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 $\mathbf{p}$ and $\mathbf{q}$:

$\mathbf{p} = [1011100111]$

$\mathbf{q} = [1001001101]$

What is the Jaccard coefficient for the bit vectors $\mathbf{p}$ and $\mathbf{q}$? Write your answer here ☐.

Note: the answer is a real-valued number.

Maximum marks: 3

4 CityBlock distance (3 marks)

There are two vectors $\mathbf{p}$ and $\mathbf{q}$:

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

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

What is the CityBlock distance between $\mathbf{p}$ and $\mathbf{q}$? 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

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

Format

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 | B I U x_2 x^2 I_x | Ω | ABC | Σ

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 | B I U x_2 x^2 I_x | Ω | ∑ | 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

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

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 $\mathbf{Entropy}(t) = -\sum_j p(j|t) \log_2 p(j|t)$, 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 $\mathbf{GAIN} = \mathbf{Entropy}(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} \mathbf{Entropy}(i) \right)$, 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

Words: 0

Maximum marks: 15