Birla Institute of Technology and Science-Pilani, Hyderabad Campus
Second Semester 2021-22
Comprehensive Examination
Course No CS F415 **Course Title: Data Mining**
Type: Closed Book **Time: 180 min** **Total Marks:120 (40%)** **Date: 10/5/22**

Attempt all the questions. Answer to the point and in the same sequence as given in the question paper.

1. (10 points) Write "T" if the statement is True and "F" otherwise.
   (a) (1 point) Given a rule $R : X \to Y$, support $s(R)$ is always less than or equal to confidence $c(R)$. **T**
   (b) (1 point) Anti-monotone property states that $X \subseteq Y \implies s(X) < s(Y)$ where X and Y are two sets and $s(X)$ is the support of X. **F**
   (c) (1 point) Given a frequent item-set L of size $|L| = k$, total number of candidate association rules is $2^k - 2$. **T**
   (d) (1 point) Closed frequent item-sets are the subsets of maximal frequent item-sets. **F**
   (e) (1 point) $< \{1\}\{2\} >$ is a contiguous subsequence of $< \{1\}\{3\}\{2\} >$. **F**
   (f) (1 point) Suppose all the points in a dataset D are the core points for DBSCAN algorithm with given ϵ and $minpts$. Applying DBSCAN on this dataset will always result in a single cluster. **F**
   (g) (1 point) Silhouette score ranges from 0 to ∞. **F**
   (h) (1 point) Clusters in the hierarchical clustering are always disjoint. **F**
   (i) (1 point) Clusters can be separated to its previous states in hierarchical clustering. **F**
   (j) (1 point) Data points in a single cluster using Complete linkage clustering forms a complete sub-graph. **T**

2. (20 points) (a) (3 points) What objective function does the K-Means algorithm minimize?
   (b) (5 points) Prove that the centroid of a cluster in the K-Means algorithm is the mean of the points in the cluster.
   (c) (12 points) Cluster the following 8 points into three clusters: $\{A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)\}$. Each point is represented with their (x, y) locations. Assume that initial cluster centers are $A1, A4$, and $A7$. Write down the points in each cluster along with the center of each cluster at the end of iteration-1 of K-Means clustering. Assume the distance between two points $p = (x_1, y_1)$ and $q = (x_2, y_2)$ is $d(p, q) = |x_2 - x_1| + |y_2 - y_1|$

Solution:

(a) The goal of K-Means clustering is to minimize the SSE which is defined as:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2$$

for one dimensional data where c_i is the mean of the i^{th} cluster C_i.

$$\begin{aligned}\frac{\partial}{\partial c_k}\text{SSE} &= \frac{\partial}{\partial c_k}\sum_{i=1}^{K}\sum_{x \in C_i}(c_i - x)^2 \\ &= \sum_{i=1}^{K}\sum_{x \in C_i}\frac{\partial}{\partial c_k}(c_i - x)^2 \\ &= \sum_{x \in C_k} 2 * (c_k - x_k) = 0\end{aligned}$$

(b) $$\sum_{x \in C_k} 2 * (c_k - x_k) = 0 \Rightarrow m_k c_k = \sum_{x \in C_k} x_k \Rightarrow c_k = \frac{1}{m_k}\sum_{x \in C_k} x_k$$

3. (30 points) Consider the following similarity matrix:

	p1	p2	p3	p4	p5
p1	1.00				
p2	0.10	1.00			
p3	0.41	0.64	1.00		
p4	0.55	0.47	0.44	1.00	
p5	0.35	0.98	0.85	0.76	1.00

(a) (15 points) Perform single link hierarchical clustering and draw the dendrogram.

(b) (15 points) Perform complete link hierarchical clustering and draw the dendrogram.

4. (20 points) Consider the following set of frequent 3-itemsets:

$$\{1,2,3\},\{1,2,4\},\{1,2,5\},\{1,3,4\},\{1,3,5\},\{2,3,4\},\{2,3,5\},\{3,4,5\}.$$

Assume that there are only five items in the data set.

(a) (7 points) List all candidate 4-item-sets obtained by a candidate generation procedure using the $F_{k-1} \times F_1$ merging strategy.

(b) (7 points) List all candidate 4-item-sets obtained by a candidate generation procedure using the $F_{k-1} \times F_{k-1}$ merging strategy.

(c) (6 points) List all candidate 4-item-sets that survive the candidate pruning step. While pruning an item-set, show the infrequent item-set that leads to pruning.

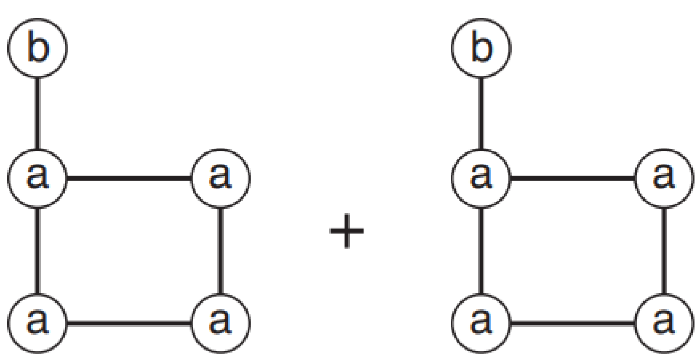
Solution:

5. (10 points) Assume that we consider only simple undirected graphs in this question.

(a) (5 points) Suppose the graph $G = (V, E)$ has n vertices $\{v_1, v_2, ..., v_n\}$ and m edges. Suppose, $d(v)$ denote the degree of the vertex v. Prove that $\sum_{i \in [1:n]} d(v_i) = 2m$.

(b) (2 points) What is the time complexity needed to determine the canonical label of a graph that contains $|V|$ vertices?

(c) (3 points) Draw all candidate subgraphs obtained from joining the pair of graphs shown below:



Solution:

6. (10 points) For each of the sequences $w =< e_1 e_2 ... e_i ... e_{last} >$ given below, determine whether they are subsequences of the sequence:

$$< \{1, 2, 3\}\{2, 4\}\{2, 4, 5\}\{3, 5\}\{6\} >.$$

subjected to the following timing constraints:

- mingap = 0
- maxgap = 3
- maxspan = 5
- ws = 1

For each of these subsequence, if one is not valid, mention which of these constraints it is violating.

(a) (2 points) $w =< \{1\}\{2\}\{3\} >$.
(b) (2 points) $w =< \{1, 2, 3, 4\}\{5, 6\} >$.
(c) (2 points) $w =< \{2, 4\}\{2, 4\}\{6\} >$.
(d) (2 points) $w =< \{1, 2\}\{3, 4\}\{5, 6\} >$.
(e) (2 points) $w =< \{1\}\{2, 4\}\{6\} >$.

7. (10 points) Consider the table below for answering the following questions:

Customer ID	Transaction ID	Items Bought
1	0001	$\{a, d, e\}$
1	0024	$\{a, b, c, e\}$
2	0012	$\{a, b, d, e\}$
2	0031	$\{a, c, d, e\}$
3	0015	$\{b, c, e\}$
3	0022	$\{b, d, e\}$
4	0029	$\{c, d\}$
4	0040	$\{a, b, c\}$
5	0033	$\{a, d, e\}$
5	0038	$\{a, b, e\}$

(a) (3 points) Compute the support for itemsets $\{e\}$, $\{b, d\}$ and $\{b, d, e\}$ by treating each transaction ID as a market basket.

(b) (4 points) Compute the confidence for the association rules $\{b, d\} \rightarrow e$ and $\{e\} \rightarrow \{b, d\}$. Is confidence a symmetric measure?

(c) (3 points) Compute the support of the itemsets $\{e\}$, $\{b, d\}$, and $\{b, d, e\}$ by treating each customer ID as a market basket.

8. (10 points) (a) (5 points) Triangle inequality (for three data points a, b, and c, $d(a, c) \leq d(a, b)+d(b, c)$) can be used in the assignment step of K-Means to avoid calculating all the distances of each point to each cluster centroid. Assume that, at some iteration of K-Means, x is a point and b and c are two different cluster centers. Prove that if $d(b, c) \geq 2d(x, b)$ then $d(x, c) \geq d(x, b)$.

(b) (5 points) Let c_1, c_2, c_3 be the confidence values of the rules $\{p\} \rightarrow \{q\}, \{p\} \rightarrow \{q, r\}$ and $\{p, r\} \rightarrow \{q\}$, respectively. If we assume that c_1, c_2, and c_3 have different values, which rule has the lowest confidence? Please show the derivation to come to a conclusion.