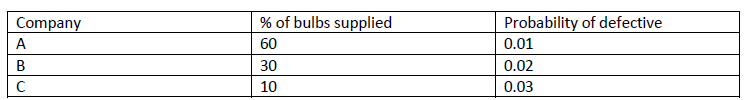
**21AI52- ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

**Question Bank on the topics covered in Machine Learning**

1. When is a dataset considered to be noisy?
2. Define Machine Learning. List with an example different types of learning.
3. What is meant by confusion matrix in a classification scenario?
4. How do we differentiate between Data mining, Machine Learning and Statistical Learning?
5. Discuss various attribute selection methods with the help of a formulae.
6. Briefly outline the major steps of decision tree classification. How are the rules formed?
7. Define and evaluate accuracy, precision and F-measure for a given confusion matrix below.
8. Why is tree pruning useful in decision tree induction? Identify different ways of pruning along with its advantages and limitations?
9. Given 4 medicines as training data points object and each medicine has 2 attributes. Each attribute represents coordinate of the object. Determine which medicines belong to cluster 1 and which medicines belong to the other cluster using K means clustering method. K = 3

|  |  |  |
| --- | --- | --- |
| Object | Feature1(X):weight index | Feature2 (Y): pH |
| Medicine A | 1 | 1 |
| Medicine B | 2 | 1 |
| Medicine C | 4 | 3 |
| Medicine D | 5 | 4 |

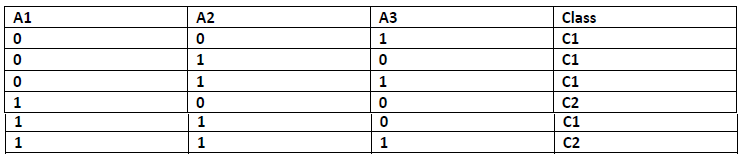
1. Differentiate single- link clustering and Complete-link clustering, with an appropriate example.
2. List various ways of fine-tuning k values in clusters.
3. How can we do hierarchical clustering with binary input vectors, for example, for text clustering using the bag-of-words representation?
4. Identify the similarities and differences between clustering and dimensionality reduction.
5. How is overfitting and underfitting scenario handled during Machine learning model building?
6. There are 3 companies that supply bulbs. The percentage of bulbs supplied by them and the probability of them being defective is given below:



Given that the bulb is defective, find the probability that it is supplied by B.

1. In a multiclass classification, Bayes classifier assigns an instance to class corresponding to which probability?

Following is the set of training examples. Each attribute can take value either 0 or 1.



1. How would a 3-NN classify the example A1=1, A2=0, A3=1, if the distance metric is Euclidean distance?
2. Compare the advantages and disadvantages of eager classification versus lazy classification with examples.
3. For the given dataset below, estimate the conditional probabilities for

P(A|+), P(B|+), P(C|+),P(A|−), P(B|−),P(C|−).



1. Briefly outline the major steps of decision tree classifier .
2. Why is Naïve Bayesian classification called naïve? Briefly outline the naïve Bayesian

Classification?

1. The support vector machine is a highly accurate classification method. However, SVM classifiers suffer from slow processing when training with a large set of data tuples. Discuss how to overcome this difficulty and develop a scalable efficient algorithm for efficient SVM classification in large data sets.
2. Write an algorithm for k-nearest neighbor classification given k and n, the number of
3. attributes describing each tuple
4. Demonstrate k-medoids algorithms for effective clustering
5. Describe each of the clustering algorithm in terms of the following criteria
   1. k-means
   2. CLARA
6. What are the major challenges faced in bringing data mining research to market?
7. List the various statistical data mining techniques.
8. List the popular trends in data mining.
9. What is cluster analysis? What are the different types of clustering?
10. What are the strengths and weakness of k-means?
11. What is the time and space complexity in DBSCAN algorithm?
12. What is data preprocessing? What are the various categories of data preprocessing?
13. Define: Noise, artifacts, outliers, precision, bias, accuracy, measurement error, sampling.
14. Define attribute. What are the various methods for expressing attribute test conditions?
15. What is rule based classifier? Discuss the direct methods for rule extraction.
16. What is cluster analysis? Discuss the various types of clustering and clusters.
17. Write a basic K-means algorithm. Discuss its time and space complexity.
18. What is cluster evaluation? Discuss
19. Suppose we find K clusters using Ward’s method, bisecting K-means, and ordinary K-means. Which of these solutions represents a local or global minimum? Explain.
20. Briefly describe the following approaches to clustering:
21. Hierarchical methods,
22. Density based methods,
23. Grid-based methods,
24. Model-based methods,
25. Methods for high-dimensional data
26. For the Bayesian network shown below, compute the following probabilities: 

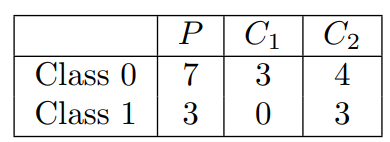




1. How does Bias Variance decomposition handled in Bagging model?
2. Using minimum spanning tree, form the clusters and draw a dendogram with height=3, for a dataset given below:



1. Summarize situations where you will use an SVM over random forest machine learning algorithm and vice versa.
2. What is the stopping criterion for Hunt’s decision tree?
3. Consider splitting a parent node P into two child nodes, C1 and C2, using some attribute test condition. The composition of labeled training instances at every node is summarized in the Table below.



(a) Calculate the Gini index and misclassification error rate of the parent node P.

(b) Calculate the weighted Gini index of the child nodes. Would you consider this attribute test condition if Gini is used as the impurity measure?

(c) Calculate the weighted misclassification rate of the child nodes. Would you consider this attribute test condition if misclassification rate is used as the impurity measure?

1. What is underfitting in the decision tree?
2. How can continuous attributes be handled? How can splitting points be chosen for binary splits and for multiway splits?
3. What are the two clusters produced by single link?
4. Which technique, K-means or single link, seems to produce the “most natural” clustering in this situation? (For K-means, take the clustering with the lowest squared error.)
5. What definition(s) of clustering does this natural clustering correspond to? (Well-separated, center-based, contiguous, or density.)
6. What well-known characteristic of the K-means algorithm explains the previous behavior?
7. How can we use cluster evaluation measures to determine the correct number of natural clusters? Do these methods always indicate the correct number of natural clusters?
8. How are single-link clustering, complete-link clustering, group average clustering, and Ward’s method clustering different?
9. Brieﬂy outline the major steps of decision tree classiﬁcation
10. Why is tree pruning useful in decision tree induction? What is a drawback of using a separate set of tuples to evaluate pruning?
11. Given a decision tree, you have the option of (a) converting the decision tree to rules and then pruning the resulting rules, or (b) pruning the decision tree and then con-verting the pruned tree to rules. What advantage does (a) have over (b)?
12. Why is naïve Bayesian classiﬁcation called “naïve”? Brieﬂy outline the major ideas of naïve Bayesian classiﬁcation.
13. Write an algorithm for k -nearest-neighbor classiﬁcation given k and n, the number of attributes describing each tuple.

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 q=3