TDT4300 Datavarehus og datagruvedrift
Spring 2020
Assignment 5

1 Datawarehousing

You are asked to create a data warehouse of traffic accidents in Norway to investigate the arterial routes that are most essential for the society to improve or set lower speed limits, etc. We will be looking at direct costs of accidents and we will not take into account injuries. The data come from various insurance companies and they contain:

- When (date) and where the accident occurred (street and city, or such road section and county).
- Driver related data (we are mostly interested in the age of the driver and whether he was drunk or not).
- Type of insurance of the car and insurance fees.

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.

(a) **Make a star or snowflake schema for this case description.**

(b) **Define two different concept hierarchies (freely chosen dimensions).**

TID	Transaction
1	A, B, C
2	A, C
4	A, D
5	B, E, F

Table 1: Market basket transactions.

2 Association Rules

Given the shopping basket in Table 1, use the Apriori algorithm to generate all possible association rules (for minimum support 0.5 and minimum confidence 0.8). **Describe thoroughly the process and the outcome of each step.**

3 Decision Trees

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}
Student: {Yes, No}
Creditworthiness: {Pass, High}
PC on Credit: {Yes, No}

Your task is to answer the following questions:

1. **Compute the Gini index for the entire training set (Table 2).**
2. **Compute the Gini index for each attribute (Customer ID, Age, Income, Student, Creditworthiness).**

3. **Which attribute should be selected as a split attribute?**

Suppose we have following two customers and we want to predict whether they should get a PC on credit or not. **Explain how would you proceed**.

- Customer # 21: A young student with medium income and "high" creditworthiness.
- Customer # 22: A young non-student with low income and "pass" creditworthiness.

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

Table 2: Sample dataset.

4 Data Types

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.

(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

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

6 Noise and Outliers

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

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

Submission Requirements

In this assignment we expect you to submit following artifacts:

- A PDF file with the report.
  - Text must not be handwritten.

- Make sure that the document follows the usual conventions (**names**, assignment/task number, etc.).

All assignment artifacts are to be delivered using ***BlackBoard***. You are allowed to **work in pairs**, however, the identical artifacts must be delivered individually.