Written Assignment 1

Benny Chen

February 13, 2023

1 Data Preprocessing

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.

1. Seat numbers assigned to passengers on a flight.
2. Hurricane intensity measured on the Saffir-Simpson scale.
3. Social Security number.

1. The seat numbers assigned to passengers on a flight would be a discrete attribute as there is only a finite number of seats on a plane and can only be integer values. This attribute would also be a qualitative attribute, specifically a ordinal attribute as the seat numbers are ordered from the first sear and passenger to the last seat and passenger.
2. Hurricane intensity measured on the Saffir-Simpson scale would also be a discrete attribute as there is only a finite number of categories on the scale and each category is a integer value scaled from 1 to 5. This attribute would also be a qualitative attribute, specifically a ordinal attribute as the categories are ordered from the lowest intensity to the highest intensity.
3. This prompt is very ambiguous as we don't know what we are supposed to do with the Social Security number. If we are supposed to use the Social Security number as an attribute, then it would be a discrete attribute as there is only a finite number of Social Security numbers that there can be in total. This attribute would also be a qualitative attribute, specifically a nominal attribute as the Social Security numbers are not ordered and are a categorical attribute.

Consider the following binary vectors:

$$x_1 = (1, 1, 1, 1, 1)$$
$$x_2 = (1, 1, 1, 0, 0)$$
$$y_1 = (0, 0, 0, 0, 0)$$
$$y_2 = (0, 0, 0, 1, 1)$$

1. According to the Jaccard coefficient, which pair of vectors (x_1; x_2) or (y_1; y_2) are more similar to each other?
2. According to the Simple Matching proximity measure, which pair of vectors (x_1; x_2) or (y_1; y_2) are more similar to each other?
3. According to the Euclidean Distance, which pair of vectors (x_1; x_2) or (y_1; y_2) are more similar to each other?

1. The Jaccard coefficient is used to find similarities between 2 sets which is

$$sim_{Jaccard}(i, j) = \frac{q}{q + r + s}$$

We first create a contingency table of then plug in the values for each set

X1/X2	1	0	Sum
1	3	0	3
0	2	0	2
Sum	5	0	5

and find that the pair of vectors (x_1; x_2) are 3/5 or 60% similar to each other. The pair of vectors (y_1; y_2) are 0/2 or 0% similar to each other. Therefore, the pair of vectors (x_1; x_2) are more similar to each other.

2. We can use simple matching which is

$$d(i, j) = \frac{p - m}{p}$$

to find for that both pairs of vectors (x_1; x_2) and (y_1; y_2) 60% similar to each other as we are finding the number of matches regardless of the value.

3. According to the Euclidean Distance, both pairs of vectors are similar to each other. The Euclidean Distance is

$$d(i, j) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$$

We then plug in the values for each set and find that the pair of vectors $(x_1; x_2)$ to get

$$d(x_1, x_2) = \sqrt{(1-1)^2 + (1-1)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2} = \sqrt{2}$$

and the pair of vectors $(y_1; y_2)$ to get

$$d(y_1, y_2) = \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (0-1)^2} = \sqrt{2}$$

Therefore, both pairs of vectors are similar to each other.

We are given the traffic accident of three drivers. We consider the three attributes are asymmetric binary ones, i.e., "Alcohol Impaired = Yes", "Traffic Violation = Yes". , and "Absence of Belt = Yes" are more severe/critical. Please find their respective similarity based on contingency tables and Distance Measure for Asymmetric Binary Variables.

Driver	Alcohol Impaired	Traffic Violation	Absence of Belt
Han	Yes	Yes	Yes
Luke	No	No	Yes
Leia	No	Yes	No

To find the Distance Measure for Asymmetric Variables and the similarities between the three people, we need a contingency table for each pair of people. We then use the following formula to find the similarity between the two people.

$$dis_{Asymmetric}(i, j) = \frac{r + s}{q + r + s}$$

For the contingency table for Luke (Column) and Han (Row) we will get:

Luke/Han	1	0	Sum
1	1	0	1
0	2	0	2
Sum	3	0	3

$$\frac{0 + 2}{1 + 0 + 2} = 0.67$$

The contingency table for Luke (Column) and Lelia (Row) will get us:

$$\frac{1 + 1}{0 + 1 + 1} = 1$$

Luke/Lelia	1	0	Sum
1	0	1	1
0	1	1	2
Sum	1	2	3

Lastly the contingency table for Leila (Column) and Han (Row):

Lelia/Han	1	0	Sum
1	1	0	1
0	2	0	2
Sum	3	0	3

$$\frac{0+2}{1+0+2} = 0.67$$

From what we can see, the similarity between Luke and Leia is 1 while the similarity between Luke and Han and Leia and Han is 0.67. This means that the pair of Luke and Han and Leia and Han are more similar to each other than Luke and Leia.

2 Association Rule Mining

Frequent Pattern Mining

Please determine whether the statement is true or false and provide your explanation.

1. For some transaction data, the number of frequent itemsets of size 3 was found out to be three. Is it safe to assume that there cannot be any frequent itemset of size 4? If no, come up with an example. If yes, justify your answer.

2. Given that {a,b}, {b,c} and {a,c} are frequent itemsets, {a,b,c} is always frequent.

1. This statement is false that we can assume that there cannot be any frequent itemset of size 4 since there is a frequent itemset of size 3. We cannot assume that there will always be 3 as the size could change due to any factor and could find frequent itemsets of 4. For example a frequent itemset of size 4 could be {a,b,c,d}, {a,b,c,e}, and {a,b,c,f} but also have a frequent itemset of size 3 with {a,b,c} as the subsets of the frequent itemsets of size 4.

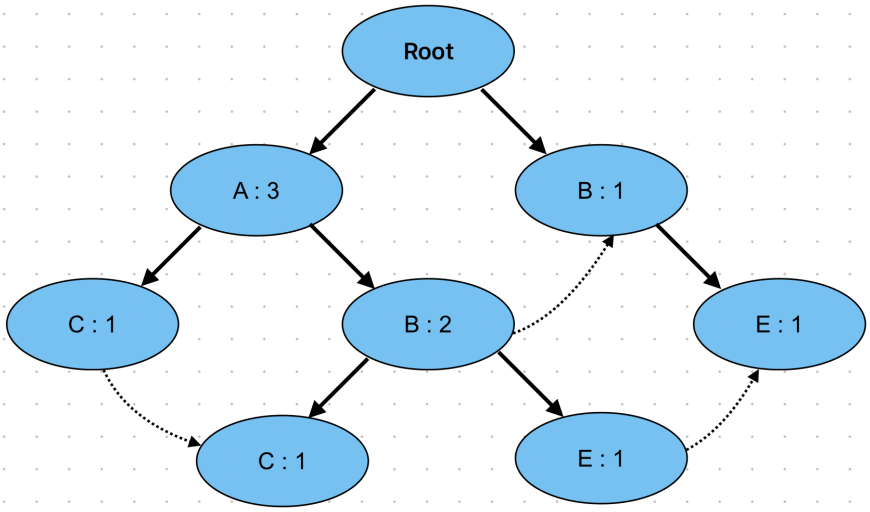
2. This statement is True as subsets of frequent itemsets would make a frequent itemset. {a,b}, {b,c}, and {a,c} are subsets of {a,b,c}. Therefore, {a,b,c} is always frequent.

FP-Tree

We are given the transaction database as:

TID	Item Bought
T1	A,C,D
T2	A,B,C
T3	A,B,E
T4	B,E

Please build the FPTree for the transaction database with the minimum support count 2. Please provide clear and readable figure or screenshot of the constructed FPTree (refer to the lecture note). We assume alphabetical order for items with the same frequency. You do not need to build the conditional FPTrees and the rules.
We first need to count and order all the items by frequency. From that we can tell A and B are the most frequent, followed by C and E with 2 and finally D with 1. We then filter out the items that do not meet the minimum support count of 2 then we build the tree to get the following:



3 Clustering

K-means

Consider the following eight two-dimensional data points:

$$x_1 : (23, 12), x_2 : (6, 6), x_3 : (15, 0), x_4 : (15, 28), x_5 : (20, 9),$$
$$x_6 : (8, 9), x_7 : (20, 11), x_8 : (8, 13)$$

Consider k-means algorithm to answer the following questions. You are required to show the information about each final cluster (including the mean of the cluster and all data points in this cluster). You can consider writing a program for this part but you are not required to submit the program.

1. If k = 2 and the initial means are (20, 9) and (8, 9), what is the output of the algorithm? In the output, you are required to show the information about each final cluster (including the mean of the cluster and all data points in this cluster).

2. If k = 2 and the initial means are (15, 0) and (15, 29), what is the output of the algorithm? In the output, you are required to show the information about each final cluster (including the mean of the cluster and all data points in this cluster).

3. What are the advantages and the disadvantages of the k-means algorithm? For each disadvantage, please also give a suggestion to enhance the k-means algorithm.

1. We first create a table for the data points and their respective distances from the initial means of (20, 9) and (8, 9). We use the Euclidean distance formula to calculate the distance between the data points and the initial means. We then are able to cluster these data points nto k = 2 clusters by finding the smallest distance between the data points and the initial means. We then get the following table:

K = 2	d_1 (20,9)	d_2 (8,9)	Cluster
x_1 (23,12)	4.24	15.30	1
x_2 (6,6)	14.32	3.61	2
x_3 (15,0)	10.30	11.40	1
x_4 (15,28)	19.65	20.25	1
x_5 (20,9)	0.00	12.00	1
x_6 (8,9)	12.00	0.00	2
x_7 (20,11)	2.00	12.17	1
x_8 (8,13)	12.65	4.00	2

   We then calculate the new means of the clusters by taking the average of the data points in each cluster. We then get the new means of

$$C_1 = (\frac{23+15+15+20+20}{5}, \frac{12+0+28+9+11}{5}) = (18.6, 12)$$

$$C_2 = (\frac{6+8+8}{3}, \frac{6+9+13}{3}) = (7.33, 9.33)$$

2. We do the same process as above but with the initial means of (15, 0) and (15, 29). We then get the following table:

K = 2	d_1 (15,0)	d_2 (15,29)	Cluster
x_1 (23,12)	14.24	18.79	1
x_2 (6,6)	10.82	24.70	1
x_3 (15,0)	0.00	29.00	1
x_4 (15,28)	28.00	1.00	2
x_5 (20,9)	10.30	20.62	1
x_6 (8,9)	11.40	21.19	1
x_7 (20,11)	12.08	18.68	1
x_8 (8,13)	14.76	17.46	1

We then calculate the new means of the clusters by taking the average of the data points in each cluster. We then get the new means of

$$C_1 = (\frac{23+6+15+20+8+20+8}{7}, \frac{12+6+0+9+9+11+13}{7})$$
$$= (14.28, 8.57)$$
$$C_2 = (\frac{15}{1}, \frac{28}{1}) = (15, 28)$$

3. K-Means clustering is a very good algorithm for clustering data points. This is especially true for larger datasets and is very simple as its a simple iterative program. However, it is not always guaranteed to find the optimal solution. This is because the algorithm is dependent on the initial means. If the initial means are not chosen well, the algorithm may not be able to find the optimal solution. One way to enhance the algorithm is to run the algorithm multiple times with different initial means and then choose the solution with the lowest cost function. This way, we can ensure that we are getting the optimal solution. This would also fix the problem of the algorithm not being able to find optimal clusters due to outliers. There also has been problems with K-Mean clustering when there are clusters of different sizes, densities, and being a non-gobular shape. To solve this problem, we can increase the amount of clusters to evenly spread out the data points.

Hierarchical Clustering

Consider the following two-dimensional data points:

Data point	x_1	x_2
1	0.1	0.2
2	0.2	0.1
3	0.4	0.8
4	0.5	1.0
5	0.7	0.35

1. Compute the Euclidean distance between every pair of points. Show your results in a 5x5 distance matrix.

2. Apply the single link (MIN) algorithm to cluster the objects. Draw the dendrogram for the clusters assuming the distance measure is Euclidean. Make sure vou label the distance axis of the dendrogram carefully.

3. Suppose we apply the complete link (MAX algorithm to cluster the objects. Draw the dendrogram for the clusters assuming the distance measure is Euclidean. Make sure you label the distance axis of the dendrogram carefully.

1. We first calculate the Euclidean distance between every pair of points. We then get the following distance matrix:

DATA POINT	1	2	3	4	5
1	0				
2	.14	0			
3	.67	.73	0		
4	.89	.95	.22	0	
5	.62	.56	.54	.68	0

2. To apply the single link (MIN) algorithm to cluster the objects we have to first find the minimum vaue of the whole table. The first cluster to appear would be the minimum value at points 1 and 2, .14. We then make a cluster of it and shifted every minimum distance. We then get the following table: The next value is in points 3 and 4 which is the value .22.

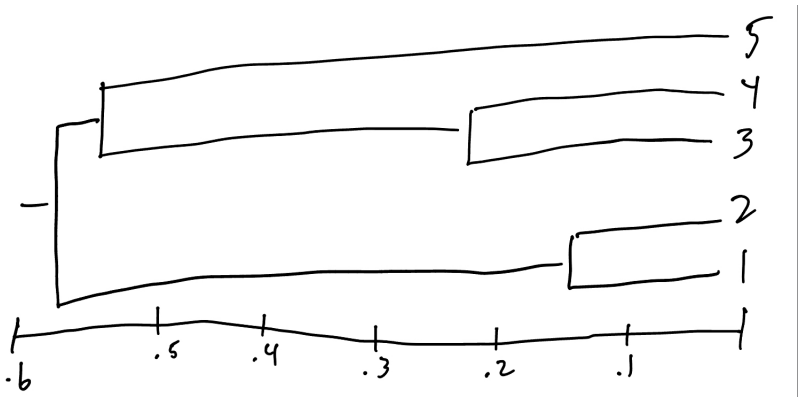
DATA POINT	(1,2)	3	4	5
(1,2)	0			
3	.67	0		
4	.89	.22	0	
5	.56	.54	.68	0

We then create the cluster and get the following table:

DATA POINT	(1,2)	(3,4)	5
(1,2)	0		
(3,4)	.89	0	
5	.56	.54	0

The last cluster is (3,4) and 5, a value of .54, and after clustering, there is a remaining value of .56. We then get the following table and lastly dendrogram:

DATA POINT	(1,2)	((3,4),5)
(1,2)	0	
((3,4),5)	**.56**	0



3. To apply the complete link (MAX) algorithm to cluster the dataset, we do the same as the previous problem, but we use the maximum value instead of the minimum value. We first find the minimum value in the table which in this case would be at points 1 and 4 with the value of .14, then move everything by the maximum instead of the minimum. We then get the following table:

DATA POINT	(1,2)	3	4	5
(1,2)	0			
3	.73	0		
4	.95	.22	0	
5	.62	.54	.68	0

The next points are 3 and 4 at a value of .22. We then get the following table:

DATA POINT	(1,2)	(3,4)	5
(1,2)	0		
(3,4)	.95	0	
5	.62	.68	0

Lastly we cluster the value .62 at (1,2) and 5 with a remainging value of .95 to get the following table and dendrogram:

DATA POINT	((1,2),5)	(3,4)
((1,2),5)	0	
(3,4)	.95	0