Bangladesh-Bharat Digital Service and Employment Training

Test-3

Total marks: 100 | Date: 14.09.2022 | Time: 2 hours

Part-I: Choose the correct option and justify your answer with one or two sentence(s) *(20 X 2 = 40)*

1. The unsupervised classification is known as
   a) Clustering
   b) Correlation
   c) Classification
   d) Logistic regression

Correct Answer: a
Explanation: The unsupervised classification is termed as clustering algorithm.

2. Under which of the following categories does the Bayesian classification method fall?
   a) Statistical-based methods
   b) Distance-based method
   c) Error-based method
   d) Decision-tree based method

Correct Answer: a
Explanation: Bayesian classifier is a statistical classifier, which performs probabilistic prediction, that is, it predicts class membership probabilities.

3. Which of the following is true about the K-Nearest Neighbour (KNN) classifier?
   a) It is considered as a lazy learner
   b) It is considered as an eager learner
   c) The learning strategy of KNN depends upon the dataset
   d) The value of k is considered as a constant for all datasets, and always equal to 1

Correct Answer: a
Explanation: KNN is a lazy learner, and not an eager learner. For a discussion on lazy/eager learning of KNN, check. The learning strategy of KNN is fixed, and it does not depend on which dataset upon which KNN is being used. Also, the value of an optimal k is not a constant.

4. Which of the following statement is true about the entropy of an n-dimensional data distributed over k distinct classes?
   a) The lowest possible value of entropy is -1.
   b) The highest possible value of entropy is ∞.
   c) The entropy can be any value between $-\infty$ and $+\infty$.
   d) The entropy is always a positive quantity.

Correct Answer: d
Explanation: The entropy is ranges from 0 to 1, hence it is always positive.

5. In the M-estimation approach of Naïve Bayes Classification, the posterior probability is given by?
   [Where, n = total number of instances from class, n_{C_i} = number of training examples from class C_i that takes the value $A_j = x$, m = equivalent sample size, p = a user-defined parameter]
   a) $P(A_j = x|C_i) = \frac{n_{C_i} - mp}{n - m}$
   b) $P(A_j = x|C_i) = \frac{n_{C_i} + mp}{n - m}$
   c) $P(A_j = x|C_i) = \frac{n_{C_i} - mp}{n + m}$
   d) $P(A_j = x|C_i) = \frac{n_{C_i} + mp}{n + m}$

Correct Answer: d
Explanation: The posterior probability in M-estimate approach is given as $P(A_j = x|C_i) = \frac{n_{C_i} + mp}{n + m}$.

6. In Naive Bayesian Classification technique, the M-estimation approach is used when,
   a) the posterior probability for one of the attributes is infinite
   b) the posterior probability for one of the attributes is zero
   c) one of the prior probabilities is zero
   d) one of the prior probabilities is infinity

Correct Answer: b
Explanation: If the posterior probability for one of the attributes is zero, then the overall class-conditional probability for that class vanishes. In this case the M-estimation technique is used. This generally happens due to the low size of training data.

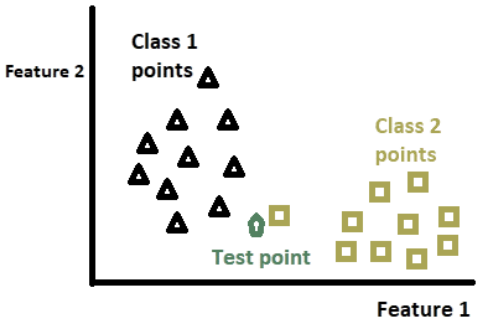
7. To study the effect of rain on the attendance of a particular school, the data for 10 days are collected. In the table below, the amount of rain of a particular day (Heavy/Medium/Low or H/M/L), and the attendance (High/Low or (H/L)) is provided. What is the value of class conditional probability of rain is High given that attendance is High (P (Rain = High | attendance = High))?

Amount of rain	H	H	L	M	L	M	H	L	L	M
Attendance	L	L	H	L	H	H	H	L	H	H

   a) $\frac{1}{3}$
   b) $\frac{1}{5}$
   c) $\frac{1}{6}$
   d) $\frac{1}{7}$

Correct Answer: c
Explanation: $P\ (Rain\ =\ High\ |\ attendance\ =\ High) = \frac{1}{6}$

8. Consider the following figure. The black triangles are the data points belonging to class 1. The grey rectangles are the data points belonging to class 2. The test point is marked as pentagon. What is the class label of the test point as per kNN classifier with k = 1 and k = 3?



   a) Class 1 and Class 1
   b) Class 1 and Class 2
   c) Class 2 and Class 1
   d) Class 2 and Class 2

Correct Answer: c
Explanation: If k=1, only one nearest neighbour of the test point is considered, which belongs to Class 2. So, the class label of the test point is Class 2. If k=3, three nearest neighbours of the test point is considered, belonging to Class 2, Class 1 and Class 1 respectively. So, the class label of the test point is Class 1, by majority.

9. Calculate Entropy for the data given in the table below-

Sample count	**Sample class**
9	1
5	2

   a) 0.94
   b) - 0.94
   c) 0.36
   d) – 0.36

Correct Answer: a
Explanation: $E = \sum_{i=1}^{m} -p_i \log_2 p_i = 0.41 + 0.53 = 0.94$

10. For two events A and B, the Bayes theorem states that
    a) P(A | B) = P(B) * P(B | A) / P(A)
    b) P(A | B) = P(A) * P(B | A) / P(B)
    c) P(A | B) = P(B) * P(A | B) / P(A)
    d) P(A | B) = P(A) * P(A | B) / P(B)

Correct Answer: b
Explanation: Check the According to Bayes theorem of reverse probability.

11. Which of the following splitting criteria is used in C4.5 algorithms?
    a) Information gain.
    b) Gain ratio.
    c) Gini index.
    d) Weighted average entropy.

Correct Answer: b
Explanation: Gain ration is used to select a splitting attribute in C4.5 algorithm.

12. In the table given below, the left column represents a classifier and right column represents the heuristic for decision tree building.

Classifier Algorithm	Heuristic
C_1. C4.5	H_1. Gini Index of Diversity
C_2. CART	H_2. Information Gain
C_3. ID3	H_3. Gain Ratio

    Which of the following mapping is appropriate?
    a) C_1 - H_1 C_2 - H_2 C_3 –H_3
    b) C_1 - H_3 C_2 – H_1 C_3 – H_2
    c) C_1 – H_1 C_2 – H_3 C_3 – H_2
    d) C_1 – H_2 C_2 – H_3 C_3 – H_1

Correct answer: b
Explanation: ID3 uses Information gain, C4.5 uses Gain ratio and CART uses Gini index of diversity.

13. A point P in an n-dimensional ($n \geq 2$) Euclidean space is denoted as $P(x_1, x_2, \ldots, x_x)$ where $\forall x_i \in R, i = 1, 2, \ldots, n$ and R is a set of all real numbers. The distance between two such points is usually NOT followed to measure following which of the form?
    a) L_0 norm
    b) L_1 norm
    c) L_2 norm
    d) L_3 norm

Correct answer: d
Explanation: Any one of the L_0, L_1 and L_2 norm can be followed to measure the similarity. There is no norm like L_3 norm.

14. Model of which classifier(s) can be expressed in the form a graph structure?
    a) Bayesian classifier
    b) Support vector machine
    c) Decision tree induction
    d) KNN.

Correct Answer: c
Explanation: The classifiers based on decision tree are in the form of a decision tree.

15. In the following Table, Column A lists some sampling classification techniques, whereas Column B lists the type of data that a classifier can handle better.

Column A		**Column B**	
(A)	k Nearest Neighbour	(W)	Numerical type
(B)	Naive Bayes′classifier	(X)	Categorical type
(C)	Support Vector Machine	(Y)	Mixed type
(D)	Decision Tree Classifier	(Z)	Ordinal type

Some matchings from Column A and Column B are given below. Select the correct matching.

a) (A)-(W), (B)-(X), (C)-(Y), (D)-(Z)
b) (A)-(Z), (B)-(Y), (C)-(W), (D)-(X)
c) (A)-(Y), (B)-(W), (C)-(Z), (D)-(X)
d) (A)-(W), (B)-(X), (C)-(W), (D)-(Y)

Correct Answer: d
Explanation:

- kNN: Consider the similarity estimation, and similarity estimation is possible with any data type; however, it works better with numerical type.
- Bayes'classifier: Bayes' classifier can be extended to work with numerical data, but it works better with categorical type.
- Support Vector Machine: It works only with numerical data.
- Decision tree classifies: It works with any type of data.

16. Mark the incorrect statement(s) in the following.
***Bayesian classifier is called Naïve*, if it**

a) Assumes all classes are mutually exclusive and exhaustive.
b) The attributes are independent, given a class.
c) It classifies provided that all attributes are categorical only.
d) It predicts class membership probabilities only.

Correct Answer: c
Explanation: The Naïve Bayes' classifier can classify a data with any type of attribute.

17. The maximum value of entropy of a training set of size *n* with *k* number of class labels such that each tuple is defined with *m* number of attributes is

a) $\log_k n$
b) $\log_2 \mathrm{k}$
c) $\log_2 m$
d) $\log_m n$

Correct Answer: b
Explanation: This will occur when n number of records are uniformly distributed over the k number of classes.

18. Which of the following decision tree induction algorithms always results a binary decision tree?

a) ID3
b) C4.5
c) CART
d) All of the above

Correct Answer: c
Explanation: CART algorithm yields a binary decision tree always.

19. Which of the following statement is true about building a decision tree?

a) ID3 always results a binary decision tree.
b) CART is applicable to any type of attribute.
c) C4.5 results an *n*-ry decision tree always.
d) All of the above.

Correct Answer: b
Explanation: ID3, C4.5 and CART are applicable to data set with any type of attribute.

20. Which of the following statement is true about the measurement of entropy?

a) It is applicable to supervised data only.
b) It is applicable to unsupervised data only.
c) It is applicable to both supervised and unsupervised data.
d) It is not applicable to supervised as well as unsupervised data.

Correct Answer: a
Explanation: A data is called supervised, if it is labeled. Entropy measurement is applicable to supervised data only.

Part-II: Solve the following problems *(10 X 6 = 60)*

1. Consider the dataset shown in the **Table A.** Using the dataset predict the record X = (Age = young, Income = Medium, Married = yes, Health = Fair") belongs to a class?

Answer:

$p_i = P(C_i) \times \prod_{j=1}^{n} P(A_j = a_j \mid C_i)$

Calculation of $P(C_i)$

$P(Select = 'Yes') = 9/14 = 0.643$

$P(Select = 'No') = 5/14 = 0.357$

Calculation of $P(X \mid C_i)$ for each class C_i

$P(Age = 'Young' \mid Select = 'Yes') = 2/9 = 0.222$

$P(Age = 'Young' \mid Select = 'No') = 3/5 = 0.6$

$P(Income = 'Medium' \mid Select = 'Yes') = 4/9 = 0.444$

$P(Income = 'Medium' \mid Select = 'No') = 2/5 = 0.4$

$P(Married = 'Yes' \mid Select = 'Yes') = 6/9 = 0.667$

$P(Married = 'Yes' \mid Select = 'No') = 1/5 = 0.2$

$P(Health = 'Fair' \mid Select = 'Yes') = 6/9 = 0.667$

$P(Health = 'Fair' \mid Select = 'No') = 2/5 = 0.4$

Thus,

$P(X \mid Select = 'Yes') = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044$

$P(X \mid Select = 'No') = 0.6 \times 0.4 \times 0.2 \times 0.4 = 0.019$

$P(C_i) \times P(X \mid C_i)$:

$P(Select = 'Yes') \times P(X \mid Select = 'Yes') = 0.643 \times 0.044 = \mathbf{0.028}$

$P(Select = 'No') \times P(X \mid Select = 'No') = 0.357 \times 0.019 = 0.007$

The test data is belonging to yes class.

2. Consider the dataset shown in **Table B** and consider the test data **[Angelina, 5, F]**, now find the class of sport for the test data for $\boldsymbol{k = 1}$, and $\boldsymbol{k = 3}$, using the **KNN** classification algorithm.

Answer:

The categorical data **[Male/Female]** converted into numerical data as **[1/2]** and we considered Euclidian distance.

Name	Age	Gender	Distance	Class
Ajay	32	1	27.02	Football
Mark	40	1	35.01	Neither
Sara	16	2	11	Cricket
Zaira	34	2	29	Cricket
Sachin	55	1	50.01	Neither

Rahul	40	1	35.01	Cricket
Pooja	20	2	15	Neither
Smith	15	1	10.05	Cricket
Laxmi	55	2	50	Football
Arun	15	1	10.05	Football

For **k=1**, the class is either **Cricket** or **Football**.
For **k=3**, the class is **Cricket**.

3. Consider a training data set as shown in the **Table C** and answer the following questions.
   a) Calculate the entropy of the data set.
   b) Suppose, you select "Gender" as the splitting attribute. Calculate the following.
      i. Information gain
      ii. Gini index
      iii. Gain ratio

Answer:

a) Entropy:

$E = \sum_{i=1}^{m} -p_i \log_2 p_i$

Here, $p_1 = \frac{5}{15} = 0.3333,\ p_2 = \frac{8}{15} = 0.5333,\ and\ p_3 = \frac{2}{15} = 0.1333$

$\therefore Entropy = \sum_{1}^{3} -p_i log_2 p_i = 0.3333 \times 0.4771 + 0.5333 \times 0.2730 + 0.1333 \times 0.8751 = 1.3996$

b) For "Gender" as the splitting attribute:

i. Information gain = $\alpha(Gender, D)$ = E(D) - $E_{Gender}(D)$

Here, $E(D) = 1.3996$ and

$E_{Gender}(D) = \frac{9}{15} \times \left(-\frac{4}{9}\log\frac{4}{9} - \frac{5}{9}\log\frac{5}{9}\right) + \frac{6}{15}\left(-\frac{1}{6}\log\frac{1}{6} - \frac{3}{6}\log\frac{3}{6} - \frac{2}{6}\log\frac{2}{6}\right) = 1.17829$

$Information\ gain = \alpha(Gender, D) = 1.3996 - 1.17829 = \mathbf{0.2213}$

ii. Gini index = $\gamma(A, D) = G(D) - G_A(D)$

$G(D) = 1 - \left(\frac{5}{15}\right)^2 - \left(\frac{8}{15}\right)^2 - \left(\frac{2}{15}\right)^2 = 0.5867$ and

$G_{Gender}(D) = \frac{9}{15} \times \left(1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2\right) + \frac{6}{15} \times \left(1 - \left(\frac{1}{6}\right)^2 - \left(\frac{3}{6}\right)^2 - \left(\frac{2}{6}\right)^2\right) = 0.5407$

$\therefore Gini\ index\ = 0.5867 - 0.5407 = \mathbf{0.046}$

iii. Gain ratio = $\beta(Gender, D) = \frac{\alpha(Gender, D)}{E^*_{Gender}(D)}$

Now, $\alpha(Gender, D) = 0.2213$ and

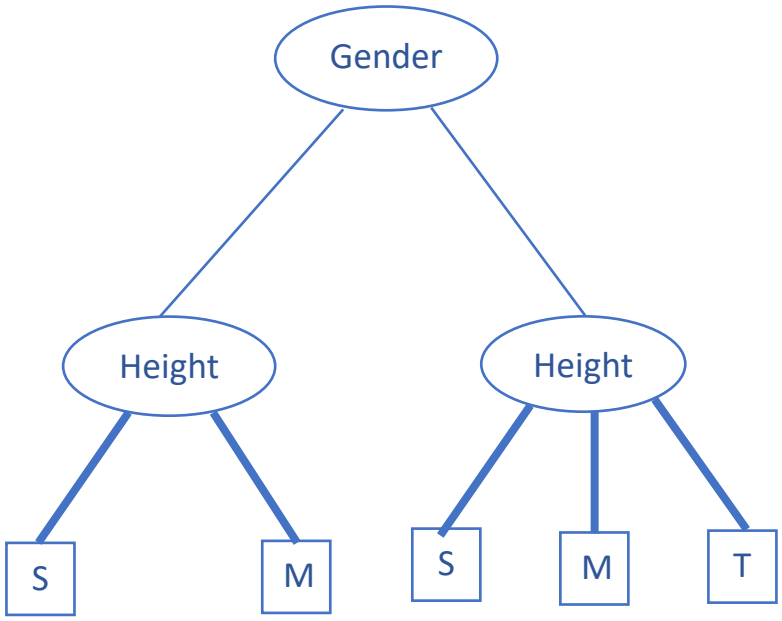
$E^*_{Gender}(D) = -\sum_{j=1}^{2} \frac{|D_j|}{|D|} . \log \frac{|D_j|}{|D|} = -\frac{9}{15} log \frac{9}{15} - \frac{6}{15} log \frac{6}{15} = 0.97$

$$\therefore Gain\ Ratio = \frac{0.2213}{0.97} = \mathbf{0.2281}$$

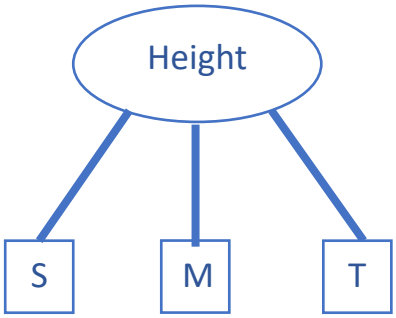
4. Consider the **Table C**, obtain the decision trees with the following splitting order and write the decision rules.
   a) Gender-Height and
   b) Height –Gender.

Answer:

a) Decision tree according to splitting order **Gender-Height**



b) Decision tree according to splitting order **Height-Gender**



5. Consider the training data set shown in the **Table D**. The x and y coordinates of the training points and their corresponding class label are provided. The test point is given as (4,5). Predict the class label of the test point using 3-NN classifier, considering Manhattan Distance as proximity measure. **Note:** The Manhattan distance between the point (x_t, y_t) and any training point (x, y) is given by, *Manhattan Distance* $= |x - x_t| + |y - y_t|$.

Answer:

Datapoint	X	Y	Distance	Class
1	1	3	5	Yes
2	2	2	5	No
3	3	2	4	Yes
4	2	4	3	Yes
5	3	4	2	Yes
6	4	4	1	Yes
7	2	6	3	Yes
8	4	7	2	No
9	5	3	3	No
10	5	7	3	No
11	6	9	6	No
12	7	7	5	No

Test point will be in **Yes** class

6. A scheme of a training data is stated as below.

Symptom	Duration	Treatment	Class
S1, S2, S3	S (Short), M (Medium), L (Large)	A (Allopathy), H (Homeopath), U (Unani)	Cure (Y), Not Cure (N)

A contingency table is prepared with 400 records of patients, which is shown below.

		Class		
		Y	N	Totals
Symptom	S1	30	10	40
	S2	14	12	26
	S3	24	14	38
Duration	S	22	18	40
	M	36	26	62
	L	32	22	54
Treatment	A	30	24	54
	H	28	18	46
	U	24	16	40
	Totals	240	160	400

A test data is given below:

S2	M	H	?

You have to classify the test data using the Naïve Bayes' classifier.

a) What is the probability that the test data is in class Cure?
b) What is the probability that the test data is in class Not Cure?
c) In which class the test data will be?
   i. The test belongs to class Cure.
   ii. The test data belongs to class Not Cure.
d) What is the entropy of the input data?

Answer:

a) $P(Y) = \frac{120}{200} \times \frac{7}{120} \times \frac{18}{120} \times \frac{14}{120} = 0.6 \times 0.058 \times 0.15 \times 0.116 = 0.0006$

b) $P(N) = \frac{80}{200} \times \frac{6}{80} \times \frac{13}{80} \times \frac{9}{80} = 0.4 \times 0.075 \times 0.162 \times 0.112 = 0.0005$

c) The predicted class will be Y i.e., it will belong to class Cure.

d) Calculation of the entropy

Here, $p_1 = \frac{240}{400} = 0.6$ and $p_2 \frac{160}{400} = 0.4$

$Entropy = -p_1 log p_1 - p_2 log p_2$

$= -0.6 \times (-0.7369) - 0.4 \times (-1.3219)$

$= 0.44214 + 0.52876$

$= \mathbf{0.9709}$

---END---