Norwegian University of Science and Technology

TDT4300 — **Assignment 5 - solution**

EXAM PREPARATION

tdt4300-undass@idi.ntnu.no

Spring 2022

1 Datawarehousing

1.1 Assignment

a. Bicycle food (SM) delivers food from restaurants to customers in several cities. Each restaurant has a set of dishes they offers, and when the customer has ordered the food online, it is delivered by bicycle courier to the customer shortly afterwards. SM wants a data warehouse that can be used to analyze and optimize the service. Examples of analyzes one should be able to do against the data warehouse:

- Total number of deliveries per day
- Total number of deliveries per restaurant per day
- Average price of each delivery
- Number of customers per city who ordered food on 12 April 2020

a.Make a star or snowflake schema for this case description.
The data are imprecisely formulated and it is part of the task to select which information is necessary to include or find a way to express the facts of the accidents. The main goal of the exercise is to practice modeling principles for data warehousing. You should mention explicitly any assumptions you may make.
b. Given a cube with dimensions and associated concept hierarchy:
-Time (day-month-quarter-year)
-Item (item_name-brand-type)
-Location (street-city-province-state-country)
Assume the following materialized cuboids:
1) {year, brand}

2) {year, item_name, street}
3) {item_name, country} where year = 2006
Given the following OLAP query: {item_name, city} with terms “year = 2006” Which of materialized cuboids can be used to process the query? Justify the answer.

1.2 Solution

Here are several possible answers based on the assumptions you make, here is a variant:
Fact table:
-Delivery (time_k, cust_k, rest_k, price)
Dimension tables:
-Time (time_k, minute, hour, day, month, year)
-Customer (cust_k, cust_name, address, city, email)
-Restaurant (rest_k, rest_name, address, city)
Note that nothing is said in the statement about (for example) price per dish etc. It is not wrong to include this, but not necessary. An important aspect is understanding what should be in the fact table, and what should be in dimension tables.
b) 1) No, location is missing
2) Yes, rollup of street to city, and selection in 2006
3) No, can not do drill-down to city on materialized cube Task.

2 Association Rules

2.1 Assignment

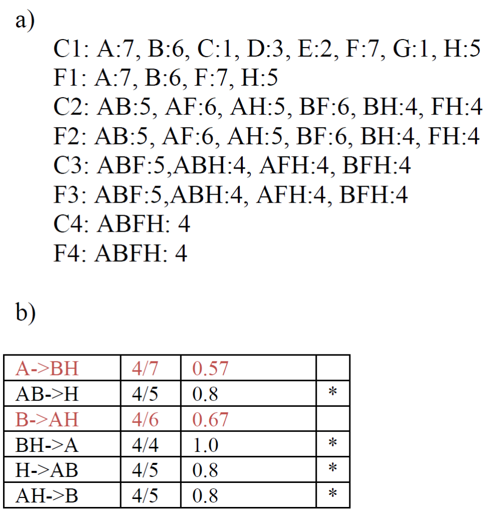
Assume the shopping cart data given below 1. Use the Apriori algorithm to find all frequent element sets with a minimum support of 50% (ie minimum support count is 4). Use the Fk-1 × Fk-1 method for candidate generation.

b. One of the frequent element sets is ABH. Find all association rules based on this set, given 75% confidence (it is not necessary to use the a priori to find the association rules, but show how confidence is calculated for each of the candidate rules based on ABH). **Describe thoroughly the process and the outcome of each step.**

TID	Transaction
T1	BF
T2	ABCDFH
T3	ABF
T4	ABFH
T5	ADEF
T6	ABFH
T7	ABDEFH
T8	AGH

Table 1: Market basket transactions.

2.2 Solution



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	*

Figure 1: Apriori.

3 Decision Trees

3.1 Assignment

A)Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?
B)What are stopping conditions in decision tree classification?

C) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

A small computer retailer, which only sells large computer equipment to youth and students (hereinafter referred to as customers), wants to predict/decide if a customer should get a PC on credit. Table 2 contains examples of the decisions the company has made in the past. Assume that each customer record has five attributes as follows:

Age:	{Young, Middle, Old}
Income:	{Low, Medium, High}
Married:	{Yes, No}
Student:	{Yes, No}
Creditworthiness:	{Pass, High}
PC on Credit:	{Yes, No}

Your task is to first draw the decision tree and then answer the following questions:

1. **D) Compute the Information gain for each attribute (Age, Income,Married, Student, Creditworthiness) in (Table 2).**
2. **E) Which attribute should be selected as a split attribute?**

3.2 Solution

A)Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?

Answer A: The decision tree built may overfit the training data. There could be too many branches, some of which may reflect anomalies in the training data due to noise or outliers. Tree pruning addresses this issue of overfitting the data by removing the least reliable branches (using statistical measures). This generally results in a more compact and reliable decision tree that is faster and more accurate in its classification of data. The drawback of using a separate set of tuples to evaluate pruning is that it may not be representative of the training tuples used to create the original decision tree. If the separate set of tuples are skewed, then using them to evaluate the pruned tree would not be a good indicator of the pruned tree's classification accuracy. Furthermore, using a separate set of tuples to evaluate pruning means there are less tuples to use for creation and testing of the tree. While this is considered a drawback in machine

Customer ID	**Age**	**Income**	**Married**	**Student**	**Creditworthiness**	**PC on Credit**
1	Young	High	No	No	Pass	No
2	Young	High	No	No	High	No
3	Middle	High	Yes	No	Pass	Yes
4	Old	Medium	No	No	Pass	Yes
5	Old	Low	Yes	No	Pass	Yes
6	Old	Low	Yes	Yes	High	No
7	Middle	Low	No	Yes	High	Yes
8	Young	Medium	No	No	Pass	No
9	Young	Low	Yes	Yes	Pass	Yes
10	Old	Medium	Yes	Yes	Pass	Yes
11	Young	Medium	Yes	Yes	High	Yes
12	Middle	Medium	Yes	No	High	Yes
13	Middle	High	Yes	Yes	Pass	Yes
14	Old	Medium	No	No	High	No
15	Middle	Medium	No	Yes	Pass	No
16	Middle	Medium	Yes	Yes	High	Yes
17	Young	Low	No	Yes	High	Yes
18	Old	High	Yes	Yes	Pass	No
19	Old	Low	Yes	Yes	High	No
20	Young	Medium	Yes	Yes	High	Yes

Table 2: Sample dataset.

learning, it may not be so in data mining due to the availability of larger data sets.

B)What are stopping conditions in decision tree classification?

Answer B:

1. **If all tuples at a given node belong to the same class, then transform that node into a leaf, labeled with that class.**
2. **If there are no more attributes left to create more partitions, then majority voting can be used to convert the given node into a leaf, labeled with the most common class among the tuples.**
3. **If there are no tuples for a given branch, a leaf is created with the majority class from the parent node.**

C) How would you modify the basic decision tree algorithm to take into consideration the count of each generalized data tuple (i.e., of each row entry)?

Answer C:

The basic decision tree algorithm should be modified as follows to take into consideration the count of each generalized data tuple.

• The count of each tuple must be integrated into the calculation of the attribute selection measure (such as information gain).

• Take the count into consideration to determine the most common class among the tuples.

Answer D:

Information gain calculated for each attribute Age 0.086

Married 0.102

Income 0.039

Student 0.02

Creditworthiness 0.0

Answer E:

The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches), *Married* attribute has the highest information gain and is therefore the best split attribute.

4 Data Types

4.1 Assignment and Solution

Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity. **Example:** Age in years. **Answer:** Discrete, quantitative, ratio.

(a) Time in terms of AM and PM. *Binary, qualitative, ordinal.*

(b) Brightness as measured by a light meter. *Continuous, quantitative, ratio.*

(c) Brightness as measured by people's judgments. *Discrete, qualitative, ordinal.*

(d) Angles as measured in degrees between 0 and 360. *Continuous, quantitative, ratio.*

(e) Bronze, Silver, and Gold medals as awarded at the Olympics. *Discrete, qualitative, ordinal.*

(f) Height above sea level. *Continuous, quantitative, interval/ratio (depends on whether sea level is regarded as an arbitrary origin).*

(g) Number of patients in a hospital. *Discrete, quantitative, ratio.*

(h) ISBN numbers for books. (Look up the format on the Web.) *Discrete, qualitative, nominal (ISBN numbers do have order information, though).*

(i) Ability to pass light in terms of the following values: opaque, translucent, transparent. *Discrete, qualitative, ordinal.*

(j) Military rank. *Discrete, qualitative, ordinal.*

(k) Distance from the center of campus. *Continuous, quantitative, interval/ratio (depends).*

(l) Density of a substance in grams per cubic centimeter. *Continuous, quantitative, ratio.*

(m) Coat check number. (When you attend an event, you can often give your coat to someone who, in turn, gives you a number that you can use to claim your coat when you leave.) *Discrete, qualitative, nominal.*

5 Autocorrelation

5.1 Assignment

Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why?

5.2 Solution

A feature shows spatial autocorrelation if locations that are closer to each other are more similar with respect to the values of that feature than locations that are farther away. It is more common for physically close locations to have similar temperatures than similar amounts of rainfall since rainfall can be very localized;, i.e., the amount of rainfall can change abruptly from one location to another. Therefore, daily temperature shows more spatial autocorrelation then daily rainfall.

6 Noise and Outliers

6.1 Assignment and Solution

Distinguish between noise and outliers. Answer following questions.

(a) Is noise ever interesting or desirable? Outliers?
No, by definition. Yes.

(b) Can noise objects be outliers?
Yes. Random distortion of the data is often responsible for outliers.

(c) Are noise objects always outliers?
No. Random distortion can result in an object or value much like a normal one.

(d) Are outliers always noise objects?
No. Often outliers merely represent a class of objects that are different from normal objects.

(e) Can noise make a typical value into an unusual one, or vice versa?
Yes.

7 Similarity Measures

7.1 Assignment

For the following vectors, x and y, calculate the indicated similarity or distance measures.

(a) $x = (1{,}1{,}1{,}1), y = (2{,}2{,}2{,}2)$ cosine, correlation, Euclidean

(b) $x = (0{,}1{,}0{,}1), y = (1{,}0{,}1{,}0)$ cosine, correlation, Euclidean, Jaccard

(c) $x = (0, -1{,}0{,}1), y = (1{,}0, -1{,}0)$ cosine, correlation, Euclidean

(d) $x = (1{,}1{,}0{,}1{,}0{,}1), y = (1{,}1{,}1{,}0{,}0{,}1)$ cosine, correlation, Jaccard

(e) $x = (2, -1{,}0{,}2{,}0, -3), y = (-1{,}1, -1{,}0{,}0, -1)$ cosine, correlation

7.2 Solution

The Pearson's product moment correlation coefficient is calculated.

(a) $cos(x,y) = 1$, $corr(x,y) = 0/0(undefined)$, $Euclidean(x,y) = 2$

(b) $cos(x,y) = 0$, $corr(x,y) = -1$, $Euclidean(x,y) = 2$, $Jaccard(x,y) = 0$

(c) $cos(x,y) = 0$, $corr(x,y) = 0$, $Euclidean(x,y) = 2$

(d) $cos(x,y) = 0.75$, $corr(x,y) = 0.25$, $Jaccard(x,y) = 0.6$

(e) $cos(x,y) = 0$, $corr(x,y) = 0$