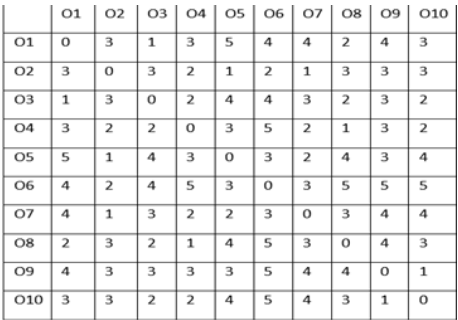
**BITS Pilani, Hyderabad Campus**

**Mid-Semester Test (Regular Exam)**

**Course No.: CSF415 Course Title: Data Mining Max. Marks: 60**

Section #1

QID: 108: Consider the Similarity matrix shown in below Table:                              **[2]**



Assume that, Data scientist A has applied k-means clustering algorithm on above Similarity matrix. The final clusters after applying k-means are given below:

***Data scientist A:***

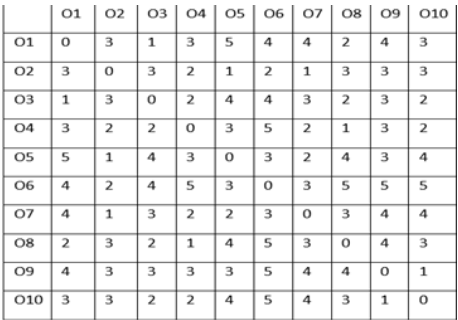
**Cluster 1=(O2,O3, O5)**

**Cluster 2=(O1, O4, O9)**

**Cluster 3=(O8,O6,O7,O10)**

**What is the single-link distance between cluster 1 and cluster 3**

QID: 109: Consider the Similarity matrix shown in below Table:                              **[2]**



Assume that, Data scientist A has applied k-means clustering algorithm on above Similarity matrix. The final clusters after applying k-means are given below:

***Data scientist A:***

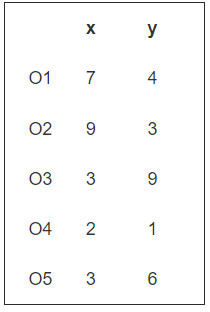
**Cluster 1=(O2,O3, O5)**

**Cluster 2=(O1, O4, O9)**

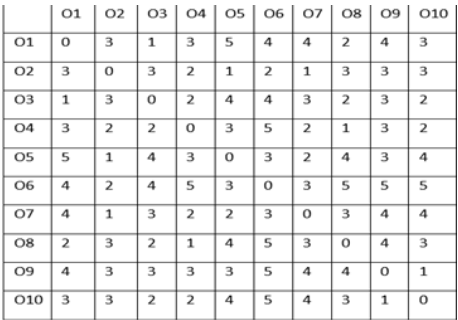
**Cluster 3=(O8, O6,O7,O10)**

**What is the complete- link distance between cluster 1 and cluster 2**

QID: 110: **Given, point points with the following attributes. Draw dendrogram for solution of hierarchical clustering [2]**



QID: 111: Consider the Similarity matrix shown in below Table:                              **[2]**



Assume that, Data scientist A has applied k-means clustering algorithm on above Similarity matrix. The final clusters after applying k-means are given below:

***Data scientist A:***

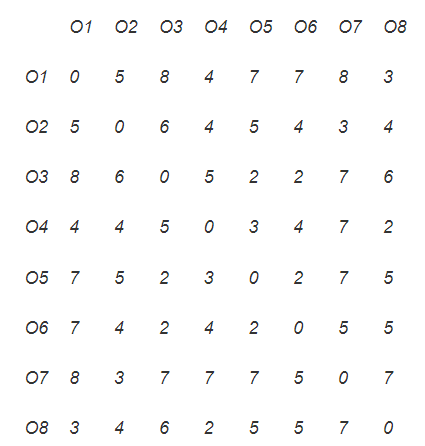
**Cluster 1=(O2,O3, O5)**

**Cluster 2=(O1, O4, O9)**

**Cluster 3=(O8, O6,O7,O10)**

***What is average link distance between cluster 1 and cluster 3***

QID: 203:  What is the number of core point after applying DBSCAN on below data with min-point=3, and eps=2 (<=).  **[2]**

**

QID: 205: Dependent or independent variable not required in  **[2]**

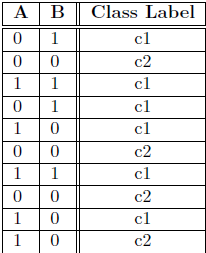
* regression analysis
* classification analysis
* cluster analysis
* None of These

QID: 206: The following data shows cyclomatic complexity of software.

***14    14    13     4     9     1     7     5     3     3     7     2     9     8    11    11    10                     1     2     5        8    10     7    13    11    15     8     5     2    10***

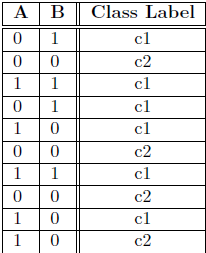
          Draw histogram to displays the above information. **[2]**

QID: 201: For the training data set given below in Table

****

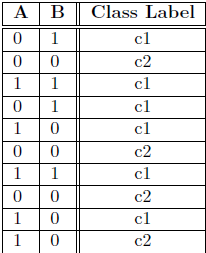
**The chi-square value of feature A is [2]**

QID: 204: For the training data set given below in Table

****

The Entropy of child of feature B when B=0 is **[2]**

QID: 202: For the training data set given below in Table

****

**The gini index of feature A is   [2]**

QID 301: Consider the transaction data shown in the below Table: ***For all of the sub-questions the minimum support is 0.3 and the minimum confidence is 0.6 [3]***

***What is the number of scans required to find all three frequent item-sets?***

***NOTE: only 3 frequent item-sets***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | O1 | O2 | O3 | O4 | O5 |
| T1 | 1 | 1 | 1 | 0 | 1 |
| T2 | 1 | 0 | 0 | 1 | 0 |
| T3 | 1 | 1 | 1 | 1 | 0 |
| T4 | 0 | 0 | 1 | 1 | 1 |
| T5 | 0 | 1 | 1 | 1 | 0 |
| T6 | 0 | 0 | 1 | 0 | 0 |
| T7 | 1 | 0 | 0 | 1 | 0 |
| T8 | 1 | 0 | 0 | 0 | 0 |
| T9 | 1 | 1 | 0 | 0 | 0 |
| T10 | 1 | 1 | 0 | 0 | 0 |

QID 302: Apply the F-P tree on the above data and write only the number of nodes present in Fp-Tree (including null node).  **[3]**

QID 303: What is number of scan (original data base) required to find only 2-frequent item sets if we are going to add three more transactions (T13: O1, O2, O3 T11: O1, O5. and T12: O1, O5,O2) **[3]**

QID 304: What is number of scan (only deleted database) required to find only 2-frequent item sets if we are going to delete T1, T2, T4  **[3]**

QID 305: What is the count for bucket with hash value 2 using hash function h(x, y ) = (15 ∗ ID(x ) + ID(y )) mod 7 **[3]**

QID 306: What is the number of 2-items candidate set using Apriori algorithm? **[3]**

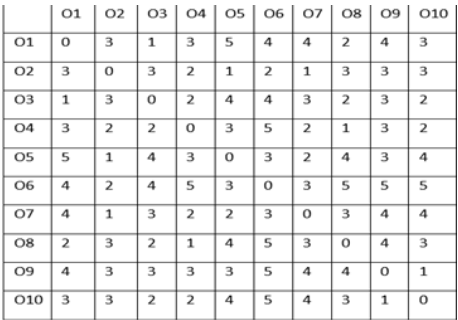
QID: 401: Consider the training data shown in below Table:                       **[4+2]**

***NOTE: Only write the value, no calculation***



1. What is the gain ratio of all features?
2. What is the rank of feature a1, and a4 using gain ratio?

QID: 103: Consider the Similarity matrix shown in below Tables:                        ***[8]***



1. What are the two clusters obtained after applying hierarchical clustering?     NOTE: only write cluster with objects present in that cluster.
2. What is the average value of intra cluster distance for above clusters?
3. What is the average value of inter cluster distance for above clusters?
4. What is the silhouette coefficient value of object 2 for above solution?

QID: 104: Assume the following dataset is given: (1,1), (2,3), (2,1), (1,4), (2,4), (7,2), (0,4), (4,0). Fuzzy C-Means clustering is used with k=2 to cluster the dataset. Moreover, Manhattan Distance is used as the distance function to compute distances between centroids and objects in the dataset. Moreover, Fuzzy C-Means clustering initial clusters centroids are as follows:               **[8]**

C1: (2,1), C2: (4,4)

Now Fuzzy C-Means clustering is run for one iteration; what are the new clusters, membership of objects, and what are their centroids?   NOTE: only write final answer

***The Manhattan distance as the sum of absolute differences***

***ManhattanDistance [{a, b, c}, {x, y, z}]***

***Abs [a − x] + Abs [b − y] + Abs [c − z]***