Department of Computer Science

Examination paper for Data warehousing and data mining

Academic contact during examination: Kjetil Nørvåg

Phone: 41440433

Examination date: May 22nd

Examination time (from-to): 1500-1900

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

Other information:

Students will find the examination results in Studentweb. Please contact the department if you have questions about your results. The Examinations Office will not be able to answer this.

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

a) What is binarization, and how should this be performed?

b) The silhouette coefficient is given by the following equation: s = (b-a)/*max*(a,b)
Explain what purpose it can be used for, and how to calculate *a* and *b*.

c) Explain the most important limitations of application of hierarchical agglomerative clustering (HAC) on large datasets.

Problem 2 – Modeling – 10 %

Ilsvika Energi (IE) supplies energy to a large number of customers in Trøndelag. All customers are now going to get smart meters installed, which once a minute send information to IE about energy consumption the last minute. With smart meters, it is possible to offer dynamic pricing, i.e., the price can change from minute to minute, so that a higher price has to be paid in periods of high consumption (e.g., afternoons when everybody is making dinner), and less in periods of lower consumption (e.g., at night). A customer can have subscription for more than one location, in this case there is one meter for each location. IE wants a data warehouse that can be used for analyzing energy consumption.

An example of analyzes you should be able to do against the data warehouse:

- Total consumption for each hour.
- Total consumption for each hour for each customer.
- Total consumption for each hour for each location.
- Total consumption per day per municipality.

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 necessary to do.

Create a star schema for the described case. Answer on paper.

Problem 3 – OLAP – 10 % (all having same weight)

a) Given a base cuboid, there are three alternative strategies for data cube materialization. Explain these, and advantages/disadvantages for each.

b) 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, brand, city*}
2) {*year, brand, street*}
3) {month*, brand, province_or_state*}
4) {*item_name, province_or_state*} where *year = 2006*

Given the following OLAP query: {*item_name, country*} with condition “*year = 2006*”
Which of the materialized cuboids can be used to process the query? Justify the answer.

Problem 4 – Clustering – 15 % (5 % on a and 10 % on b)

X	Y
2	4
2	5
2	6
2	10
2	11
3	3
3	11
4	12
4	13
4	16
7	2
7	2

a) Assume a *d*-dimensional dataset with 1000 points that is to be clustered using DBSCAN, explain how you can find suitable values for the parameters *MinPts* and *Eps*.

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

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

Nr	A	B	C	D	E	*Class*
1	L	K	R	J	2	*J*
2	H	F	S	N	4	*J*
3	H	T	S	N	4	*J*
4	L	F	S	J	2	*N*
5	L	F	G	N	5	*N*
6	H	T	G	N	2	*N*
7	L	F	S	N	6	*N*
8	L	K	G	N	4	*N*
9	H	T	H	N	2	*J*
10	L	F	S	J	5	*N*
11	L	K	B	N	7	*N*
12	H	F	B	N	9	*J*
13	L	K	R	J	2	*N*
14	L	F	H	J	1	*N*
15	L	F	H	N	7	*N*

a) Explain two techniques for reducing the negative impact of overfitting in decision trees. Which of those are usually preferred?

b) As part of a larger application we want to be able to predict class (*J* or *N*) based on input data where each record contains a sequence number and the attributes A, B, C, D, and E, cf. the table above.

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:

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

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

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

Problem 6 – Association rules (1) – 15 % (10 % on a and 5 % on B)

TransaksjonsID	Items
T1	ACDK
T2	ADK
T3	CBDJK
T4	CEF
T5	BDEJK
T6	ADK
T7	ABDEJK
T8	BDFJK

a) Assume the market basket data given above. 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.

b) Assume the following closed frequent itemsets: C:3, AC:2, BE:3, BCE:2
(Format: itemset:supportcount)
Find all frequent itemsets and their support counts.

Problem 7 – Association rules (2) – 10 %

TransaksjonsID	**Items**
T1	ABG
T2	ABCD
T3	ACJ
T4	BC
T5	ACH
T6	BCL
T7	ABCD
T8	ABCDE
T9	ABK

Assume the market basket data above. You are now going to use the *FP-growth-algorithm* in order to find all frequent itemsets with minimum support of 22 % (i.e., *minimum support count* is 2).

1) Construct a FP tree based on the dataset. Submit this on paper as problem 8.

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

Problem 8 – FP-tree for problem 7 – 5 %

Answer on paper.