

TDT4300 - Assignment 5

Andreas B. Berg

28.04.2021

1 Data Warehouse Modeling

In this task you have to illustrate a data warehouse with the vehicle insurance data from Scandinavian countries. The task is to discover relationships between vehicle insurance claim costs and a host of other attributes at different levels of a hierarchy. We will be looking at direct costs of accidents and we will not take into account injuries.

a - Make a star or snowflake schema for this case description.

I have opted to use a star schema. The way I see it, there is little need for complicated sub-dimension tables, so a star schema makes later queries easier. While I don't have the means to actually draw up the schema (with connected IDs/foreign keys, etc.), I hope the following tables will be sufficient to explain the schema.

Fact Table

accident_info_key	<i>Foreign key to "Accident Info"</i>
insurance_info_key	<i>Foreign key to "Insurance Info"</i>
vehicle_class	
claim_info_key	<i>Foreign key to "Claim Info"</i>
driver_info_key	<i>Foreign key to "Driver Info"</i>

Accident Info

Accident ID
minute*
hour*
day
weekday*
month
quarter*
year
street
road section
city
county

*) Can be dropped, depending on the granularity of the information coming in, and how it enters the system (i.e.: is it automatic, then this can be calculated automatically, but this takes more time if it is entered manually).

Insurance Info

Insurance ID
Insurance type
Insurance fees*

*) Note: Depending on how this information comes in, it might need to be a foreign key to some other container. I assume this contains information on how much the driver's insurance costs per month (and is therefore a single number), but if it contains multiple values (such as how much the driver's insurance has costed for their whole time with the company), it needs to be modelled to deal with this.

Claim Info

Claim ID
Claim type

Claim History

Claim history ID	
Driver ID	Foreign key to Driver Info
day	
month	
year	
Claim ID	Foreign key to Claim Info

Driver Info

Driver ID
Age
Blood alcohol content (BAC)
Other information*

*) Information such as name, driver's licence number, etc.

b - Define two different concept hierarchies (freely chosen dimensions)

Time

Year	
Quarter	
Month	
Day	Weekday
Hour	
Minute	

Location

County
City
Street
Road section

2 Association Rules

You are given a dataset with market basket transactions (see Table 1) and your task is to mine association rules.

Use the Frequent Pattern Growth (FP-Growth) algorithm to discover all frequent itemsets considering the support threshold $minsup = 0.6$. Construct an FP-tree and mine the frequent itemsets by creating conditional (sub-)pattern bases. Use the table notation with columns: item, conditional pattern base, conditional FP-tree, frequent patterns generated. **The recursive steps of the FP-Growth algorithm must be clearly captured using the aforementioned table notation.** Sort items alphabetically in case of ties in the item support.

Note that with $minsup = 0.6$ and 5 transactions, an itemset needs a support count of at least $0.6 * 5 = 3$ to be considered frequent.

Step 1 - Obtain the transaction database

Table 1: Market basket transactions

TID	Items
1	f, a, c, d, g, i, m, p
2	a, b, c, f, l, m, o
3	b, f, h, j, o
4	b, c, k, s, p
5	a, f, c, e, l, p, m, n

Step 2 - Sort the items in the transaction database by their support

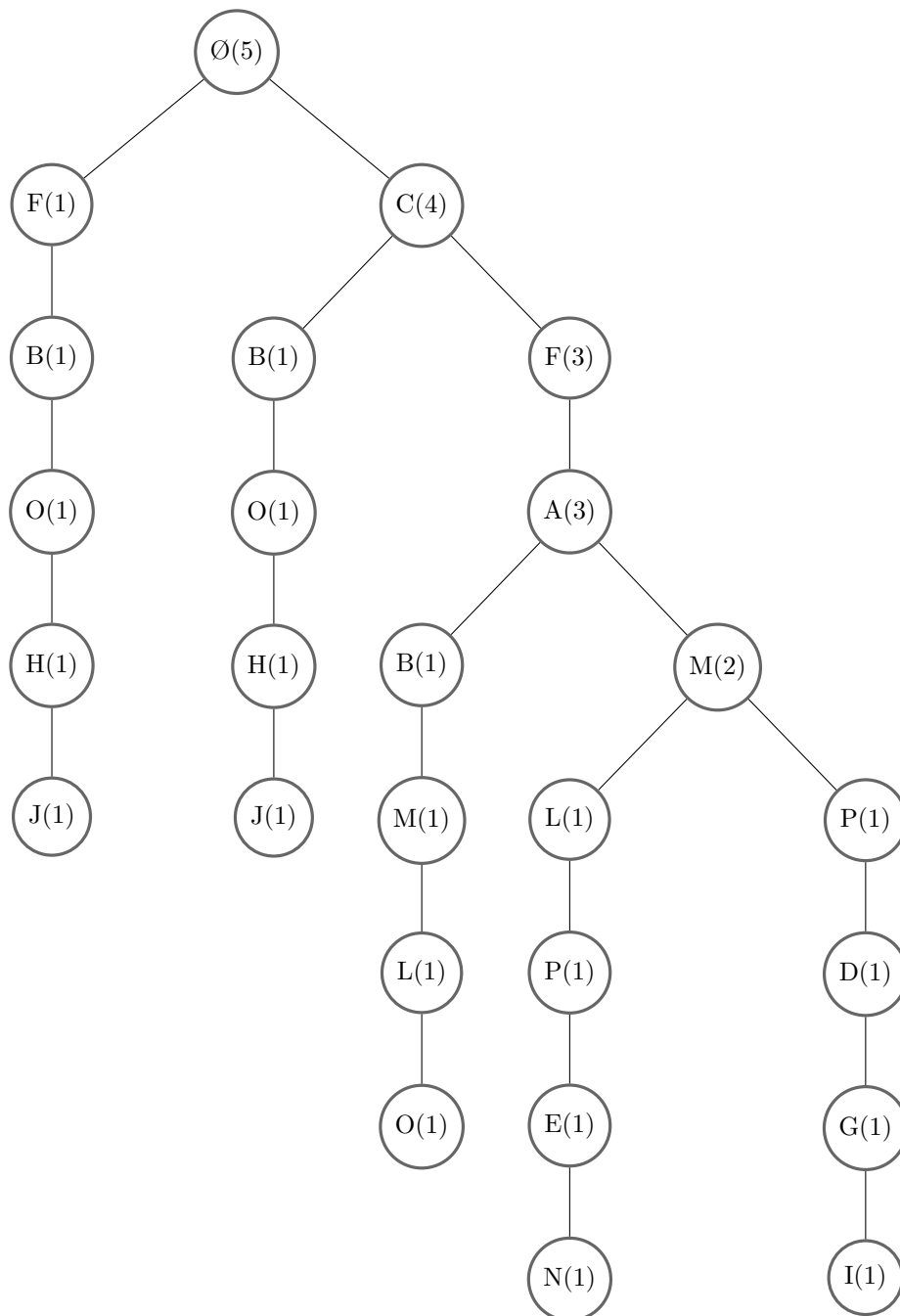
Item	Support count
c	4
f	4
a	3
b	3
m	3
l	2
o	2
p	2
d	1
e	1
g	1
h	1
i	1
j	1
k	1
n	1
s	1

Step 3 - Sort the items in each transaction in descending order

TID	Items
1	c, f, a, m, p, d, g, i
2	c, f, a, b, m, l, o
3	f, b, o, h, j
4	c, b, p, k, s
5	c, f, a, m, l, p, e, n

Step 4 - Construct FP-tree step-by-step by adding transactions

We start with the empty set as root, and then add paths found in the transactions. To avoid very a very long paper, I will only show the resulting (final) FP-tree.



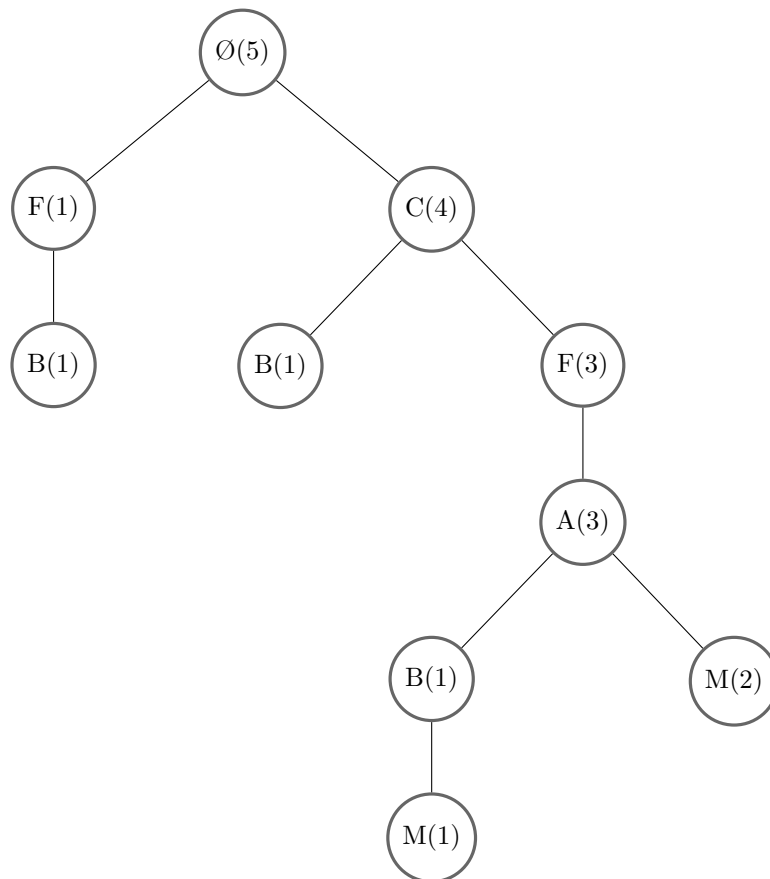
Step 5 - Build conditional FP-trees for each frequent item in increasing order of support

Step 5.1 - Remove infrequent items from the FP-tree

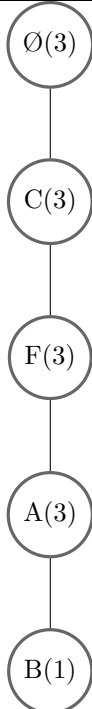
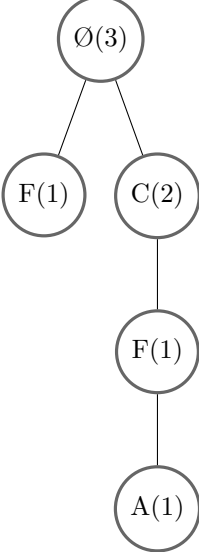
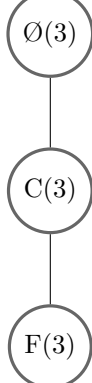
To be frequent, an item needs a support count of at least 3. We are left with:

Item	Support count
c	4
f	4
a	3
b	3
m	3

TID	Items
1	c, f, a, m
2	c, f, a, b, m
3	f, b
4	c, b
5	c, f, a, m



Step 5.2 - Build projected FP-trees for each frequent item in R in increasing order of support

Item	Conditional pattern base	Conditional FP-tree	Frequent patterns generated
m	$\{c, f, a\} : 2$ $\{c, f, a, b\} : 1$		$\{c, f, a, m\} : 3$ $\{c, a, m\} : 3$ $\{c, f, m\} : 3$ $\{f, a, m\} : 3$ $\{a, m\} : 3$ $\{c, m\} : 3$ $\{f, m\} : 3$ $\{m\} : 3$
b	$\{f\} : 1$ $\{c\} : 1$ $\{c, f, a\} : 1$		$\{b\} : 1$
a	$\{c, f\} : 3$		$\{c, f, a\} : 3$ $\{c, a\} : 3$ $\{f, a\} : 3$ $\{a\} : 3$

f	{c} : 3		{c, f} : 3 {f} : 4
c			{c} : 4

Result

Itemset	Support count
{c, f, a, m}	3
{a, c, f}	3
{a, c, m}	3
{a, f, m}	3
{c, f, m}	3
{a, c}	3
{a, f}	3
{a, m}	3
{c, m}	3
{c, f}	3
{f, m}	3
{a}	3
{b}	3
{c}	4
{f}	4
{m}	3

I have some questions about the FP-algorithm, in order to fully prepare for my finals. I hope you have time to answer (at least some of) them:

- 1. In the conditional FP-trees, I have included nodes with count less than minsup. According to the FP-Growth algorithm in the slides, this is correct, as it never compares with minsup during computation. Is my way of doing it correct?*
- 2. I have included (\emptyset) as a node, and have therefore found one-item frequent itemsets using this algorithm. In the FP-Growth algorithm in the slides, this is done in line 11, where $X = (\emptyset \cup \{i\})$ is added to F . Is it okay to include as I have done here?*

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}

1 - Compute the Gini index for the entire training set

We calculate the Gini index using the formula

$$Gini\ index = 1 - \sum_{i=0}^{c-1} [p_i(t)]^2 \quad (1)$$

where $p_i(t)$ is the frequency of class i at node t , and c is the total number of classes. I assume that to calculate this for the entire training set, we look only at the split between customers who got a PC on credit and not, meaning we have $t = \text{"PC on Credit"}$ and $c = 2$, as the result in column "PC on Credit" can be one of two classes (Yes (1)/No (0)). We then have $p_1(t) = 12/20 = 0.6$ and $p_0(t) = 8/20 = 0.4$

$$\begin{aligned} Gini\ index &= 1 - \sum_{i=0}^{c-1} [p_i(t)]^2 \\ &= 1 - ([0.4]^2 + [0.6]^2) \\ &= 0.48 \end{aligned}$$

2 - Compute the Gini index for each attribute (Customer ID, Age, Income, Student, Creditworthiness)

Customer ID

We know that the Gini index is 0 where all records belong to one class, meaning when all records in a Gini-calculation has the same result. For Customer ID, as it is a unique ID for each line, it is obvious that each key leads to the same result - there can never be two results for one key. As such, we know that **the Gini index is 0 for all Customer IDs.**

Age

Here, we must calculate the Gini index three times - one for each state of "Age", then calculate the weighted sum of the indices to find the total Gini index for the attribute.

Age	Calculation	Probability
Young	7 / 20	0.35
Middle	6 / 20	0.3
Old	7 / 20	0.35

We find the relative probability of getting a PC on Credit for each category:

Age	PC on Credit	Probability
Young	Yes	4/7
Young	No	3/7
Middle	Yes	5/6
Middle	No	1/6
Old	Yes	3/7
Old	No	4/7

We can use this to calculate the Gini index for each category:

Age	Calculation	Gini index
Young	$1 - ((4/7)^2 + (3/7)^2)$	0.49
Middle	$1 - ((5/6)^2 + (1/6)^2)$	0.28
Old	$1 - ((3/7)^2 + (4/7)^2)$	0.49

To find the combined Gini index for Age, we find the weighted sum of the indices:

$$\begin{aligned} \text{Gini index}_{age} &= (7/20) * 0.49 + (6/20) * 0.28 + (7/20) * 0.49 \\ &= \mathbf{0.427} \end{aligned}$$

Income

We follow the same procedure as for "Age" - calculate Gini index for each state of Income, then calculate the weighted sum of the indices to find the total Gini index for the attribute.

Income	Calculation	Probability
Low	6 / 20	0.3
Medium	9 / 20	0.45
High	5 / 20	0.25

We find the relative probability of getting a PC on Credit for each category:

Income	PC on Credit	Probability
Low	Yes	4/6
Low	No	2/6
Medium	Yes	6/9
Medium	No	3/9
High	Yes	2/5
High	No	3/5

We can use this to calculate the Gini index for each category:

Income	Calculation	Gini index
Low	$1 - ((4/6)^2 + (2/6)^2)$	0.44
Medium	$1 - ((6/9)^2 + (3/9)^2)$	0.44
High	$1 - ((2/5)^2 + (3/5)^2)$	0.48

To find the combined Gini index for Age, we find the weighted sum of the indices:

$$\begin{aligned} \text{Gini index}_{age} &= (6/20) * 0.44 + (9/20) * 0.44 + (5/20) * 0.48 \\ &= \mathbf{0.45} \end{aligned}$$

Student

We one again follow the same procedure as before - calculate Gini index for each state of Student, then calculate the weighted sum of the indices to find the total Gini index for the attribute.

Student	Calculation	Probability
Yes	10 / 20	0.5
No	10 / 20	0.5

We find the relative probability of getting a PC on Credit for each category:

Student	PC on Credit	Probability
Yes	Yes	8/10
Yes	No	2/10
No	Yes	4/10
No	No	6/10

We can use this to calculate the Gini index for each category:

Student	Calculation	Gini index
Yes	$1 - ((8/10)^2 + (2/10)^2)$	0.32
No	$1 - ((4/10)^2 + (6/10)^2)$	0.48

To find the combined Gini index for Age, we find the weighted sum of the indices:

$$\begin{aligned} \text{Gini index}_{age} &= (10/20) * 0.32 + (10/20) * 0.48 \\ &= \mathbf{0.4} \end{aligned}$$

Creditworthiness

We one again follow the same procedure as before - calculate Gini index for each state of Student, then calculate the weighted sum of the indices to find the total Gini index for the attribute.

Creditworthiness	Calculation	Probability
High	10 / 20	0.5
Pass	10 / 20	0.5

We find the relative probability of getting a PC on Credit for each category:

Creditworthiness	PC on Credit	Probability
High	Yes	6/10
High	No	4/10
Pass	Yes	6/10
Pass	No	4/10

We can use this to calculate the Gini index for each category:

Student	Calculation	Gini index
Yes	$1 - ((6/10)^2 + (4/10)^2)$	0.48
No	$1 - ((6/10)^2 + (4/10)^2)$	0.48

To find the combined Gini index for Age, we find the weighted sum of the indices:

$$\begin{aligned} \text{Gini index}_{age} &= (10/20) * 0.48 + (10/20) * 0.48 \\ &= \mathbf{0.48} \end{aligned}$$

3 - Which attribute should be selected as a split attribute?

When selecting a split attribute, one should pick the attribute that gives the biggest gain, which is equivalent with the lowest impurity measure (Gini index) after splitting. We have calculated the following table of post-split Gini indices:

Attribute	Gini index after split
Customer ID	0
Age	0.427
Income	0.45
Student	0.4
Creditworthiness	0.48

Based on this, it is tempting to select Customer ID as our split attribute. This would, however, be a bad idea, as we have no way of generalizing Customer ID to estimate new, incoming customers. We therefore ignore it, and instead look at what other attribute - that can more easily be generalized - has the lowest Gini index after splitting. This is the **Student** attribute, which has a post-split index of **0.4**. I would therefore select the Student attribute as a split attribute.

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

Customer #21: A young student with medium income and "high" creditworthiness

Customer #22: A young student with low income and "pass" creditworthiness

If we assume that we have a completed decision tree, then this choice is easy - simply start at the split attribute (Student) and follow the tree down until you get a leaf node, telling you what the most likely choice would be.

If we assume that we have no finished decision tree, and rather the information we have here, then we can look at similar customers amongst previous customers in Table 2.

We can find two customers with the same attributes as customer #21 - Customer 11 and 20 are both young students with medium income and high creditworthiness. Both of them got PC on Credit, so the probability of a customer with these attributes getting a PC on Credit is, based on our dataset, 100%. It is therefore safe to predict that Customer #21 will also get a PC on Credit.

We can find 1 customer with the same attributes as customer #22 - Customer 9 is a young student with low income and pass creditworthiness. Customer 9 did not get a PC on Credit, so the probability of a customer with these attributes getting a PC on Credit is, based on our dataset, 0%. It is therefore safe to predict that customer #22 will not get a PC on Credit.

4 Noise and Outliers

Distinguish between noise and outliers. Answer following questions

a - Is noise ever interesting or desirable? Outliers?

Outliers can be interesting, while noise - by definition - is not. Outliers are objects with characteristics that differ from the majority of the dataset, and can be investigated to further understand the data. Say, for instance, you work Quality Assurance for a major production company. It is safe to assume that the majority of products produced are good enough to sell, so the interesting part of the data will be the outliers - understanding why some products end up outside of the margins for acceptable products.

Noise, however, is simply a distortion of the data, and is never interesting - except if the study is on the actual sensors (or something like that), but that is a production issue, not data science.

b - Can noise objects be outliers?

Noise objects can be outliers. Say you are measuring temperatures, in order to predict how hot your sun-filled room gets in the summer. You have a sensor that produces some noise, but you are generally within ± 1 degree of the true data. If the sensor, for one measurement, measures a degree far off from the true data, then this is noise - and also an outlier, as it differs from the majority of the data.

c - Are noise objects always outliers?

No. In the example above, most of the measurements will reside within ± 1 degree of the correct data. In these cases, when noise is assumed and present and the data still is within the majority of measurements, then the noise is not an outlier.

d - Are outliers always noise objects?

Not necessarily. Assume that this class is very hard, leaving most students with a C or D in the class. If one day someone really smart takes the class and achieves an A, then this is an outlier - it is different from the majority of the data - but it is not noise, as the data is correct and accurate towards the real world.

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

Yes. In the temperature example above, I mentioned noise taking a typical value (meaning a normal room temperature) and turning it into an outlier, which is (usually) an unusual value.

It can also, in certain cases, go the other way. Assume that the room in the temperature example has a drastic temperature change (assume, for instance, that its resident spills a huge bucket of dry ice into the room, but gets it cleaned up almost immediately - it only has time to stay there for one measurement). This provides an unusual value - far off from the majority of measurements. If the sensor fails to catch this due to noise, and instead measures a "normal" temperature, then the noise has made an unusual value into a typical one.

5 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

I assume this means a measurement of whether it is AM or PM, making it a binary attribute. (If I misunderstand and it instead means time in general, only measured in 12-hour blocks, then it is discrete - if looking at hours, minutes, seconds, etc. Due to Planck time being the smallest time measurement possible, one can argue that time in general is discrete, but I will say that the general perception of time is continuous if you are not using a watch).

In lecture 5, we learned that nominal values are just different names for an attribute, that only provides information needed to distinguish objects from each other. As AM and PM only distinguishes different times (such as 1AM and 1PM) without helping to sort it, I consider time in terms of AM and PM to be nominal.

Times in terms of AM and PM		
Binary	Qualitative	Nominal

b - Brightness as measured by a light meter

A light meter will always provide numerical values, meaning it is quantitative. As light can never fall below zero - the darkest you can ever get is the complete absence of light - it is a ratio variable. Light in itself is continuous, in that there are no "steps" of brightness. However, if we assume the light meter provides a digital signal, then the signal will be discrete, as one cannot provide a continuous signal using bits.

Brightness as measured by a light meter		
Discrete	Quantitative	Ratio

c - Brightness as measured by people's judgement

Most (if not all) people will struggle with quantifying brightness, making this data qualitative. We can, however, compare different levels of brightness, making it ordinal. As mentioned above - there are no "steps" of brightness, meaning this measurement is continuous.

Brightness as measured by people's judgement		
Continuous	Qualitative	Ordinal

d - Angles measured in degrees between 0 and 360

Because it is specified that the angles are measured in degrees, then I assume we are talking about whole degrees - making this data discrete (with 360 steps). Because degrees are measurable and countable, this measurement is quantitative. Because we specify degrees between 0 and 360, we can never go below 0 degrees. This makes the attribute a ratio.

Angles measured in degrees between 0 and 360		
Discrete	Quantitative	Ratio

e - Bronze, Silver and Gold medals as awarded at the Olympics

I assume this means the count of them - how many medals each nation got. If this assumption is true, then the data is countable - making it quantitative - and discrete - as each country can only receive either one whole medal or none, no partial medals are awarded. Because no country can receive negative medals, the attribute is a ratio.

Medals awarded at the Olympics		
Discrete	Quantitative	Ratio

f - Height above sea level

While we usually measure height in meters, the actual measurement of length (if we look away from Planck lengths) is continuous, as you can split a meter into (almost) as small lengths as you want. Because this is something measurable, the data is quantitative. Because we can also go below sea level, then the attribute is an interval.

Height above sea level		
Continuous	Quantitative	Interval

g - Number of patients in a hospital

As a hospital can never have half a patient (in which case they are no longer considered patients), the data is discrete. The data is countable, meaning that it is quantitative. It can never be a negative number, making it a ratio.

Number of patients in a hospital		
Discrete	Quantitative	Ratio

h - ISBN numbers for books

ISBN numbers categorizes books, making them qualitative. While some numbers in an ISBN numbers can order the book somewhat, such as stating its language and publisher, it does not provide enough information to order books based on it. It is therefore nominal. Because ISBN are integers, it is discrete.

ISBN numbers for books		
Discrete	Qualitative	Nominal

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

With only three options, this attribute is discrete. As the options are not comparable numerically (at least not without further information), the attribute is qualitative. The attribute can be used to order objects (opaque objects are less transparent than translucent objects, etc.), so the attribute is ordinal.

Ability to pass light		
Discrete	Qualitative	Ordinal

j - Military rank

As you can never be half a sergeant, and there are a limited number of military ranks, the attribute is discrete. Military rank is not countable or measurable, so the attribute is qualitative. It can, however, be used to sort objects, as a major has a higher rank than a private. It is therefore ordinal.

Note: While some could argue that this gets complicated with the new system recently implemented in Norway, with staff sergeants, sergeant first class, etc., the ordering is still there. These are sublevels, but there are still a limited number of them, so nothing changes.

Military rank		
Discrete	Qualitative	Ordinal

k - Distance from the center of campus

Distance is measurable, making it quantitative. We once again ignores Planck lengths, which makes length (and this attribute) continuous. Because the distance from the center of campus can never be negative, this attribute is a ratio.

Distance from the center of campus		
Continuous	Quantitative	Ratio

l - Density of a substance in grams per cubic centimeter

One can always split a gram in multiple, uncountable (within reason) submeasurements, making this continuous. It is measurable, meaning it is quantitative, and it can never be less than zero, meaning it is a ratio.

Note: If density is measured in whole grams, then one can argue that there is a maximum density, meaning the attribute is discrete. I have assumed that one can go below one gram, or that there is no roof for the maximum density (which there is, but not within reason). It is therefore continuous.

Density of a substance		
Continuous	Quantitative	Ratio

m - Coat check number

While this can be measured (in that it is a number), it is reasonable to call this a quantitative attribute - in that it simply describes the coat, and provides no relevant measurable data except the order in which you handed your coat to the worker. While you can also order by it, it is simply a different "name" given to your coat - it provides no other information for the coat. Because this is - I assume, based on experience - a whole number, limited by the number of coat hangers the event has, the attribute is discrete.

Coat check number		
Discrete	Qualitative	Nominal

6 Similarity Measures

For the following vectors, x and y , calculate the indicated similarity / distance measure as well as their correlation coefficient.

We can find the cosine similarity of the vectors using the formula

$$\begin{aligned} \text{cosine} = \cos(\theta) &= \frac{x \cdot y}{|x| * |y|} \\ &= \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} * \sqrt{\sum_{i=1}^n y_i^2}} \end{aligned} \quad (2)$$

I assume that by "correlation coefficient", we mean the Pearson correlation coefficient. We can find the Pearson correlation coefficient of two vectors using the formula

$$\text{pearson} = \frac{x^c \cdot y^c}{|x^c| * |y^c|} \quad (3)$$

where $x^c = (x_1 - \bar{x}, x_2 - \bar{x}, \dots)$, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the mean of all elements in x , and likewise for y .

We can find the Euclidean distance of two vectors using the formula

$$\text{euclidean} = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (4)$$

We can find the Jaccard similarity of two vectors using the formula

$$\text{jaccard} = \frac{|x \cap y|}{|x \cup y|} \quad (5)$$

I assume that for vectors, $x \cap y$ simply means all elements that are the same in x and y , and that likewise $x \cup y$ means all elements that are in both x and y , and that the order they appear (or their index) does not matter. This is based on the information I could find about Jaccard online, *please let me know if I am mistaken in this*.

a

$$\mathbf{x} = (1,1,1,1), \mathbf{y} = (3,3,3,3)$$

Cosine, correlation, Euclidean

Cosine	=	1
Euclidean	=	4
Correlation	=	No answer

We can insert our vectors into (2), and get the cosine similarity

$$\begin{aligned} similarity_{cosine} &= \frac{1*3 + 1*3 + 1*3 + 1*3}{\sqrt{1+1+1+1} * \sqrt{9+9+9+9}} \\ &= \frac{12}{12} = 1 \end{aligned}$$

We can insert our vectors into (4) and get the Euclidean distance

$$\begin{aligned} distance_{Euclidean} &= \sqrt{(3-1)^2 + (3-1)^2 + (3-1)^2 + (3-1)^2} \\ &= \sqrt{16} = 4 \end{aligned}$$

To find the correlation coefficient, we first find the mean of both vectors.

$$\begin{aligned} \bar{x} &= \frac{1}{4} \sum_{i=1}^4 x_i = \frac{4}{4} = 1 \\ \bar{y} &= \frac{1}{4} \sum_{i=1}^4 y_i = \frac{12}{4} = 3 \end{aligned}$$

We can then use these to find x^c and y^c :

$$\begin{aligned} x^c &= (1-1, 1-1, 1-1, 1-1) = (0, 0, 0, 0) \\ y^c &= (3-3, 3-3, 3-3, 3-3) = (0, 0, 0, 0) \end{aligned}$$

It is clear that we cannot compute the correlation coefficient for these numbers, as we will end up with $\frac{0}{0}$.

b $x = (0, 2, 0, 2)$, $y = (2, 0, 2, 0)$

Cosine, correlation, Euclidean, Jaccard

Cosine	=	0
Euclidean	=	4
Jaccard	=	1
Correlation	=	-1

We can insert our vectors into (2), and get the cosine similarity

$$\begin{aligned} similarity_{cosine} &= \frac{0 * 2 + 2 * 0 + 0 * 2 + 2 * 0}{\sqrt{0 + 4 + 0 + 4} * \sqrt{4 + 0 + 4 + 0}} \\ &= \frac{0}{8} = \mathbf{0} \end{aligned}$$

We can insert our vectors into (4) and get the Euclidean distance

$$\begin{aligned} distance_{Euclidean} &= \sqrt{(0 - 2)^2 + (2 - 0)^2 + (0 - 2)^2 + (2 - 0)^2} \\ &= \sqrt{16} = \mathbf{4} \end{aligned}$$

We can insert our vectors into (5) and get the Jaccard similarity

$$similarity_{Jaccard} = \frac{|(0, 2)|}{|(0, 2)|} = \frac{2}{2} = \mathbf{1}$$

To find the correlation coefficient, we first find the mean of both vectors.

$$\begin{aligned} \bar{x} &= \frac{1}{4} \sum_{i=1}^4 x_i = \frac{4}{4} = 1 \\ \bar{y} &= \frac{1}{4} \sum_{i=1}^4 y_i = \frac{4}{4} = 1 \end{aligned}$$

We can then use these to find x^c and y^c :

$$\begin{aligned} x^c &= (0 - 1, 2 - 1, 0 - 1, 2 - 1) = (-1, 1, -1, 1) \\ y^c &= (2 - 1, 0 - 1, 2 - 1, 0 - 1) = (1, -1, 1, -1) \end{aligned}$$

We can then use (3) to find the Pearson correlation coefficient

$$\begin{aligned} pearson &= \frac{x^c \cdot y^c}{|x^c| * |y^c|} \\ &= \frac{(-1) * 1 + 1 * (-1) + (-1) * 1 + 1 * (-1)}{\sqrt{(4)} * \sqrt{4}} \\ &= \frac{-4}{4} = \mathbf{-1} \end{aligned}$$

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

Cosine, correlation, Euclidean

Cosine	=	0
Euclidean	=	2
Correlation	=	0

We can insert our vectors into (2), and get the cosine similarity

$$\begin{aligned} similarity_{cosine} &= \frac{0 * (-1) + 1 * 0 + 0 * 1 + (-1) * 0}{\sqrt{0 + 1 + 0 + 1} * \sqrt{1 + 0 + 1 + 0}} \\ &= \frac{0}{2} = \mathbf{0} \end{aligned}$$

We can insert our vectors into (4) and get the Euclidean distance

$$\begin{aligned} distance_{Euclidean} &= \sqrt{(0 + 1)^2 + (1 - 0)^2 + (0 - 1)^2 + (-1 - 0)^2} \\ &= \sqrt{4} = \mathbf{2} \end{aligned}$$

To find the correlation coefficient, we first find the mean of both vectors.

$$\begin{aligned} \bar{x} &= \frac{1}{4} \sum_{i=1}^4 x_i = \frac{0}{4} = 0 \\ \bar{y} &= \frac{1}{4} \sum_{i=1}^4 y_i = \frac{0}{4} = 0 \end{aligned}$$

We then have $x^c = x$ and $y^c = y$

We can then use (3) to find the Pearson correlation coefficient

$$\begin{aligned} pearson &= \frac{x^c \cdot y^c}{|x^c| * |y^c|} \\ &= \frac{0 * (-1) + 1 * 0 + 0 * 1 + (-1) * 0}{\sqrt{(2)} * \sqrt{2}} \\ &= \frac{0}{2} = \mathbf{0} \end{aligned}$$

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

Cosine, correlation, Jaccard

Cosine	=	0.75
Jaccard	=	1
Correlation	=	0.25

We can insert our vectors into (2), and get the cosine similarity

$$\begin{aligned} similarity_{cosine} &= \frac{1*1 + 1*1 + 0*1 + 1*0 + 0*0 + 1*1}{\sqrt{1+1+0+1+0+1} * \sqrt{1+1+1+0+0+1}} \\ &= \frac{3}{4} = \mathbf{0.75} \end{aligned}$$

We can insert our vectors into (5) and get the Jaccard similarity

$$similarity_{Jaccard} = \frac{|(0,1)|}{|(0,1)|} = \frac{2}{2} = \mathbf{1}$$

To find the correlation coefficient, we first find the mean of both vectors.

$$\begin{aligned} \bar{x} &= \frac{1}{4} \sum_{i=1}^4 x_i = \frac{2}{3} \\ \bar{y} &= \frac{1}{4} \sum_{i=1}^4 y_i = \frac{2}{3} \end{aligned}$$

We can then use these to find x^c and y^c :

$$\begin{aligned} x^c &= (1 - \frac{2}{3}, 1 - \frac{2}{3}, 0 - \frac{2}{3}, 1 - \frac{2}{3}, 0 - \frac{2}{3}, 1 - \frac{2}{3}) = (\frac{1}{3}, \frac{1}{3}, -\frac{2}{3}, \frac{1}{3}, -\frac{2}{3}, \frac{1}{3}) \\ y^c &= (1 - \frac{2}{3}, 1 - \frac{2}{3}, 1 - \frac{2}{3}, 0 - \frac{2}{3}, 0 - \frac{2}{3}, 1 - \frac{2}{3}) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, -\frac{2}{3}, -\frac{2}{3}, \frac{1}{3}) \end{aligned}$$

We can then use (3) to find the Pearson correlation coefficient

$$\begin{aligned} pearson &= \frac{x^c \cdot y^c}{|x^c| * |y^c|} \\ &= \frac{\frac{1}{3} * \frac{1}{3} + \frac{1}{3} * \frac{1}{3} + \frac{-2}{3} * \frac{1}{3} + \frac{1}{3} * \frac{-2}{3} + \frac{-2}{3} * \frac{-2}{3} + \frac{1}{3} * \frac{1}{3}}{\sqrt{\frac{4}{3}} * \sqrt{\frac{4}{3}}} \\ &= \frac{\frac{1}{9} + \frac{1}{9} - \frac{2}{9} - \frac{2}{9} + \frac{4}{9} + \frac{1}{9}}{\frac{4}{3}} = \frac{\frac{3}{9}}{\frac{4}{3}} = \frac{1}{4} = \mathbf{0.25} \end{aligned}$$