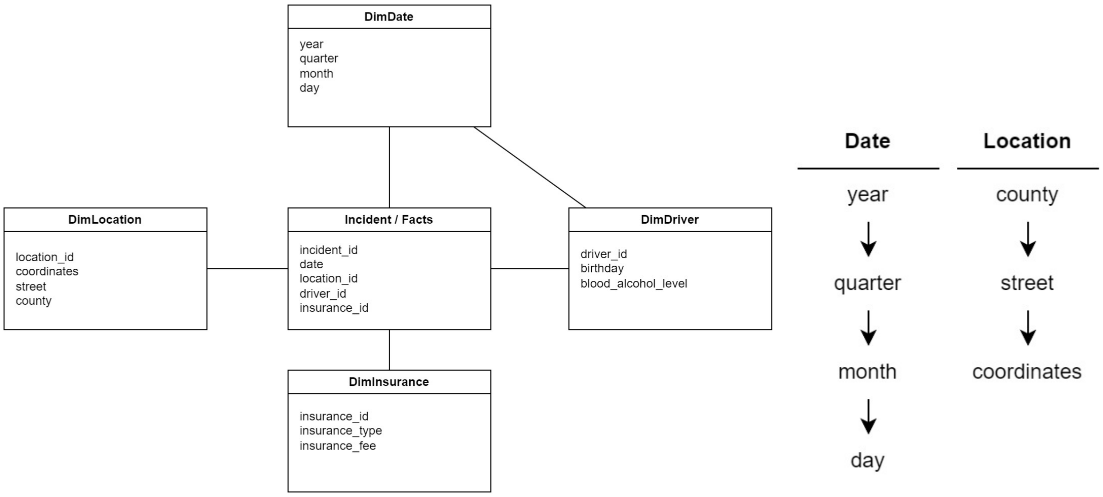
Assignment 5 - TDT4300

Hermann Owren Elton, Olaf Rosendahl

March 28, 2023

1 Datawarehousing



(a) Star schema

(b) Concept hierarchy

2 Association Rules

1-itemset:

Item	Support count σ
A	3
B	2
C	2
D	1
E	1
F	1

The assignment specifies that we should find all frequent itemsets with minimum support of 0.5 (50%), being a support count of 2 (since 4 * 0.5 = 2). Therefore, D, E, and F can be removed. We continue and generate the 2-itemset:

Item		Support count σ
A	B	1
A	C	2
B	C	0

Once again, the itemsets with support count less than 2 are removed, with the itemsets of 2 items being [A,C]. None 3-itemsets exists.

The first step in the process is finding all the combinations of this set, before we calculate the confidence of each association rule. The following function is used to calculate confidence: $c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$. Accepted if confidence is above the minimum confidence of 0.8.

Rule $X \rightarrow Y$	$\sigma(X \cup Y)$	Confidence	Accepted
$\{A\} \rightarrow \{C\}$	3	$\frac{2}{3} = 0.67\overline{7}$	No
$\{C\} \rightarrow \{A\}$	3	$\frac{2}{2} = 1$	Yes

From this table we can se that there are 1 assicuation rule that will be generated.

3 Decision Trees

PC on credit/Age	Young	Middle	Old
Yes	4	5	3
No	3	1	4

$GINI_{Age}(Young) = 1 - (\frac{4}{7})^2 - (\frac{3}{7})^2 = 0.490$

$GINI_{Age}(Middle) = 1 - (\frac{5}{6})^2 - (\frac{1}{6})^2 = 0.278$

$GINI_{Age}(Old) = 1 - (\frac{3}{7})^2 - (\frac{4}{7})^2 = 0.490$

$GINI_{Age} = \frac{7}{20} * 0.490 + \frac{6}{20} * 0.278 + \frac{7}{20} * 0.490 = 0.426$

PC on credit/Income	Low	Medium	High
Yes	4	6	2
No	2	3	3

$GINI_{Income}(Low) = 1 - (\frac{4}{6})^2 - (\frac{2}{6})^2 = 0.445$

$GINI_{Income}(Medium) = 1 - (\frac{6}{9})^2 - (\frac{3}{9})^2 = 0.445$

$GINI_{Income}(High) = 1 - (\frac{2}{5})^2 - (\frac{3}{5})^2 = 0.480$

$GINI_{Income} = \frac{6}{20} * 0.445 + \frac{9}{20} * 0.445 + \frac{5}{20} * 0.480 = 0.454$

PC on credit/Student	No	Yes
Yes	4	8
No	6	2

$GINI_{Student}(No) = 1 - (\frac{4}{10})^2 - (\frac{6}{10})^2 = 0.48$

$GINI_{Student}(Yes) = 1 - (\frac{8}{10})^2 - (\frac{2}{10})^2 = 0.320$

$GINI_{Student} = \frac{10}{20} * 0.480 + \frac{10}{20} * 0.320 = 0.400$

PC on credit/Creditworthiness	Pass	High
Yes	6	6
No	4	4

$GINI_{Creditworthiness}(Pass) = 1 - (\frac{6}{10})^2 - (\frac{4}{10})^2 = 0.48$
$GINI_{Creditworthiness}(High) = 1 - (\frac{6}{10})^2 - (\frac{4}{10})^2 = 0.480$
$GINI_{Creditworthiness} = \frac{10}{20} * 0.480 + \frac{10}{20} * 0.480 = 0.480$
Student should be used as the split attribute since its GINI index is the lowest.

- Customer #21 - Split on student and see that the value is among the lowest so give the credit
- Customer #22 - Split on student and then recompute the attributes when Student=No to see which has the lowest GINI index and split on that

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.

(a) Time in terms of AM and PM. - Continuous, ordinal, interval

(b) Brightness as measured by a light meter. - Continuous, ordinal, ratio

(c) Brightness as measured by people's judgments. - Discrete, ordinal, interval

(d) Angles as measured in degrees between 0 and 360. - Discrete, ordinal, interval

(e) Bronze, Silver, and Gold medals as awarded at the Olympics. - Binary, nominal, interval

(f) Height above sea level. - Discrete, ordinal, ratio

(g) Number of patients in a hospital. - Discrete, ordinal, ratio

(h) ISBN numbers for books. (Look up the format on the Web.) - Discrete, nominal, interval

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

(j) Military rank. - Discrete, nominal, interval

(k) Distance from the center of campus. - Continuos, ordinal, ratio

(l) Density of a substance in grams per cubic centimeter. - Continuous, ordinal, 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, ordinal, interval

5 Noise and Outliers

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

- If there is a lot of noise, this can tell something about the accuracy of the data and how certain you can be of results. But usually noise isn't interesting. Outliers on the other hand can be quite interesting and the part of date which helps define patterns.

(b) Can noise objects be outliers?

- It depends on how far from the "normal" data-points you define that points must be in order to be outliers and not just noise.

(c) Are noise objects always outliers?

- Noise objects are per definition outside the "normal" data, but it's first when they are significantly different that they are outliers, so no.

(d) Are outliers always noise objects?

- No

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

- It can change how the data appears, so yes

6 Similarity Measures

Calculating correlation with https://www.mathsisfun.com/data/correlation-calculator.html

(a) $x = (1, 1, 1, 1), y = (2, 2, 2, 2)$

- Cosine: $\frac{2*1+2*1+2*1+2*1}{\sqrt{1^2+1^2+1^2+1^2}+\sqrt{2^2+2^2+2^2+2^2}} = 1.33\bar{3}$
- Correlation: 1
- Euclidean: $\sqrt{(1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2} = 2$

(b) $x = (0, 1, 0, 1), y = (1, 0, 1, 0)$

- Cosine: $\frac{0*1+1*0+0*1+1*0}{\sqrt{1^2+1^2}+\sqrt{1^2+1^2}} = 0$
- Correlation: -1
- Euclidean: $\sqrt{(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2} = 2$
- Jaccard: $\frac{0}{8} = 0$

(c) $x = (0, -1, 0, 1), y = (1, 0, -1, 0)$

- Cosine: $\frac{0}{\sqrt{(2)+\sqrt{2}}} = 0$
- Correlation: 0
- Euclidean: $\sqrt{(0-1)^2 + (-1-0)^2 + (0-(-1))^2 + (1-0)^2} = 2$

(d) $x = (1, 1, 0, 1, 0, 1), y = (1, 1, 1, 0, 0, 1)$

- Cosine: $\frac{3}{\sqrt{4}+\sqrt{4}} = 0.75$
- Correlation: 0.25
- Jaccard: $\frac{3}{11} = 0.272\bar{}7$

(e) $x = (2, -1, 0, 2, 0, -3), y = (-1, 1, -1, 0, 0, -1)$

- Cosine: $\frac{0}{\sqrt{18}+\sqrt{4}} = 0$
- Correlation: 0