Hall Ticket Number:

Code No. : 22958

VASAVI COLLEGE OF ENGINEERING *(AUTONOMOUS)*, HYDERABAD

Accredited by NAAC with A++ Grade

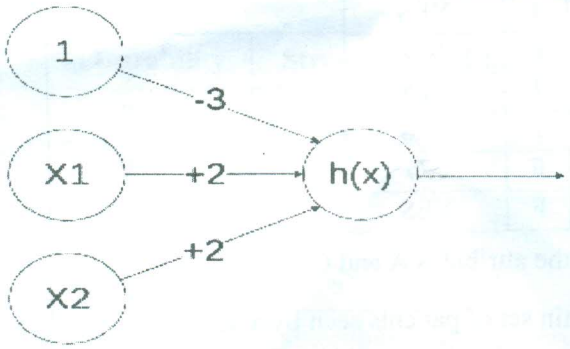
M.Tech. (C.S.E.) II-Semester Main Examinations, September-2022

Data Mining

Time: **3 hours** Max. Marks: **60**

*Note: Answer all questions from **Part-A** and any **FIVE** from **Part-B***

Part-A (10× 2 = 20 Marks)

Q. No.	Stem of the question	M	L	CO	PO
1.	For the following vectors x and y, Compute Cosine similarity. x = (0, 1, 2,1) y = (1, 2, 0, 3)	2	1	1	1,2
2.	Draw a diagram to show the steps of KDD.	2	1	1	1,2
3.	Compute Entropy at a node with 10 samples and C1=2, C2=8.	2	1	2	1,2
4.	Compare the role of Training data and Test data in a Classification model.	2	1	2	1,2
5.	Given below is a simple ANN with 2 inputs X1, X2 $\in \{0, 1\}$ and edge weight -3, +2, +2 $h(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & otherwise \end{cases}$  Which logical function does the ANN compute?	2	2	3	1,2
6.	Differentiate linearly separable and non-linearly separable problems. Give one example for each.	2	1	3	1,2
7.	Find the support count of {a}, {b}, {a, c}, {a, b}. based on the following table: Transaction ID | Itemsets 1 | {a, b, c} 2 | {a, c} 3 | {a, d} 4 | {b, e, f}	2	2	4	1,2
8.	What is Association Rule Mining? List two real-world applications of Association Rule Mining.	2	1	4	1,2
9.	Define Minpts and Epsilon used in DBSCAN algorithm.	2	1	5	1,2
10.	What is Dendrogram. Give an example.	2	1	5	1,2

Contd... 2

Code No. : 22958

Part-B (5 × 8 = 40 Marks)

11. a)	Why we need to perform data pre-processing? Explain data pre-processing techniques.	4	2	1	1,2
b)	Suppose we have the following 2-D data set:	4	2	1	1,2

	A1	A2
x1	1.5	1.7
x2	2	1.9
x3	1.6	1.8
x4	1.2	1.5
x5	1.5	1.0

Consider the data as 2-D data points. Given a new data point, x = (1.4,1.6) as a query, rank the database points based on similarity with the query using Euclidean distance, and Manhattan distance.

12. a)	Explain Sequential covering algorithm to create a Rule-set used in Rule-based classification.	3	2	2	1,2,3
b)	Assume a domain with three attributes A, B, and C. Each attribute has two possible values T and F. Given below is a set of instances.	5	3	2	1,2,3

A	B	C	TARGET
T	T	T	YES
T	T	F	NO
T	F	T	YES
F	T	T	YES
F	T	F	NO
F	F	F	YES

Calculate the information gain for the attributes A and C.

13. a)	Given the following data on a certain set of patients seen by a doctor, can the doctor conclude that a person having chills, fever, mild headache and without running nose has the flue? (Use Navie Bayes algorithm for prediction)	5	3	3	1,2,3

chills	running nose	headache	fever	has flu
Y	N	mild	Y	N
Y	Y	no	N	Y
Y	N	strong	Y	Y
N	Y	mild	Y	Y
N	N	no	N	N
N	Y	strong	Y	Y
N	Y	strong	N	N
Y	Y	mild	Y	Y

b)	Describe Ensemble methods for Classification.	3	2	3	1,2,3

Contd... 3

Code No. : 22958

14. a)	Construct the Fp-tree from the below transaction database. TID | items bought 100 | {f, a, c, d, g, i, m, p} 200 | {a, b, c, f, l, m, o} 300 | {b, f, h, j, o} 400 | {b, c, k, s, p} 500 | {a, f, c, e, l, p, m, n}	5	3	4	1,2,3
b)	Explain how Association rule mining is performed on the data with continuous attributes?	3	2	4	1,2,3
15. a)	Use single link to find the distance between the clusters given below. C1= {A1=(2,8), A2=(3,5), A3=(8,4), A4=(5,8)}, C2= { A5=(7,5), A6=(4,4), A7=(1,2), A8=(4,9)}	5	3	5	1,2,3
b)	Explain the steps in DBSCAN algorithm to perform density-based clustering.	3	2	5	1,2,3
16. a)	Normalize the following data using Z-score normalization. 200,300,400,600,1000	4	3	1	1,2
b)	Explain the steps of Decision tree induction algorithm.	4	2	2	1,2
17.	Answer any ***two*** of the following:				
a)	Consider the following data set to predict class with k nearest neighbor for Acid Durability=3 and strength =7. **Name** | **Acid Durability** | **Strength** | **Class** Type-1 | 7 | 7 | F Type-2 | 7 | 4 | F Type-3 | 3 | 4 | T Type-4 | 1 | 4 | T	4	3	3	1,2,3
b)	Trace the results of using the Apriori algorithm on the grocery store example with support threshold S=33.34% and Confidence threshold c=60%. Show the candidate and frequent itemsets for each database scan. Enumerate all the final frequent itemsets Transaction ID | Items T1 | HotDogs, Buns, Ketchup T2 | HotDogs, Buns T3 | HotDogs, Coke, Chips T4 | Chips, Coke T5 | Chips, Ketchup T6 | HotDogs, Coke, Chips	4	3	4	1,2,3
c)	Explain K-means algorithm with an example.	4	2	5	1,2,3

M : Marks; L: Bloom's Taxonomy Level; CO; Course Outcome; PO: Programme Outcome

i)	Blooms Taxonomy Level – 1	20%
ii)	Blooms Taxonomy Level – 2	40%
iii)	Blooms Taxonomy Level – 3 & 4	40%

*****