



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: 73596755

Examination date: May 26th 2016

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): 4

Number of pages enclosed: 0

Checked by:

Date

Signature

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig 2-sidig

sort/hvit farger

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

- a) List four usages for clustering validation/evaluation.
- b) Explain four techniques for data cleaning in the context of web usage data.
- c) Assume two bit vectors p and q :

$$p = 1\ 0\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 1$$
$$q = 1\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 1$$

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

Problem 2 – Modeling – 15 %

In this task we ask you to model a data warehouse for car damages for the insurance company Lillebrand. Lillebrand wants a data warehouse in order to be able to analyze events that have resulted in insurance payments.

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

- Number of damages in 2015.
- Average number of damages per month.
- Number of damages for each quarter in 2015.
- Total amount paid for each car type.
- Number of damages of type “collision” for each city.

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.

Make a star schema for the described case.

Problem 3 – OLAP – 15 % (5 % on a and 10 % on b)

a) Given a cube with dimensions:

Time(day-month-quarter-year)
 Item(item_name-brand-type)
 Location(street-city-province_or_state-country)

Assume the following materialized cuboids:

- 1) {*year, item_name, city*}
- 2) {*year, brand, country*}
- 3) {*year, brand, province_or_state*}
- 4) {*item_name, province_or_state*} where *year = 2004*

Given the following OLAP query: {*item_name, province_or_state*} with condition “*year = 2006*”

Which of the materialized cuboids can be used to process the query? Justify the answer.

b) Given a data warehouse with three tables Location/Item/Sales, where Sales is the fact table and the two others are dimension tables. We want to use join indexes in order to process queries more efficiently. Show the structure and contents of the join indexes Location/Sales and Item/Sales based on the contents of the three tables below.

Location	
LocKey	CityName
L1	Oslo
L2	Athen
L3	Trondheim

Item	
ItemKey	ItemName
I1	Sony-TV
I2	Rolex
I3	Lexus

Sales			
TransID	LocKey	ItemKey	Price
T1	L1	I1	5
T2	L2	I2	8
T3	L1	I1	6
T4	L3	I1	5
T5	L3	I3	9
T6	L1	I2	8
T7	L1	I1	4

Problem 4 – Clustering – 10 %

Assume a two-dimensional dataset as shown in the table to the right. Cluster this dataset using DBSCAN, given MinPts=4 (incl. own point) and Eps=3 (incl. points having distance 3). Use Manhattan distance.

X	Y
4	8
4	9
4	10
4	13
4	14
5	3
5	7
5	14
6	15
6	16
6	19
7	11
7	16
7	17
7	18
7	19

Problem 5 – Classification – 20 % (5 % on a and 15 % on b)

- a) Explain cross validation and what this technique is used for.
- b) A car insurance company has for existing customers stored information that include customer number, age (L/M/H, e.g., 18-25/26-70/71-100), car type, driving length per year (4000/8000/20000/Unlimited), bonus (Low/Medium/High) and if they have had damage on their car that has been covered by the insurance. When new customers ask for a price offer, the company want to set the price to normal or high depending on whether they believe the customer is going to get a damage on the car or not, i.e., they want to predict the attribute "Damage".

Cust.nr.	Age	Car type	Lenght per year	Bonus	Damage
1	L	Ferrari	8000	Low	Yes
2	M	BMW	8000	High	No
3	H	Lexus	Unlimited	High	Yes
4	L	Audi	8000	High	No
5	H	Opel	8000	Low	Yes
6	M	Toyota	8000	Low	No
7	M	Honda	8000	High	No
8	M	Nissan	8000	High	No
9	M	Audi	Unlimited	High	No
10	M	BMW	8000	Low	Yes
11	H	Toyota	Unlimited	High	No
12	L	Nissan	4000	Low	Yes
13	L	Opel	Unlimited	High	Yes
14	M	Audi	8000	High	No
15	M	Opel	8000	High	No
16	M	Toyota	4000	Low	No

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:

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

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

Task: The goal of the classification is to be able to predict “Damage”. Compute the $GAIN_{split}$ for splitting by attribute (1) “Age” and (2) “Bonus”. Which of these splits would you chose to start building your decision tree? Justify your answer.

Problem 6 – Association rules – 25 % (10 % on a, 5 % on b, and 10 % on c)

- a) 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). Use the $F_{k-1} \times F_{k-1}$ method for candidate generation.

TransactionID	Item
T1	ABCDFGH
T2	DKM
T3	FK
T4	ACGH
T5	ACDDGH
T6	BM
T7	DFKM
T8	ABCDGH

- b) Assume the market basket data below. You are now going to use the *FP-growth-algorithm* in order to find all frequent itemsets with minimum support of 60 % (i.e., *minimum support count* is 3).

1) Construct a FP tree based on the dataset.

2) Find frequent itemsets using the FP-growth-algorithm. Use table notation with the following columns in order to show the result:

- Item
- ”Conditional pattern base”
- ”Conditional FP-tree”
- Frequent itemsets

TransactionID	Item
T1	f, a, c, d, g, i, m, p
T2	a, b, c, f, l, m, o
T3	b, f, h, j, o
T4	b, c, k, s, p
T5	a, f, c, e, l, p, m, n