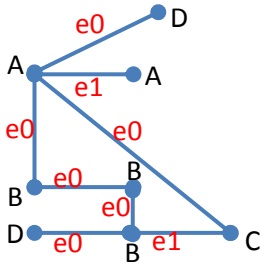
Practice Problems for Intelligent Data Analysis Final Exam.

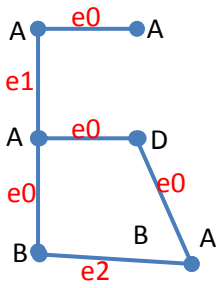
1. Consider the dataset shown here. Choose best test values for attributes X, Y, and Z to split the dataset in two partitions for each possible test. Use Gain-ratio to find the best attribute-value test that should be included at the top of the decision tree. Show your work for arriving at the answer.
2. What is the accuracy of classification plane given by X+Y+Z+1=0. What points lie on each side of this plane?

Point Id	X	Y	Z	Class
P1	3	2	8	0
P2	4	2	3	0
P3	3	5	9	0
P4	8	-12	2	0
P5	9	-19	7	1
P6	8	-3	-11	1
P7	10	-16	4	1

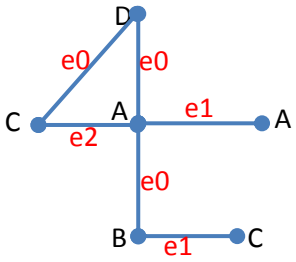
3. Consider the following set of transaction: (T1: ABCEGH, T2: ACFHJ; T3: ABCJ; T4: BCDEF; T5: BCEJ). Determine the frequent 3-itemsets in this set. Find the 3-itemset with the highest support. Use this item-set to find the rule that has the maximum confidence level. Find the rule that has the maximum Lift value.
4. What does it mean when an itemset is closed but not maximal. Illustrate your answer with an example.
5. Consider the following set of graphs.
   a. Determine the Canonical labels for graphs G2 and G3
   b. Find the frequent subgraphs of these three graphs using the minimum support level of 3.



G1

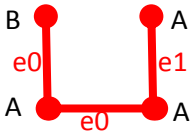


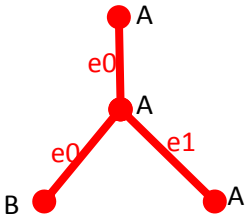
G2



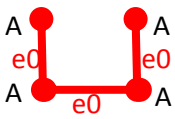
G3

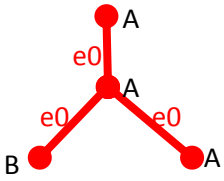
6. Consider the following sequences: (s1: ABACAD, s2: BABCAD, s3:ACABADE, s4:ACBAD). Find the most frequent subsequences using a minimum support threshold of 3.
7. Consider the following two subgraphs of size 3 each (number of edges is 3). They share a subgraph of size 2 as their parent. Generate candidate subgraph(s) of size 4 from these two subgraphs.





8. Repeat problem #7 for the following subgraphs.
9. Consider the following points on a number line: (-2, -4, 28, -7, 16, 4, 5, 9, 12, 2, -10). An iteration of K-means algorithm includes the phases of determining the nearest center for each data point





and the recalculating the centers.

a. Assume starting centers to be -15, 0, and 15 and show the results two iterations of the k-means algorithm. After each iteration show the cluster center to which each data point belongs and the new values of the revised cluster centers.
b. Repeat part(a) above with the difference that the starting locations of the cluster centers are: (14, 18, and 22)

	C1	C2	C3	C4	C5
C1	0	8	9	11	10
C2		0	12	5	16
C3			0	14	15
C4				0	8
C5					0

10. The clusterings produced by two different algoritms are: ((A B C) (D E) (F G H)) and ((A B F) (C D E G H)). What is the Rand Index for these two clusterings?
11. Consider an intermediate stage in a hierarchical clustering algorithm given by the following distance matrix between various clusters. The number of data points in each clusters is: (C1:20, C2:8, C3:80, C4:24, C5:10)
    a. Show the distance matrix at the next iteration when we use the single link algorithm.
    b. Show the distance matrix at the next iteration when we use the complete link algorithm.
    c. Show the distance matrix at the next iteration when we use the UPGMA algorithm.
12. The difference between Interval and Ratio type of variables is as follows:

Interval variables are variables for which their central characteristic is that they can be measured along a continuum and they have a numerical value (for example, temperature measured in degrees Celsius or Fahrenheit). So the difference between 20C and 30C is the same as 30C to 40C. However, temperature measured in degrees Celsius or Fahrenheit is NOT a ratio variable.

Ratio variables are interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable. So, temperature measured in degrees Celsius or Fahrenheit is not a ratio variable because 0C does not mean there is no temperature. However, temperature measured in Kelvin is a ratio variable as 0 Kelvin (often called absolute zero) indicates that there is no temperature whatsoever. Other examples of ratio variables include height, mass, distance and many more. The name "ratio" reflects the fact that you can use the ratio of measurements. So, for example, a distance of ten metres is twice the distance of 5 metres.

Give three examples of each type of variable. These should be different from the examples given in the above discussion.