Register Number:……..

KUMARAGURU college of technology
character is life

B.E/B.TECH DEGREE EXAMINATIONS: APRIL /MAY 2024

(Regulation 2018)

Fourth Semester

ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

U18AII4203: Data Mining and Modeling

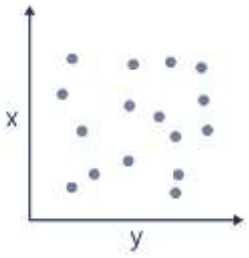
COURSE OUTCOMES

CO1: Understand about data mining basics, issues and the working principle of classification technique.
CO2: Explain the basic concepts of Association Rule Mining and evaluate the working of various Association Rule Mining algorithms.
CO3: Implement classification and prediction techniques.
CO4: Analyze the working of different clustering algorithms.

Time: Three Hours **Maximum Marks: 100**

Answer all the Questions:-
PART A (10 x 2 = 20 Marks)
(Answer not more than 40 words)

1.	Recall the concept of "confidence" in the context of association rules.	CO1	[K_1]
2.	Infer when z-score is preferred over min-max normalization?	CO1	[K_2]
3.	Identity the correlation from the given graph.	CO2	[K_2]



4.	Tell the significance of lift in market basket analysis using Apriori.	CO2	[K_1]
5.	Spell one limitation of traditional association rule mining methods.	CO2	[K_1]
6.	Tell about Laplacian correction in the context of probability estimation.	CO3	[K_1]
7.	Identify why naive Bayesian classification is called "naïve" ?	CO3	[K_1]
8.	How do kernel tricks enhance the capabilities of Support Vector Machines?	CO3	[K_2]
9.	Describe the shape, input parameters and limitations of CHAMELEON clustering algorithm.	CO4	[K_2]
10.	Why is it that BIRCH encounters difficulties in finding clusters of arbitrary shape but OPTICS does not?	CO4	[K_2]

Answer any FIVE Questions:-

PART B (5 x 16 = 80 Marks)

(Answer not more than 400 words)

11.	a)	Design a data preprocessing pipeline for a real-world dataset, considering the stages of data cleaning, transformation, reduction, and discretization.	8	CO1	[K_3]
	b)	Discuss the potential challenges in each stage of data preprocessing pipeline.	8	CO1	[K_2]
12.	a)	Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70. (a) What is the mean of the data? What is the median? (b) What is the mode of the data? Comment on the datas' modality (i.e., bimodal, trimodal, etc.). (c) What is the midrange of the data? (d) Find the first quartile (Q1) and the third quartile (Q3) of the data? (e) Give the five-number summary of the data.	8	CO1	[K_4]
	b)	Normalize the following group of data: 200,300,400,600,1000 using (a) min-max normalization by setting min = 0 and max = 1 (b) z-score normalization	8	CO1	[K_4]
13.	a)	Find all frequent itemsets for the given training set using Apriori (Minimum Support count is 2).	8	CO2	[K_3]

TID	items bought
T100	{ M, O, N, K, E, Y}
T200	{ D, O, N, K, E, Y }
T300	{ M, A, K, E}
T400	{ M, U, C, K, Y}
T500	{ C, O, O, K, I ,E}

	b)	Explain the FP growth algorithm to determine frequent patterns in a dataset?	8	CO2	[K_2]
14.	a)	Explain how a decision tree is generated from a set of training tuples.	8	CO3	[K_2]

	b)	Compare and contrast the rule-based classification approach with classification by backpropagation in neural networks, highlighting their strengths and weaknesses.	8	CO3	[K_2]
15.	a)	Summarize the limitations of KNN algorithm and explain how the value of k is determined.	8	CO3	[K_2]
	b)	A Patient takes a cancer test and result is positive. The test returns a correct positive results in only 98% of the cases in which disease is actually present, and a correct negative result is only 97% of the cases in which disease is not present. Furthermore 0.008 of the entire population have this cancer. Calculate the probability of the patient having cancer?	8	CO3	[K_3]
16.	a)	Apply hierarchical clustering algorithm in the following data and construct a dendrogram to cluster the states.	10	CO4	[K_3]

STATES	x	y
Bihar	0.40	0.53
Karnataka	0.22	0.38
Andhra	0.35	0.32
Tamil Nadu	0.26	0.19
Gujarat	0.08	0.41
Orissa	0.45	0.30

	b)	Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8): (a) Compute the Euclidean distance between the two objects. (b) Compute the Manhattan distance between the two objects. (c) Compute the Minkowski distance between the two objects, using p = 3.	6	CO3	[K_3]

****