BRAINWARE UNIVERSITY

Term End Examination 2023
Programme – B.Tech.(CSE)-2019/B.Tech.(CSE)-2020
Course Name – Data Mining
Course Code - PEC-602C
(Semester VI)

Full Marks : 60 **Time : 2:30 Hours**

[The figure in the margin indicates full marks. Candidates are required to give their answers in their own words as far as practicable.]

Group-A

(Multiple Choice Type Question) 1 x 15=15

1. *Choose the correct alternative from the following :*

(i) Margin of a hyperplane is explained as:

a) The angle it makes with the axes
b) The intercept it makes on the axes
c) Perpendicular distance from its closest point
d) Perpendicular distance from origin

(ii) Identify the following statement is true for a multi-layered perceptron

a) Output of all the nodes of a layer is input to all the nodes of the next layer
b) Output of all the nodes of a layer is input to all the nodes of the same layer
c) Output of all the nodes of a layer is input to all the nodes of the previous layer
d) Output of all the nodes of a layer is input to all the nodes of the output layer

(iii) Data mining is the process of finding _____, novel, useful, and actionable, patterns in large volume of data. Identify the following terms best fills the gap above.

a) voluminous
b) heterogeneous
c) valid
d) Non-valid

(iv) Consider a fuzzy set A defined on the interval X = [0, 10] of integers by the membership Junction: $\mu A(x) = x / (x+2)$
Then the α cut corresponding to $\alpha = 0.5$ will be asked

a) {0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
b) {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}
c) {2, 3, 4, 5, 6, 7, 8, 9, 10}
d) None of the above

(v) Subject orientation defines

a) The science of collecting, organizing, and applying numerical facts
b) Measure of the probability that a certain hypothesis is incorrect given certain observations.
c) One of the defining aspects of a data warehouse, which is specially built around all the existing applications of the operational data
d) None of these

(vi) Indicate the Apriori Algorithm work to mine frequent itemsets and learn association rule over databases

a) The algorithm sorts the items into pairs or triplets and eliminates all but itemsets that appear together the least.

b) The algorithm reduces the number of items being considered by eliminating all itemsets whose support count is less than the minimum support count.

c) The algorithm orders the items in alphabetical order and then picks the itemset that starts with "A" or is closest to the beginning of the alphabet.

d) The algorithm picks out 2 items randomly until they find an itemset with 0% confidence count.

(vii) 'Height of a person', can be recalled as an attribute of type.

a) Nominal
b) Ordinal
c) Interval
d) Ratio

(viii) The Apriori algorithm match with the

a) Positive definiteness property of support

b) Positive semi-definiteness property of support

c) Monotone property of support

d) Antimonotone property of support

(ix) Suppose you are working on stock market prediction. You would like to predict whether the US Dollar will go up against the Euro tomorrow (i.e., whether a dollar will be worth more euros tomorrow than it is worth today). Compare this as a classification or a regression problem

a) Regression
b) Classification
c) All of the above
d) None of these

(x) Indicate the following is NOT a valid association rule

a) A -> DB
b) D -> AB
c) AD -> B
d) DB -> A

(xi) Observe the following criteria is not used to decide which attribute to split next in a decision tree

a) Gini index
b) Information gain
c) Entropy
d) Scatter

(xii) A time series prediction problem is often examine using

a) Multivariate regression
b) Autoregression
c) Logistic regression
d) Sinusoidal regression

(xiii) An outlier is defined as

a) description of records in the data

b) data point which is considerably different from other data points

c) record with missing attributes

d) duplicate record

(xiv) Show the goal of shunting activation model.

a) to make system dynamic

b) to keep operating range of activation value to a specified range

c) to make system static

d) can be either for dynamic or static, depending on inputs

(xv) Transparency defines

a) The large set of candidate solutions possible for a problem

b) The information stored in a database that can be retrieved with a single query

c) Worth of the output of a machine learning program that makes it understandable for humans

d) None of these

Group-B

(Short Answer Type Questions) 3 x 5=15

2. Define Knowledge Discovery in Databases with proper examples. (3)
3. Explain Apriori Algorithm. (3)
4. Illustrate the necessity of activation function and categorize the various types of activation functions (3)
5. Discover the K-Means Clustering Algorithm with proper example (3)
6. Evaluate the multivariant linear regression is more optimal in respect to single variant linear regression. (3)

OR

Reframe an auto associative network to store vector [-1, 1, 1, 1]. Use iterative auto associative network to test the vector with three missing elements. (3)

Group-C

(Long Answer Type Questions) 5 x 6=30

7. Demonstrate the "Snowflake Schema" with proper diagram involving fact and dimension table. (5)
8. Illustrate the basic model of Artificial Neural Network and Illustrate XOR function using McCulloch-Pitts neuron (Consider Binary Data). (5)
9. Construct an auto associative network to store the vectors X1 = [1, 1, 1, 1, 1], X2 = [1, -1, -1, 1, -1], X3 = [-1, 1, -1, -1, -1]. Classify weight matrix with no self-connection. Calculate the energy of the stored patterns. Using discrete Hopfield network test patterns if the test pattern are given as X1 = [1, 1, 1-1, 1], X2 = [1, -1, -1, -1, -1 and X3 = [1,1,-1, -1, -1]. Compare the test patterns energy with the stored patterns energy. (5)
10. Identity the proper combination between the products based on the below mentioned transaction data using Association Rule Mining where support=22% and confidence = 70% (5)

Transaction ID	Item Purchased
1	I1, I2, I5
2	I2, I4
3	I2, I3
4	I1, I2, I4
5	I1, I3
6	I2, I3
7	I1, I3
8	I1, I2, I3, I5
9	I1, I2, I3

11. Calculate the Gain value, considering the below mentioned Data Set (5)

Day	Outlook	Temperature	Humidity	Wind	Play Ball
1	Sunny	Hot	High	Weak	NO
2	Sunny	Hot	High	Strong	NO
3	Overcast	Hot	High	Weak	YES
4	Rain	Mild	High	Weak	YES
5	Rain	Cool	Normal	Weak	YES
6	Rain	Cool	Normal	Strong	NO
7	Overcast	Cool	Normal	Strong	YES
8	Sunny	Mild	High	Weak	NO
9	Sunny	Cool	Normal	Weak	YES
10	Rain	Mild	Normal	Weak	YES
11	Sunny	Mild	Normal	Strong	YES
12	Overcast	Mild	Normal	Strong	YES
13	Overcast	Hot	Normal	Weak	YES
14	Rain	Mild	High	Strong	NO

12. Discriminate between the various types of clustering procedure i.e., Partitional and Hierarchical clustering Algorithm. (5)

OR

Decide the utility of the dimensionality reduction Principle in classification approaches. (5)

****