FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERING

Department of Computer Engineering

Class Test 1

(2019-20)

Class/Sem./Branch -TE/VI/COMP **Course code: CSC603**

Subject: Data warehousing and mining (DWM) **Date: 28/03/2020**

Total marks: 20

Name: __SHUBHAM SACHIN BHATE___ **Roll no._____8318_____**

A. Bayesian classifiers is

A.	**A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory.**
B.	**Any mechanism employed by a learning system to constrain the search space of a hypothesis**
C.	**An approach to the design of learning algorithms that is inspired by the fact that when people encounter new situations, they often explain them by reference to familiar experiences, adapting the explanations to fit the new situation.**
D.	**None of these**

Ans : Option A

A class of learning algorithm that tries to find an optimum classification of a set of examples using the probabilistic theory

----------X----------X----------

1. Decision Trees are built using: Select one:
a. heuristics
b. greedy algorithms
c. dynamic programming
d. divide and conquer strategy
Ans : Option B : Greedy Algorithms

----------X----------X----------

2. Give some alternative terms for data mining
Ans:
Alternative terms for data mining are as follows

- **data archaeology**
- **information harvesting**
- **information discovery**
- **knowledge extraction**

----------X----------X----------

3. What is meant by pattern?
Ans:

Frequent patterns are patterns (e.g., item sets, subsequence, or substructures) that appear frequently in a data set. For example, a set of items, such as milk and bread,that appear frequently together in a transaction data set is a frequent item set. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a (frequent) sequential pattern. A substructure can refer to different structural forms, such as sub graphs, sub trees, or sub lattices, which may be combined with itemsets or sub sequences. If a substructure occurs frequently, it is called a (frequent) structured pattern.

Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data classification, clustering, and other data mining tasks. Thus, frequent pattern mining has become an important data mining task and a focused theme in data mining research.

----------X----------X----------

4. The problem of Curse of Dimensionality is associated with: Select one:
a. increasing data points
b. increasing noise in data
c. increasing dimensions
d. increasing users
Ans : Option C : Increasing Dimensions

----------X----------X----------

5. Which of following function selects a median as centroid-
k-means
k-medoids
optics
Ans : Option B : K Medoids

----------X----------X----------

6. Give few techniques to improve the efficiency of Apriori algorithm.
Ans.

Many variations of the Apriori algorithm have been proposed that focus on improving the efficiency of the Apriori algorithm. Several of these variations are summarized as follows:

1. Hash-based technique can be used to reduce the size of the candidate k-itemsets, Ck, for k>1. For example when scanning each transaction in the database to generate the frequent 1-itemsetes, L1, from the candidate 1-itemsets in C1, we can generate all of the 2-itemsets for each transaction, hash them into a different buckets of a hash table structure and increase the corresponding bucket counts: a. H(x,y)=((order of x)X10+(order of y)) mod 7 b. A 2-itemset whose corresponding bucket count in the hash table is below the threshold cannot be frequent and thus should be removed from the candidate set.

2. Transaction reduction – a transaction that does not contain any frequent itemset cannot contain any frequent k+1 itemsets. Therefore, such a transaction can be marked or removed from further consideration because subsequent scans of the database for j-itemsets, where j>k, will not require it.

3. Partitioning (partitioning the data to find candidate itemsets): A partitioning technique can be used that requires just two database scans to mine the frequent itemsets. It consists of two phases. In phase I, the algorithm subdivides the transactions of D into n non-overlapping partitions. If the minimum support threshold for transactions in D is min_sup, then the minimum support count for a partition is min_sup X the number of transactions in that partition. For each partition, all frequent itemsets within the partition are found. These are referred to as local frequent itemsets. A local frequent itemset may or may not be frequent with respect to the entire database, D. Any itemset that is potentially frequent with respect to D must occur as a frequent itemset in at least one of the partitions. Therefore, all local frequent itemsets are candidate itemsets with respect to D. The collection of frequent itemsets from all partitions forms the global candidate itemsets with respect to D. In phase II, a second 10 scan of D is conducted in which the actual support of each candidate is assessed in order to determine the global frequent itemsets

4. Sampling (mining on a subset of a given data): The basic idea of the sampling approach is to pick a random sample S of the given data D, and then search for frequent itemsets in S instead of D. In this way, we trade off some degree of accuracy against efficiency. The sample size of S is such that the search for frequent itemsets in S can be done in main memory, and so only one scan of the transactions in S in required overall. Because we are searching for frequent itemsets in S rather than in D, it is possible that we will miss some of the global frequent itemsets. To lessen this possibility, we use a lower support threshold than minimum support to find the frequent itemsets local to S.

5. Dynamic itemset counting (adding candidate itemsets at different points during a scan): A dynamic itemset counting technique was proposed in which the database is partitioned into blocks marked by start points. In this variation new candidate itemsets can be added at any start point, which determines new candidate itemsets only immediately before each complete database scan. The resulting algorithm requires fewer database scan than Apriori.

----------X----------X----------

7. What are the things suffering the performance of Apriori candidate generation technique?

Ans: -

Apriori, while historically significant, suffers from a number of inefficiencies or trade-offs, which have spawned other algorithms. Candidate generation generates large numbers of subsets (the algorithm attempts to load up the candidate set with as many as possible before each scan).

Bottom-up subset exploration (essentially a breadth-first traversal of the subset lattice) finds any maximal subset S only after all (2^(|S|) -1) of its proper subsets.

The algorithm scans the database too many times, which reduces the overall performance. Due to this, the algorithm assumes that the database is Permanent in the memory.

Also, both the time and space complexity of this algorithm are very high: O(2^|D|), thus exponential, where |D| is the horizontal width (the total number of items) present in the database. Later algorithms such as Max-Miner try to identify the maximal frequent item sets without enumerating their subsets, and perform "jumps" in the search space rather than a purely bottom-up approach.

----------X----------X----------

8. Describe the method of generating frequent item sets without candidate generation.
Ans)
The candidate generate and test method

1. **It may need to generate a huge number of candidate sets**
2. **It may need to repeatedly scan the database and check a large set of candidates by pattern matching**

Frequent-pattern growth method(FPgrowth) – frequent pattern tree(FP-tree)
Transforming the horizontal data format of the transaction database D into a vertical data format

----------X----------X----------

9. Define Post Pruning and pre pruning.
Ans.

Post-pruning

Post-pruning is also known as backward pruning. In this, first Generate the decision tree and then remove non-significant branches. Post-pruning a decision tree implies that we begin by generating the (complete) tree and then adjust it with the aim of improving the classification accuracy on unseen instances. There are two principal methods of doing this.

Pre-pruning

Pre-pruning is also called forward pruning or online-pruning. Pre-pruning prevent the generation of non-significant branches. Pre-pruning a decision tree involves using a 'termination condition' to decide when it is desirable to terminate some of the branches prematurely as the tree is generated. When constructing the tree some significant measures can be used to assess the goodness of a split. If partitioning the tuples at a node would result the split that falls below a pre specified threshold, then further partitioning of the given subset is halted otherwise it is expanded. High threshold result in oversimplified trees, whereas low threshold result in very little simplification.

----------X----------X----------

10. Define CLARA and CLARANS?

Ans:

CLARA :

To deal with larger data sets, a sampling-based method called CLARA (Clustering LARge Applications) can be used. Instead of taking the whole data set into consideration, CLARA uses a random sample of the data set. The PAM algorithm is then applied to compute the best medoids from the sample

CLARANS:

A randomized algorithm called CLARANS (Clustering Large Applications based upon RANdomized Search) presents a trade-off between the cost and the effectiveness of using samples to obtain clustering. First, it randomly selects k objects in the data set as the current medoids. It then randomly selects a current medoid x and an object y that is not one of the current medoids It checks if replacing x by y improve the absolute-error criterion. If yes, the replacement is made. CLARANS conducts such a randomized search l times. The set of the current medoids after the l steps is considered a local optimum. CLARANS repeats this randomized process m times and returns the best local optimal as the final result

----------X----------X----------

11. Define Chameleon method?

Ans.

Chameleon is a hierarchical clustering algorithm that overcomes the limitations of the existing models and the methods present in the data warehousing. This method operates on the sparse graph having nodes that represent the data items and edges represent the weights of the data items. The representation of it allows large data set to be created and operated on successfully. The method finds the clusters that are used in the data set using the two phase algorithm. The first phase consists of the graph partitioning that allows the clustering of the data items into large number of sub-clusters. Second phases use an agglomerative hierarchical clustering algorithm to search for the clusters that are genuine and can be combined together with the sub-clusters that are produced

----------X----------X----------

12. Define Wave Cluster?

Ans: -

The WAVE clustering algorithm is a grid-based clustering algorithm. It depends on the relation between spatial dataset and multidimensional signals. The idea is that the cluster in a multidimensional spatial dataset turns out to be more distinguishable after a wavelet transformation, that is, after applying wavelets to the input data or the preprocessed input dataset. The dense part segmented by the sparse area in the transformed result represents clusters.

The characteristics of the WAVE cluster algorithm are as follows:

- **Efficient for a large dataset**
- **Efficient for finding various shapes of clusters**
- **Insensitive to noise or outlier**
- **Insensitive with respect to the input order of a dataset**
- **Multiresolution, which is introduced by wavelet transforms**
- **Applicable to any numerical dataset**

----------X----------X----------

13. What is the use of Regression?

Ans.

Regression is a data mining technique used to predict a range of numeric values (also called *continuous values*), given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.
Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modelling and analysis of trends.

----------X----------X----------

14. What is Wave Cluster
Ans.

Wave Cluster is a multiresolution clustering algorithm that first summarizes the data by imposing a multidimensional grid structure onto the data space. It then uses a wavelet transformation to transform the original feature space, finding dense regions in the transformed space.

----------X----------X----------

15. What is apex cuboid?
Ans.

Data cubes provide fast access to pre computed, summarized data, thereby benefiting online analytical processing as well as data mining. A cube at the highest level of abstraction is the apex cuboid.

----------X----------X----------

16. Give some data mining tools
Ans: Some predominant data mining tools are as follows

A. **Rapid Miner**
B. **Weka**
C. **Orange**
D. **R**
E. **Knime**
F. **Rattle**
G. **Tanagra**
H. **XL Miner**

----------X----------X----------

17. Discuss the requirements of clustering in data mining.
Ans : The main requirements that a clustering algorithm should satisfy are:

- **scalability;**
- **dealing with different types of attributes;**
- **discovering clusters with arbitrary shape;**
- **minimal requirements for domain knowledge to determine input parameters;**
- **ability to deal with noise and outliers;**
- **insensitivity to order of input records;**
- **high dimensionality;**
- **interpretability and usability.**

----------X----------X----------

18. Explain data mining applications for Biomedical and DNA data analysis

Ans.

The unique combination of complexity, richness, size, and importance of biological and biomedical data warrants special attention in data mining. Mining DNA and protein sequences, mining high dimensional microarray data, and biological pathway and network analysis are just a few topics in this field. Other areas of biological data mining research include mining biomedical literature, link analysis across heterogeneous biological data, and information integration of biological data by data mining. Different biological processes involve different sets of genes acting together in precisely regulated patterns. Thus, to understand a biological process we need to identify the participating genes and their regulators. This requires the development of sophisticated data mining methods to analyze large biological data sets for clues about regulatory influences on specific genes, by finding DNA segments ("regulatory sequences") mediating such influence.

----------X----------X----------

19. Give Various forms of visualizing the discovered patterns

Ans.

- **Data mining result visualization:**

  Visualization of data mining results is the presentation of the results or knowledge obtained from data mining in visual forms. Such forms may include scatter plots and boxplots,as well as decision trees, association rules, clusters, outliers, and generalized rules
- **Data mining process visualization:**

  This type of visualization presents the various processes of data mining in visual forms so that users can see how the data are extracted and from which database or data warehouse they are extracted, as well as how the selected data are cleaned, integrated, preprocessed, and mined.
- **Interactive visual data mining:**

  In (interactive) visual data mining, visualization tools can be used in the data mining process to help users make smart data mining decisions. For example, the data distribution in a set of attributes can be displayed using colored sectors

----------X----------X----------

20. Give the name of proximity function in k-medoid.

Ans: Voronoi Iteration

----------X----------X----------

21. Write support and confidence formula
Ans.

LET
T: (A set of) All transactions that customers make and are recorded in the stores system. (Since most of the customers use credit/debit cards, there is a unique number associated to their shopping list.)
Basket: A set of all items bought by a customer.
Item-set: A set of items that we are interested in.

Now, let's assume that we analyzed customers' transactions and realized that "many of them", if they had *wine* in their baskets, they also have *cheese*! Why can't we put *cheese* at the beginning of an aisle and *wine* at the end of the next one and put all tempting items that a customer with a bottle of *wine* may need a nudge to buy, in between!!
So, there has to be a way to evaluate the importance of a discovered rule. Here comes the support and confidence.

Suppose the rule we discovered is as follows

Wine -> Cheese (Support: 9% Confidence: 65%)

Support: is the percentage of transactions in T that contain both *wine* and *Cheese* together. (9% of all baskets had these 2 items together.)

$$support(A \rightarrow B) = P(A \cup B)$$

Confidence: is the percentage of transactions in T, containing *wine*, that also contain *Cheese*. In other words, the probability of having *Cheese*, given that *wine* is already in the basket. (65% of all those who bought *Wine*, also bought *Cheese*.)

$$confidence(A \rightarrow B) = P(B|A)$$

----------X----------X----------

22. Explain over fitting and under fitting in classification
Ans :
<u>Overfitting in Machine Learning</u>

Overfitting refers to a model that models the training data too well.
Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data. This means that the noise or random fluctuations in the training data is picked up and learned as concepts by the model. The problem is that these concepts do not apply to new data and negatively impact the models ability to generalize.

<u>Underfitting in Machine Learning</u>

Underfitting refers to a model that can neither model the training data nor generalize to new data.
An underfit machine learning model is not a suitable model and will be obvious as it will have poor performance on the training data. Underfitting is often not discussed as it is easy to detect given a good performance metric. The remedy is to move on and try alternate machine learning algorithms. Nevertheless, it does provide a good contrast to the problem of overfitting.

----------X----------X----------

23. What is Meant by Confusion Matrix
Ans:

Confusion Matrix:
A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made

----------X----------X----------

24. How to calculate accuracy from confusion matrix.
Ans:

Definition of the Terms:
Positive (P) : Observation is positive (for example: is an apple).
Negative (N) : Observation is not positive (for example: is not an apple).
True Positive (TP) : Observation is positive, and is predicted to be positive.
False Negative (FN) : Observation is positive, but is predicted negative.
True Negative (TN) : Observation is negative, and is predicted to be negative.
False Positive (FP) : Observation is negative, but is predicted positive.

Classification Rate/Accuracy:
Classification Rate or Accuracy is given by the relation:

Accuracy = (TP+TN)/(TP+TN+FP+FN)

However, there are problems with accuracy. It assumes equal costs for both kinds of errors. A 99% accuracy can be excellent, good, mediocre, poor or terrible depending upon the problem.

----------X----------X----------