USN											CMRIT CMR INSTITUTE OF TECHNOLOGY, BENGALURU. ACCREDITED WITH A++ GRADE BY NAAC

Internal Assessment Test II– July-2024---Scheme

Sub :	**DATA MINING AND DATA WAREHOUSING**				Sub Code:	**21IS643**	Branch:	**ISE**
Date:	**31/07/2024**	Duration:	**90 min**	Max Marks: **50**	Sem/Sec:	**VI / C**		**OBE**

	Answer any FIVE FULL Questions	MARKS	CO	RBT
1	Illustrate the concept of rule based classifier along with example & sequential covering algorithm;	[10]	CO4	L3
2	Apply the Naïve bayes theorem to predict If a patient has stiff neck, what's the probability he/she has meningitis? P(M|S),explain the concept of Naïve bayes classifier and its components.	[10]	CO4	L3
3	Explain K Nearest Neighborhood classifier in detail. Write the algorithm and explain	[10]	CO4	L2

1. Illustrate the concept of rule based classifier along with example & sequential covering algorithm;

Solution : Rule based Classifier ---Explanation with Coverage and Accuracy terms

Sequential Covering Algorithm

```
Let E be the training records and A be the set of attribute-value pairs, {(A_j, v_j)}.
Let Y_o be an ordered set of classes {y_1, y_2, ..., y_k}.
Let R = { } be the initial rule list.
for each class y ∈ Y_o − {y_k} do
    while stopping condition is not met do
        r ← Learn-One-Rule (E, A, y).
        Remove training records from E that are covered by r.
        Add r to the bottom of the rule list: R ⟶ R ∨ r.
    end while
end for
Insert the default rule, {} ⟶ y_k, to the bottom of the rule list R.
```

2. Apply the Naïve bayes theorem to predict If a patient has stiff neck, what's the probability he/she has meningitis? P(M|S),explain the concept of Naïve bayes classifier and its components.

Given : A doctor knows that meningitis causes stiff neck 50% of the time P(S|M) = 0.5 – Prior probability of any patient having meningitis is 1/50,000 P(M) = 1/50,000 – Prior probability of any patient having stiff neck is 1/20 P(S) = 1/20.

Solution :

P(A) is **prior probability** (**unconditional probability**) of event A.

•P(A|B) is **posterior probability** (**conditional probability**) of event A given that event B holds.

•P(A,B) is the **jointprobability** of two events A and B.

–The (unconditional) probability of the events A and B occurring together.

–P(A,B) = P(B,A).

If a patient has stiff neck, what's the probability he/she has meningitis? **P(M|S)** ?

$$\mathbf{P(M|S) = \frac{P(S|M)\ P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002}$$

3. Explain K Nearest Neighborhood classifier in detail. Write the algorithm and explain

Solution :

Algorithm 5.2 The k-nearest neighbor classification algorithm.

```
Let k be the number of nearest neighbors and D be the set of training examples.
for each test example z = (x', y') do
    Compute d(x', x), the distance between z and every example, (x, y) ∈ D.
    Select D_z ⊆ D, the set of k closest training examples to z.
    y' = argmax_v Σ_{(x_i, y_i) ∈ D_z} I(v = y_i)
end for
```

<table>
<tr><td>4</td><td>Categorize the various Clustering methods. and explain the hierarchical clustering method.
Solution : Two main types of hierarchical clustering
◦Agglomerative:
□Start with the points as individual clusters
□at each step, merge the closest pair of clusters. This requires defining a notion of cluster proximity.
◦Divisive:
□Start with one, all-inclusive cluster
□at each step, split a cluster until only singleton clusters of individual points remain. In this case we need to decide which cluster to split at each step and how to do the splitting</td><td>[10]</td><td>CO5</td><td>L3</td></tr>
<tr><td>5</td><td>Illustrate the concept of K-means clustering algorithm with example.
Solution :
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change</td><td>[10]</td><td>CO5</td><td>L3</td></tr>
<tr><td>6</td><td>Explain the Density-based clustering methods with example
Algorithm: DBSCAN: a density-based clustering algorithm.
Input:
▪ D: a data set containing n objects,
▪ ϵ: the radius parameter, and
▪ MinPts: the neighborhood density threshold.
Output: A set of density-based clusters.
Method:
mark all objects as unvisited;
do
    randomly select an unvisited object p;
    mark p as visited;
    if the ϵ-neighborhood of p has at least MinPts objects
        create a new cluster C, and add p to C;
        let N be the set of objects in the ϵ-neighborhood of p;
        for each point p' in N
            if p' is unvisited
                mark p' as visited;
                if the ϵ-neighborhood of p' has at least MinPts points, add those points to N;
            if p' is not yet a member of any cluster, add p' to C;
        end for
        output C;
    else mark p as noise;
until no object is unvisited;
Solution :</td><td>[10]</td><td>CO5</td><td>L2</td></tr>
</table>

Faculty Signature CCI Signature HOD Signature