Norwegian University of Science and Technology

TDT4300 — **Assignment 5**

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.

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.

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?**

4 Data Types

4.1 Assignment

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.

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.

(b) Brightness as measured by a light meter.

(c) Brightness as measured by people's judgments.

(d) Angles as measured in degrees between 0 and 360.

(e) Bronze, Silver, and Gold medals as awarded at the Olympics.

(f) Height above sea level.

(g) Number of patients in a hospital.

(h) ISBN numbers for books. (Look up the format on the Web.)

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

(j) Military rank.

(k) Distance from the center of campus.

(l) Density of a substance in grams per cubic centimeter.

(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.)

5 Autocorrelation

5.1 Assignment

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

6 Noise and Outliers

6.1 Assignment

Distinguish between noise and outliers. Answer following questions.

(a) Is noise ever interesting or desirable? Outliers?

(b) Can noise objects be outliers?

(c) Are noise objects always outliers?

(d) Are outliers always noise objects?

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

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