** AUTUMN END SEMESTER EXAMINATION-2022**

*Semester: 8th Semester*

*Programme: B.Tech*

*Branch/Specialization: Minor*

**8thSemester, Minor**

**DATA MINING AND DATA WAREHOUSING**

**IT 3031**

**(For 2020 Admitted Batches)**

**Time: 2 Hours Full Marks: 50**

**KIIT Deemed to be University**

**End Semester Examination(Autumn Semester-2022)**

**SECTION-A(Answer All Questions. Each question carries 2 Marks)**

**Time:30 Minutes (7×2=14 Marks)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Question No** | **Question Type (MCQ/SAT)** | **Question** | **CO Mapping** | **Answer Key**  **(For MCQ Questions only)** |
| **Q.No:1** | Let a configuration of the k means clustering algorithm correspond to the k way partition (on the set of instances to be clustered) generated by the clustering at the end of each iteration. Is it possible for the k-means algorithm to revisit a configuration? Justify how your answer proves that the k means algorithm converges in a finite number of steps. | Question -1 on concept 1 | CO5 |  |
|  | Consider a dataset with 200 samples. Out of these, 120 samples belong to class A and 80 samples belong to class B. Calculate the entropy of this dataset using the formula for entropy. | Question -2 on concept 1 | CO5 |  |
|  | Give one advantage of hierarchicalclustering over K-means clustering, and one advantage of K-means clustering over hierarchical clustering. | Question -3 on concept 1 | CO5 |  |
|  | Differentiate between classification and clustering | Question -4 on concept 1 | CO5 |  |
| **Q.No:2** | Explain the difference between overfitting and underfitting in a supervised learning setup. | Question -1 on concept 2 | CO4 |  |
|  | Compute the outcome for the Sigmoid function for input x=1.5 | Question -2 on concept 2 | CO4 |  |
|  | Compute the outcome for the Sigmoid function for input x=2.8 | Question -3 on concept 2 | CO4 |  |
|  | Compute the outcome for the Sigmoid function for input x=3.1 | Question -4 on concept 2 | CO4 |  |
| **Q.No:3** | Define normal, ordinal variables with suitable examples | Question -1 on concept 3 | CO1 |  |
|  | Compute the Jaccard similarity between the following two binary vectors:  X = {0011001100}  Y = {1110001110} | Question -2 on concept 3 | CO1 |  |
|  | Compute the Jaccard similarity between the following two binary vectors:  X = {1100001100}  Y = {1110001110} | Question -3 on concept 3 | CO1 |  |
|  | Compute the Jaccard similarity between the following two binary vectors:  X = {0011001100}  Y = {0010111110} | Question -4 on concept 3 | CO1 |  |
| **Q.No:4** | What is a Data Mart? Explain itsusefulness in a Data Warehouse | Question -1 on concept 4 | CO2 |  |
|  | Explain difference between OLAP and OLTP | Question -2 on concept 4 | CO2 |  |
|  | Explain the difference between rollup and drill down in data mining | Question -3 on concept 4 | CO2 |  |
|  | Explain the difference between slicing and dicing in data mining | Question -4 on concept 4 | CO2 |  |
| **Q.No:5** | Explain the difference between mean, median and mode with an example. | Question -1 on concept 5 | CO1 |  |
|  | How a similarity measure is useful in data mining? Explain any 2 similarity measures. | Question -2 on concept 5 | CO1 |  |
|  | Explain the min-max normalization | Question -3 on concept 5 | CO1 |  |
|  | What are some applications of data mining? | Question -4 on concept 5 | CO1 |  |
| **Q.No:6** | What is entropy? For a series {0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.50} calculate the entropy. | Question -1 on concept 6 | CO2&CO3 |  |
|  | Find the normalized data after performing min-max normalization by setting min = -1 and max = +1 for the data values 100, 200, 300, 400, 50, 600, 700. | Question -2 on concept 6 | CO2&CO3 |  |
|  | Find the five point summary of the following data:  2,4,6,7,8,8,8,10,12,14,14,15 | Question -3 on concept 6 | CO2&CO3 |  |
|  | Explain Euclidean distance and Manhattan distance. What is the significant difference between them? | Question -4 on concept 6 | CO2&CO3 |  |
| **Q.No:7** | How to calculate information gain in decision tree? | Question -1 on concept 7 | CO5 |  |
|  | Write any four limitations of hierarchical clustering | Question -2 on concept 7 | CO5 |  |
|  | If a classifier correctly classify 80 tuples out of 130 tuples of test data, what will be the error rate of the classifier? | Question -3 on concept 7 | CO5 |  |
|  | In K-nearest neighbour classification method, why do we select an odd value for *k*? | Question -4 on concept 7 | CO5 |  |

**SECTION-B(Answer Any Three Questions. Each Question carries 12 Marks)**

**Time: 1 Hour and 30 Minutes** **(3×12=36 Marks)**

|  |  |  |
| --- | --- | --- |
| **Question No** | **Question** | **CO Mapping**  **(Each question should be from the same CO(s))** |
| **Q.No:8** | Suppose you have the following training set with three boolean input x, y and z, and a boolean output U.    ﻿Suppose you have to predict U using a naive Bayes classifier, answer the following:   1. After learning is complete what would be the predicted probability ﻿P (U = 0|x = 0, y = 1, z = 0)? 2. Using the probabilities obtained during the Bayes Classifier training, what would be the predicted probability P (U = 0|x = 0)? | CO4 |
| Draw a fully connected Neural Network with the following specifications:  • 4 Inputs x1, x2, x3, x4  • 2 Hidden Layers  • Hidden Layer 1 is made up of    • Hidden Layer 2 is made up of    • The output layer is made up of    • Include the bias terms in all layers marking them as *β*m with *m* as their layer number. Make sure you label your drawing appropriately. |
| ISRO wants to be able to discriminate between Martians (M) and Humans (H) based on the below characteristics: Green ∈{N, Y}, Legs ∈ {2 legs, 3 legs}, Height ∈ {S, T}, Smelly ∈ {N, Y}. Our available training data is as follows:    ﻿Learn a decision tree using the ID3 algorithm and draw the tree. |
| **Q.No:9** | Consider the dataset: {0, 4, 5, 20, 25, 39, 43, 44}  Suppose we want the *two* top level clusters from this dataset. What will be the output of single link, complete link, and average link clustering methods? If single link and complete link give the same 2 clusters, will average link clustering method give same 2 cluster? Explain your answer. | CO5 |
| Consider a dataset with the following 8 data points in 2-dimensional space:  {(2,3), (4,5), (5,8), (7,3), (9,5), (6,2), (8,8), (10,4)}  Perform **K-means clustering** with K=3 (i.e., partition the data into 3 clusters) using **Euclidean distance** as the distance metric. Assume the initial centroids are (3,4), (6,6), and (8,3). Show **2 iteration** of the K-means algorithm including the assignment of points to clusters and the calculation of new centroids. |
| Explain the algorithm for constructionof k-nearestneighbour classifier. From the given data, usingk-nearest neighbour classifier find the class of the datatuple (38,45), where **k value is 3**.Mention about advantage and disadvantage of k-nearest neighbour classifier over any other classification techniques.   |  |  |  |  | | --- | --- | --- | --- | | # | A | B | Class | | 1 | 26 | 30 | L | | 2 | 30 | 32 | L | | 3 | 36 | 42 | L | | 4 | 36 | 52 | P | | 5 | 40 | 62 | P | | 6 | 43 | 70 | P | |
| **Q.No:10** | Suppose a group of 12 sales price records has been sorted as follows: 5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215. Partition them into 3 bins by each of the following methods:   1. Equal frequency partitioning 2. Equal width partitioning   For each type of partitioning, smooth out the bin values by **boundary**. | CO1 & CO2 |
| What is a Data Warehouse? How is it different from traditional database? Briefly explain the different stages of a Data Warehouse architecture with suitable diagrams. |
| While working with real world data, we encounter the problem of missing values. Explain the following 3 types of missing values (if possible, with suitable examples):   1. Missing completely at random (MCAR) 2. Missing at Random (MAR) 3. Missing Not at Random (MNAR) |
| **Q.No:11** | Consider a retail dataset containing 1000 transactions with items A, B, C, and D. The support threshold is set to 0.2 (20%). Use the Apriori algorithm to find **frequent itemsets** and association rules. The itemsets are as follows:  Transaction 1: A, B, C  Transaction 2: A, C, D  Transaction 3: A, B  Transaction 4: B, C, D  Transaction 5: A, D  Transaction 6: B, C  Transaction 7: A, C  Transaction 8: B, D  Transaction 9: C, D | CO3 & CO6 |
| Briefly, explain the following:  I) Web Mining  II) Spatial Mining  III)Temporal Mining  IV)Multimedia Mining |
| Construct the confusion matrix for the following actual vs predicted data and compute the following performance matrices (in terms of %)   1. Accuracy 2. Precision 3. Specificity 4. Recall  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | Actual | N | N | Y | Y | Y | N | Y | N | | Predicted | Y | Y | N | N | Y | N | Y | N | |