficient for the bit vectors p and q.

c) In many datasets there can be values missing for attributes in some of the objects, often because some attributes are not relevant for everybody (for example, children usually don't have any salary). Give three methods/strategies that can be used to handle missing values.

Problem 2 – Modeling – 20 % (17 % on a, 3 % on b)

In this task we ask you to model a data warehouse for Netflix. Netflix provides streaming of TV series and films, and want a data warehouse in order to be able to analyze viewings of TV series (for simplicity, we ignore movies in this problem). A *viewing* in this context is defined as the event of a user watching a TV episode or part of a TV episode.

In order to simplify the modeling you can assume that time of a viewing is the time it starts, and that finest granularity of viewing is *chapter* (e.g., you don't have to model start- and end-time), where it is assumed that one episode consists of one or several chapters.

Examples of analysis one should be able to perform using the data warehouse:

- Average duration (time) of each viewing.
- Method of viewing (e.g., Android app, web browser, etc.) per quarter.
- Number of viewings for each chapter of a particular TV series for each country.

The description is somewhat imprecisely 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 schema for the described case.

b) A concept hierarchy for time can for example be *year-quarter-month-day*. Can *week* be part of this hierarchy? Justify your answer.

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

a) Explain advantages and disadvantages of k-means.

b) 1) Explain hierarchical agglomerative clustering.
2) Assume a two-dimensional dataset as given in the table to the right. Perform hierarchical agglomerative clustering on this dataset using MIN (single link) and Manhattan-distance. Show the resulting dendrogram.

X	Y
2	3
4	5
6	4
6	5
7	5
7	12
8	2
8	10

Problem 4 – Classification – 25 % (5 % on a and 20 % on b)

a) Explain *confusion matrix,* its contents, and how to calculate *accuracy* based on this.

b) Rosenborg and Vålerenga will tomorrow (Saturday) play a football match (Elite League) at Ullevaal (home stadium for Vålerenga). The teams have played against each other many times before, and we want to use the results and information from previous matches in order to predict the result of tomorrow's match. This information is given in the table below (matches that ended in a draw are not included, and H/A means Rosenborg played at home or away).

Weekday	**Tournament**	**Location**	**Time**	**Result**
Friday	Elite League	H	Afternoon	R
Sunday	Cup	H	Evening	R
Sunday	Elite League	A	Afternoon	R
Sunday	Elite League	H	Evening	R
Saturday	Elite League	A	Afternoon	V
Sunday	Elite League	H	Afternoon	R
Sunday	Elite League	H	Evening	R
Saturday	Elite League	A	Afternoon	R
Sunday	Elite League	H	Evening	R
Sunday	Elite League	H	Afternoon	R
Friday	Elite League	H	Evening	R
Sunday	Elite League	A	Evening	V
Saturday	Elite League	H	Afternoon	R
Sunday	Elite League	A	Afternoon	R
Sunday	Elite League	A	Evening	V
Saturday	Elite League	H	Afternoon	V

Assume that we will use *decision tree* as the classification method. We will use the above dataset as our training data. We use the *Gini index* as measure for impurity, and the following two equations might be of help for solving the problem:

$$GINI(t) = 1 - \sum_j [p(j \mid t)]^2$$

$$GAIN_{split} = GINI(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} GINI(i) \right)$$

Task: The goal of the classification is to be able to predict the outcome of tomorrows (Saturday) match between Rosenborg and Vålerenga. Compute the $GAIN_{split}$ for splitting by attribute (1) "*Location*" and (2) "*Weekday*". Which of these splits would you chose to start building your decision tree? Justify your answer.

Problem 5 – Association rules – 20 %

Assume the market basket data below. Use the apriori-algorithm to find all frequent itemsets with minimum support of 50 % (i.e., *minimum support count* is 4). Show how the candidate sets are generated.

BDE is one of the frequent itemsets. Find all association rules based on this set, given confidence of 75 % (it is not necessary to use apriori to find the association rules, but show how confidence is calculated for each of the candidate rules that are based on BDE).

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