

FINAL EXAMINATION

TDT4300

AUTUMN 2021

INFORMATION

- Academic contact during examination: Dhruv Gupta
- E-mail: dhruv.gupta@ntnu.no
- Examination date: 09-August-2021
- Examination time (from-to): 09:00-13:00
- Permitted examination support material: Open book
- Language: English

- Checked By:
- Date:
- Signature:

1 DATAWAREHOUSES AND OLAP OPERATIONS

Exercise 1. Book Franchise. Shops belonging to well-known book franchise sell various kinds of printed materials (e.g., recipe books, children books, novels etc.). Items in each shop are provided by publishers located around the world. The franchise currently has bookshops only in locations across Europe. The franchise utilizes an out-dated method of maintaining its sales across stores. Concretely, each shop-owner sends out a summary of everyday's sales to the headquarters for analysis. The CEO of the franchise wants to adopt the data warehousing approach for data analytics. As a new data scientist at the franchise headquarters you are tasked with the implementation of their data warehouses. Answer the questions below and state any assumptions you have made to model the data warehouse.

1. Create the concept hierarchies for the different dimensions that are part of the above problem statement.
2. Create a star schema to implement the data warehouse.
3. Additionally, create a snowflake schema to implement the data warehouse. What is the one key feature that would make snowflake schema more useful for implementing the data warehouse as compared to star schema.

Solution 1

1. Concept Hierarchies.

- a) Printed Material: Name → Author → Type → Genre → ALL.
- b) Location: City → Country → Continent → ALL.
- c) Time: Day → Week → Month → Year → ALL.
- d) Shop: Location → Type → ALL.

2. Star Schema.

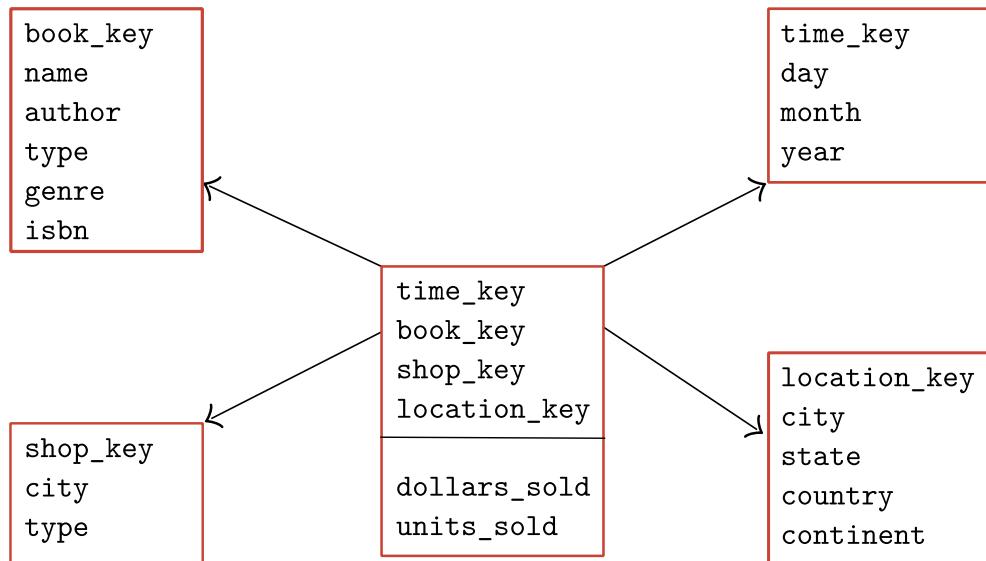


Figure 1: Star Schema.

3. Snowflake Schema.

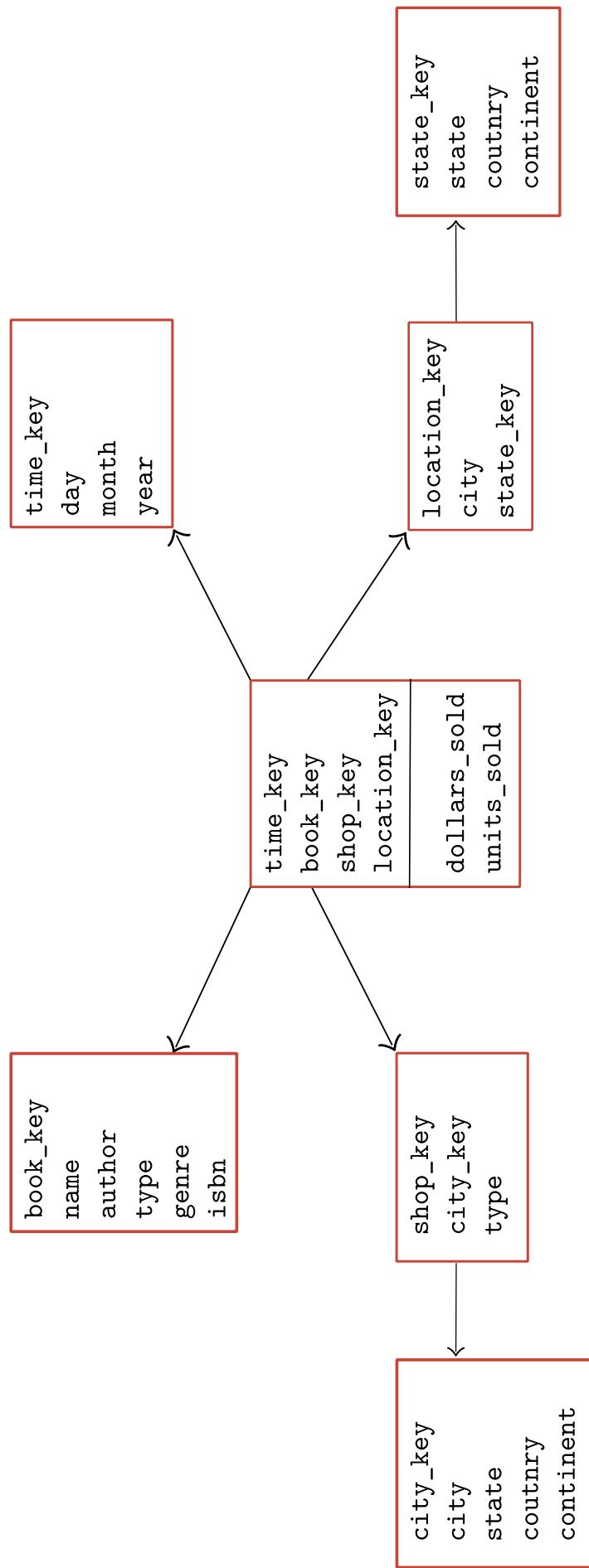


Figure 2: Snowflake Schema.

Exercise 2. Oracle has multiple software engineering teams across Norway. The following table details of some of the employees in the country. Due to the ongoing Covid-19 pandemic, Oracle would like to implement the hybrid-workplaces concept. Specifically, to gain insights into the working habits of its employees Oracle wants to utilize data warehouses which store its employee records. To speedup the query processing Oracle utilizes the concept of bitmap indexes in its data warehouse solution. How would the bitmap indexes look like for the sample data below? Having constructed the bitmap indexes answer the following questions to help Oracle.

Employee ID	Name	Gender	Dependents	Office	Title
1	Magnus	Male	Yes	Oslo	Developer
2	Kjetil	Male	Yes	Bergen	Tester
3	Anna	Female	No	Trondheim	Developer
4	Charlotte	Female	No	Oslo	Tester
5	John	Male	Yes	Oslo	Project manager
6	Birgit	Female	Yes	Bergen	Developer
7	Robert	Male	No	Trondheim	Developer

Show the bitmap indices on Gender, Office, and Title attributes and use them to:

1. Identify developers that have dependents.
2. Identify employees that are situated outside Trondheim.
3. Identify male employees that reside outside Trondheim.
4. Identify testers working in Oslo or Bergen.

Solution 2

- Bit Map Indexes for Gender.

	1	2	3	4	5	6	7
Male	1	1	0	0	1	0	1
Female	0	0	1	1	0	1	0

- Bit Map Indexes for Dependents.

	1	2	3	4	5	6	7
Male	1	1	0	0	1	1	0
Female	0	0	1	1	0	0	1

- Bit Map Indexes for Office.

	1	2	3	4	5	6	7
Oslo	1	0	0	1	1	0	0
Bergen	0	1	0	0	0	1	0
Trondheim	0	0	1	0	0	0	1

- Bit Map Indexes for Title.

	1	2	3	4	5	6	7
Developer	1	0	1	0	0	1	1
Tester	0	1	0	1	0	0	0
Project Manager	0	0	0	0	1	0	0

- Answer to part 1.

`Title = "Developer" AND Dependents = "Yes".`

	1	2	3	4	5	6	7
<code>Title = "Developer"</code>	1	0	1	0	0	1	1
<code>Dependents = "Yes"</code>	1	1	0	0	1	1	0
<code>AND</code>	1	0	0	0	0	1	0

Answer = {Magnus, Birgit}.

- Answer to part 2.

`¬(Office = "Trondheim").`

	1	2	3	4	5	6	7
<code>Office = "Trondheim"</code>	0	0	1	0	0	0	1
<code>¬</code>	1	1	0	1	1	1	0

Answer = {Magnus, Kjetil, Charlotte, John, Birgit}.

- Answer to part 3.

`Gender = "Male" AND (¬(Office = "Trondheim")).`

	1	2	3	4	5	6	7
<code>Gender = "Male"</code>	1	1	0	0	1	0	1
<code>¬Office = "Trondheim"</code>	1	1	0	1	1	1	0
<code>AND</code>	1	1	0	0	1	0	0

Answer = {Magnus, Kjetil, John}.

- Answer to part 4.

`Title = "Tester" AND ((Office = "Oslo") OR (Office = "Bergen")).`

	1	2	3	4	5	6	7
<code>Office = "Oslo"</code>	1	0	0	1	1	0	0
<code>Office = "Bergen"</code>	0	1	0	0	0	1	0
<code>OR</code>	1	1	0	1	1	1	0
	1	2	3	4	5	6	7
<code>Title = "Tester"</code>	0	1	0	1	0	0	0
<code>Office = "Oslo" OR "Bergen"</code>	1	1	0	1	1	1	0
<code>AND</code>	0	1	0	1	0	0	0

Answer = {Kjetil, Charlotte}.

2 DATA

Exercise 3. Consider the data presented in the table below. What are some of the steps that you will take to clean the data before applying any classification methods (supervision is done based on the Class label) and why?

Instance	Type	A	B	C	D	Class
1	High	3.5	4.0	7.0	-2.00	H
2	High	2.0	-4.0	4.0	2.00	H
3	Low	9.1	4.5	18.2	-2.25	L
4	High	2.0	-6.0	4.0	3.00	H
5	High	1.5	7.0	3.0	-3.50	H
6	High	7.0	-6.5	14.0	3.25	H
7	Low	2.1	2.5	4.2	-1.25	L
8	Low	8.0	-4.0	16.0	2.00	L

Table 1: Data for pre-processing.

Solution 3

1. Remove Type as it is correlated and same as Class attribute for training.
2. Similarly, attribute A and C are positively correlated. Remove A or C.
3. Similarly, attribute B and D are negatively correlated. Remove B or D.

Exercise 4. Consider a dataset that has 1×10^6 features and 1×10^9 instances. We need to apply a data mining algorithm to this massive dataset and measure its performance. Given its size, we would ideally select the most relevant subset of features to reduce computation cost. How many times would we need to run the algorithm for this ideal feature selection procedure? Do you think this is feasible? What would be some alternative methods of solving this problem (only name the methods)?

Solution 4

We need $2^n = 2^{10^6}$ iterations as there are 2^n subset of features. Other ways of feature selection or dimensionality reduction: PCA, LDA etc.

Exercise 5. Consider the following employee database. What is the attribute types for each of the attributes that are recorded for the employees.

Employee ID	Name	Age	Joining Date	Title
1	Magnus	23	01-09-2021	Developer
2	Kjetil	25	01-04-2020	Tester
3	Anna	30	01-01-2016	Developer
4	Charlotte	25	01-02-2019	Tester
5	John	35	01-06-2014	Project manager
6	Birgit	60	01-03-2000	Developer
7	Robert	55	01-07-2005	Developer

Solution 5

1. Name - nominal.
2. Age - ratio.
3. Joining Date - interval.
4. Title - Ordinal.

3 ASSOCIATION RULE ANALYSIS

Exercise 6. Compute the frequent itemsets for the transaction database given in table below using the Apriori algorithm with minimum support equal to 3. While answering the question, also write down step-by-step procedure that would entail by applying the Apriori algorithm.

A	B	C	D	E
1	2	1	1	2
3	3	2	6	3
5	4	3		4
6	5	5		5
	6	6		

Solution 6

- Candidate 1-itemsets and frequent 1-itemsets.

C_1	Support
A	4
B	5
C	5
D	2
E	4

L_1	Support
A	4
B	5
C	5
E	4

- Candidate 2-itemsets and frequent 2-itemsets.

C_2	Support
AB	3
AC	4
AE	2
BC	4
BE	4
CE	3

L_2	Support
AB	3
AC	4
BC	4
BE	4
CE	3

- Candidate 3-itemsets and frequent 3-itemsets.

C_3	Support
ABC	3
BCE	3

L_3	Support
ABC	3
BCE	3

There is no need to scan the transaction database for frequent 4-itemsets as there are no candidates.

Exercise 7. Compute the frequent itemsets for the transaction database given in the table below using the FP-Growth algorithm with minimum support equal to 3. Show the FP-Growth procedure step-by-step including the building of the FP-Tree and the projected FP Trees.

A	B	C	D	E
1	2	1	1	2
3	3	2	6	3
5	4	3		4
6	5	5		5
	6	6		

Solution 7 • 1-itemset support values.

1-Itemset	Support
A	4
B	5
C	5
D	2
E	4

- 1-itemset reordered based on support values.

1-Itemset	Support
B	5
C	5
A	4
E	4
D	2

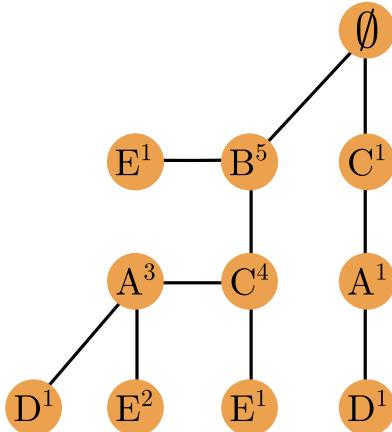
- Transaction database.

tid	Transaction
1	ACD
2	BCE
3	ABCE
4	BE
5	ABCE
6	ABCD

- Reordered transaction database.

tid	Transaction
1	CAD
2	BCE
3	BCAE
4	BE
5	BCAE
6	BCAD

- FP-Tree for the entire transaction database.



1. Project on D.

We can remove D from the FP-Tree as its support (2) is less than min. support (3).

Frequent Path = {Ø}

2. Project on E.

Path	Count
BCAE	2
BCE	1
BE	1

– Projected FP-Tree for E: Ø → B⁴ → C³ → A².

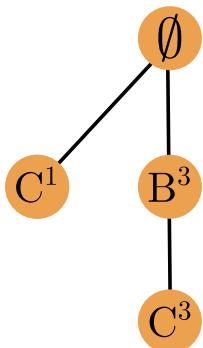
Since, support for A is less than minimum support we can remove it for frequent itemset generation.

Frequent Path = {EB(4), EC(3), EBC(3)}

4. Project on A.

Path	Count
CA	1
BCA	3

– Projected FP-Tree for A:



2.1 Project on AC.

Path	Count
BC	3
B	1

– Projected FP-Tree for AC: $\emptyset \rightarrow B^3$.
Frequent Pattern = {AC(4), ACB(3)}.

2.2 Project on AB.

Path	Count
B	3

Frequent Pattern = {AB(3)}.

4. Project on C.

Path	Count
C	1
BC	4

– Projected FP-Tree for C: $\emptyset \rightarrow B^4$.
Frequent Path = {CB(4)}

5. Project on B.

Path	Count
B	5

Frequent Path = { \emptyset }

4 CLUSTERING

Exercise 8. K-Means Clustering. Consider the dataset given below. Apply the k-Means Clustering algorithm assuming $k = 2$ and initial cluster assignments are: $C_1 = \{x_1, x_2, x_4\}$ and $C_2 = \{x_3, x_5\}$. Show the iterations of the k-Means Clustering algorithm until it converges. Utilize the L_1 norm (also known as the Manhattan Distance) for distance computations.

Instance	X ₁	X ₂
x ₁	0	2
x ₂	0	0
x ₃	1.5	0
x ₄	5	0
x ₅	5	2

Table 2: Table for K-means based exercise.

Solution 8 1 Iteration 1.

$$k = 2$$

$$C_1 = \{x_1, x_2, x_4\}$$

$$C_2 = \{x_3, x_5\}$$

Initial Centroids:

$$C_1 = (\text{median}(0, 0, 5), \text{median}(0, 0, 2)) = (0, 0).$$

$$C_2 = (\text{median}(1.5, 5), \text{median}(0, 2)) = (3.25, 1).$$

Distances of the data points to the two centroids:

id	X ₁	X ₂	d ₁	d ₂
x ₁	0	2	2	4.25
x ₂	0	0	0	4.25
x ₃	1.5	0	1.5	2.75
x ₄	5	0	5	2.75
x ₅	5	2	7	2.75

$$C_1 = \{x_1, x_2, x_3\}$$

$$C_2 = \{x_4, x_5\}$$

Updated Centroids:

$$C_1 = (\text{median}(0, 0, 1.5), \text{median}(0, 0, 2)) = (0, 0).$$

$$C_2 = (\text{median}(5, 5), \text{median}(0, 2)) = (5, 1).$$

2 Iteration 2.

Distances of the data points to the two centroids:

$$C_1 = \{x_1, x_2, x_3\}$$

$$C_2 = \{x_4, x_5\}$$

id	X ₁	X ₂	d ₁	d ₂
x ₁	0	2	2	6
x ₂	0	0	0	6
x ₃	1.5	0	1.5	4.5
x ₄	5	0	5	1
x ₅	5	2	7	2

Updated Centroids:

$$C_1 = (\text{median}(0, 0, 1.5), \text{median}(0, 0, 2)) = (0, 0).$$

$$C_2 = (\text{median}(5, 5), \text{median}(0, 2)) = (5, 1).$$

Centroids and cluster assignments repeat hereafter.

Exercise 9. DBScan Algorithm. DBScan algorithm allows the discovery of clusters of arbitrary shapes. We would like to process the given set of points in the figure below with the DBScan algorithm. To do so, use the following parameters: $\text{eps} = 1$ and $\text{minpts} = 6$. Also, consider the following distance function L_{\min} ,

$$L_{\min}(x, y) = \min_{i=1}^d \{|x_i - y_i|\}.$$

As an example computation of distances, consider two points $x = \langle 1, 2 \rangle$ and $y = \langle 2, 4 \rangle$. Then, L_{\min} is computed as:

$$\begin{aligned} L_{\min}(x, y) &= \min_{i=1}^2 \{|x_i - y_i|\} \\ &= \min \{|1 - 2|, |2 - 4|\} \\ &= \min \{1, 2\} \\ &= 1. \end{aligned}$$

Specifically, identify the noise points, border points, and core points of the clusters while answering the question.

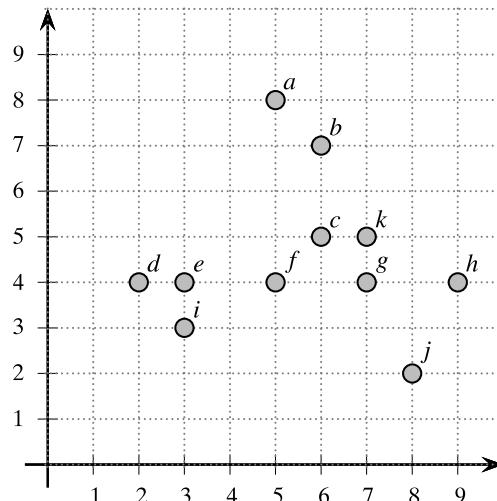


Figure 3: Figure for the hierarchical agglomerative clustering and density based clustering.

Solution 9 Distance Matrix

	a	b	c	d	e	f	g	h	i	j	k
a											
b	1										
c	1	0									
d	3	3	1								
e	2	3	1	0							
f	0	1	1	0	0						
g	2	1	1	0	0	0					
h	4	3	1	0	0	0	0				
i	2	3	2	1	0	1	1	1			
j	3	2	2	2	2	1	1	1			
k	2	1	1	1	1	0	1	2	1	0	

Density

	a	b	c	d	e	f	g	h	i	j	k
Density	4	6	6	8	8	10	10	9	7	5	9

- Core Points = {b,c,d,e,f,g,h,i,k}, as their density (number of points within their eps radius) is ≥ 6 .
- Border Points = {a, j}, as they are in the vicinity (within eps radius) of a core point but their density is < 6 .
- Noise Points = {}.

5 CLASSIFICATION

Exercise 10. To build decision trees, the simplest algorithm to use is the Hunt's algorithm. For the dataset given in Table 3, create a decision tree using the Hunt's algorithm. For the evaluation criteria utilize the Gini Index. Having constructed the decision tree, what class does the instance $\langle \text{Age} = 27, \text{Car} = \text{Vintage} \rangle$ belong to?

Instance	Age	Car	Risk
1	25	Sports	L
2	20	Vintage	H
3	25	Sports	L
4	45	SUV	H
5	20	Sports	H
6	25	SUV	H

Table 3: Table for decision tree based exercise.

Solution 10 1. Determination of root node split.

- Split on Age:

	2×H		2×L/H		H			
Data Point	20		25		45			
Split Point	10		22.5		35		50	
	\leq	$>$	\leq	$>$	\leq	$>$	\leq	$>$
Class='L'	0	2	0	2	2	0	2	0
Class='H"	0	4	2	2	3	1	4	0
Gini Index	0.4		0.3		0.4		0.4	

– Split on comparing Age with 10:

Age ≤ 10	
L	0
H	0

$$\text{Gini} = 1 - 0 - 0 = 1.$$

Age > 10	
L	2
H	4

$$\begin{aligned}\text{Gini} &= 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{2}{6}\right)^2 \\ &= 1 - \frac{1}{9} - \frac{4}{9} \\ &= \frac{9-5}{9} = \frac{4}{9} \\ &= 0.\bar{4}.\end{aligned}$$

$$\begin{aligned}\text{Gini} &= \frac{0}{6} \cdot 1 + \frac{6}{6} \cdot \frac{4}{9} \\ &= 0.\bar{4}.\end{aligned}$$

– Split on comparing Age with 22.5:

Age ≤ 10	
L	0
H	2

$$\text{Gini} = 1 - 0 - 1 = 0.$$

Age > 10	
L	2
H	2

$$\begin{aligned}\text{Gini} &= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 \\ &= 1 - \frac{1}{4} - \frac{1}{4} \\ &= \frac{1}{2} \\ &= 0.5.\end{aligned}$$

$$\begin{aligned}\text{Gini} &= \frac{2}{6} \cdot 0 + \frac{4}{6} \cdot \frac{1}{2} \\ &= \frac{2}{6} = \frac{1}{3} = 0.\bar{3}.\end{aligned}$$

– Split on comparing Age with 35:

Age ≤ 35	
L	2
H	3

$$\begin{aligned}\text{Gini} &= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 \\ &= 1 - \frac{4}{25} - \frac{9}{25} \\ &= \frac{25-13}{25} = \frac{12}{25} \\ &= 0.48.\end{aligned}$$

Age > 35	
L	0
H	1

$$\text{Gini} = 1 - 0 - 1 = 0$$

$$\begin{aligned}\text{Gini} &= \frac{5}{6} \cdot \frac{12}{25} + \frac{1}{6} \cdot 0 \\ &= \frac{2}{5} = 0.4.\end{aligned}$$

– Split on comparing Age with 50:

Age ≤ 50	
L	2
H	4

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2 \\ &= 1 - \frac{1}{9} - \frac{4}{9} \\ &= \frac{9-5}{9} = \frac{4}{9} \\ &= 0.\bar{4}. \end{aligned}$$

Age > 50	
L	0
H	0

$$\text{Gini} = 1 - 0 - 0 = 1.$$

$$\begin{aligned} \text{Gini} &= \frac{6}{6} \cdot \frac{4}{9} + \frac{0}{6} \cdot 1 \\ &= 0.\bar{4}. \end{aligned}$$

• Split on Car:

– Multi-way split on Car:

Car = Sports	
L	2
H	1

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \\ &= 1 - \frac{1}{9} - \frac{4}{9} \\ &= \frac{9-5}{9} = \frac{4}{9} \\ &= 0.\bar{4}. \end{aligned}$$

Car = Vintage	
L	0
H	1

$$\text{Gini} = 1 - 0 - 1 = 0$$

Car = SUV	
L	0
H	2

$$\text{Gini} = 1 - 0 - 1 = 0$$

$$\text{Gini} = \frac{3}{6} \cdot \frac{4}{9} + 0 + 0 = \frac{2}{9} = 0.\bar{2}$$

– Best binary split on Car:

Car = Sports	
L	2
H	1

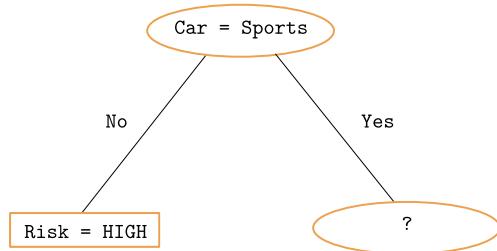
$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \\ &= 1 - \frac{1}{9} - \frac{4}{9} \\ &= \frac{9-5}{9} = \frac{4}{9} \\ &= 0.\bar{4}. \end{aligned}$$

Car $\in \{\text{SUV, Vintage}\}$	
L	0
H	3

$$\text{Gini} = 1 - 0 - 1 = 0$$

$$\text{Gini} = \frac{3}{6} \cdot \frac{4}{9} + 0 = \frac{2}{9} = 0.\bar{2}$$

- Decision tree based on splitting on Car.



2. Determination of next node split.

- Split on Age:

	H	$2 \times L$				
Data Point	20	25				
Split Point	10	22.5	30			
	\leq	>	\leq	>	\leq	>
Class='L'	0	2	0	2	2	0
Class='H'	0	1	1	0	1	0
Gini Index	0.4		0.3		0.4	

– Split on comparing Age with 10:

Age ≤ 10	
L	0
H	0

$$\text{Gini} = 1 - 0 - 0 = 1.$$

Age > 10	
L	2
H	1

$$\begin{aligned}
 \text{Gini} &= 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \\
 &= 1 - \frac{4}{9} - \frac{1}{9} \\
 &= \frac{9-5}{9} = \frac{4}{9} \\
 &= 0.4.
 \end{aligned}$$

$$\begin{aligned}
 \text{Gini} &= \frac{0}{3} \cdot 1 + \frac{3}{3} \cdot \frac{4}{9} \\
 &= 0.4.
 \end{aligned}$$

– Split on comparing Age with 22.5:

Age ≤ 22.5	
L	0
H	1

Age > 22.5	
L	2
H	0

$$\text{Gini} = 1 - 0 - 1 = 0.$$

$$\text{Gini} = 1 - 1 - 0 = 0.$$

$$\text{Gini} = 0.$$

- Split on comparing Age with 30:

Age ≤ 35	
L	2
H	1

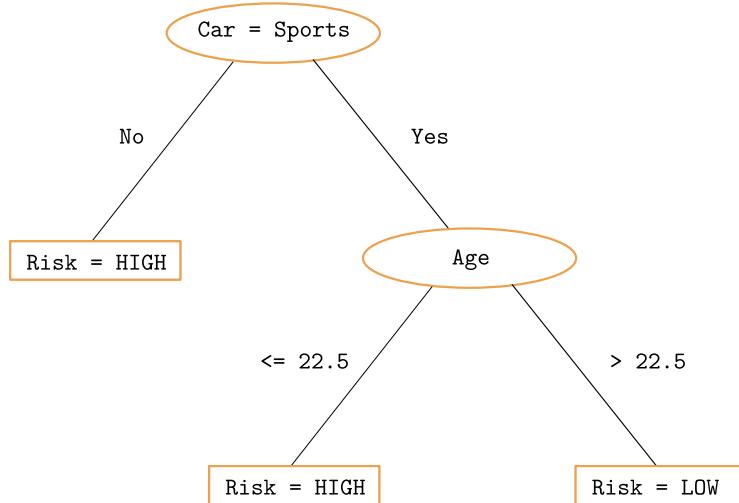
$$\begin{aligned}
 \text{Gini} &= 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 \\
 &= 1 - \frac{4}{9} - \frac{1}{9} \\
 &= \frac{9-5}{9} = \frac{4}{9} \\
 &= 0.\bar{4}.
 \end{aligned}$$

Age > 35	
L	0
H	0

$$\text{Gini} = 1 - 0 - 0 = 1$$

$$\begin{aligned}
 \text{Gini} &= \frac{3}{3} \cdot \frac{4}{9} + \frac{0}{3} \cdot 1 \\
 &= \frac{4}{9} = 0.\bar{4}.
 \end{aligned}$$

- Final decision tree based on splitting on Age.



- The given instance $\langle \text{Age} = 27, \text{Car} = \text{Vintage} \rangle$ would be classified with the risk label of High.

FINAL EXAMINATION

TDT4300

SPRING 2021

INFORMATION

- Academic contact during examination: Dhruv Gupta
- E-mail: dhruv.gupta@ntnu.no
- Examination date: 28-May-2021
- Examination time (from-to): 09:00-13:00
- Permitted examination support material: Open book
- Language: English

- Checked By:
- Date:
- Signature:

1 DATAWAREHOUSES AND OLAP OPERATIONS

Exercise 1. Total Marks: 12 Marks

You are hired as a data analyst at YouTube. On your first day, you are given the task of creating a data warehousing based analytical solution for their platform. YouTube, hosts content in the form videos hosted on channels. YouTube organizes its videos (which are contained in channels) in various entertainment categories (e.g., music, sports, news etc.). Users can access YouTube from across the world and watch videos. For each watched video YouTube tracks the following aspects of the user's visit: their location (e.g., Europe, North America, Asia etc.), the device type (e.g., smartphone, PC, laptop, TV etc.) used for the visit, watched duration of their videos, and the time of visit. YouTube would like you to design a data warehousing solution that can answer the following kind of analytical queries:

- How many users watched the "Euronews" channel in the location set to Europe.
- What is the duration that each channel was watched in country set to Norway, per month.
- What is the watch duration in Norway per week for the channels that belong to the Music category.

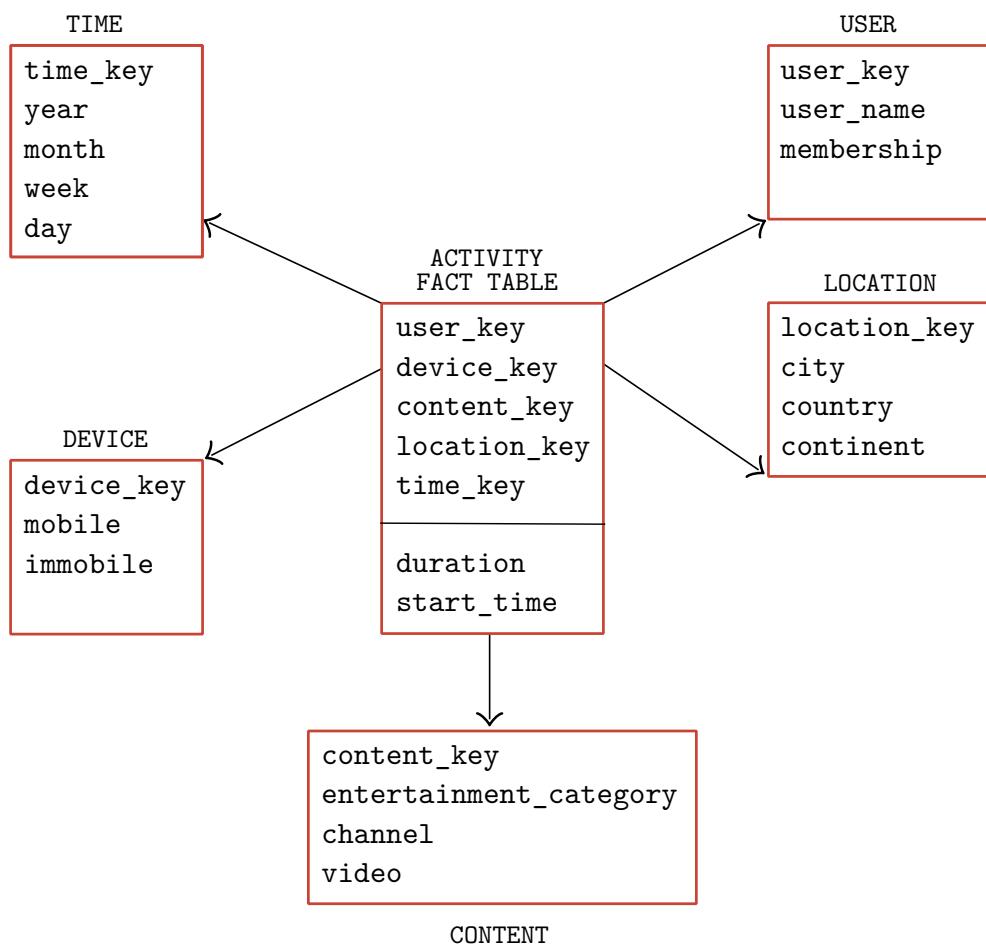
To help YouTube create their data warehouse answer the following questions. For the idea to be implementable, explain any assumptions you have made. Also, explain how the dimensions created as part of the modelling process are related to the quantities being measured and tracked.

1. Create the concept hierarchies for the different dimensions that are part of the above problem statement.
2. Create a Star schema to implement the data warehouse.
3. Specify the OLAP operations for each of the three example analytical query scenarios described above.

Solution 1

1. Concept Hierarchies.
 - a) Content: Video → Channel → Entertainment_Category → ALL.
 - b) Location: City → Country → Continent → ALL.
 - c) Time: Day → Week → Month → Year → ALL.
 - d) Device: Immobile (PC, TV) → Mobile (Laptop, Smartphone) → ALL.
 - e) User/ Membership: Free → Premium → ALL.

2. Star Schema.



3. OLAP Operations.

a)

ROLL-UP content : video → channel;
ROLL-UP location : city → continent;
DICE : channel = "Euronews" AND
: continent = "Europe";

b)

ROLL-UP location : city → country;
ROLL-UP time : day → month;
SLICE : country = "Norway"

c)

Exercise 2. Total Marks: 8 Marks

The Norwegian Directorate of Health has created a data warehouse in order to assist the vaccination of the general population. To that end, consider the vaccine dimension that is part of their data warehouse schema:

Vaccine ID	Name	Origin	Type	Doses	Storage
1	Oxford-AstraZeneca	UK	Vector	Two	Fridge
2	Pfizer-BioNTech	US	RNA	Two	Freezer
3	Sputnik V	Russia	Vector	Two	Freezer
4	BBIBP-SorV	China	Inactivated	Two	Fridge
5	Johnson & Johnson	US	Vector	One	Fridge
6	CoronaVac	China	Inactivated	Two	Fridge
7	BBV152	India	Inactivated	Two	Fridge
8	Ad5-nCoV	China	Vector	One	Fridge
9	EpiVacCorona	Russia	Subunit	Two	Fridge
10	ZF2001	China	Subunit	Two	Fridge
11	CoviVac	Russia	Inactivated	Two	Fridge

To speed up the analytics processing engine, the Directorate has created bitmap indexes over the attributes of Origin , Type , Doses , and Storage . To answer the questions below, show the bitmap indexes (with contents) that the Directorate may have created. Then, using these bitmap indexes answer the following questions:

1. Find all the vaccines that can be stored in a fridge and require two doses.
2. Find all the vaccines that do not utilize the RNA based method.
3. Find all the vaccines that can be stored in fridge and are produced by China.
4. Find all the vaccines that are vector based and are either produced by the US or India.

In order for the Directorate to understand your answers to the questions above, explain the operations that you have applied to arrive at the results.

Solution 2

- Bit Map Indexes for Origin.

	1	2	3	4	5	6	7	8	9	10	11
UK	1	0	0	0	0	0	0	0	0	0	0
US	0	1	0	0	1	0	0	0	0	0	0
Russia	0	0	1	0	0	0	0	0	1	0	1
China	0	0	0	1	0	1	0	1	0	1	0
India	0	0	0	0	0	0	1	0	0	0	0

- Bit Map Indexes for Type.

	1	2	3	4	5	6	7	8	9	10	11
Vector	1	0	1	0	1	0	0	1	0	0	0
RNA	0	1	0	0	0	0	0	0	0	0	0
Inactivated	0	0	0	1	0	1	1	0	0	0	1
Subunit	0	0	0	0	0	0	0	0	1	1	0

- Bit Map Indexes for Doses.

	1	2	3	4	5	6	7	8	9	10	11
One	0	0	0	0	1	0	0	1	0	0	0
Two	1	1	1	1	0	1	1	0	1	1	1

- Bit Map Indexes for Storage.

	1	2	3	4	5	6	7	8	9	10	11
Freezer	0	1	1	0	0	0	0	0	0	0	0
Fridge	1	0	0	1	1	1	1	1	1	1	1

- Answer to part 1.

Storage = "Fridge" AND Doses = "Two".

	1	2	3	4	5	6	7	8	9	10	11
Doses = "Two"	1	1	1	1	0	1	1	0	1	1	1
Storage = "Fridge"	1	0	0	1	1	1	1	1	1	1	1
AND	1	0	0	1	0	1	1	0	1	1	1

Answer = {1, 4, 6, 7, 9, 10, 11}

Answer = {Oxford-AstraZeneca, BBIBP-SorV, CoronaVac,
BBV152, EpiVacCorona, ZF2001, CoviVac}

- Answer to part 2.

¬(Type = "RNA").

	1	2	3	4	5	6	7	8	9	10	11
RNA	0	1	0	0	0	0	0	0	0	0	0
NOT	1	0	1	1	1	1	1	1	1	1	1

Answer = {All except 2}

Answer = {All except Pfizer BioNTech}

- Answer to part 3.

Storage = "Fridge" AND Origin = "China".

	1	2	3	4	5	6	7	8	9	10	11
Storage = "Fridge"	1	0	0	1	1	1	1	1	1	1	1
Origin = "China"	0	0	0	1	0	1	0	1	0	1	0
AND	0	0	0	1	0	1	0	1	0	1	0

Answer = {4,6,8,10}

Answer = {BBIBP-SorV, CoronaVac, Ad5-nCov, ZF2001}

- Answer to part 4.

Type = "Vector" AND (Origin = "US" OR Origin = "India").

	1	2	3	4	5	6	7	8	9	10	11
Origin = "US"	0	1	0	0	1	0	0	0	0	0	0
Origin = "India"	0	0	0	0	0	0	1	0	0	0	0
OR	0	1	0	0	1	0	1	0	0	0	0

	1	2	3	4	5	6	7	8	9	10	11
Origin = "US" OR "India"	0	1	0	0	1	0	1	0	0	0	0
Type = "Vector"	1	0	1	0	1	0	0	1	0	0	0
AND	0	0	0	0	1	0	0	0	0	0	0

Answer = {5}

Answer = {Johnson & Johnson}

2 DATA

Exercise 3. Total Marks: 5 Marks

Consider that we are trying to construct a sample of size 5 from dataset of 50 records that have identifiers numbered from 1, 2, ..., 50. What is the probability that we obtain the records with ids 30, 15, 7, 5, and 14 if:

1. The sampling is done without replacement?
2. The sampling is done with replacement?

While solving write down the key concept behind each type of sampling and the steps to arrive at the answer probabilities.

Solution 3

1. Sampling without replacement: $p = \frac{1}{50} \cdot \frac{1}{49} \cdot \frac{1}{48} \cdot \frac{1}{47} \cdot \frac{1}{46}$.
2. Sampling with replacement: $p' = \frac{1}{50} \cdot \frac{1}{50} \cdot \frac{1}{50} \cdot \frac{1}{50} \cdot \frac{1}{50}$.

Exercise 4. Total Marks: 5 Marks

Facebook allows its users to leave reactions on posts. These reactions can be from the following set: Love, Care, Haha, Wow, Sad, and Angry. Consider that we have a dataset containing reactions as an attribute. Perform binarization of this attribute for association analysis. Justify the number of attributes that you utilize to binarize.

Solution 4

Reaction	x_1	x_2	x_3	x_4	x_5	x_6
Love	1	0	0	0	0	0
Care	0	1	0	0	0	0
Haha	0	0	1	0	0	0
Wow	0	0	0	1	0	0
Sad	0	0	0	0	1	0
Angry	0	0	0	0	0	1

Six variables are required (say, as opposed to $\lceil \log_2(6) \rceil = 3$) due to requirement of asymmetric attributes for association analysis.

Exercise 5. Total Marks: 5 Marks

What is the attribute type for the following cases:

1. Years (e.g., 2014, 2015, 2016 ...).
2. Years or time is computationally recorded as UNIX epochs (i.e., number of milliseconds elapsed since 01-January-1970). Example: 1612813881000 milliseconds \equiv Monday, February 8, 2021 7:51:21 PM. What is the attribute type for UNIX epochs?
3. Consider two timestamps recorded as UNIX epochs t_1 and t_2 , where $t_2 \geq t_1$. Consider an attribute "run-time" that records values calculated from $t_2 - t_1$. What is the attribute type for "run-time"?

Solution 5

1. Interval — zero point is chosen manually and it can be shifted.
2. Interval — zero point is simply shifted; still does not correspond to an "absolute zero".
3. Ratio — recorded value is duration; ratios are meaningful.

3 ASSOCIATION RULE ANALYSIS

Exercise 6. Total Marks: 10 Marks

Compute the frequent itemsets for the transaction database given in table below using the Apriori algorithm with minimum support equal to 3. Having computed the frequent itemsets, was it necessary to scan the data database in order to determine if the final candidate 4-itemset(s) are frequent or not in this case? Explain why or why not.

tid	itemset
1	ABCD
2	ACDF
3	ACDEG
4	ABDF
5	BCG
6	DFG
7	ABG
8	CDFG

While answering the question, also write down step-by-step procedure that would entail by applying the Apriori algorithm.

Solution 6

- Candidate 1-itemsets and frequent 1-itemsets.

C ₁	Support
A	5
B	4
C	5
D	6
E	1
F	4
G	5

L ₁	Support
A	5
B	4
C	5
D	6
F	4
G	5

- Candidate 2-itemsets and frequent 2-itemsets.

C ₂	Support
AB	3
AC	3
AD	4
AF	2
AG	2
BC	2
BD	2
BF	1
BG	2
CD	4
CF	2
CG	3
DF	4
DG	3
FG	2

L ₂	Support
AB	3
AC	3
AD	4
CD	4
CG	3
DF	4
DG	3

- Candidate 3-itemsets and frequent 3-itemsets.

C ₃	Support
ABC	1
ABD	2
ACD	3
CDG	2
DFG	2

L ₃	Support
ACD	3

There is no need to scan the transaction database for frequent 4-itemsets.

Exercise 7. Total Marks: 15 Marks

Compute the frequent itemsets for the transaction database given in the table below using the FP-Growth algorithm with minimum support equal to 2.

tid	itemset
1	ABCD
2	ACDF
3	ACDEG
4	ABDF
5	BCG
6	DFG
7	ABG
8	CDFG

Show the FP-Growth procedure step-by-step including the building of the FP-Tree and the projected FP Trees.

Solution 7 • Vertical database representation for easy counting.

	1	2	3	4	5	6	7	8
A	1	1	1	1	0	0	1	0
B	1	0	0	1	1	0	1	0
C	1	1	1	0	1	0	0	1
D	1	1	1	1	0	1	0	1
E	0	0	1	0	0	0	0	0
F	0	1	0	1	0	1	0	1
G	0	0	1	0	1	1	1	1

- 1-itemset support values.

1-Itemset	Support
A	5
B	4
C	5
D	6
E	1
F	4
G	5

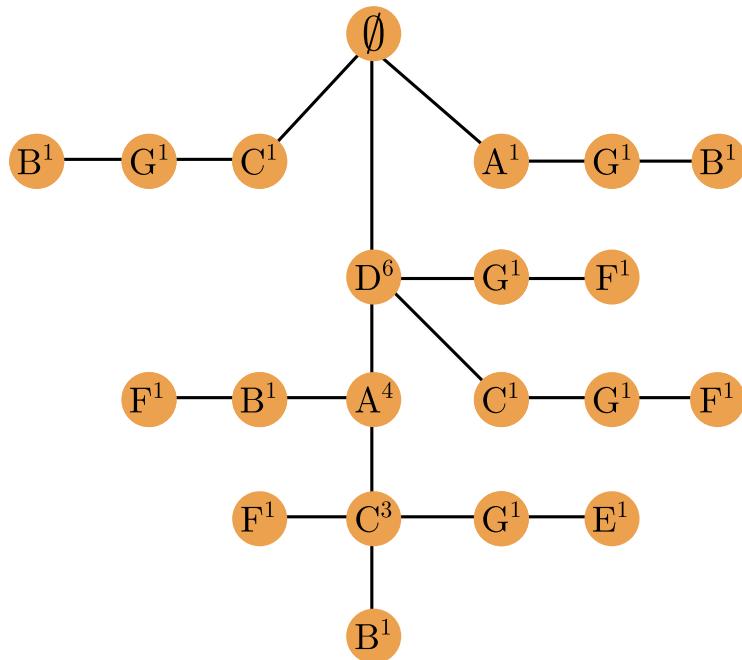
- 1-itemset reordered based on support values.

1-Itemset	Support
D	6
A	5
C	5
G	5
B	4
F	4
E	1

- Reordered transaction database.

tid	Transaction
1	DACB
2	DACF
3	DACGE
4	DABF
5	CGB
6	DGF
7	AGB
8	DCGF

- FP-Tree for the entire transaction database.



1. Project on E.

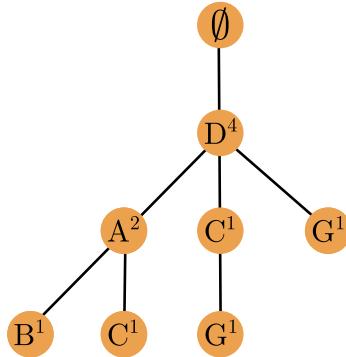
Path	Count
DACGE	1

Frequent Path = { \emptyset }

2. Project on F.

Path	Count
DCGF	1
DACF	1
DGF	1
DABF	1

– Projected FP-Tree for F.



2.1 Project on FB.

Path	Count
DAB	1

Frequent Pattern = $\{\emptyset\}$.

2.2 Project on FG.

Path	Count
DG	1
DCG	1

– Projected FP-Tree for FG: $\emptyset \rightarrow D^2 \rightarrow C^1$.

Frequent Pattern = {FG, FGD}.

2.3 Project on FC.

Path	Count
DC	1
DAC	1

– Projected FP-Tree for FC: $\emptyset \rightarrow D^2 \rightarrow A^1$.

Frequent Pattern = {FC, FCD}.

2.4 Project on FA.

Path	Count
DA	2

– Projected FP-Tree for FA: $\emptyset \rightarrow D^2$.

Frequent Pattern = {FA, FAD}.

2.5 Project on FD.

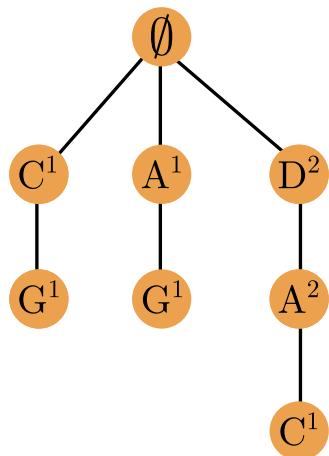
Path	Count
D	4

Frequent Pattern = {FD}.

3 Project on B.

Path	Count
AGB	1
CGB	1
DAB	1
DACB	1

– Projected FP-Tree for B.



3.1 Project on BC.

Path	Count
C	1
DAC	1

– Projected FP-Tree for BC: $\emptyset \rightarrow D^1 \rightarrow A^1$.
Frequent Patterns = {BC}.

3.2 Project on BG.

Path	Count
CG	1
AG	1

Frequent Patterns = {BG}.

3.3 Project on BD.

Path	Count
D	2

Frequent Patterns = {BD}.

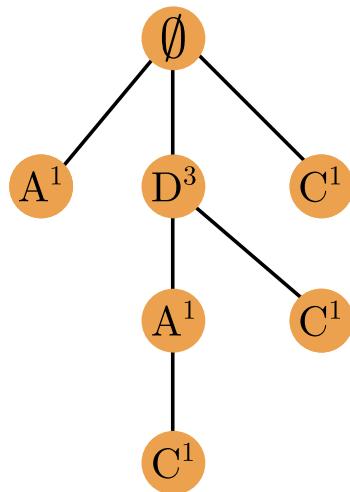
3.4 Project on BA.

Path	Count
A	1
DA	2

– Projected FP-Tree for BC: $\emptyset \rightarrow D^2$.
Frequent Patterns = {BA, BAD}.

4 Project on G.

Path	Count
CG	1
AG	1
DG	1
DCG	1
DACG	1



4.1 Project on GA.

Path	Count
A	1
DA	1

– Projected FP-Tree for BC: $\emptyset \rightarrow D^1$.
Frequent Patterns = {GA}.

4.2 Project on GC.

Path	Count
C	1
DC	1
DAC	1

– Projected FP-Tree for GC: $\emptyset \rightarrow D^2 \rightarrow A^1$.
Frequent Patterns = {GC, GCD}.

4.3 Project on GD.

Path	Count
D	3

Frequent Patterns = {GD}.

5 Project on C.

Path	Count
DAC	3
C	1
DC	1

– Projected FP-Tree for GC: $\emptyset \rightarrow D^4 \rightarrow A^3$.
Frequent Patterns = {CA, CD, CAD}.

6 Project on A.

Path	Count
DA	4
A	1

– Projected FP-Tree for A: $\emptyset \rightarrow D^4$.
Frequent Patterns = {AD}.

7 Project on D.

Path	Count
D	6

Frequent Patterns = {D}.

- Tabular summary of the FP-Growth Process.

Item	Conditional Pattern Base	Conditional FP-Tree	Frequent Itemsets
E	{D,A,C,G,E:1}	$\langle D:1, A:1, C:1, G:1, E:1 \rangle$	\emptyset
F	{D,C,G,F:1}, {D,A,C,F:1}, {D,G,F:1}, {D,A,B,F:1}	$\langle D:4, G:1 \rangle, \langle D:4, C:1, G:1 \rangle, \langle D:4, A:2, C:1 \rangle, \langle D:4, A:2, B:1 \rangle$	-
FB	{D,A,B:1}	$\langle D:1, A:1, B:1 \rangle$	\emptyset
FG	{D,G:1}, {D,C,G:1}	$\langle D:2, C:1 \rangle$	{FG, FGD}
FC	{D,C:1}, {D,A,C:1}	$\langle D:2, A:1 \rangle$	{FC, FCD}
FA	{D,A:2}	$\langle D:2 \rangle$	{FA, FAD}
FD	{D:4}	$\langle \emptyset : 4 \rangle$	{FD}
B	{AGB:1}, {CGB:1}, {DAB:1}, {DACB:1}	$\langle C:1, G:1 \rangle, \langle A:1, G:1 \rangle, \langle D:2, A:2, C:1 \rangle$	-
BC	{C:1}, {DAC:1}	$\langle D:1, A:1 \rangle$	{BC}
BG	{CG:1}, {AG:1}	$\langle C:1 \rangle, \langle A:1 \rangle$	{BG}
BD	{D:2}	$\langle \emptyset : 2 \rangle$	{BD}
BA	{A:1, DA:2}	$\langle D:2 \rangle$	{BA, BAD}
G	{CG:1}, {AG:1}, {DG:1}, {DCG:1}, {DAGC:1}	$\langle C:1 \rangle, \langle D:3, C:1 \rangle, \langle D:3, A:1, C:1 \rangle, \langle A:1 \rangle$	-
GA	{A:1}, {DA:1}	$\langle D:1 \rangle$	{GA}
GC	{C:1}, {DC:1}, {DAC:1}	$\langle D:2, A:1 \rangle$	{GC, GCD}
GD	{D:3}	$\langle D:3 \rangle$	{GD}
C	{DAC:3}, {C:1}, {DC:1}	$\langle D:4, A:3 \rangle$	{CA, CD, CAD}
A	{DA:4}, {A:1}	$\langle D:4 \rangle$	{AD}
D	{D:6}	$\langle \emptyset : 6 \rangle$	{D}

Table 1: Summarizing the FP-Growth algorithm results.

4 CLUSTERING

Exercise 8. Total Marks: 5 Marks

Apply the K-means algorithm in one-dimension for the data points: 2, 4, 10, 12, 3, 20, 30, 11, 25.

As parameters for the algorithm, take $k = 3$ and the initial centroids as follows:

$$\text{centroid}_1 = 2$$

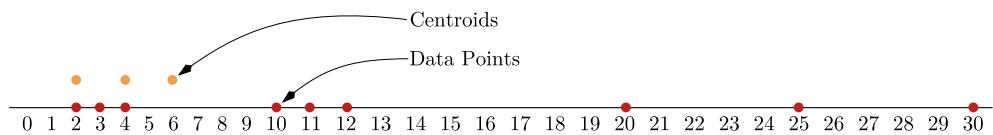
$$\text{centroid}_2 = 4$$

$$\text{centroid}_3 = 6$$

In your answer, compute the clusters after one iteration of K-means. Also compute new centroids (or means) for the next iteration of the algorithm. While writing your answers explain how the cluster memberships are decided and the new centroids (or means) calculated.

Solution 8

Dataset = {2, 4, 10, 12, 3, 20, 30, 11, 25}.



Distances of the data points to the three centroids:

D	d_1	d_2	d_3
2	0	2	4
3	1	1	3
4	2	0	2
10	8	6	4
11	9	7	5
12	10	8	6
20	18	16	14
25	23	21	19
30	28	26	24

$$C_1 = \{2, 3\}$$

$$C_2 = \{4\}$$

$$C_3 = \{10, 11, 12, 20, 25, 30\}$$

$$\text{centroid}'_1 = \frac{2+3}{2} = \frac{5}{2} = 2.5$$

$$\text{centroid}'_2 = \frac{4}{1} = 4$$

$$\text{centroid}'_3 = \frac{33+75}{6} = \frac{108}{6} = 18$$

Exercise 9. Total Marks: 20 Marks

Apply single-link hierarchical agglomerative clustering for the dataset given in Figure 1. To compute the clusters use the Manhattan distance (i.e., L_1 -norm) as the distance measure. Provide the answer in the form of a dendrogram as well as show the full distance matrix at each step. Break ties by prioritizing lexicographically smaller point labels. Terminate the clustering process when you have 4 clusters.

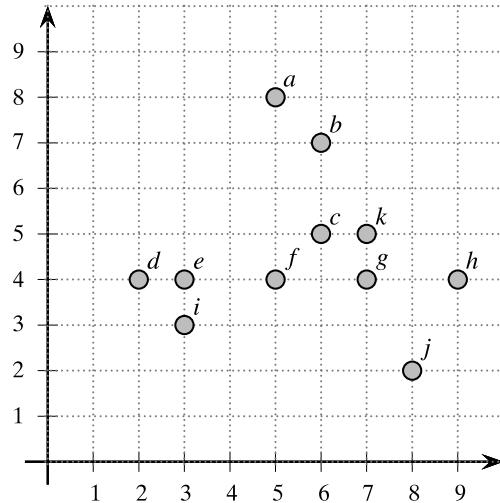


Figure 1: Figure for the hierarchical agglomerative clustering.

Solution 9

- Iteration 1:
Distance Matrix

	a	b	c	d	e	f	g	h	i	j	k
a											
b	2										
c	4	2									
d	7	7	5								
e	6	6	4	1							
f	4	4	2	3	2						
g	6	4	2	5	4	2					
h	8	6	4	7	6	4	2				
i	7	7	5	2	1	3	5	7			
j	9	7	5	8	7	5	3	3	6		
k	5	3	1	6	5	3	1	3	6	4	

- Iteration 2: Merge {c} and {k}.

	a	b	c,k	d	e	f	g	h	i	j
a										
b										
c,k	4	2								
d			5							
e			4							
f			2							
g			1							
h			3							
i			5							
j			4							

- Iteration 3: Merge {c, k} and {g}.

	a	b	c,k,g	d	e	f	h	i	j
a									
b									
c,k,g	4	2							
d			5						
e			4						
f			2						
h			2						
i			5						
j			3						

- Iteration 4: Merge {d} and {e}.

	a	b	c,k,g	d,e	f	h	i	j
a								
b								
c,k,g								
d,e	6	6	4					
f				2				
h				6				
i				1				
j				7				

- Iteration 5: Merge {d, e} and {i}.

	a	b	c,k,g	d,e,i	f	h	j
a							
b							
c,k,g							
d,e,i	6	6	4				
f				2			
h				6			
j				6			

- Iteration 6: Merge {a} and {b}.

	a,b	c,k,g	d,e,i	f	h	j
a,b						
c,k,g	2					
d,e,i	6					
f	4					
h	6					
j	7					

- Iteration 7: Merge {a, b} and {c, k, g}

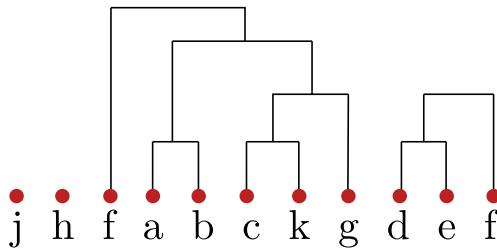
	a,b,c,k,g	d,e,i	f	h	j
a,b,c,k,g					
d,e,i	4				
f	2				
h	2				
j	3				

- Iteration 8: Merge {a, b, c, k, g} and {f}.

	a,b,c,k,g,f	d,e,i	h	j
a,b,c,k,g,f				
d,e,i	2			
h	2			
j	3			

End of algorithm as 4 clusters exist.

- Dendogram.



Note the following in the dendrogram for ease of understanding. First, the data point labels are reorganized for better visibility of merging. Second, several merge points at same distance are separated so as to clearly visualize the steps. For instance, clusters {c} and {k} and {c, k} and {g} both merge at a distance of 1 but have their branches separated for better representation.

5 CLASSIFICATION

Exercise 10. Total Marks: 15 Marks

To build decision trees, the simplest algorithm to use is the Hunt's algorithm. An important aspect of Hunt's algorithm is the selection of attributes and determination of split points. For the dataset given in Table 2, we want to identify an attribute that should be used at the root of the decision tree. To make the selection we would like to use Gini index. Determine which attribute amongst a_1 , a_2 , and a_3 is the best to split the records at the root node using Gini index measures. To make the choice compute all split points for attributes a_1 , a_2 , and a_3 .

Is it wise to utilize "Instance ID" as attribute for splitting in the decision tree construction? Why or why not?

Instance	a_1	a_2	a_3	Class
1	T	T	5.0	Y
2	T	T	7.0	Y
3	T	F	8.0	N
4	F	F	3.0	Y
5	F	T	7.0	N
6	F	T	4.0	N
7	F	F	5.0	N
8	T	F	6.0	Y
9	F	T	1.0	N

Table 2: Table for decision tree based exercise.

Solution 10 • Split on a_1 :

$a_1 = \text{TRUE}$	
Y	3
N	1

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 \\ &= 1 - \frac{9}{16} - \frac{1}{16} \\ &= \frac{16 - 10}{16} = \frac{6}{16} \\ &= 0.375. \end{aligned}$$

$a_1 = \text{FALSE}$	
Y	1
N	4

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 \\ &= 1 - \frac{1}{25} - \frac{16}{25} \\ &= \frac{25 - 17}{25} = \frac{8}{25} \\ &= 0.32. \end{aligned}$$

$$\begin{aligned} \text{Gini} &= \frac{4}{9} \cdot \frac{6}{16} + \frac{5}{9} \cdot \frac{8}{25} \\ &= \frac{1}{6} - \frac{8}{45} \\ &= 0.34. \end{aligned}$$

- Split on a_2 :

$a_2 = \text{TRUE}$	
Y	2
N	3

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 \\ &= 1 - \frac{4}{25} - \frac{9}{25} \\ &= \frac{25-13}{25} = \frac{12}{25} \\ &= 0.48. \end{aligned}$$

$a_2 = \text{FALSE}$	
Y	2
N	2

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 \\ &= 1 - \frac{1}{4} - \frac{1}{4} \\ &= \frac{2-1}{2} = \frac{1}{2} \\ &= 0.5. \end{aligned}$$

$$\begin{aligned} \text{Gini} &= \frac{5}{9} \cdot \frac{12}{25} + \frac{4}{9} \cdot \frac{1}{2} \\ &= \frac{4}{15} + \frac{2}{9} = \frac{22}{45} \\ &= 0.48. \end{aligned}$$

- Split on a_3 :

	N	Y	N	N/Y	Y	N/Y	N			
Data Point	1	3	4	5	6	7	8			
Split Point	$\frac{1}{2}$	2	3.5	4.5	5.5	6.5	7.5	9		
	\leq	$>$	\leq	$>$	\leq	$>$	\leq	$>$	\leq	$>$
Class='Y'	0	4	0	4	1	3	1	3	2	2
Class='N'	4	5	1	4	1	4	2	3	3	2
Gini Index	0.49	0.4	0.4921	0.481	0.48	0.481	0.4	0.49		

– Split on comparing a_3 with $\frac{1}{2}$:

$a_3 \leq \frac{1}{2}$	
Y	0
N	0

$$\text{Gini} = 1 - 0 - 0 = 1.$$

$a_3 > \frac{1}{2}$	
Y	4
N	5

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2 \\ &= 1 - \frac{16}{81} - \frac{25}{81} \\ &= \frac{81-16-25}{81} = \frac{40}{81} \\ &= 0.4938. \end{aligned}$$

$$\begin{aligned} \text{Gini} &= \frac{0}{9} \cdot 1 + \frac{9}{9} \cdot \frac{40}{81} \\ &= 0.4938. \end{aligned}$$

– Split on comparing a_3 with 2:

$a_3 \leq 2$	
Y	0
N	1

$$\text{Gini} = 1 - 1 = 0.$$

$a_3 > 2$	
Y	4
N	4

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{4}{8}\right)^2 - \left(\frac{4}{8}\right)^2 \\ &= 1 - 2 \cdot \frac{1}{4} \\ &= \frac{1}{2} \\ &= 0.5. \end{aligned}$$

$$\begin{aligned} \text{Gini} &= \frac{1}{9} \cdot 0 + \frac{8}{9} \cdot \frac{1}{2} \\ &= \frac{4}{9} \\ &= 0.4. \end{aligned}$$

– Split on comparing a_3 with 3.5:

$a_3 \leq 3.5$	
Y	1
N	1

$$\text{Gini} = 1 - \frac{1}{2} = \frac{1}{2}.$$

$a_3 > 3.5$	
Y	3
N	4

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 \\ &= \frac{49 - 9 - 16}{49} \\ &= \frac{24}{49} \\ &= 0.4897. \end{aligned}$$

$$\begin{aligned} \text{Gini} &= \frac{2}{9} \cdot \frac{1}{2} + \frac{7}{9} \cdot \frac{24}{49} \\ &= \frac{1}{9} + \frac{8}{21} = \frac{31}{63} \\ &= 0.4921. \end{aligned}$$

– Split on comparing a_3 with 4.5:

$a_3 \leq 4.5$	
Y	1
N	2

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 \\ &= 1 - \frac{5}{9} \\ &= \frac{4}{9} \\ &= 0.4. \end{aligned}$$

$a_3 > 4.5$	
Y	3
N	3

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 \\ &= 1 - 2 \cdot \frac{1}{4} \\ &= \frac{1}{2} \\ &= 0.5. \end{aligned}$$

$$\begin{aligned} \text{Gini} &= \frac{3}{9} \cdot \frac{4}{9} + \frac{6}{9} \cdot \frac{1}{2} \\ &= \frac{4}{3 \cdot 9} + \frac{3^2}{3 \cdot 9} = \frac{4+9}{27} = \frac{13}{27} \\ &= 0.481. \end{aligned}$$

– Split on comparing a_3 with 5.5:

$a_3 \leq 5.5$	
Y	2
N	3

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 \\ &= 1 - \frac{13}{25} \\ &= \frac{12}{25} \\ &= 0.48. \end{aligned}$$

$a_3 > 5.5$	
Y	2
N	2

$$\begin{aligned} \text{Gini} &= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 \\ &= 1 - 2 \cdot \frac{1}{4} \\ &= \frac{1}{2} \\ &= 0.5. \end{aligned}$$

$$\begin{aligned} \text{Gini} &= \frac{5}{9} \cdot \frac{12}{25} + \frac{4}{9} \cdot \frac{1}{2} \\ &= \frac{4}{3 \cdot 5} + \frac{2}{9} = \frac{4 \cdot 9 + 2 \cdot 3 \cdot 5}{3 \cdot 5 \cdot 9} = \frac{36 + 30}{3 \cdot 5 \cdot 9} = \frac{22}{45} \\ &= 0.48. \end{aligned}$$

Repetition of calculations from here.

Split on a_1 as it has the minimum Gini index.

- It is not wise to utilize Instance ID as a test / splitting attribute. Splitting on the Instance ID will lead to leaf nodes equal to the number of data points each having Gini index of value 0. However, when evaluating a new test instance the split attribute is useless as it will not have seen the new ID.

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 73596755

Eksamensdato: 9. august 2016

Eksamensstid (fra-til): 09.00-13.00

Hjelpekode/Tillatte hjelpeemidler: D: Ingen trykte eller håndskrevne hjelpeemiddel tillatt.
Bestemt, enkel kalkulator tillatt.

Annен informasjon:

Målform/språk: Bokmål

Antall sider (uten forside): 2

Antall sider vedlegg: 0

Kontrollert av:

Dato

Sign

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig 2-sidig

sort/hvit farger

Oppgave 1 – Diverse – 20 % (alle deler teller likt)

- Hva er en *outlier*?
- Forklar *web-bruk-gruvedrift* (Web usage mining), hva som er målet, og hva som er typiske data man bruker i denne prosessen.
- Forklar hvordan man kan gjøre kontinuerlige attributtverdier om til kategorier.
- Anta to bit-vektorer p og q :

$$p = 1010100111$$

$$q = 1000101101$$

Regn ut Jaccard-koeffisienten for bitvektorene p og q .

Oppgave 2 – Datavarehus og OLAP – 30 % (alle deler teller likt)

- Forklar hva som menes med ordene som er understreket i følgende definisjon av datavarehus: "A data warehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data."
- Forklar *enterprise-varehus*, *data mart*, og *virtuelle varehus*.
- Forklar *stjerne-skjema* og *snøflak-skjema*.
- Forklar *konsept-hierarki*.
- Forklar *materialisering av kuboider*, hensikten med materialisering, og hvordan man kan velge hvilke kuboider som skal materialiseres.

Oppgave 3 – Klynging – 15 % (alle deler teller likt)

- Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør hierarkisk agglomerativ klynging på dette datasettet ved å bruke MIN (single link) og Manhattan-distanse. Vis det resulterende dendrogrammet.
- Forklar "Bisecting K-means". Hva er en viktig fordel med denne sammenlignet med ordinær K-means?

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

Oppgave 4 – Klassifisering – 10 % (alle deler teller likt)

- a) Forklar *klassifisering*.
- b) Forklar hvordan man konstruerer et *beslutningstre*.

Oppgave 5 – Assosiasjonsregler – 25 % (alle deler teller likt)

- a) Anta handlekorg-data som er gitt under. Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Bruk $F_{k-1} \times F_{k-1}$ -metoden for kandidat-generering.

TransaksjonsID	Element
T1	ABGH
T2	ADGHK
T3	ABC
T4	ACD
T5	ACGHK
T6	ACGHK
T7	BD
T8	ADGHK

- b) Konstruer et FP-tre basert på datasettet ovenfor.
- c) Vis hvordan man kan bruke ECLAT for å finne støtte (-tall) for elementsettet AG.

LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG

TDT4300 – AUGUST 2016

NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element, hovudsakleg laga for at vi skal ha oversikt over arbeidsmengda på eksamen, og som huskeliste under sensurering (i og med at dette er kontinuasjonseksamen er den heller ikkje fullstendig, for dei to "grafiske" oppgåvene ser vi lett om studentane har rett eller ikkje :). Det er også viktig å være klar over at det også kan vere andre svar enn dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving eller dei gjev meining utifrå kva det vert spurd om.

Oppgåve 1 – Diverse

- a) "Outliers" er dataobjekter som er signifikant forskjellige fra de fleste andre objekter i datasettet.
- b) **Web-bruk-gruvedrift** (Web usage mining): automatisk oppdaging av mønstre i *clickstreams* og tilhøyrande data som har vorte samla inn eller generert som følgje av brukarar sine interaksjonar med ein eller fleire nettstader
Mål: Analysere åferdsmønster og profilar av brukarar i interaksjon med ein nettstad
Data i web-bruk-gruvedrift:
 - Loggar frå web-tenrarar
 - Innhold frå nettstadane
 - Data om brukarar frå eksterne kjelder
 - Andre applikasjonsdata
- c) I hovedsak å bestemme antallet kategorier (n) og finne split-points ($n-1$)
Et sett av intervaller som mappes til kategorier
Unsupervised :
 - Tar ikke hensyn til klassebetegnelse
 - Kan være brukerdefinerte intervaller, eller basert på klyngeanalyser (eksempel K-means)
 - Lik fordeling eller lik frekvens
- d) $M11/(M11+M01+M10)=4/(4+1+2)=4/7=0.57$

Oppgåve 2 – Datavarehus og OLAP

- a) Sjå læreboka.
- b) Enterprise-varehus
 - Inneholder data om alle emner, for hele virksomheten
 - Data mart
 - Hver mart (marked) er et subsett av virksomhetens totale varehus, og inneholder bare det som er av interesse for en subgruppe, underavdeling etc.
 - Virtuelle varehus
 - Verktøy som gir datavarehus data/funksjonalitet på toppen av operasjonelle data (relasjons-databaser) vha. views
 - Kan være begrenset hva som kan gjøres (begrensninger på hvilke views som kan være materialisert)
 - Avhengig av overskudds-kapasitet på operasjonelle database-tjenere
- c) Sjå læreboka.
- d) Sjå læreboka.
- e) Sjå læreboka.

Oppgåve 3 – Klynging

- a) ...

```
1: Initialize the list of clusters to contain the cluster containing all points.  
2: repeat  
3:   Select a cluster from the list of clusters  
4:   for  $i = 1$  to number_of_iterations do  
5:     Bisect the selected cluster using basic K-means  
6:   end for  
7:   Add the two clusters from the bisection with the lowest SSE to the list of clusters.  
8: until Until the list of clusters contains  $K$  clusters
```

- b) Viktig fordel: Mindre følsom for initialisering

Oppgåve 4 – Klassifisering

- a) Sjå læreboka.
- b) Sjå læreboka.

Oppgåve 5 – Assosiasjonsreglar

- a) C1: a7,b3,c4,d4,g5,h5,k4
F1: a,c,d,g,h,k
C2: ac4,ad3,ag5,ah5,ak4,cd1,cg2,ch2,ck2,dg2,dh2,dk2, gh5,gk4,hk4
F2:ac,ag,ah,ak,gh,gk,hk
C3:acg*,ach*,ack*,agh5,agk4,ahk4,ghk4
F3:agh,agk,ahk,ghk
C4: aghk4
F4: aghk

*: Inneheld ikkje-frekvente subsett og kan prunast umiddelbart

- b) Forventa at ein har sortert 1-elementsett og transaksjonar med re-ordna element, og deretter sett inn i FP-tre.
- c) A:T1,T2,T3,T4,T5,T6,T8
B: T1,T3,T7
...
G: T1,T2,T5,T6,T8
...

Finn støttetal for AG ved å utføre snitt på listene for A og G som gir T1,T2,T5,T6,T8, resultatet har 5 element dvs. støttetal 5 (støtte 5/8).

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 73596755

Eksamensdato: 12. august 2017

Eksamensstid (fra-til): 09.00-13.00

Hjelpemiddelkode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrevne hjelpemiddel tillatt.
Bestemt, enkel kalkulator tillatt.

Annен informasjon:

Målform/språk: Bokmål

Antall sider (uten forside): 2

Antall sider vedlegg: 0

Kontrollert av:

17.07.2017

Jon Olav Hauglid

Dato

Sign

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig 2-sidig

sort/hvit farger

Oppgave 1 – Diverse – 20 % (alle deler teller likt)

- a) I kontekst av web-bruk-gruvedrift, hva er sesjonering? Hvorfor kan dette være vanskelig? Forklar to heuristikker som kan brukes til å utføre sesjonering.
- b) Forklar *støy* ("noise") og *outlier*.
- c) *Curse of Dimensionality* kan føre til problem når man skal utføre klynging eller klassifisering på høy-dimensjonale datasett. Forklar minst to data-preprosessering-metoder som kan redusere problemene.
- d) Forklar hvordan en likhetsmatrise ("similarity matrix") kan brukes til klyngingsvalidering.

Oppgave 2 – OLAP – 25 % (alle deler teller likt)

- a) Forklar hva som menes med termene som er understreket i følgende definisjon av datavarehus: "data-warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data."
- b) Forklar begrepene OLTP ("Online Transaction Processing") og OLAP ("Online Analytical Processing"). Legg vekt på å få frem forskjeller mhp. egenskaper og bruk.
- c) Forklar *stjerne-skjema* og *snøflak-skjema*.
- d) Forklar bitmap-indeks. I hvilke tilfeller er en bitmap-indeks egnet?
- e) Forklar OLAP-operasjonene *slice* og *dice*.

Oppgave 3 – Klynging – 10 %

Forklar algoritmen for *DBSCAN*.

Oppgave 4 – Klassifisering – 15 % (5 % på a og 10 % på b)

- a) Forklar *kryss-validering* ("cross validation").
- b) Forklar *overtilpasning* ("overfitting"). Hva kan forårsake overtilpasning? Hva kan man gjøre for å redusere overtilpassing når man bruker beslutningstre?

Oppgave 5 – Assosiasjonsregler – 30 % (5 % på a, 10 % på b og 15 % på c)

a) Definer *maksimale* ("maximal") og *lukkede* ("closed") frekvente elementsett.

b) Anta handlekorg-data som er gitt under:

TransaksjonsID Element

T1	ACD
T2	BCE
T3	ABCE
T4	BE

- 1) Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 2). Bruk $F_{k-1} \times F_{k-1}$ -metoden for kandidat-generering.
- 2) Bruk apriori-algoritmen til å generere alle 3-elements assosiasjonsregler basert på resultatet i (1), gitt minimum konfidens på 100 %. Vis hvordan regler evt. kan "prunes".

c) Anta handlekorg-data som er gitt under:

TransaksjonsID Element

T1	ABCD
T2	ABCD
T3	ABCEF
T4	ABEF
T5	ABDEFG
T6	AEF

Du skal nå bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støttetall (*minimum support count*) på 2.

- 1) Konstruer et FP-tre basert på datasettet.
- 2) Finn frekvente elementsett ved å bruke FP-growth-algoritmen. Bruk tabell-notasjon med følgende kolonner for å vise resultatet:
 - Element
 - "Conditional pattern base"
 - "Conditional FP-tree"
 - Frekvente elementsettForklar rekursivitet der dette er nødvendig.

LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – AUGUST 2017

NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element, hovudsakleg laga for at vi skal ha oversikt over arbeidsmengda på eksamen, og som huskeliste under sensurering. Det er også viktig å være klar over at det også kan vere andre svar enn dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving eller dei gjev meining utifrå kva det vert spurd om.

Oppgave 1

- a) 1) Identifisere sesjonar, dvs. *aktivitetar utført av ein brukar frå ho/han først aksesserer nettstaden og til han/ho forlet nettstaden.*
2) Vanskeleg å få pålitelege data pga.:
 - Proxy-tenrarar og anonymisatorar
 - Dynamiske IP adresser
 - Manglende referansar pga. caching
 - Cookies kan koplast ut3) Teknikkar: tidsbasert og referent-basert, sjå læreboka.
- b) Støy er tilfeldige unøyaktigheter/feil i målingene.
“Outliers” er dataobjekter som er signifikant forskjellige frå dei fleste andre objekt i datasettet.
- c) Forklar minst to av følgjande (jfr. læreboka):
Dimensjonsreduksjon
Feature subset selection
Feature creation
- d) Jfr. læreboka.

Oppgave 2

- a) Jfr. læreboka.
b) Jfr. læreboka.
c) Jfr. læreboka.
d) Jfr. læreboka. Eigna for attributtar med få distinkte verdiar.
e) Slicing er seleksjon på ein dimensjon slik at vi får ei subkube.
Dicing er å velge ut ei subkube med to eller fleire dimensjonar.

Oppgave 3

Jfr. algoritme 8.4 i læreboka. Kjerne/grense/støy-punkt, Eps og MinPts må forklarast, i tillegg til korleis sjølve klynginga vert gjort med desse som utgangspunkt.

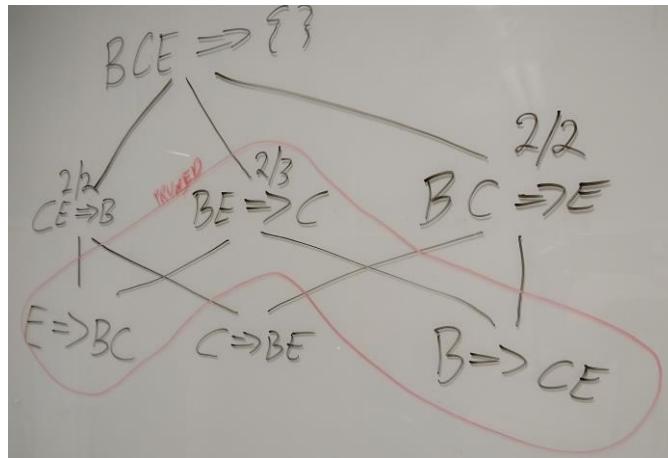
Oppgave 4

- a) Partisjonerer data i k disjunkte subsett.
k-fold: tren på k-1 partisjonar, teste på den gjenverande, etc.
Metrikk er gjennomsnittleg effektivitet
- b) Overtilpasning (overfitting): om modellen er for komplisert (altfor tilpassa treningsdata).
Pga støy, manglende representative eksempel.
pre-pruning og post-pruning, jfr. læreboka

Oppgåve 5

- a) Elementsett er maksimalt frekvente om ingen av dei umiddelbare supersetta er frekvente. Eit elementsett er lukka om ingen av dei umiddelbare supersetta har same støtte som elementsettet.
- b) Vesentleg poeng at join er brukt for å generere BCA og at ein skal bruke apriori-metoden (som spesifisert i oppgaven), og ikkje berre basere seg på brute-force (teste mot alle mogleg variantar).

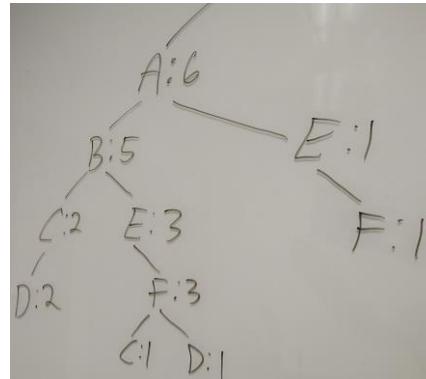
A	2
B	3
C	3
D	1
E	3
AB	1
AC	2
AE	1
BC	2
BE	3
CE	2
BCE	2



- c) Støttetal: A:6, B:5, C:3, D:3, E:4, F:4, G:1

Viktig at G ikke er med i trelet og at rekursivitet er vist/forstått.

tid	Itemset	(Ordered) frequent items
T1	ABCD	ABCD
T2	ABCD	ABCD
T3	ABCEF	ABEFC
T4	ABEF	ABEF
T5	ABDEFG	ABEFD
T6	AEF	AEF



Item	Conditional sub-database	Conditional FP-tree	Frequent itemsets
D	{(ABC:2), (ABEF:1)}	A3B3C2	D:3, AD:3, BD:3, CD:2, ABD:3, ACD:2, BCD:2, ABCD:2
C	{(AB:2), (ABEF:1)}	AB3	C:3, AC:3, BC:3, ABC:3
F	{(ABE:3), (AE:1)}	A4(BE:3,E:1)	F:4, EF:4, BF:3, AF:4
EF	{(AB:3), (A:1)}	A:4B:3	AFE:4, BFE:3, ABFE:3
BF	A:3	A:3	FBA:3
E	{(AB:3), (A:1)}	A:4B:3	E:4, AE:4, BE:3, ABE:3
B	{(A:5)}	A:5	B:5, AB:5
A	{(null)}	{(null)}	A:6

(OK om 1-elementsett ikkje er med i tabellen om sagt tidlegare kva som er frekvente element, eller ved figur har vist at dette er forstått).

i Cover Page

Department of Computer Science
Examination paper for TDT4300 Data warehousing and data mining

Academic contact during examination:
Tárik Saleh Salem, phone: 948 58 617

Examination date: 08-08-2019
Examination time (from-to): 09.00-13.00
Permitted examination support material: D: No tools allowed except an 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.

1 Attribute Type (3 marks)

Which type of attribute is Fahrenheit temperature?

Select one alternative:

- Ratio
- Interval
- Nominal
- Ordinal

Maximum marks: 3

2 Dimensionality reduction (3 marks)

What are the purposes of dimensionality reduction?

Select one or more alternatives:

- Allow data to be more easily visualized
- May help to eliminate irrelevant features or reduce noise
- Remove outliers in data
- Reduce amount of time and memory required by data mining algorithms

Maximum marks: 3

3 Jaccard coefficient (3 marks)

There are two bit vectors **p** and **q**:

$$\mathbf{p} = [1110100011]$$

$$\mathbf{q} = [1101001011]$$

What is the Jaccard coefficient for the bit vectors **p** and **q**? Write your answer here .

Note: the answer is a real-valued number.

Maximum marks: 3

4 Cosine similarity (3 marks)

There are two vectors **p** and **q**:

$$\mathbf{p} = [3, 1, 5]$$

$$\mathbf{q} = [6, 7, 2]$$

What is the cosine similarity between **p** and **q**? Write your answer here .

Note the answer is a real-valued number (up to two decimals).

Maximum marks: 3

5 Modeling (10 marks)

Design the data warehouse for an electronics company. The data warehouse has to allow to analyze the company's situation at least with respect to the Product, Customers and Time. Moreover, the company needs to analyze:

- the product with respect to its name and type (office, home, accessories, wearable, photography, networking, etc.);
- the customers with respect to their spatial location, by considering at least cities, regions and states.

The company is interested in learning at least the quantity, income and discount of its sales.

One should be able to perform the following example analysis against the data warehouse:

- Total quantity for each year.
- Total quantity for each month.
- Average income for every day for each product type.

Create a star schema for the described case **and define a concept hierarchy for each dimension.**

Note: You have two options for drawing: 1) use a separate paper that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Words: 0

B I U \times_1 \times^2 | \mathcal{I}_x | | | | Ω | | Σ |

Maximum marks: 10

6 OLAP (10 marks)

Given a cube with dimensions:

- Time(Day-Month-Year)
- Item(ItemName-Brand)
- Location(Street-City-ProvinceOrState-Country)

Assume the following materialized cuboids:

- {Month, ItemName, City}
- {Month, Brand, Country}
- {Year, Brand, ProvinceOrState}
- {ItemName, City} where year = 2016

Given the following OLAP query: {Brand, City} with condition Month = June 2016, which cuboid(s) should be used? Explain your answer below.

Fill in your answer here

Format ABC

 | B I U x_1 x_2 x^2 | I_x | □ □ | ↶ ↷ ↶ | ⋮⋮ | Ω grid | ✎ | Σ | ☒

Words: 0

Maximum marks: 10

7 Apriori Algorithm (15 marks)

Assume the market basket data below. Use the Apriori algorithm to find all frequent itemsets with minimum support **33.33%** (i.e. minimum support count is 2).

Transaction ID	Items
T1	a, b
T2	a, b, c
T3	a, d, e
T4	a, d, e, f
T5	c, e
T6	d, e

1. Show how the **frequent itemsets** are generated.
2. $\{a, d, e\}$ is one of the frequent itemsets. Find all association rules based on this set, given confidence threshold **60%** (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 $\{a, d, e\}$).

Note: you have two options for drawing: 1) use a separate paper that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | **B** *I* U \times_2 \times^2 | \mathbb{T}_x | | | \approx $\approx\approx$ | Ω | | Σ | ABC |

Words: 0

Maximum marks: 15

8 FP-growth Algorithm (15 marks)

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 22% (i.e., minimum support count is 2).

Transaction ID	Items
T1	D, E, H
T2	D, H
T3	D, F, G, H
T4	B, G, H
T5	C, G, H
T6	D, G, I
T7	D, F, G, H
T8	G, H, J
T9	A, D, F, G, H

- 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

Note: you have two options for drawing: 1) use a separate paper that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | B I U x₁ x² | T_x | D L | ← → ⟳ | ≡ :: | Ω ■ | ↶ ↷ | Σ | ABC | X

Words: 0

Maximum marks: 15

9 K-means Clustering (15 marks)

Do three iterations of the Lloyd's algorithm for K-means clustering on the 2-dimensional data below, using Euclidean distances. Use $K = 2$ clusters and the initial prototype vectors (i.e. mean vectors) $\mathbf{m}_1 = (3.0, 5, 0)$, $\mathbf{m}_2 = (2.0, 1.0)$. Write down calculation procedure and the cluster memberships as well as mean vectors after each iteration. Draw the data points, cluster means and cluster boundary after each iteration.

t	$\mathbf{x}^{(t)}$
1	(5.0, 2.0)
2	(4.0, 1.0)
3	(2.0, 4.0)
4	(1.0, 3.0)
5	(0.0, 4.0)

Note: you have two options for drawing: 1) use a separate paper, that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | **B** *I* U \mathbf{x}_1 \mathbf{x}_2 \mathbf{x}^2 | $\mathbf{I}_{\mathbf{x}}$ | | \vdash $\vdash \vdash$ | Ω | | Σ | ABC |

Words: 0

Maximum marks: 15

10 Hierarchical clustering pros and cons (3 marks)

When does single linkage hierarchical clustering probably not perform well?

Select one or more alternatives:

- Data contains outliers.
- Clusters have non-elliptical shapes
- Data is high-dimensional.
- There are large amounts of data points

Maximum marks: 3

11 Cross validation (5 marks)

Explain cross validation and what this technique is used for.

Note: if needed, you have two options for drawing: 1) use a separate paper, that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | **B** *I* U \times_1 \times^2 | \int_x | | | $\frac{1}{z}$ $\frac{a}{z}$ | Ω | | Σ | ABC |

Words: 0

Maximum marks: 5

12 Decision tree (15 marks)

You are going to predict whether mushrooms are edible. You have the following data:

Example	Smooth	Spotted	Smelly	NotHeavy	Edible
A	1	1	0	1	0
B	1	0	0	0	0
C	1	0	1	1	0

D	0	0	1	0	0
E	0	1	1	1	0
F	1	0	1	0	1
G	0	0	0	1	1
H	0	1	0	1	1
U	0	0	1	1	?
V	1	1	1	0	?
W	1	0	1	1	?

For mushrooms A through H, you know whether it is edible (1) or not edible (0), but you do not know about U through W.

You should use decision tree as a classification method. You will use the examples A through H as the training data. To decide the best split, you need to use **Entropy** for a node t , given by

$\text{Entropy}(t) = - \sum_j p(j|t) \log_2 p(j|t)$, where $p(j|t)$ is the probability for class j given node t (i.e. the portion of class j in the node t). For each split, the "information gain" is defined by

$\text{GAIN} = \text{Entropy}(p) - \left(\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) \right)$, where n_i is the number of element in node i and n is the total number of elements in the parent node p .

For Tasks 1 and 2, consider only mushrooms A through H. Tasks:

1. Which attribute should you choose as the root of a decision tree? Justify your choice by calculating the information gains of the attributes.
2. Build an ID3 decision tree to classify mushrooms as edible or not.
3. Classify mushroom U, V, and W using the decision tree to be edible or not edible.

Note: if needed, you have two options for drawing: 1) use a separate paper that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | **B** *I* U x_1 x_2 x^2 | Σ | Text Image | Left Right Center | Horizontal Vertical | Equation | Grid | Pen | Text | Sum | ABC | X

Words: 0

Maximum marks: 15

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagrivedrift

Faglig kontakt under eksamen: Heri Ramampiaro

Tlf.: 7359 3440

Eksamensdato: 7. august

Eksamenstid (fra-til): 09.00-13.00

Hjelpekode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrivne hjelpemiddel tillatt. Bestemt, enkel kalkulator tillatt.

Annен informasjon:

Målform/språk: Bokmål

Antall sider: 3

Antall sider vedlegg: 0

Kontrollert av:

Dato

Sign

Oppgave 1 – Diverse – 15 % (alle deler teller likt)

- a) Forklar hensikt og teknikker for diskretisering og dimensjonalitetsreduksjon.
- b) Forklar prinsippene bak join-indekser.
- c) Gitt settet med verdier under, finn 10 %, 25 %, 50 %, 75 % og 100 % persentiler
(10th,25th,50th,75th and 100th percentiles)
 Verdier: 1,2,3,5,5,7,7,10,12,15,18,18,19,20,22,23,24,27,28,32

Oppgave 2 – Modellering – 20 % (15 % på a, 5 % på b)

I denne oppgaven ser vi på en kjede av butikker som selger aviser og blader. Kjeden selger mange forskjellige typer publikasjoner (for eksempel knyttet til mote, barn, biler, sport) fra mange forskjellige utgivere. Butikkene spenner fra små nærbutikker til større butikker med samlokaliserte kafeer. Imidlertid er kjeden litt gammeldags, der hver butikksjef på slutten av hver dag legger inn informasjon i et regneark om hvor mange eksemplarer som ble solgt av hver publikasjon. Dette regnearket blir så sendt til hovedkvarteret, og for tiden er dette den eneste måten hovedkvarter kan samle inn og analysere salgsdata fra butikkene. Ledelsen ønsker nå å lage et datavarehus for å få mer innsikt i salget av de ulike publikasjonene (og typer publikasjoner) fra hver butikk.

Beskrivelsen er litt upresist formulert og det er en del av oppgaven å velge ut hva som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

- a) Lag et snøflak-skjema for denne case-beskrivelsen.
- b) Lag to forskjellige konsepthierarkier (fritt valgte dimensjoner).

Oppgave 3 – Klassifisering – 20 % (8 % på a og b, 4 % på c)

- a) Forklar hvordan man kan bestemme hvilke attributter man skal splitte på når man lager et beslutningstre.
- b) Forklar hvordan man kan splitte på kontinuerlige attributter.
- c) Forklar metrikkene *Accuracy* og *Error rate* i forbindelse med klassifisering.

Oppgave 4 – Klynging – 25 % (5 % på a, 10 % på b og c)

- a) Silhuett-koeffisienten kan beskrives som $s = (b-a)/\max(a,b)$. Forklar innholdet i formelen, og hva Silhuett-koeffisienten kan brukes til.
- b) Forklar algoritmen for *bisecting k-means*. Hva er fordelen med denne framfor vanlig *k-means*?
- c) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør klynging ved hjelp av hierarkisk klynging på dette datasettet. Bruk MIN (single link) som interklynende distanse.

X	Y
2	2
3	2
4	8
5	4
5	7
7	4
9	17
13	4
17	4

Oppgave 5 – Assosiasjonsregler – 20 %

Anta handlekurv-data til høyre. Bruk apriori-algoritmen for å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Finn alle assosiasjonsregler basert på dette settet, gitt konfidens på 75 %.

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

Norges teknisk-naturvitenskapelige universitet
Institutt for datateknikk og informasjonsvitenskap



EKSAMENSOPPGAVE I FAG TDT4300 – DATAVAREHUS OG DATAGRUVEDRIFT

Faglig kontakt under eksamen: Kjetil Nørvåg og Trond Aalberg

Tlf.: 41440433/97631088

Eksamensdato: 9. juni 2012

Eksamensstid: 09.00-13.00

Tillatte hjelpe middel: D: Ingen trykte eller håndskrivne hjelpe middel tillatt. Bestemt, enkel kalkulator tillatt.

Språkform: Bokmål

Sensurdato: 30. juni 2012

Oppgave 1 – Datavarehus – 20% (alle deler teller likt)

- Forklar begrepene OLTP (Online Transaction Processing) og OLAP (Online Analytical Processing). Legg vekt på å få fram hva som er forskjellig mellom systemer for det ene eller det andre mht. egenskaper og bruk.
- Forklar hva som menes med ordene som er understrekket i følgende definisjon av datavarehus: "A datawarehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data."
- Forklar begrepet datakube og hva som menes med cuboids.
- Gitt en datakube med 5 dimensjoner, hvor mange cuboids kan vi generere fra denne (eller si at den inneholder)?
- Hva menes med datakube-operasjonene slice, dice, rollup og drill-down?

Oppgave 2 – Modellering – 20%

Du skal lage et datavarehus over trafikkulykker i Norge for å kunne undersøke hvilke veistrekninger som det er mest samfunnsnyttig å utbedre eller sette ned farten på etc. Vi skal bare se på direkte kostnader ved ulykkene og ikke bry oss om personskader etc. Datagrunnlaget kommer fra forskjellige forsikringsselskap og inneholder:

- hvor (dato) og når ulykken skjedde (gate og by, eller for eksempel veistrekning og fylke).
- data om fører (vi er mest interessert i alder til fører og om vedkommende var beruset eller ikke)
- type forsikring på bilen og forsikringsutbetalingen.

Datagrunnlaget er litt upresist formulert og det er en del av oppgaven å velge ut det som er mulig å få med, eller tenke ut en måte å uttrykke det som fakta om ulykkene. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

- Lag et stjerne- eller snøflak-skjema for denne case-beskrivelsen.
- Lag to forskjellige konsept hierarkier (fritt valgte dimensjoner).
- Skriv et eksempel-query i mdx hvor du genererer en todimensjonal tabell (som gir et to-dimensjonalt svarsett) som viser forsikringsutbetalinger for månedene i 2011 for forskjellige fylker (her antar vi at du har konsept hierarkier som lar deg bruke disse kategoriene).

Oppgave 3 – Klynging – 30 % (10% på a og b, 5% på c og d)

- a) Forklar hierarkisk klynging, og forskjellen på MIN-link og MAX-link.
- b) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør hierarkisk agglomerativ klynging på dette datasettet, og vis det resulterende dendrogrammet. Spesifiser om du bruker MIN-link eller MAX-link.
- c) Gi 4 årsaker til at vi ønsker å evaluere klynginger (klynge-validitet).
- d) Forklar hvordan man kan finne hva som er passende antall klynger K ved klynging vha. K -means.

X	Y
1	11
1	9
1	5
1	2
6	7
11	7

Oppgave 4 – Assosiasjonsregler – 20 %

Anta handlekurv-data til høyre. Bruk apriori-algoritmen for å finne hvilke assosiasjonsregler som gjelder, gitt at minimum støtte er 50 % (dvs. *minimum support count* er 3) og konfidens er 70 %.

TransaksjonsID	Element
T1	A,C,D
T2	B,C,E
T3	A,B,C,E
T4	B,E
T5	B,C
T6	A,C,D,E

Oppgave 5 – Klassifisering – 10 %

- a) Beskriv Hunt's algoritme.
- b) Forklar *underfitting* og *overfitting* i kontekst av beslutningstre.

LØSNINGSSKISSE TIL DELAR AV EKSAMENSOPPGAVE I FAG TDT4300 – JUNI 2012

(NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element)

Oppgave 3 – Assosiasjonsregler – 15 %

A	3
B	4
C	5
D	2
E	4

AB	1
AC	3
AE	2
BC	3
BE	3
CE	3

ABC	(remove AB not frequent)
ACE	(remove AE not frequent)
BCE	2

Forventar at også kandidatsett vert brukt.

A->C	3/3	1
C->A	3/5	0.6
B->C	3/4	0.75
C->B	3/5	0.6
B->E	3/4	0.75
E->B	3/4	0.75
C->E	3/5	0.6
E->C	3/4	0.75

Oppgave 4 – Klynging – 20 %

- a) Se læreboka s. 515-> Forventa både agglomerativ og divisiv.

- b) MIN-link: 1,11 med 1,9

1,2 med 1,5

1,11,1,9 med

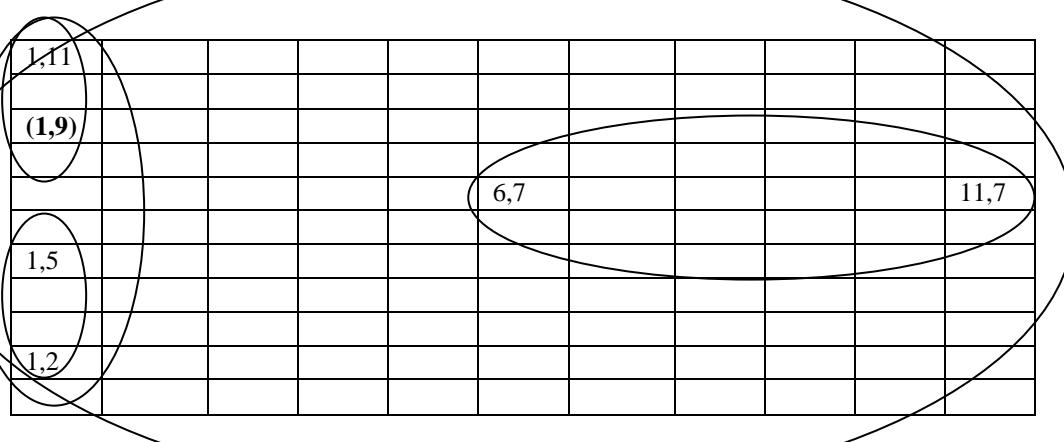
1,2,1,5

6,7 med 11,7. Godtek
både utrekna og
"visuell" løysing"

- c) Legg merke til
at oppgåva
bed om
årsaker og
ikkje metoder.

- Unngå å finne
mønster i støy
- For å samanlikne klyngings-algoritmer
- For å samanlikne to sett med klynger
- For å samanlikne to klynger

- d) Ein metode: Finn "knekkt" i SSE vs. K, jfr. s. X



Oppgave 5 – Klassifisering – 10 %

- a) Beskriv Hunt's algoritme. Læreboka s. 152->.

- b) Forklar *underfitting* og *overfitting* i kontekst av beslutningstre. Læreboka s. 174
Forventar også forklaring på korleis situasjonen oppstår.

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagravedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 735 96755

Eksamensdato: 29. mai

Eksamensstid (fra-til): 09.00-13.00

Hjelpekode/Tillatte hjelpeemidler: D: Ingen trykte eller håndskrivne hjelpeemidler tillatt. Bestemt, enkel kalkulator tillatt.

Annен informasjon:

Målform/språk: Bokmål

Antall sider: 4

Antall sider vedlegg: 0

Kontrollert av:

Dato

Sign

Oppgave 1 – Diverse – 15% (alle deler teller likt)

- a) Gitt to bit-vektorer p og q :

$$\begin{aligned} p &= 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \\ q &= 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \end{aligned}$$

Regn ut Jaccard-koeffisienten for bitvektorene p og q . Hva er fordelen med Jaccard i forhold til "simple matching"?

- b) Forklar 3 teknikker som kan brukes til pre-prosessering av numeriske data.
- c) Forklar prinsippene bak bitmap-indekser. For hva slags data er denne type indeks egnet, og når er den ikke egnet?

Oppgave 2 – Modellering – 20% (15% på a, 5% på b)

I denne oppgaven skal dere modellere et datavarehus for en regional værvarslingstjeneste. Denne har ca. 1000 målestasjoner, som er spredt over ulike land- og hav-områder i regionen for å samle inn grunnleggende værdata, herunder luftrykk, temperatur og nedbør for hver time. Alle data blir sendt til hovedsentralen, som har samlet inn slike data i over 10 år. Ditt design bør legge til rette for effektive spørreninger og on-line analytisk behandling, og utlede generelle værmønstre.

Beskrivelsen er litt upresist formulert og det er en del av oppgaven å velge ut det som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

- a) Lag et stjerne- eller snøflak-skjema for denne case-beskrivelsen.
- b) Lag to forskjellige konsept hierarkier (fritt valgte dimensjoner).

Oppgave 3 – Klynging – 20 % (5% på a, 15% på b)

- a) Forklar potensielle ulemper med hierarkisk klynging.
- b) 1) Forklar DBSCAN-algoritmen.
 2) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt MinPts=3 og 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

Oppgave 4 – Assosiasjonsregler – 20 %

Anta handlekurv-data til høyre. Bruk apriori-algoritmen for å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Velg deretter et av de frekvente 3-elementsettene og finn alle assosiasjonsregler basert på dette settet, gitt konfidens på 75 %.

TransaksjonsID	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

Oppgave 5 – Klassifisering – 25 % (5% på a og b, 15% på c)

- a) Forklar hva som er hensiktene med *klassifisering*. Gi tre eksempler på typiske oppgaver som kan løses ved hjelp av klassifisering.
- b) Forklar kort prinsippene bak *nærmeste-nabo-klassifisering* (*nearest neighbour classification*).
- c) Magnus Carlsen og Vishy Anand skal spille VM-finale mot hverandre senere i år. Det er bestemt at finalen skal gå i India. Se for deg at disse to har truffet hverandre flere ganger tidligere, og at vi har fått tak i data og informasjon om møtene. Vi får også vite at resultatene til Carlsen tidligere har vært avhengig av hvor mye innsats han har lagt i å forberede møtene. I følgende tabell er et datasett som viser verdien 1 hvis Carlsen har ytt full innsats mens verdien 0 betyr innsatsen hans ikke har vært helt topp. Typen kamp og kamptidspunkt, samt sted for kampene tas også hensyn til her.

Tid	Kamptype	Sted	Innsats	Resultat
Morgen	Master	Indoor	1	C
Ettermiddag	Grand Tour	Indoor Crowded	1	C
Kveld	Show	Mall	0	C
Ettermiddag	Show	Mixed	0	A
Ettermiddag	Master	Indoor Crowded	1	A
Ettermiddag	Grand Tour	Indoor	1	C
Ettermiddag	Grand Tour	Mall	1	C
Ettermiddag	Grand Tour	Mall	1	C
Morgen	Master	Indoor	1	C
Ettermiddag	Grand Tour	Indoor Crowded	1	A
Kveld	Show	Mall	0	C
Kveld	Master	Mixed	1	A
Ettermiddag	Master	Indoor Crowded	1	A
Ettermiddag	Master	Indoor	1	C
Ettermiddag	Grand Tour	Mall	1	C
Ettermiddag	Grand Tour	Indoor Crowded	1	C

Kamper som har endt uavgjort (remis) er ikke med i datasettet.

Anta at vi skal bruke *beslutningstre* som klassifiseringsmetode. Vi bruker da dataene i tabellen over som treningsdata. For å avgjøre den beste splitten trenger vi å bruke **Entropy** for en node t som følger:

$$\text{Entropy}(t) = -\sum_j p(j|t) \log p(j|t), \text{ hvor } p(j|t) \text{ er sannsynligheten for klasse } j \text{ gitt node } t \text{ (dvs. andelen}$$

av klasse j i node t) . For hver splitting er “information gain” angitt som

$$GAIN_{split} = \text{Entropy}(p) - \left(\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) \right), \text{ hvor } n_i \text{ er antall elementer i node } i \text{ og } n \text{ total elementer i}$$

forelder-noden p .

Oppgave: Målet med klassifiseringen er å kunne predikere utfallet av fremtidige kamper mellom Carlsen og Anand. Beregn GAIN for splitting på (1) ”Tid” og (2) ”Kamptype”. Hvilken av disse splittingene ville du valgt for å starte opprettelsen av beslutningstreet? Begrunn svaret ditt.



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 :

$$\begin{aligned} p &= 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \\ q &= 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \end{aligned}$$

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 $\text{Entropy}(t) = -\sum_j p(j|t) \log p(j|t)$, 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 $\text{GAIN}_{\text{split}} = \text{Entropy}(p) - \left(\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) \right)$, 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.

LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG

TDT4300 – JUNI 2013

(NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element)

Oppgave 1

- a) $1/(1+1+1)=1/3$.
Jaccard egnet i kontekst av asymmetriske attributter (ignorerer attributter der begge er 0). Eksempel: For likhet mellom handlekorger er kjøpte varer det som oftest er interessant. Rett formel men feil på resten: 7p.
- b) Sjå boka (Tan kap. 2.3). NB! Ikke berre opplisting, men også forklaring. Gjev 4p for (rett) oppramsing.
I ein del lærebøker er det som i Tan er omtala som datakvalitet/datavasking presentert som preprosessering, dette gjeld mellom anna outlier removal og duplikatfjerning, så desse vert også godtekne som svar.
- c) Prinsipp: Sjå boka (Han kap. 4.4.2)..
Type data egnet for: " It is especially useful for low-cardinality domains because comparison, join, and aggregation operations are then reduced to bit arithmetic, which substantially reduces the processing time. Bitmap indexing leads to significant reductions in space and input/output (I/O) since a string of characters can be represented by a single bit. For higher-cardinality domains, the method can be adapted using compression techniques."

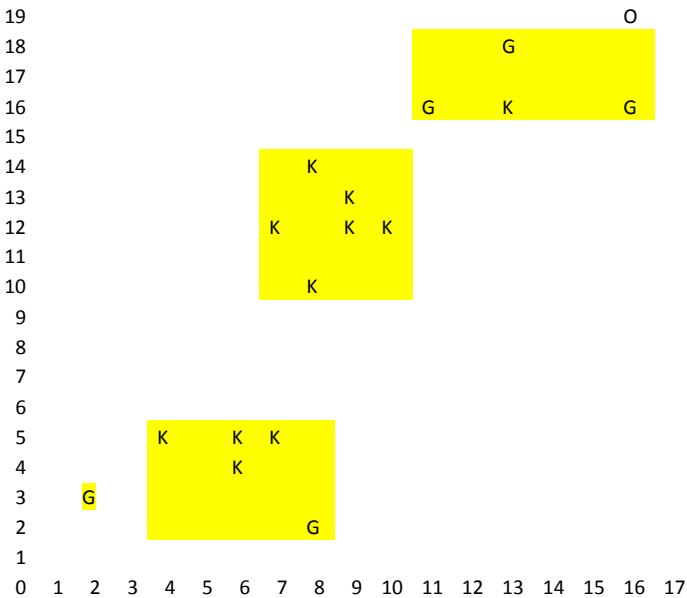
Oppgave 2

- a) Temperatur og nedbør representert som fakta (attributt men ikkje separat tabell) i fakta-tabell.
- b) T.d. år-kvartal-månad-dag og provins-region-county-by

Oppgave 3 – Klynging – 20 %

- a) Forklar potensielle ulemper med hierarkisk klynging.
 - Når avgjerd er teken om samanslåing av to klynger kan den ikke gjerast om.
 - Høg tidskompleksitet og minneforbruk.
 - Dei forskjellige metodane har problem med eit eller fleire aspekt: Sensitivitet mht. støy og outliers, problem med handsaming av klynger med forskjellig størrelse, oppdeling av store klynger

- b) 1) DBSCAN-algoritmen: se læreboken/foilene. Forventar også forklaring på kva som er core/border/noise-point, og også sjølve algoritma (mange har berre forklaring på punkttypen og teikna ein sirkel rundt visuelle klynger i del 2)..
 2) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt MinPts=3 og Eps=3. Det kan forekomme andre variantar som er korrekte, avhengig av om ein reknar MinPts som inkludert punktet eller ikkje, og om ein reknar Ept som inklusive eller ikkje (det siste kan føre til at øverste klynge vert støy). Det er også studentar som har brukt Manhattan-distanse for å gjere avstands-utrekning enklare. Kreativt, men rett. :)



1,414	1 opp og 1 til sida
2,236	1 opp og to til side
3,162	1 opp og 3 til side
2,828	2 opp og 2 til side

Oppgave 4 – Assosiasjonsregler

6p på elementsett og 4p på assosiasjonsreglar. 4p max. på elementsett om ein ikkje har 4-elementsettet (det er ikkje noko i oppgåva som tilseier at dei skal stoppe på 3-elementsett!).

A	6
B	4
C	7
D	2
E	4
F	4
G	3

AB	3
AC	5
AE	4
AF	4
BC	4
BE	2
BF	2
CE	4
CF	4
EF	4

ACE	4
ACF	4
AEF	4
CEF	4

Kun eit 4-elementsett mogleg: ACEF | 4

A->C	5/6	0.83		A->EF	4/6	.67
C->A	5/7	.71		AE->F	4/4	1*
A->E	4/6	.67		E->AF	4/4	1*
E->A	4/4	1		F->AE	4/4	1*
A->F	4/6	.67		EF->A	4/4	1*
F->A	4/4	1		AF->E	4/4	1*
B->C	4/4	1		C->EF	4/7	.57
C->B	4/7	.57		CE->F	4/4	1*
C->E	4/7	.57		E->CF	4/4	1*
E->C	4/4	1		F->CE	4/4	1*
C->F	4/7	.57		EF->C	4/4	1*
F->C	4/4	1		CF->E	4/4	1*
E->F	4/4	1				
F->E	4/4	1				
A->CE	4/6	.67				
AC->E	4/5	.8*				
C->AE	4/7	.57				
CE->A	4/4	1*				
E->AC	4/4	1*				
AE->C	4/4	1*				
A->CF	4/6	.67				
AC->F	4/5	.8*				
C->AF	4/7	.57				
CF->A	4/4	1*				
F->AC	4/4	1*				
AF->C	4/4	1*				

Oppgave 5 – Klassifisering – 25 %

- a) Klassifisering er prosessen med å identifisere hvilken av et sett av kategorier (eller klasser) en ny observasjon hører til (predikseringsaspektet er viktig, mange studentar har ikkje med dette) Dette blir ofte gjort på grunnlag av et sett av treningsdata som inneholder observasjoner (eller instanser) hvor kategorimedlemskap er kjent. Basert på denne definisjonen er hensiktene med klassifisering å kunne kategorisere tidligere ukjente data basert på en modell er bygd basert på tidligere observasjoner eller treningsdata. Eksempler på dette er kategorisering av søkeresultater, kategorisering av sykdomsrisiko i en befolkning, predikere været basert på tidligere værdata.
- b) Kap. 5.2 i Tan. Når det gjelder nærmeste-nabo-klassifisering er det viktig at de viser at de forstår basisideene med avstandsmål mellom dataene og ant. naboen definert av avstandsmålet. For å klassifiser bruker man dette avstandsmålet mot andre data i treningssettet. Deretter kan man identifisere k nærmest nabo innenfor dette målet. For å avgjøre en klasse bruker ”etiketten” til naboen som klassen til en ukjent data.
- c) **Svar:** NB! Utrekningane under er basert på log₂, i oppgåveteksta er gjeve **log** som gjev andre numeriske verdiar. Log₁₀-varianten gjeve med tal i kursiv (omtrentlege tal, orka ikkje rekne dei ut :). Også nokon som har bruk ln. Har vore relativt fleksibel med små reknefeil.

Entropy i rootnode:

$$p(C|Parent) = 11/16$$

$$p(A|Parent) = 5/16$$

$$I(C, A) = (11, 5) = -11/16 \cdot \log_2(11/16) - 5/16 \cdot \log_2(5/16) = 0.896 \quad 0.27$$

1) **Splitting på tid A1:**

S1="Morgen"

$$C1=2, A1=0, I(C1, A1)=I(2,0)=0 \quad 0$$

S2="Ettermiddag"

$$C2=7, A2=4, I(C2, A2)=I(7,4)=0.946 \quad 0.28$$

S3="Kveld"

$$C3=2, A3=1, I(C3, A3)=I(2,1)=0.918 \quad 0.276$$

$$\text{Dermed GAIN}(A1) = 0.896 - (2/16 \cdot I(2,0) + 11/16 \cdot I(7,4) + 3/16 \cdot I(2,1)) = \underline{\underline{0.074}} \quad \underline{\underline{0.022}}$$

2) **Splitting på kamptype A2**

S1="Master"

$$C1=3, A1=3, I(C1, A1)=I(3,3)=1 \quad 0.3$$

S2="Grand tour"

$$C2=6, A2=1, I(C2, A2)=I(6,1)=0.591 \quad 0.177$$

S3="Show"

$$C3=2, A3=1, I(C3, A3)=I(2,1)=0.918 \quad 0.276$$

$$\text{Dermed GAIN}(A2) = 0.896 - (6/16 \cdot I(3,3) + 7/16 \cdot I(6,1) + 3/16 \cdot I(2,1)) = \underline{\underline{0.09}} \quad \underline{\underline{0.037 \text{ (eller } 0.028 :)}}$$

Vi velger attributtet med høyeste GAIN (eller laveste weightet avg.entropy). Derfor foretrekkes A2: Kamptype for første splitting av treet.

Institutt for dатateknikk og informasjonsvitenskap (IDI)

Eksamensoppgave i TDT4300 Datavarehus og datagruverdrift - Vår 2014 (Sensurveiledning)

Faglig kontakt under eksamen: Heri Ramampiaro

Tlf.: 99027656

Eksamensdato: 03. juni 2014

Eksamensstid (fra-til): 09:00 - 13:00

Hjelphemiddelkode/Tillatte hjelphemidler: D – Ingen trykte eller håndskrevne tillatt. Kun typegodkjent kalkulator er tillatt

Annен informasjon: Svar **kort** og **konsist** på alle spørsmålene. Stikkord foretrekkes fremfor lange forklaringer.

Målform/språk: Bokmål

Antall sider: 5

Antall sider vedlegg: 0

Kontrollert av:

Dato

Sign

Oppgave 1 (15%):

1. Forklar hva er datavarehus er.
2. Beskriv de viktigste forskjellene mellom et datavarehus og et operasjonelt databasesystem.
3. Hvilke prosesser inngår typisk i det som forkortes ETL.
4. Beskriv hva en datakube er.
5. Hvilke data representerer en 0-D cuboid (også kalt apex cuboid)?

Oppgave 2 (25%):

Du er i et prosjekt og skal lage et datavarehus med karakterdata som skal brukes for å undersøke hvordan ulike institusjoner og studieprogram benytter karakterskalaen. Dere skal kun støtte høyere utdanning hvor alle bruker samme karakterskala.

Statistikk som skal genereres er typisk karakterfordelingen (andel % for hver karakter), snittkarakter og lignende. Dere ønsker å undersøke i forhold til forskjellige grupper av studenter (alder, kjønn og annen informasjon som er relatert til person). Alle høyere utdanningsinstitusjoner i Norge (universiteter og høgskoler) tilbyr forskjellige studieprogram som kan kategoriseres med fagdisiplin (økonomi, samfunnsvitenskap, medisin, psykologi etc) og med nivå (bachelor, master, phd). Et studieprogram består bestandig av en samling emner og studentene får karakterer i enkeltemner. Andre aspekter som kan være av interesser er karakterfordeling over tid og karakterfordeling i forhold til landsdeler og byer.

Beskrivelsen er litt upresist formulert og det er en del av oppgaven å velge ut hva som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

1. Lag et snøflak eller stjerneskjema for dette caset og beskriv hva som er forskjellen mellom disse to skjematypene.
2. Lag to forskjellige konsept hierarkier. Beskriv forskjellen mellom et hierarkisk og gitter-basert konsept hierarki (hierarchical vs. lattice).
3. Lag en tabell som eksemplifiserer hvilke data som kan genereres fra datavarehuset. Legg vekt på å få fram aggregeringsprinsippet.

Oppgave 3 (15%):

1. Forklar hva som hovedforskjellene mellom datavarehus (data warehouse) og datagravedrift (data mining).

Svar:

Datavarehus: Analyser, rapporter og spørninger.

Delvis å hente ut forhåndsdefinert “**kjent informasjon**” (*men ikke kjente resultater*)

Datagravedrift: *oppdage ny informasjon (knowledge discovery)*

2. Forklar hvorfor assosiasjonsregler ikke kan regnes som en prediksjon (prediction) mens klassifisering er det.
- Svar:** AR kan ikke regnes som prediksjon fordi den bygger regler basert på eksisterende data og har ingen kausalitet. Klassifisering er derimot prediskjon da man prøver å "spå" klassesettihørighet basert på en nåværende kunnskap (feks. trenet klassifiseringsmodell).
3. Hva er hovedforskjellene mellom klassifisering og klynging (clustering)? Forklar.
- Svar:** Her er det nok at studentene fokuserer forklaringene på "supervised learning" (klassifisering) vs. "unsupervised learning" (clustering).
4. Forklar hovedfordelene med flat klynging (flat clustering) som feks. K-means mot hierarkisk klyning (hierarchical clustering) og omvendt (d.v.s hierarkisk mot flat).
- Svar:** Her er kreves at studentene får med seg hvilke aspekter relater til hastighet, valg av sentroider, sensitivitet for støy og outliers, osv.

Oppgave 4 (20%):

Amazon Inc. er kjent for å analysere hva kundene deres kjøper. De er opptatt av å ha balanse mellom etterspørsel av varene og lagerbeholdningen. Defor bestemmer de seg for å bruke assosiasjonsregelanalyse og -mining til dette formålet.

Før de setter igang bestemmer de seg for å gjøre noen enkle lavskala-analyser på noen få varer som kundene har kjøpt sammen i det siste:

```
t1: {Headphones, iPod}
t2: {Headphones, iPad, iTunes-giftcard, Running shoes }
t3: {iPod, iPad, iTunes-giftcard, Bike-computer}
t4: {Headphones, iPod, iPad, iTunes-giftcard}
t5: {Headphones, iPod, iPad, Bike-computer}
```

1. Hva menes med **støtteantall** (support count) og **støtte** (support)? Hvor mye blir støtteantall og støtte for {Headphones, iPod}?

Svar: Støtteantall: Frekvens av forekomst av et elementsett. Støtte: Andel av transaksjoner som inneholder et elementsett. $sc(\{Headphones, iPod\}) = 3$, $s(\{Headphones, iPod\}) = 3/5 = 60\%$.

2. Anta at minimum support $minsup = 0.5$. Hvilke elementsett er **frekvente elementsett**?

Svar: $minsup=0.5$ betyr min sup.count på 3.

Disse oppyller dette kravet på $minsup=0.5$

1-sett elementsett: {Headphones}, {iPod}, {iPad}, {iTunes-giftcard}

2-sett elementsett: {Headphones, iPod}, {Headphones, iPad}, {iPod, iPad}, {iPad, iTunes-giftcard}

3-sett elementsett: ingen oppfyller minsup-kravet.

3. Anta regelen {iPod, iPad} => iTunes-giftcard

a) Hva blir verdien av **støtte** (s) og **konfidens** (c) her?

Svar: Støtte, $s = \text{støtte}(\{\text{iPod}, \text{iPad}, \text{iTunes-giftcard}\}) = 2/5 = 0.4$. Konfidens, $c = \text{støtte}(\{\{\text{iPod}, \text{iPad}, \text{iTunes-giftcard}\}\}) / \text{støtte}(\{\text{iPod}, \text{iPad}\}) = 2/3 = 0.67$

b) Hvorfor er “brute-force” regelgenereringsmetoden generelt uegnet til assosiasjonsregeloppdagelser?

Svar: Hovedproblemet er tidskompleksitet siden en brute-force algoritmer ville se etter alle mulige kombinasjoner, også de som ikke er nødvendigvis interessante jobbe videre med.

c) Hvordan kan “apriori-algoritmen” utnyttes til assosiasjonsregeloppdagelser? Tips: vi er ute etter apriori-prinsippet.

Svar: Her kreves forklaring på anti-monoton-egenskapene for å “prune” ikke frekvente elementsett.

d) Forklar hvordan aprori-algoritmen er bygd opp ved å bruke eksempel-transaksjonene over som utgangspunkt. Anta fortsatt at $\text{minsup} = 0.5$.

Svar: Her skal studentene lage noe tilsvarende dette:

Element	Tal	Element (1-elementsett)
Bred	4	
Cola	2	
Mjølk	4	
Øl	3	
Potetg.	4	
Egg	1	

Elementsett	Tal	Par (2-elementsett) (Ikke nødvendig å generere kandidater som inneholder Cola eller Egg)
{Bred, Mjølk}	3	
{Bred, Øl}	2	
{Bred, Potetg.}	3	
{Mjølk, Øl}	2	
{Mjølk, Potetg.}	3	
{Øl, Potetg.}	3	

Elementsett	Tal	Tripletter (3-elementsett)
{Bred, Mjølk, Potetg.}	3	

4. Nevn *minst tre ting* som påvirker kompleksiteten på regelgenerering. Nevn deretter *minst en ting* man kan gjøre for å møte/løse utfordringene forbundet med denne kompleksiteten.

Svar:

- **Valg av minimum støtte (minsup)**
 - lavere terskel resulterer i flere frekvente elementsett
 - fører typisk til høyre antall på kandidater og maksstørrelse på frekvente elementsett
- **Dimensjonalitet** (antall av element) på datasettet
 - mer plass nødvendig for å lagre støtteantall (support count) til elementene
 - hvis antall på frekvente element også øker kan beregning og IO-kostnader også øke
- **Størrelse på databasen**

- Apriori gir flere pass gjennom databasen
→ køyretid vil øke med antall på transaksjoner
- **Gjennomsnittlig transaksjonsbredde**
 - Økt transaksjonsbredde
→ (typisk) økt maksstørrelse på frekvente elementsett

Oppgave 5 (25%):

Skandiabanken ASA vil finne ut om hvilke av deres kunder kommer til å kjøpe bil i fremtiden. Tabellen under viser informasjon om noen av kundene deres som allerede har kjøpt bil.

ID	Alder	Inntekt	Arbeidstype	Kredittverdighet	Bilkjøper
1	<= 30	Høy	Fulltid	Passe	Nei
2	<= 30	Høy	Fulltid	Høy	Nei
3	31 - 40	Høy	Fulltid	Passe	Ja
4	> 40	Middels-høy	Fulltid	Passe	Ja
5	> 40	Lav	Deltid	Passe	Ja
6	> 40	Lav	Deltid	Høy	Nei
7	31 - 40	Lav	Deltid	Høy	Ja
8	<= 30	Middels-høy	Fulltid	Passe	Nei
9	<= 30	Lav	Deltid	Passe	Ja
10	> 40	Middels-høy	Deltid	Passe	Ja
11	<= 30	Middels-høy	Deltid	Høy	Ja
12	31 - 40	Middels-høy	Fulltid	Høy	Ja
13	31 - 40	Høy	Deltid	Passe	Ja
14	> 40	Middels-høy	Fulltid	Høy	Nei

1. Bruk tabellen over og **Hunt's algoritmen** som utgangspunkt til å indusere et beslutningstre. Det forventes ikke at du skal lage et komplett tre. Det holder at du viser du har forstått prinsippet.

Svar: Her må studentene vise at de har forstått prinsippet bak Hunt's algoritmen. Jo mer detaljer de har med, desto mer poeng får de.

2. For å kunne finne ut om et splitt (split) er bra bruke en ofte å måle homogenitet av kandidatnoden for splittingen. Hvordan måles homogeniteten?

Svar: Homogenitet måles ved hvordan klassene er fordelt på en (kandidat)splitnode. En node som har lik fordeling på klassen er ikke homogen (50:50 vs. 100:00 (homogen)).

3. Anta at du starter splittingen din på “Arbeidstype”. Finn **GINI** og **Entropy**-verdiene for denne noden. Gitt GINI og Entropy som følgende:

$$GINI(t) = 1 - \sum_j p(j|t)$$

$$Entropy(t) = - \sum_j p(j|t) \log p(j|t)$$

Her er $p(j|t)$ sannsynligheten for klassen j gitt noden t (d.v.s andelen av klassen j i noden t).

Svar: Ved noden $t=$ “Arbeidstype” er klassene fordelt på 5 av 14 “nei”-klasse og 9 av 14 “ja”-klasse. Dette betyr $GINI(t) = 1 - ((5/14)^2 + (9/14)^2) = 0.46$, $ENTROPY(t) = -((5/14)\log(5/14) + (9/14)\log(9/14)) = 0.28$.

4. Gitt $GINI_{split}$ som følgende:

$$GINI_{split} = \sum_{i=0}^k \frac{n_i}{n} GINI(i)$$

Her er n_i antall elementer i node i og n total elementer i foreldernoden p . Bruk dette til å finne ut om “Arbeidstype” eller “Kredittverdiget” er best å starte splittingen med.

Svar:

For noden “Arbeidstype” kan vi splitte på “Fulltid” og “Deltid” som gir $GINI(t)$ som følgende:

$$GINI(Fulltid) = 1 - ((4/7)^2 - (3/7)^2) = 1 - 0.327 - 0.187 = 0.486$$

$$GINI(Deltid) = 1 - ((1/7)^2 - (6/7)^2) = 1 - 0.02 - 0.735 = 0.245$$

$$GINI_{split} = 7/14 * 0.486 + 7/14 * 0.245 = 0.366$$

For noden “Kredittverdiget” kan vi gjøre tilsvarende beregning:

$$GINI(Passe) = 1 - ((2/8)^2 - (6/8)^2) = 1 - 0.0625 - 0.5625 = 0.375$$

$$GINI(Høy) = 1 - ((3/6)^2 - (3/6)^2) = 1 - 0.25 - 0.25 = 0.5$$

$$GINI_{split} = 8/14 * 0.375 + 6/14 * 0.5 = 0.429$$

“Kredittverdiget” har høyre $GINI_{split}$ og er dermed dårligere valg enn “Arbeidstype”.

Institutt for dømteknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagrudevdrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 41440433

Eksamensdato: 5. juni 2015

Eksamenstid (fra-til): 09.00-13.00

**Hjelpemiddelkode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrivne hjelpeiddel
tillatt. Bestemt, enkel kalkulator tillatt.**

Annen informasjon:

Målform/språk: Bokmål

Antall sider (uten forside): 3

Antall sider vedlegg: 0

Kontrollert av:

Dato

Sign

Oppgave 1 – Diverse – 15 % (alle deler teller likt)

- Forklar *asymmetriske attributt*. Gi et eksempel på et slikt attributt.
- Anta to bit-vektorer p og q :

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

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

Regn ut Jaccard-koeffisienten for bitvektorene p og q .

- I mange datasett kan verdier mangle for attributt i noen av objektene, ofte fordi noen attributt ikke er relevante for alle (f.eks. barn har typisk ikke inntekt). Gi tre metoder/strategier man kan bruke for å håndtere manglende verdier.

Oppgave 2 – Modellering – 20 % (17 % på a, 3 % på b)

I denne oppgaven skal dere modellere et datavarehus for Netflix. Netflix tilbyr strømming av TV-serier og filmer, og ønsker et datavarehus for å kunne analysere visninger av TV-serier (for enkelhets skyld kan dere se bort fra filmer i denne oppgaven). En *visning* er i denne sammenheng definert som hendelsen at en bruker ser på en TV-episode eller deler av en TV-episode.

For å forenkle modelleringen kan dere anta at tidspunktet for en visning er tidspunktet den starter, og at laveste granularitet for visning er *kapittel* (dvs. dere trenger ikke modellere start- og slutt-tidspunkt), der man antar at en episode består av ett eller flere kapitler.

Eksempel på analyser man skal være i stand til å gjøre mot datavarehuset:

- Gjennomsnittlig lengde (tid) på hver visning.
- Visningsmetode (f.eks. Android-app, nettleser, etc.) per kvartal.
- Antall visninger for hvert kapittel av en bestemt TV-serie for hvert land.

Beskrivelsen er litt upresist formulert og det er en del av oppgaven å velge ut det som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

- Lag et stjerne-skjema for denne case-beskrivelsen.
- Konsepthierarki for tid kan f.eks. være *år-kvartal-måned-dag*. Kan *uke* være en del av dette hierarkiet? Begrunn svaret.

Oppgave 3 – Klynging – 20 % (5 % på a, 15 % på b)

- Forklar fordeler og ulemper med k-means.
- 1) Forklar hierarkisk agglomerativ klynging.
2) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør hierarkisk agglomerativ klynging på dette datasettet ved å bruke MIN (single link) og Manhattan-distanse. Vis det resulterende dendrogrammet.

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

Oppgave 4 – Klassifisering – 25 % (5 % på a og 20 % på b)

- a) Forklar *forvekslingsmatrise* ("confusion matrix"), innholdet i denne, og hvordan man regner ut *nøyaktighet* ("accuracy") basert på denne.
- b) Rosenborg og Vålerenga skal i morgen (lørdag) spille tippeliga-kamp på Ullevaal (som er hjemmestadion for Vålerenga). Disse har spilt mot hverandre mange ganger tidligere, og vi ønsker å bruke resultat og informasjon fra tidligere kamper til å predikere morgendagens resultat. Denne informasjonen er vist i tabellen under (kamper som har endt uavgjort er ikke med i datasettet, H/B betyr Rosenborg hjemme/borte).

Dag	Turnering	Sted	Tidspunkt	Resultat
Fredag	Tippeligaen	H	Ettermiddag	R
Søndag	NM	H	Kveld	R
Søndag	Tippeligaen	B	Ettermiddag	R
Søndag	Tippeligaen	H	Kveld	R
Lørdag	Tippeligaen	B	Ettermiddag	V
Søndag	Tippeligaen	H	Ettermiddag	R
Søndag	Tippeligaen	H	Kveld	R
Lørdag	Tippeligaen	B	Ettermiddag	R
Søndag	Tippeligaen	H	Kveld	R
Søndag	Tippeligaen	H	Ettermiddag	R
Fredag	Tippeligaen	H	Kveld	R
Søndag	Tippeligaen	B	Kveld	V
Lørdag	Tippeligaen	H	Ettermiddag	R
Søndag	Tippeligaen	B	Ettermiddag	R
Søndag	Tippeligaen	B	Kveld	V
Lørdag	Tippeligaen	H	Ettermiddag	V

Anta at vi skal bruke *beslutningstre* ("decision tree") som klassifiseringsmetode. Vi bruker da data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for urenhet ("impurity"), og følgende to formler kan være til hjelp for å løse oppgaven:

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

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

Oppgave: Målet med klassifiseringen er å kunne predikere utfallet av morgendagens kamp mellom Rosenborg og Vålerenga. Regn ut $GAIN_{split}$ for splitting på (1) "Sted" og (2) "Dag". Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Oppgave 5 – Assosiasjonsregler – 20 %

Anta handlekorg-data som er gitt under. Bruk apriori-algoritmen til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Vis hvordan kandidatsettene blir generert.

Et av de frekvente elementsettene er BDE. Finn alle assosiasjonsregler basert på dette settet, gitt konfidens på 75 % (det er ikke nødvendig å bruke apriori til å finne assosiasjonsreglene, men vis hvordan konfidens blir regnet ut for hver av kandidatreglene som er basert på BDE).

TransaksjonsID Element

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



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:

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

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

Task: The goal of the classification is to be able to predict the outcome of tomorrow's (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

LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG

TDT4300 – JUNI 2015

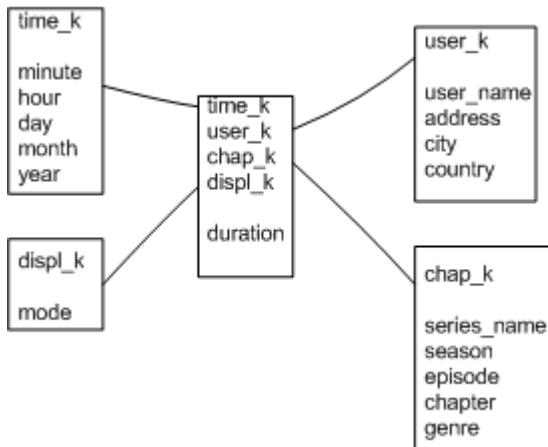
NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element, hovudsakleg laga for at vi skal ha oversikt over arbeidsmengda på eksamen, og som huskeliste under sensurering. Det er også viktig å være klar over at det også kan vere andre svar ein dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving.

Oppgåve 1

- Kun tilstedevarsel (attributtverdi ulik 0) viktig (alternativt: nokre verdiar er viktigare enn andre). Eksempel (forventar forklaring som under, ikkje berre ”varer i handlekorg”): Objekt er student, eitt attributt for kvart fag på universitetet Handlekorg, en attributt for kvar var, kun varer som er kjøpt er interessante.
- $2/(1+1+2)=2/4=1/2$
Poengtrekk der det manglar noko som forklarar talet.
- (Minst 3) Eliminere (eller overser) objektet, Estimere manglende verdi, Interpolere manglende verdi, Ignorere det aktuelle attributtet.

Oppgåve 2

- Eit viktig poeng er at fakta-attributt må gje mening utifrå oppgåva, og også gje mening ved aggregering.

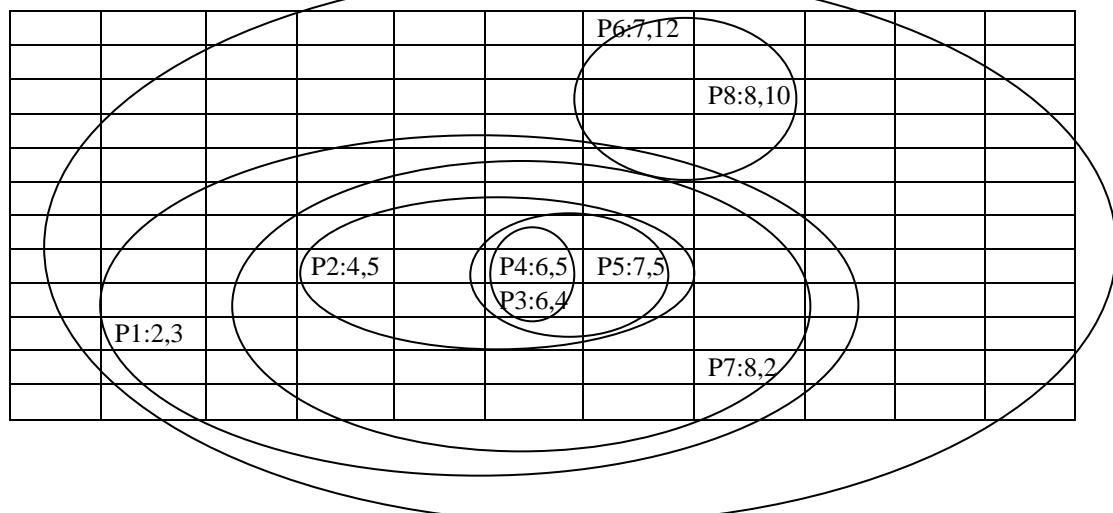


- Ja, men hierarkiet må i såtilfelle ordnast som eit gitter (lattice), jfr. figure 4.10 i læreboka (Han kap. 4). Godtek også svaret ”Nei” med fornuftig forklaring (litt diffust forklart i boka).

Oppgåve 3

- Pro: Lav kostnad (både mht. utrekning og minne).
Con (minst 4): 1) Finn kun globulære klynger. 2) Problem med forskjellig størrelse og tettleik. 3) Problem med data med outliers. 4) Kun for data der ein kan definere ”senter”. 5) Sensitiv for val av initial-sentroide 6) må velje tal på klynger.

- MIN-link:
P4 med P5 eller P3, deretter vert hhv. P3 eller P5 lagt til. Deretter P2 med P3/P4/P5. P6 og P8.
P2/P3/P4/P5 med enten P1 eller P7. Deretter P1 eller P7 til denne. Til slutt, P6/P8 med resten. Er OK om



ein løyser den grafisk (med forklaring på korleis avstand kan reknast ut).

Dendrogram: basert på desse samanslåingane, figur forventa (også viktig for å vise rekkefølgje av samanslåingane, sjølv om nokre av samanslåingane her vil ende opp "på same nivå" fordi ein har fleire par med same distanse og det då er vilkårleg kva par som vert slått saman).

Oppgåve 4

- a) Forvekslingsmatrise er matrise med korrekt klasse-merkelapp (class label) versus estimert klasse-merkelapp for kvar klasse (to rader og to kolonner som viser tall på falske positive, falske negative, ekte positive og ekte negative, jfr. figuren under). Nøyaktigheit rekna ut som vist i formelen under.

		PREDICTED CLASS		
		Class=Yes	Class>No	a: TP (true positive) b: FN (false negative) c: FP (false positive) d: TN (true negative)
ACTUAL CLASS	Class=Yes	a	b	
	Class>No	c	d	
Accuracy		$\frac{a+d}{a+b+c+d}$	$\frac{TP+TN}{TP+TN+FP+FN}$	

b)

Gini i rotnode:

$$p(R|Parent) = 12/16 = 0.75, p(V|Parent) = 4/16 = 0.25. GI(C, A) = 1 - 0.75 * 0.75 - 0.25 * 0.25 = 0.375$$

1) **Splitting på dag D:**

D1="Fredag"

$$R1=2, V1=0, GI(R1,V1)=GI(2,0)=0$$

D2="Laurdag"

$$R2=2, V2=2, GI(R2,V2)=GI(2,2)=0.5$$

D3="Søndag"

$$R3=8, V3=2, GI(R3, V3)=GI(8,2)=0.32$$

$$GAIN(D) = 0.375 - 2/16 * 0 - 4/16 * 0.5 - 10/16 * 0.32 = 0.375 - 0.125 * 0 - 0.25 * 0.5 - 0.625 * 0.32 = 0.05$$

2) **Splitting på stad S**

S1="H"

$$R1=9, V1=1, GI(R1,V1)=1 - 0.9 * 0.9 - 0.1 * 0.1 = 0.18$$

S2="B"

$$R2=3, V2=3, GI(R2,V2)=0.5$$

$$GAIN(S) = 0.375 - 10/16 * 0.18 - 6/16 * 0.5 = 0.075$$

Vi vel attributtet med høgste GAIN, dvs. Stad vert føretrekt for første splitting av treeet.

Oppgåve 5 – Assosiasjonsreglar

A	6
B	8
C	3
D	4
E	4
F	4
G	2
H	1

AB	6
AD	2
AE	2
AF	2
BD	4
BE	4
BF	4
DE	4
DF	4
EF	4

BDE	4
BDF	4
BEF	4
DEF	4

Kun eit 4-elementsett mogleg: BDEF | 4

B->DE	4/8	0.5	
BD->E	4/4	1.0	*
D->EB	4/4	1.0	*

DE->B	4/4	1.0	*
E->BD	4/4	1.0	*
BE->D	4/4	1.0	*

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 73596755

Eksamensdato: 26. mai 2015

Eksamensstid (fra-til): 09.00-13.00

Hjelpekode/Tillatte hjelpeemidler: D: Ingen trykte eller håndskrevne hjelpeemiddel tillatt.
Bestemt, enkel kalkulator tillatt.

Annен informasjon:

Målform/språk: Bokmål

Antall sider (uten forside): 4

Antall sider vedlegg: 0

Kontrollert av:

Dato

Sign

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig 2-sidig

sort/hvit farger

Oppgave 1 – Diverse – 15 % (alle deler teller likt)

- Beskriv kort fire formål med klyngevalidering/evaluering.
- Forklar fire teknikker for data-vasking i kontekst av web-bruk-data.
- Anta to bit-vektorer p og q :

$$\begin{aligned} p &= 1 0 1 0 0 0 0 1 1 1 \\ q &= 1 0 0 0 0 0 1 1 0 1 \end{aligned}$$

Regn ut Jaccard-koeffisienten for bitvektorene p og q .

Oppgave 2 – Modellering – 15 %

I denne oppgaven skal du modellere et datavarehus for bilskader i forsikringsselskapet Lillebrand. Lillebrand ønsker et datavarehus for å kunne analysere hendelser som har medført forsikringsutbetalinger.

Eksempel på analyser man skal være i stand til å gjøre mot datavarehuset:

- Antall skader i 2015.
- Gjennomsnittlig antall skader per måned.
- Antall skader for hvert kvartal i 2015.
- Totalt beløp utbetalt for hver biltype.
- Antall skader av type ”kollisjon” per by.

Beskrivelsen er litt upresis og det er en del av oppgaven å velge ut det som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

Lag et stjerne-skjema for denne case-beskrivelsen.

Oppgave 3 – OLAP – 15 % (5 % på a og 10 % på b)

- a) Gitt en kube med dimensjoner:

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

Anta følgende materialiserte kuboider:

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

Gitt følgende OLAP-spørring: {item_name, province_or_state} med vilkår “year = 2006”
Hvilke(n) materialiserte kuboider kan brukes for å prosessere spørringen? Begrunn svaret.

- b) Gitt et datavarehus med tre tabeller Location/Item/Sales, der Sales er fakta-tabellen og de to andre er dimensjonstabeller. Vi ønsker å bruke *join-indeks* for å kunne utføre spørringer mer effektivt.
Vis struktur og innhold for join-indeksene Location/Sales og Item/Sales med utgangspunkt i innholdet i de tre tabellene under.

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

Oppgave 4 – Klynging – 10 %

Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt MinPts=4 (inkl. eget punkt) og Eps=3 (inkl. punkt som har distanse 3). Bruk Manhattan –distanse som avstandsmål.

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

Oppgave 5 – Klassifisering – 20 % (5 % på a og 15 % på b)

- a) Forklar *kryssvalidering* ("cross validation") og hva denne teknikken brukes til.
- b) Et bilforsikringsselskap har for eksisterende kunder lagret informasjon som inkluderer kundenr, alder (L/M/H, dvs. 18-25/26-70/71-100), biltype, kjørelengde per år (4000/8000/20000/Ubegrenset), bonus (Lav/Middels/Høy) og om de har hatt skade på bilen som ble dekket av forsikringen. Når nye kunder ber om tilbud på forsikring, ønsker selskapet å sette prisen til normal eller høy basert på om de tror kunden kommer til å få skade på bilen eller ikke, dvs. de ønsker å predikere attributtet "Skade".

Kundenr	Alder	Biltype	Kjørelengde per år	Bonus	Skade
1	L	Ferrari	8000	Lav	Ja
2	M	BMW	8000	Høy	Nei
3	H	Lexus	Ubegrenset	Høy	Ja
4	L	Audi	8000	Høy	Nei
5	H	Opel	8000	Lav	Ja
6	M	Toyota	8000	Lav	Nei
7	M	Honda	8000	Høy	Nei
8	M	Nissan	8000	Høy	Nei
9	M	Audi	Ubegrenset	Høy	Nei
10	M	BMW	8000	Lav	Ja
11	H	Toyota	Ubegrenset	Høy	Nei
12	L	Nissan	4000	Lav	Ja
13	L	Opel	Ubegrenset	Høy	Ja
14	M	Audi	8000	Høy	Nei
15	M	Opel	8000	Høy	Nei
16	M	Toyota	4000	Lav	Nei

Anta at vi skal bruke *beslutningstre* ("decision tree") som klassifiseringsmetode. Vi bruker da data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for urenhet ("impurity"), og følgende to formler kan være til hjelp for å løse oppgaven:

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

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

Oppgave: Målet med klassifiseringen er å kunne predikere "Skade". Regn ut $GAIN_{split}$ for splitting på (1) "Alder" og (2) "Bonus". Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Oppgave 6 – Assosiasjonsregler – 25 % (10 % på a, 5 % på b, og 10 % på c)

- a) Anta handlekorg-data som er gitt under. Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Bruk $F_{k-1} \times F_{k-1}$ -metoden for kandidat-generering.

TransaksjonsID	Element
T1	ABCDEFGH
T2	DKM
T3	FK
T4	ACGH
T5	ACDDGH
T6	BM
T7	DFKM
T8	ABCDGH

- b) Anta handlekorg-data som er gitt under. Du skal nå bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 60 % (dvs. *minimum support count* er 3).
- 1) Konstruer et FP-tre basert på datasettet.
 - 2) Finn frekvente elementsett ved å bruke FP-growth-algoritmen. Bruk tabell-notasjon med følgende kolonner for å vise resultatet:
 - Element
 - "Conditional pattern base"
 - "Conditional FP-tree"
 - Frekvente elementsett

TransaksjonsID	Element
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



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:

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig X 2-sidig

sort/hvit X farger

Date

Signature

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 :

$$\begin{aligned} p &= 1 0 1 0 0 0 0 1 1 1 \\ q &= 1 0 0 0 0 0 1 1 0 1 \end{aligned}$$

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	ABCDEFGH
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

LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG

TDT4300 – MAI 2016

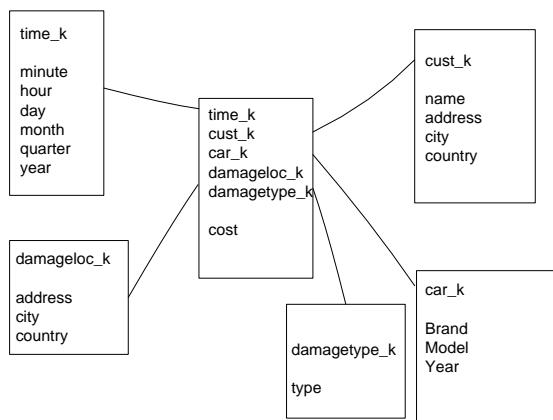
NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element, hovudsakleg laga for at vi skal ha oversikt over arbeidsmengda på eksamen, og som huskeliste under sensurering. Det er også viktig å være klar over at det også kan vere andre svar enn dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving eller dei gjev meining utifrå kva det vert spurd om.

Oppgåve 1

- a) Unngå å finne mønster i støy (Finne klyngings-tendens til eit datasett, dvs. om ikkje-tilfeldige strukturar faktisk finnest)
- For å samanlikne klyngings-algoritmer
 - For å samanlikne to sett med klynger
 - For å samanlikne to klynger
- (Også andre er OK, t.d. "finne tal på klynger", "velge riktige parametre", etc.)
- b) Fjerne irrelevant referansar og felt i loggar
- Fjerne referansar som er resultat av aksessar frå søkerobotar
 - Fjerne feilaktige side-referansar (404 etc)
 - Legge til manglante referansar forårsaka av caching (etter sesjonering)
- c) $M11/(M11+M01+M10)=3/(2+1+3)=3/6=1/2$

Oppgåve 2

- a) Eit viktig poeng er at fakta-attributt må gje meining utifrå oppgåva, og også gje meining ved aggregering. Det er eigentleg ikkje nødvendig med "Antall" sidan kvar "event" er "ein skade", men sidan det har vore liknande attributt på eksempel i undervisninga trekk vi ikkje for det. Det er ikkje sagt noko om kunde i oppgåva så OK om dette ikkje er med. Med tanke på at oppgåve er relativt triviell er vi tilsvarende strenge, t.d. trekk ved manglante dimensjonstabell (typisk lokasjon), manglante skadetype, om fakta (beløp) er i ein av dimensjonstabellane men ikkje i det heile i faktatabell.



Oppgåve 3

- a) 1) `{year, item_name, city}` // Ja (rollup frå city til province_or_state)
 2) `{year, brand, country}` // Nei (materialisert, drill-down ikkje mogleg)
 3) `{year, brand, province_or_state}` // Nei (materialisert, drill-down ikkje mogleg)
 4) `{item_name, province_or_state}` where `year = 2004` // Nei. År 2016 ikkje med i denne

Det er eit viktig poeng at oppgåva seier ”Kven av dei *materialiserte* kuboidane kan brukast til å prosessere spørjinga”, dvs. ein kan ikkje bruke drill-down (men ein kan gjere roll-up, dvs. vidare aggregering).

- b) Join indexen skal helst vere på tabell-form, ein tuppel for kvar nøkkel/nøkkel. På eksempel på ein av foilane er det brukte ”liste” for sekundærattributt så dette vert også godteke. Om tabell er ferdig join-resultat er det teikn på at studenten ikkje veit kva ein join indeks er, og null poeng.

Location/Sales	
LocKey	TransID
L1	T1
L1	T3
L1	T6
L1	T7
L2	T2
L3	T4
L3	T5

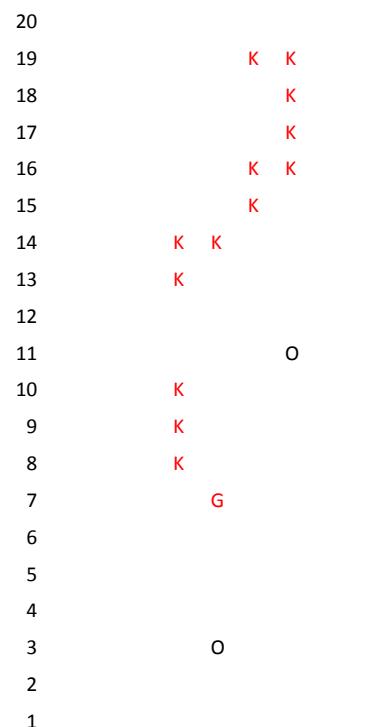
Item/Sales	
LocKey	TransID
I1	T1
I1	T3
I1	T4
I1	T7
I2	T2
I2	T6
I3	T5

Oppgåve 4

	X	Y	
A	P1	4	8
B	P2	4	9
C	P3	4	10
D	P4	4	13
E	P5	4	14
F	P6	5	3
G	P7	5	7
H	P8	5	14
I	P9	6	15
J	P10	6	16
K	P11	6	19
L	P12	7	11
M	P13	7	16
N	P14	7	17
O	P15	7	18
P	P16	7	19

Det er forventa at ein 1) har klassifisert punkt i K/G/O, og 2) har identifisert en klynge.

C1= P1,P2,P3,P4,P5,P7,P8,P9,P10,P11,P13,P14,P15,P16,
 Støypunkt er P6 og P12.



0 1 2 3 4 5 6 7 8 9

Oppgave 5

a) *Cross validation:*

- 1) Partisjonerer data i k disjunkte subsett
 - 2) k -fold: tren på $k-1$ partisjonar, teste på den gjenverande
 - 3) Metrikk er gjennomsnittleg effektivitet
- Brukes til estimering av effektivitet til klassifiseringsmodell.

Legg merke til at oppgåva inneholdt to spørsmål.

b)

Entropi i rotnode:

$$p(J|Parent) = 6/16 = 0.375, p(N|Parent) = 10/16 = 0.625. GI(J,N) = 1 - 0.375 * 0.375 - 0.625 * 0.625 = 0.468750$$

1) **Splitting på alder A1:**

S1="L"

$$J_1=3, N_1=1, GI(J_1, N_1)=GI(3,1)=0.375$$

S2="M"

$$J_2=1, N_2=8, GI(J_2, N_2)=GI(1,8)=0.197$$

S3="H"

$$J_3=2, N_3=1, GI(J_3, N_3)=GI(2,1)=0.444$$

$$GAIN(A1) = 0.469 - 4/16 * 0.375 - 9/16 * 0.197 - 3/16 * 0.444 = 0.181$$

2) **Splitting på bonus A2**

S1="L"

$$J_1=4, N_1=2, GI(J_1, N_1)=GI(4,2)=0.444$$

S2="H"

$$J_2=2, N_2=8, GI(J_2, N_2)=GI(2,8)=0.32$$

$$GAIN(A2) = 0.468750 - 6/16 * 0.444 - 10/16 * 0.32 = 0.102$$

Vi vel attributtet med høgste GAIN, dvs. **alder** vert føretrekt for første splitting av treeet.

NB! Viktig å ha med GAIN inkl. $p(J|Parent)$, det kan skje at ein får negativ verdi for begge dei alternative splittingane, som betyr at ein ikkje bør velje nokon av dei.

Oppgåve 6 – Assosiasjonsreglar

NB! I tittelen til oppg. 6 er det referert til ei oppgåve c. Dette er ein feil (kun a og b), det korrekt3 er 10% på a, 5% på b.1, og 10% på b.2. I T5 i a er det ein D for mykje. Studentane vart bede om å slette den eine, men med tanke på dei som allereie hadde gjort oppgåva godtek vi også at dei har rekna med begge D'ane og sett støtte til D til $s(D)=6$.

a)

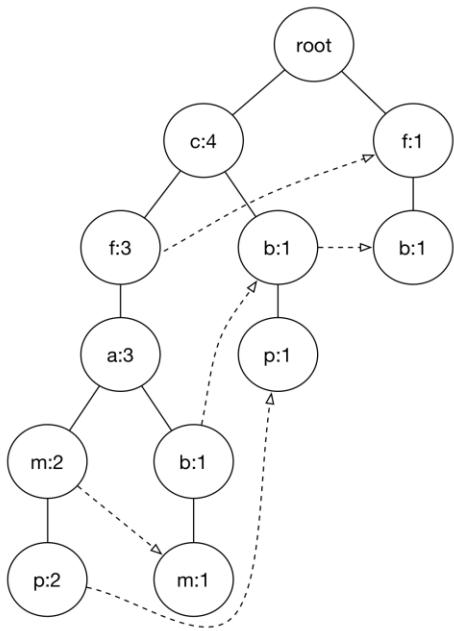
A	4
B	3
C	4
D	5
F	3
G	4
H	4
K	3
M	3
AC	4
AD	3
AG	4
AH	4
CD	3
CG	4
CH	4
DG	3
DH	3
GH	4
ACG	4
ACH	4
AGH	4
CGH	4

Kun eit 4-elementsett mogleg: ACGH | 4

ACGH	4
------	---

b)

tid	Itemset	(Ordered) frequent items
100	f,a,c,d,g,i,m,p	c,f,a,m,p
200	a,b,c,f,l,m,	c,f,a,b,m
300	b,f,h,j,o	f,b
400	b,c,k,s,p	c,b,p
500	a,f,c,e,l,p,m,n	c,f,a,m,p



Legg merke til om ein brukar anna sortering på element med same støttetal kan ein få andre (korrekte) tre. Typisk eksempel er cfabmp ("sortert rekkefølgje").

Item	Conditional sub-database	Conditional FP-tree	Frequent itemsets
<i>p</i>	$\{(cfam:2), (cb:1)\}$	$\{(c:3)\} p$	p, cp
<i>m</i>	$\{(cfa:2), (cfab:1)\}$	$\{(cfa:3)\} m$	m, cm, fm, am, cam, fam, cfm, cfam
<i>b</i>	$\{(cfa:1), (f:1), (c:1)\}$	\emptyset	b
<i>a</i>	$\{(cf:3)\}$	$\{(cf:3)\} a$	a, ca, fa, cfa
<i>f</i>	$\{(c:3)\}$	$\{(c:3)\} c$	f, cf
<i>c</i>	\emptyset	\emptyset	c

Legg merke til at andre rekkefølgjer og delvis anna mellom-resultat kan oppstå sidan fleire element har same støttetal (typisk vel mange å bruke abmpcf ("sortert rekkefølgje")).

Institutt for datateknikk og informasjonsvitenskap

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 73596755

Eksamensdato: 1. juni 2017

Eksamensstid (fra-til): 09.00-13.00

Hjelpekode/Tillatte hjelpemidler: D: Ingen trykte eller håndskrevne hjelpekode tillatt.
Bestemt, enkel kalkulator tillatt.

Annен informasjon:

Målform/språk: Bokmål

Antall sider (uten forside): 3

Antall sider vedlegg: 0

Kontrollert av:

Dato

Sign

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig 2-sidig

sort/hvit farger

Oppgave 1 – Diverse – 10 % (alle deler teller likt)

- Forklar *sidevisning* (pageview) i kontekst av web-bruk-gruvedrift.
- En form for preprosessering i web-bruk-gruvedrift er *sti-fullføring* (path completion). *Hvorfor* må man gjøre dette, og *hvordan*?

Oppgave 2 – Modellering – 15 %

Miljøbomringen AS vil om kort tid få ansvaret for bomstasjonene i alle de store byene i Norge, og ønsker et datavarehus som kan brukes til å analysere trafikk, dvs. passering av bomstasjoner. Som en del av denne reorganiseringen, skal alle biler ha AutoPass (brikke) for automatisk registrering av passering. En kunde kan ha flere biler, og må da ha en brikke for hver bil. Prisen for hver passering endres dynamisk/kontinuerlig for hver stasjon uavhengig av andre, basert på tid på døgnet, forurensing, kø-dannelse, etc.

Eksempel på analyser man skal være i stand til å gjøre mot datavarehuset:

- Antall bom-passeringer for hvert kvartal for hver stasjon.
- Antall bom-passeringer for hvert kvartal for hver bil.
- Gjennomsnittlig antall passeringer per måned.
- Gjennomsnittspris per bil for en bestemt stasjon.

Beskrivelsen er litt upresis og det er en del av oppgaven å velge ut det som skal være med. Vi er først og fremst ute etter at du skal vise modelleringssprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

Lag et stjerne-skjema for denne case-beskrivelsen.

Oppgave 3 – OLAP – 15 % (5 % på a og 10 % på b)

- Forklar *roll-up* og *drill-down*.
- Gitt en dimensjonstabell *Book* i et datavarehus, der vi ønsker å bruke *bitmap-indekser* på attributtene *Language* og *Binding* for å kunne utføre spørninger mer effektivt. Vis struktur og innhold for bitmap-indeksene med utgangspunkt i innholdet i tabellen under.

Book				
RowID	BookID	Title	Language	Binding
1	45	The Hobbit	English	Hardcover
2	63	À la recherche du temps perdu	French	Hardcover
3	88	For Whom the Bell Tolls	English	Paperback
4	143	Madame Bovary	French	Paperback
5	236	La Peste	French	Hardcover
6	463	The Grapes of Wrath	English	Hardcover
7	768	The Great Gatsby	English	Paperback

Oppgave 4 – Klynging – 10 %

Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør klynging ved hjelp av K-means, med $k=3$ og initialsentroider $S1=(4,4)$, $S2=(5,8)$ og $S3=(5,11)$. Bruk Manhattan-distanse som avstandsmål.

	X	Y
P1	4	8
P2	4	10
P3	4	13
P4	5	3
P5	5	7
P6	7	11

Oppgave 5 – Klassifisering – 25 % (10 % på a og 15 % på b)

- a) Anta et datasett med samplene $P1 = (4,8)$, $P2 = (8,8)$, $P3 = (8,4)$, $P4 = (6,7)$, $P5 = (1,10)$, $P6 = (3,6)$, $P7 = (2,4)$, $P8 = (1,7)$, $P9 = (6,4)$, $P10 = (6,2)$, $P11 = (6,3)$, $P12 = (4,3)$, og $P13 = (4,4)$. Samplene hører til de tre klyngene $C1 = \{P1, P2, P3, P4\}$, $C2 = \{P5, P6, P7, P8\}$ og $C3 = \{P9, P10, P11, P12, P13\}$. Anta at klyngen de tilhører er klasse-merkelapp (class label). Klassifiser samplene $A = (6,6)$, $B = (4,6)$, $C = (4,5)$, og $D = (2,6)$ ved å bruke k-nærmeste-nabo-metoden (k-nearest neighbor, k-NN). Bruk Manhattan-distanse og $k = 3$. Forklar hvordan du kommer fram til klassifiseringen av de fire punktene.
- b) Som en del av en større applikasjon ønsker vi å kunne predikere klasse (J eller N) basert på inndata der hver post består av et sekvensnummer og attributtene A , B , C , og D :

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

Anta at vi skal bruke *beslutningstre* ("decision tree") som klassifiseringsmetode. Vi bruker da data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for urenhet ("impurity"), og følgende to formler kan være til hjelp for å løse oppgaven:

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

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

Oppgave: Målet med klassifiseringen er å kunne predikere "Klasse". Regn ut $GAIN_{split}$ for splitting på (1) "A" og (2) "B". Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Oppgave 6 – Assosiasjonsregler – 25 % (10 % på a og 15 % på b)

- a) Anta handlekorg-data som er gitt under. Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Bruk $F_{k-1} \times F_{k-1}$ -metoden for kandidat-generering.

TransaksjonsID Element

TransaksjonsID	Element
T1	ABCDEG
T2	CDFH
T3	AFG
T4	DF
T5	BDEG
T6	BDEG
T7	BCDEGH
T8	ACF

- b) Anta handlekorg-data som er gitt under. Du skal nå bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 40 % (dvs. *minimum support count* er 2).
- 1) Konstruer et FP-tre basert på datasettet.
 - 2) Finn frekvente elementsett ved å bruke FP-growth-algoritmen. Bruk tabell-notasjon med følgende kolonner for å vise resultatet:
 - Element
 - ”Conditional pattern base”
 - ”Conditional FP-tree”
 - Frekvente elementsett

TransaksjonsID Element

TransaksjonsID	Element
T1	ACE
T2	BCE
T3	BCDE
T4	CDE
T5	DE



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: June 1st 2017

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:

Informasjon om trykking av eksamensoppgave

Originalen er:

1-sidig X 2-sidig

sort/hvit X farger

Date

Signature

Problem 1 – Various – 10 % (all having same weight)

- a) Explain *pageview* in context of web usage mining.
- b) One type of pre-processing in web usage mining is *path completion*. *Why* is this necessary, and *how* can it be done?

Problem 2 – Modeling – 15 %

Miljøbomringen AS will soon be responsible for the tollbooths in all major cities in Norway, and want a data warehouse that can be used to analyze traffic, i.e. the toll passages. As part of this reorganization, all cars must have AutoPass (transponder) for automatic registration of passages. A customer may have several cars and must have one transponder for each car. The price for each passage changes dynamically/continuously for each station independently of others, based on time of day, pollution, traffic jams, etc.

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

- Number of passages for each quarter for each station.
- Number of passages for each quarter for each car.
- Average number of passages per month.
- Average price for cars for one particular station.

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.

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

- a) Explain *roll-up* and *drill-down*.
- b) Given a dimension table *Book* in a data warehouse, we want to use *bitmap indexes* on the attributes Language and Binding in order to be able to perform queries more efficiently. Show structure and contents of the bitmap indexes based on the contents in the table below.

Book				
RowID	BookID	Title	Language	Binding
1	45	The Hobbit	English	Hardcover
2	63	À la recherche du temps perdu	French	Hardcover
3	88	For Whom the Bell Tolls	English	Paperback
4	143	Madame Bovary	French	Paperback
5	236	La Peste	French	Hardcover
6	463	The Grapes of Wrath	English	Hardcover
7	768	The Great Gatsby	English	Paperback

Problem 4 – Clustering – 10 %

Assume a two-dimensional dataset as shown in the table to the right. Perform clustering using K-means, with k=3 and initial centroids S1=(4,4), S2=(5,8) and S3=(5,11). Use Manhattan distance.

	X	Y
P1	4	8
P2	4	10
P3	4	13
P4	5	3
P5	5	7
P6	7	11

Problem 5 – Classification – 25 % (10 % on a and 15 % on b)

- a) You are given a dataset of samples P1 = (4,8), P2 = (8,8), P3 = (8,4), P4 = (6,7), P5 = (1,10), P6 = (3,6), P7 = (2,4), P8 = (1,7), P9 = (6,4), P10 = (6,2), P11 = (6,3), P12 = (4,3), and P13 = (4,4). The samples belong to three clusters C1 = {P1,P2,P3,P4}, C2 = {P5,P6,P7,P8} and C3 = {P9,P10,P11,P12,P13}. Consider associated clusters as class labels. Classify the samples A = (6,6), B = (4,6), C = (4,5), and D = (2,6) by employing the k-nearest neighbor (k-NN) method. Use the Manhattan distance metric and k = 3. Describe how the results of the classification are achieved.
- 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, and D:

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

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 “Class”. Compute the $GAIN_{split}$ for splitting by attribute (1) ”A” and (2) ”B”. Which of these splits would you choose to start building your decision tree? Justify your answer.

Problem 6 – Association rules – 25 % (10 % on a and 15 % on b)

- 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	ABCDEG
T2	CDFH
T3	AFG
T4	DF
T5	BDEG
T6	BDEG
T7	BCDEGH
T8	ACF

- 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 40 % (i.e., *minimum support count* is 2).
- 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	ACE
T2	BCE
T3	BCDE
T4	CDE
T5	DE

LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG

TDT4300 – JUNI 2017

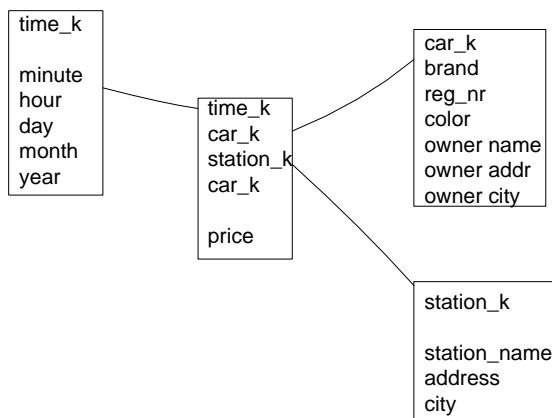
NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element, hovudsakleg laga for at vi skal ha oversikt over arbeidsmengda på eksamen, og som huskeliste under sensurering. Det er også viktig å være klar over at det også kan vere andre svar enn dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving eller dei gjev meining utifrå kva det vert spurde om.

Oppgåve 1

- a) Sidevisning: Samla framstilling av eit sett med Web-objekt som bidrar til ei visning i en nettlese som følgje av ei enkelt brukarhandling (for eksempel ein "clickthrough")
Konseptuelt: Kvar sidevisning kan verte sett på som ein samling av Web-objekt eller ressursar som representerer ein bestemt "brukarhending"
T.d.: Lese ein artikkel, vise ei produktseite eller å legge til eit produkt i handlevogna
- b) Klient- eller proxy-caching kan medføre at ein ikkje har alle referansar i loggen. Kan gjerast vha. referent-informasjon i logg i kombinasjon med kunnskap om lenke-struktur.

Oppgåve 2

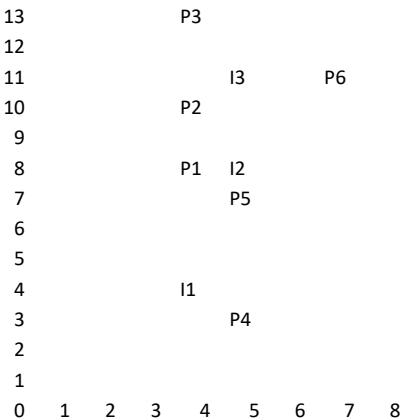
- a) Eit viktig poeng er at fakta-attributt må gje meining utifrå oppgåva, og også gje meining ved aggregering. Med tanke på at oppgåve er relativt triviell er vi tilsvarende strenge, t.d. trekk ved manglende dimensjonstabell, om fakta (pris) er i ein av dimensjonstabellane men ikkje i det heile i faktatabell. Legg merke til at dette skjemaet har meir informasjon enn strengt tatt nødvendig utifrå oppgåveteksta, for car_k er det også OK om dei andre attributtane ikkje er med. Eit viktig poeng er at ein må ha pris og ikkje berre antal i faktatabell. Kan også ha kunde som separat dimensjonstabell.



Oppgåve 3

- a) Roll-up: Roll-up betyr å bevege seg oppover et konsept-hierarki, eller foreta reduksjon av dimensjoner. I praksis aggregering av data, vi går fra detaljerte data til mindre detaljerte.
Drill-down er det motsatte av roll-up: Går et trinn ned i konsephierarkiet eller inkluderer en dimensjon til. Går fra overordnede data til mer spesifikke data.
- b) Binding:
 - Hardcover: 1100110
 - Paperback: 0011001Language:
 - English: 1010011
 - French: 0101100Eit viktig poeng er at dette er indeksar, dvs. ikke en del av "data-tabellene".

Oppgåve 4



- 1) C1=P4, C2=P1, P5, C3=P2, P3, P6
 - 2) S1=(5,3), S2=(4.5,7.5), S3=(5,11,33)
 - 3) C1=P4, C2=P1, P5, C3=P2, P3, P6
- Ingen endring, dvs. terminert!

Oppgåve 5

- a) A=C3 (3 nærmeste er P4/P9/P11 som hører til hhv. C1/C3/C3, flertall for C3) evt. C1 eller C2 om 3 nærmeste er P4/P9/P6
B=C1, C2, eller C3 (3 nærmeste er P1/P6/P13 som hører til hhv. C1/C2/C3, ikke flertall for noen). Også korrekt om vekting på distanse er brukt, da C1 eller C2.
C=C3 (3 nærmeste er P6/P12/P13, hhv. C2/C3/C3).
D=C2 (3 nærmeste er P6/P7/P8, hhv. C2/C2/C2).

b)

Gini i rotnode:

$$p(J|Parent) = 4/15 = 0.26667, p(N|Parent) = 11/15 = 0.73333$$
$$GI(J, N) = 1 - 0.26667 * 0.26667 - 0.73333 * 0.73333 = 0.39111$$

1) Splitting på A:

S1="L"

$$J1=1, N1=6, GI(J1, N1)=GI(1, 6)=1-1/7*1/7-6/7*6/7=0.244898$$

S2="H"

$$J2=3, N2=5, GI(J2, N2)=GI(3, 5)=1-3/8*3/8-5/8*5/8=0.46875$$

$$GAIN(A1) = 0.39111 - 7/15 * 0.244898 - 8/15 * 0.46875 = 0.0268242$$

2) Splitting på B

S1="T"

$$J1=3, N1=1, GI()=1-3/4*3/4-1/4*1/4=0.3750$$

S2="K"

$$J2=0, N2=3, GI()=1-3/3*3/3=0$$

S3="F"

$$J3=1, N3=7, GI()=1-1/8*1/8-7/8*7/8=0.218750$$

$$GAIN (A2) = 0.39111 - 4/15 * 0.3750 - 3/15 * 0 - 8/15 * 0.218750 = 0.17444$$

Vi vel attributtet med høgste GAIN, dvs. **B** vert føretrekt for første splitting av treet.

NB! Viktig å ha med GAIN inkl. $p(J|Parent)$, det kan skje at ein får negativ verdi for begge dei alternative splittingane, som betyr at ein ikkje bør velje nokon av dei.

Oppgåve 6

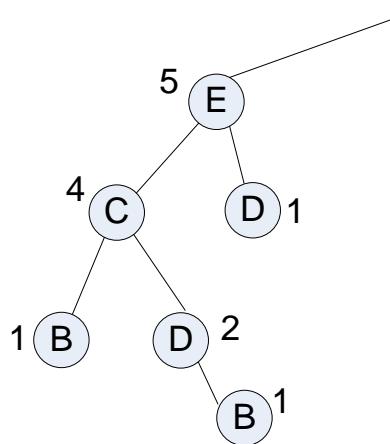
a)

A	3
B	4
C	4
D	6
E	4
F	4
G	5
H	2
BC	2
BD	4
BE	4
BF	0
BG	4
CD	3
CE	2
CF	2
CG	2
DE	4
DF	2
DG	4
EF	0
EG	4
FG	1
BDE	4
BDG	4
BEG	4
DEG	4
BDEG	4

Kun eit 4-elementsett mogleg: BDEG | 4

BDEG	4
------	---

b) Støttetal: A:1, B:2, C:4, D:3, E:5



tid	Itemset	(Ordered) frequent items
T	ACE	EC
T2	BCE	ECB
T3	BCDE	ECDB

T4	CDE	ECD
T5	DE	ED

Item	Conditional sub-database	Conditional FP-tree	Frequent itemsets
B	$\{(EC:1), (ECD:1)\}$	$EC:2$	B, BC, BE, BCE
D	$\{(EC:2), (E:1)\}$	$E3C2$	D, DE, DC, DEC
C	$\{(E:4)\}$	$E4$	C, CE
E	\emptyset	\emptyset	E

(OK om 1-elementsett ikkje er med i tabellen).

Eksamensoppgave i TDT4300 Datavarehus og datagrutedrift

Faglig kontakt under eksamen: Kjetil Nørvåg

Tlf.: 41440433

Eksamensdato: 22. mai 2018

Eksamenstid (fra-til): 1500-1900

Hjelpekode/Tillatte hjelpeemidler: D: Ingen trykte eller håndskrevne hjelpeemiddel tillatt. Bestemt, enkel kalkulator tillatt.

Annен informasjon:

Merk! Studenter finner sensur i Studentweb. Har du spørsmål om din sensur må du kontakte instituttet ditt. Eksamenskontoret vil ikke kunne svare på slike spørsmål.

Oppgave 1 – Diverse – 15 % (alle deler teller likt)

- Hva er binærising, og hvordan bør man gjøre dette?
- Silhouett-koeffisienten er gitt ved følgende formel: $s = (b-a)/\max(a,b)$
Forklar hva denne kan brukes til, og hvordan man regner ut a og b i denne.
- Forklar viktigste begrensninger for bruk av hierarkisk agglomerativ klynging (HAC) på store datasett.

Oppgave 2 – Modellering – 10 %

Ilsvika Elektrisitetsverk (IE) leverer strøm til et stort antall beboere i Trøndelag. Alle abonnenter skal nå få montert "smarte strømmålere", som en gang i minuttet sender en melding til IE om strømforbruk siste minuttet. Med smarte strømmålere er det mulig å tilby dynamisk prising, dvs. prisen kan endre seg fra minutt til minutt, slik at man f.eks. må betale mer for strømmen i perioder med høyt strømforbruk (for eksempel når alle lager middag på ettermiddagen), og mindre når det er lavt strømforbruk (f.eks. om natten). En kunde kan ha strøm-abonnement for mer enn en lokasjon, det er da en strømmåler for hver lokasjon. IE ønsker et datavarehus som kan brukes til å analysere strømforbruk.

Eksempel på analyser man skal være i stand til å gjøre mot datavarehuset:

- Total-forbruk for hver time.
- Total-forbruk for hver time for hver kunde.
- Total-forbruk for hver time for hver lokasjon.
- Totalt-forbruk per døgn per kommune.

Beskrivelsen er litt upresis og det er en del av oppgaven å velge ut det som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

Lag et stjerne-skjema for denne case-beskrivelsen. Svar på papir.

Oppgave 3 – OLAP – 10 % (alle deler teller likt)

- Gitt en base-kuboid har man tre alternative strategier for datacube-materialisering. Forklar disse, og eventuelle fordeler/ulemper for hver av dem.
- Gitt en kube med dimensjoner:

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

Anta følgende materialiserte kuboider:

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

Gitt følgende OLAP-spørring: $\{item_name, country\}$ med vilkår “ $year = 2006$ ”
 Hvilke(n) materialiserte kuboider kan brukes til å prosessere spørringen? Begrunn svaret.

Oppgave 4 – Klynging – 15 % (5 % på a og 10 % på 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) Gitt et d -dimensjonalt datasett med 1000 punkt som man ønsker å klynge vha. DBSCAN, forklar hvordan man kan finne passende verdier for parameterne $MinPts$ og Eps .
- b) Gitt et to-dimensjonalt datasett som vist i tabellen ovenfor. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt $MinPts=4$ (inkl. eget punkt) og $Eps=3$ (inkl. punkt som har distanse 3). Bruk Manhattan-distanse som avstandsmål.

Oppgave 5 – Klassifisering – 20 % (5 % på a og 15 % på b)

Nr	A	B	C	D	E	Klasse
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) Forklar to teknikker for å redusere problem med overtilpasning ("overfitting") i beslutningstre ("decision tree"). Hvilken av disse er vanligvis foretrukket?

- b) Som en del av en større applikasjon ønsker vi å kunne predikere klasse (J eller N) basert på inndata der hver post består av et sekvensnummer og attributtene A, B, C, D, og E, jfr tabellen ovenfor.

Anta at vi skal bruke *beslutningstre* som klassifiseringsmetode. Vi bruker da data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for urenhet ("impurity"), og følgende to formler kan være til hjelp for å løse oppgaven:

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

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

Oppgave: Målet med klassifiseringen er å kunne predikere "Klasse". Regn ut $GAIN_{split}$ for splitting på (1) "A" og (2) "B". Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Oppgave 6 – Assosiasjonsregler (1) – 15 % (10 % på a og 5% på b)

TransaksjonsID Element

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

- a) Anta handlekorg-data som er gitt ovenfor. Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Bruk $F_{k-1} \times F_{k-1}$ -metoden for kandidat-generering.
- b) Gitt følgende lukkede frekvente elementsett (closed frequent itemsets): C:3, AC:2, BE:3, BCE:2 (Format: elementsett:støttetall)
 Finn alle frekvente elementsett og deres støttetall.

Oppgave 7 – Assosiasjonsregler (2) – 10 %

TransaksjonsID Element

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

Anta handlekorg-data som er gitt ovenfor. Du skal nå bruke *FP-growth-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 22 % (dvs. *minimum support count* er 2).

- 1) Konstruer et FP-tre basert på datasettet. Lever dette på papir som oppgave 8.

2) Finn frekvente elementsett ved å bruke FP-growth-algoritmen. Bruk tabell-notasjon med følgende kolonner for å vise resultatet:

- Element
- ”Conditional pattern base”
- ”Conditional FP-tree”
- Frekvente elementsett

Oppgave 8 – FP-tre til oppgave 7 – 5 %

FP-tre til oppgave 7. Svar på papir.

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 $\text{MinPts}=4$ (incl. own point) and $\text{Eps}=3$ (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:

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

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

Task: The goal of the classification is to be able to predict 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.

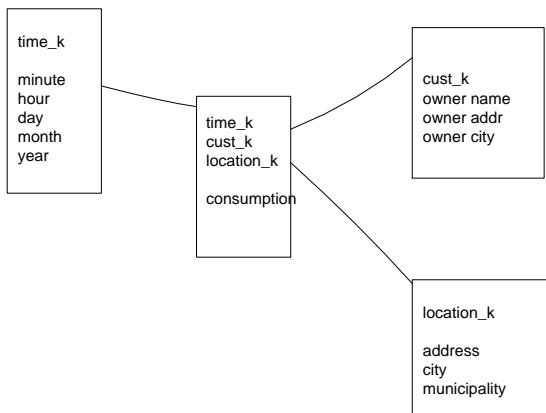
LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – MAI 2018

NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element, hovudsakleg laga for at vi skal ha oversikt over arbeidsmengda på eksamen, og som huskeliste under sensurering. Det er også viktig å være klar over at det også kan vere andre svar enn dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving eller dei gjev meining utifrå kva det vert spurd om.

Oppgåve 1

- Gjere kategoriske attributter til binære.
Ein asymmetrisk binær attributt for kvar moglege verdi av kategorisk attributt
- Brukast til å evaluere godheita til eit *punkt* i ei klynge, og er metrikk for både kor likt eit punkt er andre punkt i klynge og kor ulikt det er punkt i andre klynger.
a = gjennomsnittleg distanse frå i til punkta i klynge til i
b = min(gjennomsnittleg distanse frå i til punkt i ei anna klynge) Reknar ut for kvar klynge j
- Minnekrev $O(N^2)$ pga. matrise, og tid $O(N^3)$ (kompleksitet kan reduserast til $O(N^2 \log(N))$ med adekvate datastrukturar).

Oppgåve 2



Det er også mogleg (men ikkje krav) å ha med pris som ekstra attributt i fakta-tabell. Eit viktig poeng er at fakta-attributt må gje meining utifrå oppgåva, og også gje meining ved aggregering. Med tanke på at oppgåve er relativt triviell er vi tilsvarende strenge, t.d. trekk ved manglende dimensjonstabell, om fakta (consumption) er i ein av dimensjonstabellane men ikkje i det heile i faktatabell.

Oppgåve 3

- Jfr. læreboka (Han) s. 159:
 - No materialization:** Do not precompute any of the “nonbase” cuboids. This leads to computing expensive multidimensional aggregates on-the-fly, which can be extremely slow.
 - Full materialization:** Precompute all of the cuboids. The resulting lattice of computed cuboids is referred to as the *full cube*. This choice typically requires huge amounts of memory space in order to store all of the precomputed cuboids.
 - Partial materialization:** Selectively compute a proper subset of the whole set of possible cuboids. Alternatively, we may compute a subset of the cube, which contains only those cells that satisfy some user-specified criterion, such as where the tuple count of each cell is above some threshold. We will use the term *subcube* to refer to the latter case, where only some of the cells may be precomputed for various cuboids.
For alternativ 3 er problemet å bestemme kva subkuber som skal materialiserast (Han s. 160).
- 1) {year, brand, city} // Nei, (materialisert drill-down frå brand ikkje mogleg)
2) {year, brand, street} // Nei, (materialisert drill-down frå brand ikkje mogleg)
3) {month, brand, province_or_state} // Nei, kan ikkje ta drilldown frå brand til item_name
4) {item_name, province_or_state} where year = 2006 // Ja, rollup til country (og same år i spørjing og kube)

Oppgåve 4

- a) 1) Sorter punkt i høve til distanse til deira k^{te} nærmeste nabo

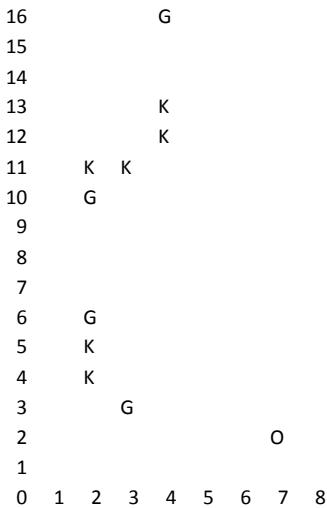
2) Plott distansane

3) Høvande Eps er Eps for "knekkpunktet" der avstand aukar drastisk.

$k(\text{MinPts})$: For to dimensjonar er erfaringsvis $k = 4$ høvande, generell tommelfingerregel er $k \geq D + 1$ der D er dimensjonalitet

Kan tenkast at andre strategiar kan gje "litt poeng", men ingen basert på SSE vil gje mening.

b)



Grensepunkt: Pkt. 3, 4, 6, 10

Støypunkt: Pkt. 11 (evt. også pkt. 12)

Klynger: 1,2,3,6 og 4,5,7,8,9,10

Punktet (7,2) førekjem to gongar i datasettet. På spørsmål under eksamen fekk studentane valet mellom å enten slette eit av dei (som gjev klynginga på figuren ovanfor), eller tolke dei som to distinkte objekt med same koordinat (som gjev to støypunkt i (7,2)).

Oppgåve 5

- a) *Pre-pruning* (Early Stopping Rule)

Stopp før ein har eit "fullgrodd" tre

Meir restriktive vilkår:

- Stopp om Gain ved splitting av noverande node er mindre enn ein gjeve brukar-spesifisert terskel
- Stopp om tal på instansar er mindre enn ein gjeve brukar-spesifisert terskel
- Stopp om klasse-distribusjon til instansar er uavhengig av tilgjenglege "features" (t.d. ved bruk av statistisk test)

Post-pruning

- Konstruer det fullstendige treet
- Fjern noder i beslutningstreet botn-opp
- Om generaliseringsfeil vert redusert, erstatt sub-tre med løvnode
- Klasse-etikett (class label) for den nye løvnoda bestemt av kva som er majoritet-klassen til instansane i sub-treet
-

Ein føretrekk vanlegvis post-pruning i staden for pre-pruning pga.:

- o Uansett relativt billig å konstruere heile treet
- o I praksis vanskeleg å vite parametre for stopp-vilkår i pre-pruning

b)

Gini i rotnode:

$$p(J|\text{Parent}) = 5/15 = 0.3333, p(N|\text{Parent}) = 10/15 = 0.6667$$

$$GI(J, N) = 1 - 0.333 * 0.3333 - 0.6667 * 0.6667 = 0.4444$$

- 1) **Splitting på A:**

S1="L"

$$J1=1, N1=9, GI(J1, N1)=GI(1, 9)=1 - 1/10 * 1/10 - 9/10 * 9/10 = 0.1800$$

S2="H"

J2=4, N2=1, GI(J2, N2)=GI(4,1)= $1-4/5*4/5-1/5*1/5=0.3200$

GAIN(A1) = $0.4444-10/15*0.1800-5/15*0.3200 = 0.2177$

2) Splitting på B

S1="T"

J1=2, N1=1, GI()= $1-2/3*2/3-1/3*1/3=0.4444$

S2="K"

J2=1, N2=3, GI()= $1-1/4*1/4-3/4*3/4=0.375$

S3="F"

J3=2, N3=6, GI()= $1-2/8*2/8-6/8*6/8=0.375$

GAIN (A2) = $0.4444-3/15*0.4444-4/15*0.375-8/15*0.375 = 0.055$

Vi vel attributtet med høgast GAIN, dvs. A vert føretrekt for første splitting av treet.

NB! Viktig å ha med GAIN inkl. p(J|Parent), det kan skje at ein får negativ verdi for begge dei alternative splittingane, som betyr at ein ikkje bør velje nokon av dei.

Oppgåve 6

a) Viktig at det går fram at 3-kandidatsett er generert med bruk av $F_{k-1} \times F_{k-1}$

A	4
B	4
C	3
D	7
E	3
F	2
J	4
K	5
AB	1
AD	4
AJ	1
AK	4
BD	4
BJ	4
BK	4
DJ	4
DK	5
JK	4
ADK	4
BDJ	4
BDK	4
BJK	4
DJK	4
BDJK	4

B) BCE:2

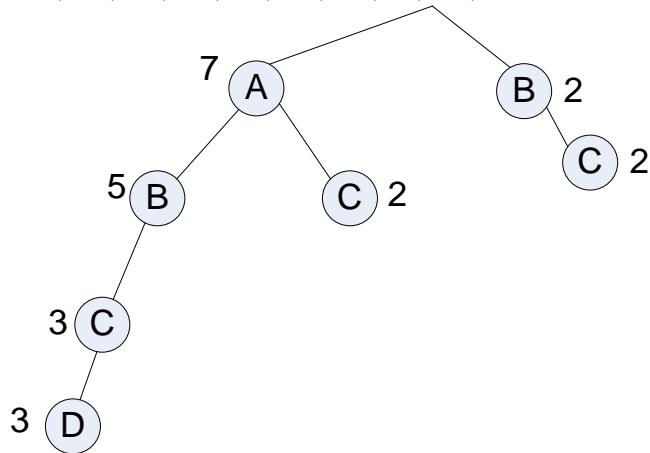
AC:2, BE:3, BC:2, CE:2

C:3, A:2, B:3, E:3

Forventa at ein viser korleis ein har kome fram til desse.

Oppgåve 7/8

Støttetal: A:7, B:7, C:7, D:3, E:1, G:1, H:1, J:1, K:1, L:1.



Item	Conditional sub-database	Conditional FP-tree	Frequent itemsets
D	{(ABC:3)}	<A:3, B:3, C:3>	D3, AD:3, BD:3, CD:3, ABD:3, ACD:3, BCD:3, ABCD:3
C	{(AB:3), (A:2), (B:2)}	<A:5, B:3>, <B:2>	C:7, BC5, AC:5
CB	{(A:3)}	A3	ABC:3
(CA	Ø	Ø)	
B	{(A:5)}	A5	B:7, AB:5
A	Ø	Ø	A:7

OK om 1-elementsett ikke er med i tabellen.

Blir godteke om støttetal på frekvente elementsett manglar.

i Cover Page

Department of Computer Science
Examination paper for TDT4300 Data warehousing and data mining

Academic contact during examination: Zhirong Yang
Phone: 90154911
Examination date: 25-05-2019
Examination time (from-to): 09.00-13.00
Permitted examination support material: D: No tools allowed except an 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.

1 Attribute Type (3 marks)

Which type of attribute is Celsius temperature?

Select one alternative:

- Nominal
- Ratio
- Interval
- Ordinal

Maximum marks: 3

2 Missing values (3 marks)

How can we handle missing values?

Select one or more alternatives:

- Estimate missing values
- Ignore missing values during data analysis
- Treat missing values as zeros
- Eliminate data objects with missing values

Maximum marks: 3

3 Jaccard coefficient (3 marks)

There are two bit vectors **p** and **q**:

$$\mathbf{p} = [1011100111]$$

$$\mathbf{q} = [1001001101]$$

What is the Jaccard coefficient for the bit vectors **p** and **q**? Write your answer here .

Note: the answer is a real-valued number.

Maximum marks: 3

4 CityBlock distance (3 marks)

There are two vectors **p** and **q**:

$$\mathbf{p} = [3, 2, 6]$$

$$\mathbf{q} = [1, 4, 5]$$

What is the CityBlock distance between **p** and **q**? Write your answer here .

Note the answer is a real-valued number.

Maximum marks: 3

5 Modeling (10 marks)

Design the data warehouse for a wholesale furniture company. The data warehouse has to allow to analyze the company's situation at least with respect to the Furniture, Customers and Time. Moreover, the company needs to analyze:

- the furniture with respect to its type (chair, table, wardrobe, cabinet, etc.), category (kitchen, living room, bedroom, bathroom, office, etc.) and material (wood, marble, etc.);
- the customers with respect to their spatial location, by considering at least cities, regions and states.

The company is interested in learning at least the quantity, income and discount of its sales.

One should be able to perform the following example analysis against the data warehouse:

- Total quantity for each month.
- Total quantity for each year.
- Average income for every day for each furniture type.
- Max discount for each category.

Create a star schema for the described case **and define a concept hierarchy for each dimension.**

Note: You have two options for drawing: 1) use a separate paper that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Words: 0

B I U \times_1 \times^2 | $\frac{I}{x}$ | | | | | | |

Maximum marks: 10

6 OLAP (10 marks)

Given a cube with dimensions:

- Time(Day-Month-Year)
- Item(ItemName-Brand)
- Location(Street-City-ProvinceOrState-Country)

Assume the following materialized cuboids:

- {Month, ItemName, City}
- {Month, Brand, Country}
- {Year, Brand, ProvinceOrState}
- {ItemName, City} where year = 2016

Given the following OLAP query: {Brand, City} with condition Month = June 2010, which cuboid(s) should be used? Explain your answer below.

Fill in your answer here

The form consists of a large text area for writing the answer. Above the text area is a toolbar with various editing icons: bold (B), italic (I), underline (U), superscript (x₂), subscript (x²), italic superscript (I_x), italic subscript (I_x), copy (C), paste (P), cut (X), undo (U), redo (R), font size (A), alignment (E), orientation (O), grid (G), and a checkmark icon (checkmark). Below the text area, in the bottom right corner, is the text "Words: 0".

Maximum marks: 10

7 Apriori Algorithm (15 marks)

Assume the market basket data below. Use the Apriori algorithm to find all frequent itemsets with minimum support **33.33%** (i.e. minimum support count is 2).

Transaction ID	Items
T1	H, B, K
T2	H, B
T3	H, C, I
T4	C, I
T5	I, K
T6	H, C, I, U

1. Show how the **frequent itemsets** are generated.
2. $\{H, C, I\}$ is one of the frequent itemsets. Find all association rules based on this set, given confidence threshold $c = 60\%$ (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 $\{H, C, I\}$).

Note: you have two options for drawing: 1) use a separate paper that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | **B** *I* U \times_2 \times^2 | \mathbb{I}_x | | | \approx $\approx\approx$ | Ω | | Σ | ABC |

Words: 0

Maximum marks: 15

8 FP-growth Algorithm (15 marks)

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 22% (i.e., minimum support count is 2).

Transaction ID	Items
T1	b,e,g
T2	b,d,i
T3	b,d,e,f
T4	a,d,e
T5	d,e
T6	b,d,j
T7	b,c,d,e,f
T8	b,d,e,f
T9	b,e,h

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

Note: you have two options for drawing: 1) use a separate paper that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | B I U x₁ x² | T_x | D L | ← → ⟳ | ≡ :: | Ω ■ | ↶ ↷ | Σ | ABC ↴ | ✖

Words: 0

Maximum marks: 15

9 K-means Clustering (15 marks)

Do three iterations of the Lloyd's algorithm for K-means clustering on the 2-dimensional data below. Use $K = 2$ clusters and the initial prototype vectors (i.e. mean vectors) $\mathbf{m}_1 = (2.0, 0, 0)$, $\mathbf{m}_2 = (3.0, 4.0)$. Write down calculation procedure and the cluster memberships as well as mean vectors after each iteration. Draw the data points, cluster means and cluster boundary after each iteration.

t	$\mathbf{x}^{(t)}$
1	(0.0, 3.0)
2	(1.0, 4.0)
3	(3.0, 1.0)
4	(4.0, 2.0)
5	(5.0, 1.0)

Note: you have two options for drawing: 1) use a separate paper, that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | **B** *I* U \mathbb{X} \mathbb{X}^2 | $\mathbb{I}_{\mathbb{X}}$ | | | \vdash \vdash | Ω | | Σ | ABC |

Words: 0

Maximum marks: 15

10 DBSCAN pros and cons (3 marks)

When does DBSCAN probably not perform well?

Select one or more alternatives:

- Clusters have different sizes and shapes.
- Data contains outliers.
- Data is high-dimensional.
- Data has varying densities.

Maximum marks: 3

11 Cross validation (5 marks)

Explain cross validation and what this technique is used for.

Note: if needed, you have two options for drawing: 1) use a separate paper, that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Format Font | **B** *I* U \times_2 \times^2 | $\frac{I}{x}$ | | | | | |

Words: 0

Maximum marks: 5

12 Decision tree (15 marks)

You are going to predict whether mushrooms are edible. You have the following data:

Example	NotHeavy	Smelly	Spotted	Smooth	Edible
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1

D	0	0	0	1	0
E	1	1	1	0	0
F	1	0	1	1	0
G	1	0	0	1	0
H	0	1	0	0	0
U	0	1	1	1	?
V	1	1	0	1	?
W	1	1	0	0	?

For mushrooms A through H, you know whether it is edible (1) or not edible (0), but you do not know about U through W.

You should use ID3 decision tree as a classification method. You will use the examples A through H as the training data. To decide the best split, you need to use **Entropy** for a node t , given by

Entropy(t) = $-\sum_j p(j|t) \log_2 p(j|t)$, where $p(j|t)$ is the probability for class j given node t (i.e. the portion of class j in the node t). For each split, the "information gain" is defined by

GAIN = $\text{Entropy}(p) - \left(\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) \right)$, where n_i is the number of element in node i and n

is the total number of elements in the parent node p .

For Tasks 1 and 2, consider only mushrooms A through H. Tasks:

1. Which attribute should you choose as the root of a decision tree? Justify your choice by calculating the information gains of the attributes.
 2. Build an ID3 decision tree to classify mushrooms as edible or not.
 3. Classify mushroom U, V, and W using the decision tree to be edible or not edible.

Note: if needed, you have two options for drawing: 1) use a separate paper that will be scanned after the exam, 2) use the "Insert Drawing" tool in the toolbar.

Fill in your answer here

Maximum marks: 15

1 Attribute Type

Answer: interval

2 Missing values

Answer:

- Eliminate data objects with missing values
- Estimate missing values
- Ignore missing values during data analysis

3 Jaccard coefficient

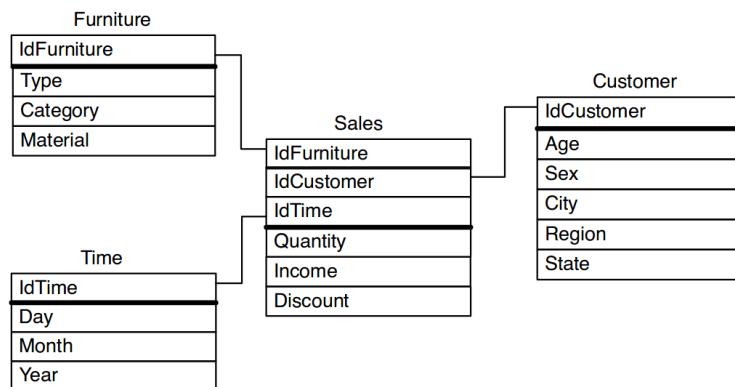
Answer: 0.5

4 CityBlock distance

Answer: 5

5 Modeling

Answer:



6 OLAP

Answer:

- {Month, ItemName, City}: Yes, after roll-up from ItemName to Brand

- {Month, Brand, Country}: No, cannot retrieve city in the query
- {Year, Brand, City}: No, cannot retrieve month and city in the query
- {ItemName, City} where year = 2016: No, cannot retrieve the specified month in the query

7 Apriori Algorithm

Answer:

- 1) Applying Apriori

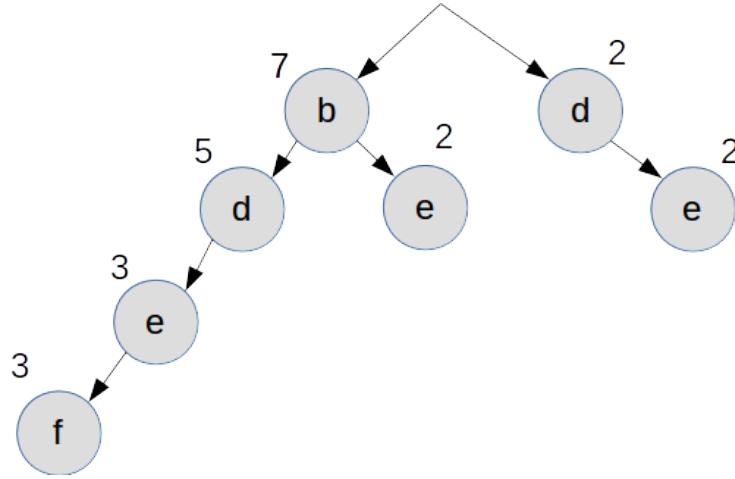
Pass(k)	Candidate k-itemsets and their support	Frequent k-itemsets
k=1	{H}(4), {B}(2), {K}(2), {C}(3), {I}(4)	{H}, {B}, {K}, {C}, {I}
k=2	{H, B}(2), {H, K}(1), {H, C}(2), {H, I}(2), {B, K}(1), {B, C}(0), {B, I}(0), {K, C}(0), {K, I}(1), {C, I}(3)	{H, B}, {H, C}, {H, I}, {C, I}
k=3	{H, C, I}(2)	{H, C, I}
k=4	{}	
2)	$\{H\} \rightarrow \{C, I\}$ (confidence=2/4=0.5) $\{C\} \rightarrow \{H, I\}$ (confidence=2/3=0.66) $\{I\} \rightarrow \{H, C\}$ (confidence=2/4=0.5) $\{H, C\} \rightarrow \{I\}$ (confidence=2/2=1) $\{H, I\} \rightarrow \{C\}$ (confidence=2/2=1) $\{C, I\} \rightarrow \{H\}$ (confidence=2/3=0.66)	

Therefore, the four qualified association rules are $\{C\} \rightarrow \{H, I\}$, $\{H, C\} \rightarrow \{I\}$, $\{H, I\} \rightarrow \{C\}$, and $\{C, I\} \rightarrow \{H\}$.

8 FP-growth

Answer:

- 1)



Item	Conditional sub-database	Conditional FP-tree	Frequent Item sets
f	{b,d,e}:3	$\langle b : 3, d : 3, e : 3 \rangle$	{f}:3, {b,f}:3, {d,f}:3, {e,f}:3, {b,d,f}:3, {b,e,f}:3, {d,e,f}:3, {b,d,e,f}:3
e	{b,d}:3, {b}:2, {d}:2	$\langle b : 5, d : 3 \rangle, \langle d : 2 \rangle$	{e}:7, {d,e}:5, {b,e}:5
ed	{b}:3	b:3	{b,d,e}:3
(eb	empty	empty)	
d	{b}:5	b:5	{d}:7, {b,d}:5
b	empty	empty	{b}:7

9 K-means clustering

Step 1a: compute squared Euclidean distances between data points and mean vectors

	$m_1 = (2, 0)$	$m_2 = (3, 4)$
$x^{(1)}$ = (0, 3)	$(0-2)^2 + (3-0)^2 = 13$	$(0-3)^2 + (3-4)^2 = 10$
$x^{(2)}$ = (1, 4)	$(1-2)^2 + (4-0)^2 = 17$	$(1-3)^2 + (4-4)^2 = 4$
$x^{(3)}$ = (3, 1)	$(3-2)^2 + (1-0)^2 = 2$	$(3-3)^2 + (1-4)^2 = 9$
$x^{(4)}$ = (4, 2)	$(4-2)^2 + (2-0)^2 = 8$	$(4-3)^2 + (2-4)^2 = 5$
$x^{(5)}$ = (5, 1)	$(5-2)^2 + (1-0)^2 = 10$	$(5-3)^2 + (1-4)^2 = 13$

Pick the smallest in each row, which results that $x^{(1)}$, $x^{(2)}$, and $x^{(4)}$ belong to Cluster 2, and $x^{(3)}$, and $x^{(5)}$ belong to Cluster 1

Step 1b: update mean vectors

$$m_1 = (x^{(3)} + x^{(5)})/2 = ((3, 1) + (5, 1))/2 = (4, 1)$$

$$m_2 = (x^{(1)} + x^{(2)} + x^{(4)})/3 = ((0, 3) + (1, 4) + (4, 2))/3 = (5/3, 3) \text{ or } (1.67, 3)$$

3)

Step 2a: compute squared Euclidean distances between data points and mean vectors

	$m_1 = (4, 1)$	$m_2 = (5/3, 3)$
$x^{(1)} = (0, 3)$	$(0-4)2+(3-1)2=20$	$(0-5/3)2+(3-3)2=2+7/9$ (or 2.78)
$x^{(2)} = (1, 4)$	$(1-4)2+(4-1)2=18$	$(1-5/3)2+(4-3)2=1+4/9$ (or 1.44)
$x^{(3)} = (3, 1)$	$(3-4)2+(1-1)2=1$	$(3-5/3)2+(1-3)2=5+7/9$ (or 5.78)
$x^{(4)} = (4, 2)$	$(4-4)2+(2-1)2=1$	$(4-5/3)2+(2-3)2=6+4/9$ (or 6.44)
$x^{(5)} = (5, 1)$	$(5-4)2+(1-1)2=1$	$(5-5/3)2+(1-3)2=15+1/9$ (or 15.11)

Pick the smallest in each row, which results that $x^{(1)}$ and $x^{(2)}$ belong to Cluster 2, and $x^{(3)}$, $x^{(4)}$, and $x^{(5)}$ belong to Cluster 1

Step 2b: update mean vectors

$$m_1 = (x(3)+x(4)+x(5))/3 = ((3, 1)+(4, 2)+(5, 1))/3 = (4, 4/3) \text{ or } (4, 1.33)$$

$$m_2 = (x(1)+x(2))/2 = ((0, 3)+(1, 4))/2 = (0.5, 3.5)$$

Step 3a: compute squared Euclidean distances between data points and mean vectors

	$m_1 = (4, 4/3)$	$m_2 = (1/2, 7/2)$
$x^{(1)} = (0, 3)$	$(0-4)2+(3-4/3)2=18+7/9$ (or 18.78)	$(0-0.5)2+(3-3.5)2=0.5$
$x^{(2)} = (1, 4)$	$(1-4)2+(4-4/3)2=16+1/9$ (or 16.11)	$(1-0.5)2+(4-3.5)2=0.5$
$x^{(3)} = (3, 1)$	$(3-4)2+(1-4/3)2=1+1/9$ (or 1.11)	$(3-0.5)2+(1-3.5)2=12.5$
$x^{(4)} = (4, 2)$	$(4-4)2+(2-4/3)2=4/9$ (or 0.44)	$(4-0.5)2+(2-3.5)2=14.5$
$x^{(5)} = (5, 1)$	$(5-4)2+(1-4/3)2=1+1/9$ (or 1.11)	$(5-0.5)2+(1-3.5)2=26.5$

Pick the smallest in each row, which results that $x^{(1)}$ and $x^{(2)}$ belong to Cluster 2, and $x^{(3)}$, $x^{(4)}$, and $x^{(5)}$ belong to Cluster 1.

Since there is no change in the cluster assignment, the algorithm ends and outputs

$$m_1 = (4, 4/3) \text{ or } (4, 1.33)$$

$$m_2 = (0.5, 3.5)$$

$$\text{Cluster}(x^{(1)})=2$$

$$\text{Cluster}(x^{(2)})=2$$

$$\text{Cluster}(x^{(3)})=1$$

$$\text{Cluster}(x^{(4)})=1$$

$$\text{Cluster}(x^{(5)})=1$$

10 DBSCAN pros and cons

Answer:

- Data is high-dimensional
- Data has varying densities

11 Cross-validation

Purpose of cross validation. One of the following answers or the like is acceptable:

- Cross-validation can be used to evaluate the performance of a supervised model (e.g. a classifier or regressor)
- Cross-validation is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.
- Cross-validation can be used to tune hyper-parameters in algorithms or models

Cross validation procedure. It requires the following parts in the answer

- has explained training/validation/test sets
- has correctly show K-fold rotation steps

12 Decision tree

Answer:

1)

$$\text{Entropy}(p) = -\frac{3}{8} \cdot \log_2 \frac{3}{8} - \frac{5}{8} \cdot \log_2 \frac{5}{8} \approx 0.9544$$

If using "NotHeavy" for root splitting,

$$\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) = \frac{5}{8} \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) + \frac{3}{8} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) \approx 0.9512$$

$$\text{GAIN}_{\text{NotHeavy}} = \text{Entropy}(p) - \sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) = 0.0032$$

If using "Smelly" for root splitting,

$$\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) = \frac{3}{8} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) + \frac{5}{8} \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) \approx 0.9512$$

$$\text{GAIN}_{\text{Smelly}} = \text{Entropy}(p) - \sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) = 0.0032$$

If using "Spotted" for root splitting,

$$\sum_{i=1}^k \frac{n_i}{n} \text{Entropy}(i) = \frac{3}{8} \left(-\frac{1}{3} \log_2 \frac{1}{3} - \frac{2}{3} \log_2 \frac{2}{3} \right) + \frac{5}{8} \left(-\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{3}{5} \right) \approx 0.9512$$

$$\text{GAIN}_{\text{Spotted}} = \text{Entropy}(p) - \sum_{i=1}^k \frac{n_i}{n} \text{ Entropy } (i) = 0.0032$$

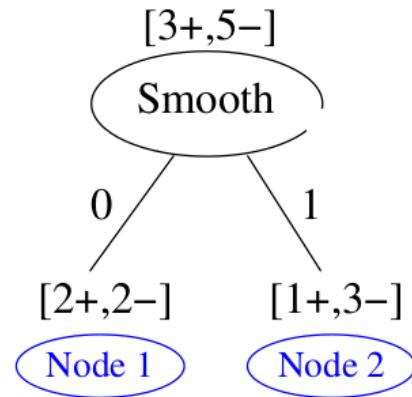
If using "Smooth" for root splitting,

$$\sum_{i=1}^k \frac{n_i}{n} \text{ Entropy } (i) = \frac{4}{8} \left(-\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} \right) + \frac{4}{8} \left(-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \log_2 \frac{2}{4} \right) \approx 0.9056$$

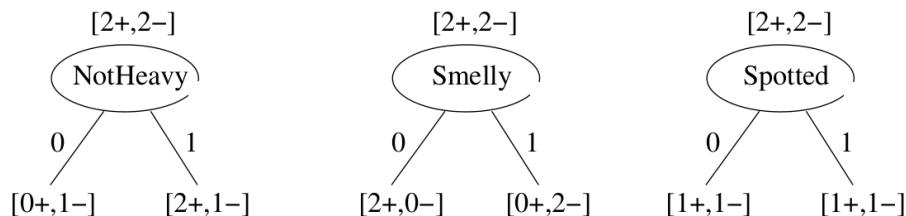
$$\text{GAIN}_{\text{Smooth}} = \text{Entropy}(p) - \sum_{i=1}^k \frac{n_i}{n} \text{ Entropy } (i) = 0.0488$$

So we should use "Smooth" for the root splitting because its GAIN is the largest.

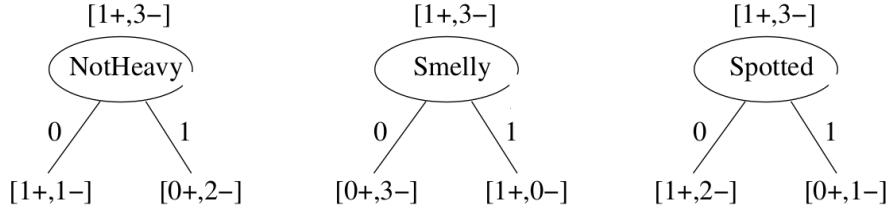
2)



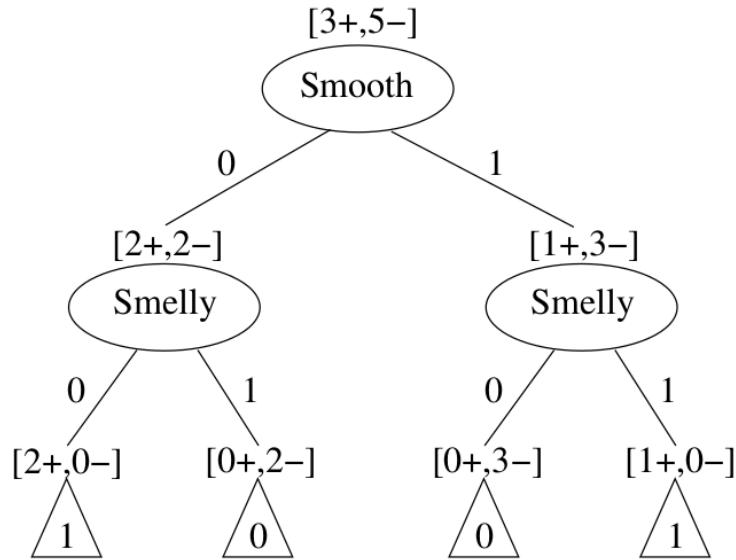
Node 1: Smooth = 0



Node 2: Smooth = 1



It can be seen that after splitting with “Smooth” and then “Smelly”, all training samples have been classified. The Entropy after the two-level decision becomes zero (i.e. giving the maximum GAIN). The GAINs using other features are smaller. Therefore the resulting decision tree is



3)

- For U: Smooth = 1, Smelly = 1 \Rightarrow Edible = 1
- For V: Smooth = 1, Smelly = 1 \Rightarrow Edible = 1
- For W: Smooth = 0, Smelly = 1 \Rightarrow Edible = 0

i Framside

Institutt for datateknologi og informatikk

Eksamensoppgave i TDT4300 Datavarehus og datagrudevdrift

Eksamensdato: 29. mai 2020

Eksamensstid (fra-til): 09:00 – 13:00

Hjelpekode/Tillatte hjelpeemidler: A / Alle hjelpeemidler tillatt

Faglig kontakt under eksamen:

Tlf.: 41 44 04 33

Teknisk hjelp under eksamen: [NTNU Orakel](#)

Tlf: 73 59 16 00

ANNEN INFORMASJON:

Gjør dine egne antagelser og presiser i besvarelsen hvilke forutsetninger du har lagt til grunn i tolkning/avgrensing av oppgaven. Faglig kontaktperson skal kun kontaktes dersom det er direkte feil eller mangler i oppgavesettet.

Lagring: Besvarelsen din i Inspera Assessment lagres automatisk. Jobber du i andre programmer – husk å lagre underveis.

Juks/plagiat: Eksamens skal være et individuelt, selvstendig arbeid. Det er tillatt å bruke hjelpeemidler. Alle besvarelser blir kontrollert for plagiat. [Du kan lese mer om juks og plagiering på eksamen her.](#)

Varslinger: Hvis det oppstår behov for å gi beskjeder til kandidatene underveis i eksamen (f.eks. ved feil i oppgavesettet), vil dette bli gjort via varslinger i Inspera. Et varsel vil dukke opp som en dialogboks på skjermen i Inspera. Du kan finne igjen varselet ved å klikke på bjella øverst i høyre hjørne på skjermen. Det vil i tillegg bli sendt SMS til alle kandidater for å sikre at ingen går glipp av viktig informasjon. Ha mobiltelefonen din tilgjengelig.

Vektning av oppgavene: Som vist i oppgavesettet. Alle deloppgaver innenfor en oppgave teller likt.

OM LEVERING:

Besvarelsen din leveres automatisk når eksamenstida er ute og prøven stenger, forutsatt at minst én oppgave er besvart. Dette skjer selv om du ikke har klikket «Lever og gå tilbake til Dashboard» på siste side i oppgavesettet. Du kan gjenåpne og redigere besvarelsen din så lenge prøven er åpen. Dersom ingen oppgaver er besvart ved prøveslutt, blir ikke besvarelsen din levert.

Trek fra eksamen: Ønsker du å levere blankt/trekke deg, gå til hamburgermenyen i øvre høyre hjørne og velg «Lever blankt». Dette kan ikke angres selv om prøven fremdeles er åpen.

Tilgang til besvarelse: Du finner besvarelsen din i Arkiv etter at sluttida for eksamen er passert.

1 1

Oppgave 1 – Modellering og OLAP – 20 %

- a. Sykkemat (SM) leverer mat fra restauranter til kunder i flere byer. Hver restaurant har et sett med retter de tilbyr, og når kunden har bestilt maten på nett blir den levert med sykkelbud til kunden kort tid etterpå. SM ønsker et datavarehus som kan brukes til å analysere og optimalisere tjenesten.

Eksempel på analyser man skal være i stand til å gjøre mot datavarehuset:

- Totalt antall leveranser per dag
- Totalt antall leveranser per restaurant per dag
- Gjennomsnittlig pris på hver levering
- Antall kunder per by som bestilte mat 12. april 2020

Beskrivelsen er litt upresis og det er en del av oppgaven å velge ut det som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

Lag et stjerne-skjema for denne case-beskrivelsen. Siden det ikke er anledning til å levere figurer, ber vi om at dere i stedet leverer tabellene og deres attributter, og angir hva som er fakta- og hva som er dimensjonstabeller. Eksempel på hvordan angi tabell: TabellA(key, attrib_a, attrib_b).

- b. Gitt en kube med dimensjoner og tilhørende konsept hierarki:

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

Anta følgende materialiserte kuboider:

- 1) {year, brand}
- 2) {year, item_name, street}
- 3) {item_name, country} where year = 2006

Gitt følgende OLAP-spørring: {item_name, city} med vilkår "year = 2006"
 Hvilke(n) materialiserte kuboider kan brukes til å prosessere spørringen? Begrunn svaret.

Skriv ditt svar her...

Words: 0

Maks poeng: 20

2

Oppgave 2 – Klynging og klyngingsvalidering – 30 %

PointID	X	Y
P1	2	4
P2	2	5

P3	2	10
P4	2	11
P5	2	16
P6	3	3
P7	3	10
P8	3	11
P9	4	3

- a. Gitt et datasett som vist i tabellen ovenfor, der første kolonne er punkt-identifikator, og kolonnene X og Y er numeriske verdier. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt $MinPts=4$ (inkl. eget punkt) og $Eps=3$ (inkl. punkt som har distanse 3). Bruk Manhattan-distanse som avstandsmål.
- b. Gitt datasettet under, der vi allerede har utført klynging (basert på X og Y) og endt opp med 3 klynger (jfr. tilhørende klynge-identifikator for hvert punkt). Regn ut Silhouett-koeffisienten for punkt P2.

PointID	X	Y	ClusterID
P1	1	1	C1
P2	2	1	C1
P3	2	4	C2
P4	2	5	C2
P5	4	1	C3
P6	5	1	C3

Skriv ditt svar her...

Format ▼ | **B** *I* U \times_2 \times^2 | \mathbb{I}_x | | | | | | Σ | |

Words: 0

Maks poeng: 30

3

Oppgave 3 – Klassifisering – 20 %

Nr	A	B	C	D	Klasse
1	L	K	S	2	J

2	H	F	S	4	J
3	H	T	H	4	J
4	L	F	S	2	N
5	L	F	H	5	N
6	H	T	G	2	N
7	L	F	S	6	N
8	L	K	G	4	N
9	H	T	S	4	J
10	L	F	S	5	N
11	L	K	H	7	N
12	H	F	G	9	J
13	L	K	G	2	N
14	L	F	H	1	J
15	L	F	H	7	N

Som en del av en større applikasjon ønsker vi å kunne predikere klasse (J eller N) basert på inndata der hver post består av et sekvensnummer og attributtene A, B, C, og D, jfr tabellen ovenfor.

Anta at vi skal bruke *beslutningstre* som klassifiseringsmetode. Vi bruker da data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for urenhet ("impurity").

Oppgave: Målet med klassifiseringen er å kunne predikere "Klasse". Regn ut $GAIN_{split}$ for splitting på (1) "A" og (2) "C". Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Skriv ditt svar her...

Format ▼ | **B** *I* U \mathbf{x}_1 \mathbf{x}_2 \mathbf{x}^2 | $\mathbf{I}_{\mathbf{x}}$ | □ □ | ◀ ▶ ✖ | := :: | Ω ■■■ | ✎ | Σ | ABC ▼ | ☒

Words: 0

Maks poeng: 20

4 4

Oppgave 4 – Assosiasjonsregler – 30 %

TransactionID	Element
T1	BF
T2	ABCDFH
T3	ABF
T4	ABFH
T5	ADEF
T6	ABFH
T7	ABDEFH
T8	AGH

- a. Anta handlekorg-data som er gitt ovenfor. Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Bruk $F_{k-1} \times F_{k-1}$ -metoden for kandidatgenerering.
- b. Et av de frekvente elementsettene er ABH. Finn alle assosiasjonsregler basert på dette settet, gitt konfidens på 75 % (det er ikke nødvendig å bruke apriori til å finne assosiasjonsreglene, men vis hvordan konfidens blir regnet ut for hver av kandidatreglene som er basert på ABH).

Skriv ditt svar her...

Format ▼ | **B** *I* U \times_a \times^a | \mathbb{I}_x | | | $=$ \approx | Ω | | Σ | |

Words: 0

Maks poeng: 30

LØSNINGSSKISSE TIL EKSAMENSOPPGAVE I FAG TDT4300 – MAI 2020

NB! Dette er ikkje fullstendige løysingar på oppgåvene, kun skisse med viktige element, hovudsakleg laga for at vi skal ha oversikt over arbeidsmengda på eksamen, og som huskeliste under sensurering. Det er også viktig å være klar over at det også kan vere andre svar enn dei som er gjeve i skissa som vert rekna som korrekt om ein har god grunngjeving eller dei gjev meining utifrå kva det vert spurd om.

Oppgave 1 – Modellering og OLAP – 20 %

- a) Her er det flere mulige svar utifra forutsetningene man gjør, her er en variant:

Faktatabell:

Delivery(time_k, cust_k, rest_k, pris)

Dimensjonstabeller:

Time(time_k, minute, hour, day, month, year)

Customer(cust_k, cust_name, address, city, email)

Restaurant(rest_k, rest_name, address, city)

Legg merke til at det ikke er sagt noe i oppgaven om (for eksempel) pris per rett etc. Det er ikke galt å ha med dette, men ikke nødvendig. Et viktig moment er forståelse for hva som skal være i faktatabell, og hva som skal være i dimensjonstabeller.

b)

- 1) Nei, lokasjon mangler
- 2) Ja, rollup av street til city, og seleksjon på 2006
- 3) Nei, kan ikke gjøre drill-down til city på materialisert kube

Oppgave 2 – Klynging og klyngingsvalidering – 30 %

- a) Svar: 2 klynger (P1,P2,P6,P9) og (P3,P4,P7,P8), en outlier som er P5

b) $a=1, b=\min((3+4)/2, (2+3)/2)=\min(3.5, 2.5) = 2.5$
 $s = (b-a)/\max(a,b)=(2.5-1)/\max(2.5,1)=1.5/2.5=0.6$

Oppgave 3 – Klassifisering – 20 %

Gini i rotnode:

$$p(J|Parent) = 6/15 = 0.4, p(N|Parent) = 9/15 = 0.6$$
$$GI(J, N) = 1 - 0.4 * 0.4 - 0.6 * 0.6 = 1 - 0.16 - 0.36 = 0.48$$

1) **Splitting på A:**

S1="L"

$$J1=2, N1=8, GI(J1, N1)=GI(2,8)=1-2/10*2/10-8/10*8/10=0.32$$

S2="H"

$$J2=3, N2=1, GI(J2, N2)=GI(4,1)=1-4/5*4/5-1/5*1/5=0.32$$

$$GAIN(A) = 0.48 - 10/15 * 0.32 - 5/15 * 0.32 = 0.16$$

2) Splitting på C

S1="G"

J1=1, N1=3, GI() $=1-1/4*1/4-3/4*3/4=0.375$

S2="H"

J2=2, N2=3, GI() $=1-2/5*2/5-3/5*3/5=0.48$

S3="S"

J3=3, N3=3, GI() $=1-3/6*3/6-3/6*3/6=0.5$

$$\text{GAIN (C)} = 0.48 - \frac{4}{15} * 0.375 - \frac{5}{15} * 0.48 - \frac{6}{15} * 0.5 = 0.02$$

Vi vel attributtet med størst GAIN, dvs. A blir foretrukket for første splitting av treet.

Oppgave 4 – Assosiasjonsregler – 30 %

a)

C1: A:7, B:6, C:1, D:3, E:2, F:7, G:1, H:5

F1: A:7, B:6, F:7, H:5

C2: AB:5, AF:6, AH:5, BF:6, BH:4, FH:4

F2: AB:5, AF:6, AH:5, BF:6, BH:4, FH:4

C3: ABF:5, ABH:4, AFH:4, BFH:4

F3: ABF:5, ABH:4, AFH:4, BFH:4

C4: ABFH: 4

F4: ABFH: 4

b)

A->BH	4/7	0.57	
AB->H	4/5	0.8	*
B->AH	4/6	0.67	
BH->A	4/4	1.0	*
H->AB	4/5	0.8	*
AH->B	4/5	0.8	*