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 answering 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.

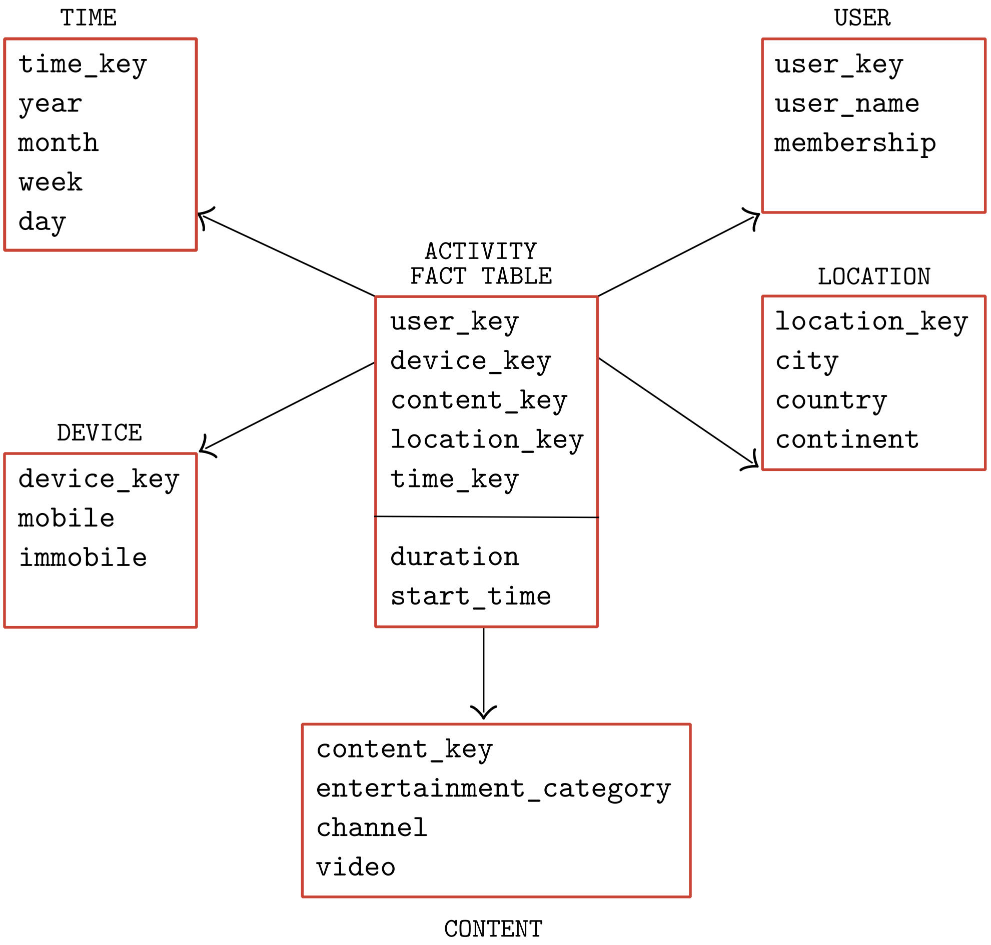
To help YouTube create their data warehouse answer the following questions. For the idea to be implementable, explain any assumptions you have made. Also, explain how the dimensions created as part of the modelling process are related to the quantities being measured and tracked.

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)

```
ROLL-UP location     : city → country;
ROLL-UP time         : day → week;
ROLL-UP content      : video → Entertainment_Category;
DICE                 : country = "Norway"AND
                     : Entertainment_Category = "Music".
```

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 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

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

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

- 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

$$Answer = \{\text{All except } 2\}$$

$$Answer = \{\text{All except } \texttt{Pfizer BioNTech}\}$$

- 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

$$Answer = \{4,6,8,10\}$$

$$Answer = \{\texttt{BBIBP-SorV}, \texttt{CoraonaVac}, \texttt{Ad5-nCov}, \texttt{ZF2001}\}$$

- Answer to part 4.

`Type = "Vector"` AND $\Big($`Origin = "US"` OR `Origin = "India"`$\Big)$.

	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

$$Answer = \{5\}$$
$$Answer = \{\texttt{Johnson \& Johnson}\}$$

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 $1, 2, \ldots, 50$. 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 a have 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 $t_2 \geqslant t_1$. 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 ASSOCATION 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 FPGrowth 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 FPGrowth 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

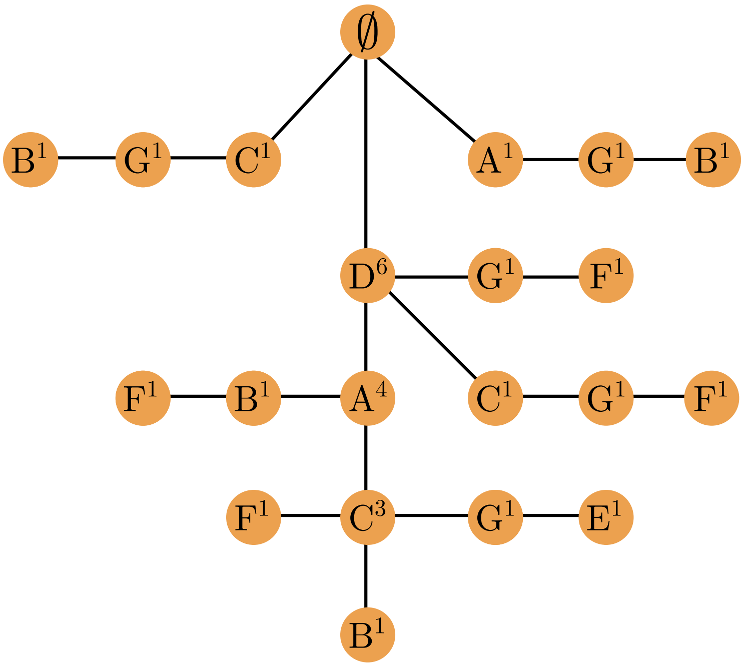
- 1-itemset reordered based on support values.

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**.**

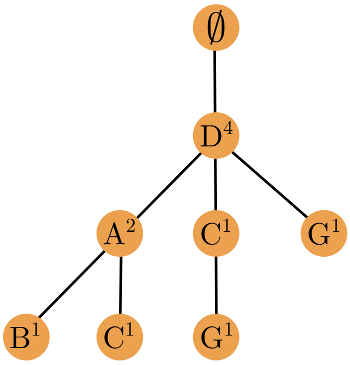
Path	Count
DACGE	1

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: $\emptyset \to D^2 \to C^1$.
Frequent Pattern = $\{FG, FGD\}$.

2.3 **Project on** FC.

Path	Count
DC	1
DAC	1

– Projected FP-Tree for FC: $\emptyset \to D^2 \to A^1$.
Frequent Pattern = $\{FC, FCD\}$.

2.4 **Project on** FA.

Path	Count
DA	2

– Projected FP-Tree for FA: $\emptyset \to D^2$.
Frequent Pattern = $\{FA, FAD\}$.

2.5 **Project on** FD.

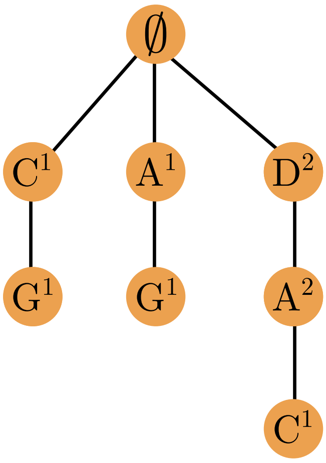
Path	Count
D	4

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 = $\{BD\}$.

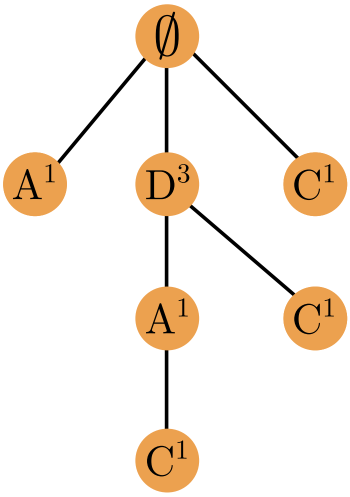
3.4 **Project on** BA.

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}	⟨D:1, A:1, C:1, G:1, E:1⟩	∅
F	{D,C,G,F:1}, {D,A,C,F:1}, {D,G,F:1}, {D,A,B,F:1}	⟨D:4, G:1⟩, ⟨D:4, C:1, G:1⟩, ⟨ D:4, A:2, C:1 ⟩, ⟨ D:4, A:2, B:1 ⟩	-
FB	{D,A,B:1}	⟨D:1, A:1, B:1⟩	∅
FG	{D,G:1}, {D,C,G:1}	⟨D:2, C:1⟩	{FG, FGD}
FC	{D,C:1}, {D,A,C:1}	⟨D:2, A:1⟩	{FC, FCD}
FA	{D,A:2}	⟨D:2⟩	{FA, FAD}
FD	{D:4}	⟨∅ : 4⟩	{FD}
B	{AGB:1}, {CGB:1}, {DAB:1}, {DACB:1}	⟨C:1, G:1⟩, ⟨A:1, G:1⟩, ⟨D:2, A:2, C:1⟩	-
BC	{C:1}, {DAC:1}	⟨ D:1, A:1 ⟩	{BC}
BG	{CG:1}, {AG:1}	⟨ C:1 ⟩, ⟨ A:1⟩	{BG}
BD	{D:2}	⟨∅ : 2⟩	{BD}
BA	{A:1, DA:2}	⟨ D:2 ⟩	{BA, BAD}
G	{CG:1}, {AG:1}, {DG:1}, {DCG:1}, {DACG:1}	⟨C:1⟩, ⟨D:3, C:1⟩, ⟨D:3,A:1,C:1⟩, ⟨A:1⟩	-
GA	{A:1}, {DA:1}	⟨D:1⟩	{GA}
GC	{C:1}, {DC:1}, {DAC:1}	⟨D:2, A:1⟩	{GC, GCD}
GD	{D:3}	⟨D:3⟩	{GD}
C	{DAC:3}, {C:1}, {DC:1}	⟨ D:4, A:3 ⟩	{CA, CD, CAD}
A	{DA:4}, {A:1}	⟨ D:4 ⟩	{AD}
D	{D:6}	⟨∅ : 6⟩	{D}

Table 1: Summarizing the FP-Growth algorithm results.

4 CLUSTERING

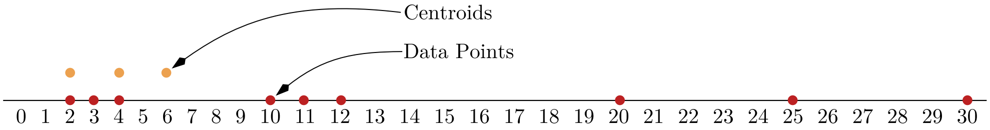
Exercise 8. **Total Marks: 5 Marks**
Apply the K-means algorithm in one-dimension for the data points: 2 , 4 , 10 , 12 , 3 , 20 , 30 , 11 , 25.

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 you 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

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



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., L1-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.

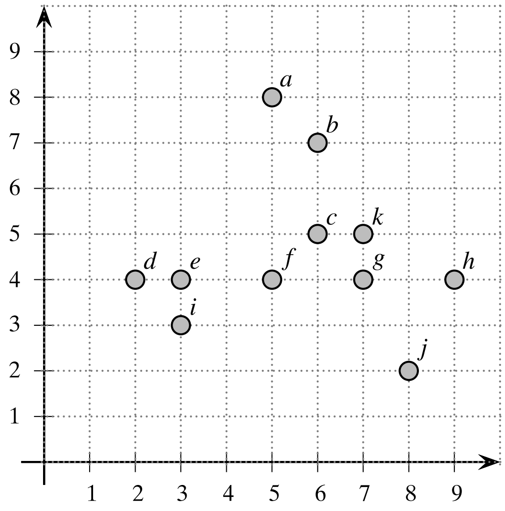


Figure 1: Figure for the hierarchical agglomerative clustering.

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 $\{c, k, g\}$

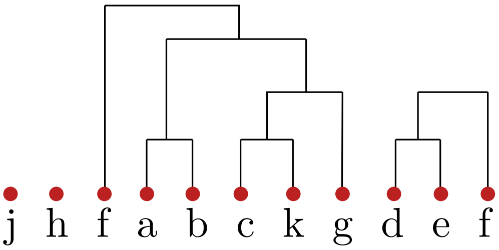
	a,b,c,k,g	d,e,i	f	h	j
a,b,c,k,g					
d,e,i	4				
f	2				
h	2				
j	3				

- 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 dendogram 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

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

$$\begin{aligned}\text{Gini} &= \frac{4}{9} \cdot \frac{6}{16} + \frac{5}{9} \cdot \frac{8}{25} \\ &= \frac{1}{6} - \frac{8}{45} \\ &= 0.3\bar{4}.\end{aligned}$$

- 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

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

$$\begin{aligned}\text{Gini} &= \frac{5}{9} \cdot \frac{12}{25} + \frac{4}{9} \cdot \frac{1}{2}\\ &= \frac{4}{15} + \frac{2}{9} = \frac{22}{45}\\ &= 0.4\bar{8}.\end{aligned}$$

- 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	
	$\leqslant$	$>$	$\leqslant$	$>$	$\leqslant$	$>$	$\leqslant$	$>$	$\leqslant$	$>$	$\leqslant$	$>$	$\leqslant$	$>$	$\leqslant$	$>$
Class=’Y’	0	4	0	4	1	3	1	3	2	2	3	1	4	0	4	0
Class=’N"	4	5	1	4	1	4	2	3	3	2	3	2	4	1	5	0
Gini Index	0.49		$0.\bar{4}$		0.4921		0.481		$0.4\bar{8}$		0.481		$0.\bar{4}$		$0.4\bar{9}$	

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

$a_3 \leqslant \frac{1}{2}$	
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:

$a_3 \leqslant 2$	
Y	0
N	1

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

$a_3 > 2$	
Y	4
N	4

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

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

- Split on comparing a_3 with 3.5:

$a_3 \leqslant 3.5$	
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:

$a_3 \leqslant 4.5$	
Y	1
N	2

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

$a_3 > 4.5$	
Y	3
N	3

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

$$\begin{aligned} \text{Gini} &= \frac{3}{9} \cdot \frac{4}{9} + \frac{6}{9} \cdot \frac{1}{2} \\ &= \frac{4}{3 \cdot 9} + \frac{3^2}{3 \cdot 9} = \frac{4 + 9}{27} = \frac{13}{27} \\ &= 0.4\bar{81}. \end{aligned}$$

– Split on comparing a_3 with 5.5:

$a_3 \leqslant 5.5$	
Y	2
N	3

$$\begin{aligned}\text{Gini} &= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 \\ &= 1 - \frac{13}{25} \\ &= \frac{12}{25} \\ &= 0.\bar{4}8.\end{aligned}$$

$a_3 > 5.5$	
Y	2
N	2

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

$$\begin{aligned}\text{Gini} &= \frac{5}{9} \cdot \frac{12}{25} + \frac{4}{9} \cdot \frac{1}{2} \\ &= \frac{4}{3 \cdot 5} + \frac{2}{9} = \frac{4 \cdot 9 + 2 \cdot 3 \cdot 5}{3 \cdot 5 \cdot 9} = \frac{36 + 30}{3 \cdot 5 \cdot 9} = \frac{22}{45} \\ &= 0.4\bar{8}.\end{aligned}$$

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.