

DS QUESTION BANK ANSWERS

! Try to Find Answers for 12..!

1. What do you mean by regression?
- ☒ 2. What is the equation of simple linear regression?
- ☒ 3. What is the equation of multiple linear regression?
- ☒ 4. Where is non-linear regression used?
- ☒ 5. What do you mean by Polynomial regression?
- ☒ 6. What is logistic regression?
- ☒ 7. What do you mean by the k-nn approach?
- ☒ 8. Why is weighted k-nn used?
- ☒ 9. What do you mean by the r-near algorithm?
- ☒ 10. What is the disadvantage of the near neighbor approach?
- ☒ 11. What is the branch and bound algorithm? What is its advantage?
- ☐ 12. What do you mean by a projection algorithm?
- ☒ 13. What is the Bayes theorem?
- ☒ 14. What is the naïve Bayes approach?
- ☒ 15. Numerical based on Naïve Bayes approach
- ☒ 16. What are the salient features of decision trees?
- ☒ 17. What are axis parallel, oblique split and non-linear tests?
- ☒ 18. State various impurities in decision trees
- ☒ 19. Numerical based on impurity calculations
- ☒ 20. What is decision tree pruning?
- ☒ 21. What are the advantages of decision trees?
- ☒ 22. What do you mean by inter-cluster and intra-cluster distance?
- ☒ 23. What are the various types of clustering approaches?
- ☒ 24. What is polythetic clustering?
- ☒ 25. What is monothetic clustering?
- ☒ 26. What are the advantages and disadvantages of agglomerative clustering?
- ☒ 27. What is the k-means clustering approach?
- ☒ 28. What do you mean by holdout, resubstitution, cross-validation and bootstrapping strategies for training and testing?
- ☒ 29. What do you mean by a confusion matrix?
30. What are various terms like accuracy, precision, recall, sensitivity, specificity, f-score?

DS QUESTION BANK ANSWERS

1. What do you mean by regression?

Ans.

In statistical modeling regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called "predictors").

Steps of Regression

- 1) Modeling - Developing Regression Model
- 2) Estimation - Using Software to estimate model
- 3) Interference - Interpreting the estimated model
- 4) Prediction - Making predictions about variable of interest

Example - There is a Sales manager of a 3 toys retail company which sells various kinds of toys in the local market. This sales manager needs to make some kind of prediction about the number of monthly units that the retail company will be able to sell of this particular toy in the coming half year. In the past she has been making such projections based on her gut feeling and now wishes to be a little more scientific about the whole process.

2. What is the equation of simple linear regression?

Ans.

Simple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and single independent variables.

Simple linear regression formula

The formula for a simple linear regression is:

$$y = \beta_0 + \beta_1 X + \epsilon$$

- **y** is the predicted value of the dependent variable (**y**) for any given value of the independent variable (**x**).
- **B₀** is the **intercept**, the predicted value of **y** when the **x** is 0.
- **B₁** is the regression coefficient – how much we expect **y** to change as **x** increases.
- **x** is the independent variable (the variable we expect is influencing **y**).
- **e** is the **error** of the estimate, or how much variation there is in our estimate of the regression coefficient.

DS QUESTION BANK ANSWERS

3. What is the equation of multiple linear regression?

Ans.

Multiple regression is a statistical technique that can be used to analyze the relationship between a single dependent variable and several independent variables.

Formula and Calculation of Multiple Linear Regression

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon$$

where, for $i = n$ observations:

y_i = dependent variable

x_i = explanatory variables

β_0 = y-intercept (constant term)

β_p = slope coefficients for each explanatory variable

ϵ = the model's error term (also known as the residuals)

4. Where is non-linear regression used?

Ans.

- I. Non-linear regression can be used to predict population growth over time.
- II. A scatterplot of changing financial prices over time shows an association between changes in prices and time. Because the relationship is nonlinear, a nonlinear regression model is the best model to use.
- III. A logistic price change model can provide the estimates of the market prices that were not measured and a projection of the future changes in market prices.
- IV. To illustrate, recessions versus expansions, bull and bear stock markets, or low versus high volatility are some of the dual regimes that require nonlinear models in economic time series data.

5. What do you mean by Polynomial regression?

Ans.

- I. Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:
$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \dots + b_nx^n$$
- II. It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.
- III. It is a linear model with some modification in order to increase the accuracy.
- IV. The dataset used in Polynomial regression for training is of non-linear nature.
- V. It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.

DS QUESTION BANK ANSWERS

6. What is logistic regression?

Ans.

- I. Logistic regression is a process of modeling the probability of a discrete outcome given an input variable.
- II. It is used to predict a binary outcome or state, such as *yes/no*, *success/failure*, and *true/false*.
- III. Logistic regression is essentially a classification algorithm. And its name comes from the linear regression
- IV. Logistic regression is a simple and more efficient method for binary and linear classification problems.
- V. It is extremely robust and flexible.

7. What do you mean by the k-nn approach?

Ans

- I. K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.
- II. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.
- III. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.
- IV. K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data.
- V. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.
- VI. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- VII. **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know whether it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

DS QUESTION BANK ANSWERS

8. Why is weighted k-nn used?

Ans.

- I. Weighted kNN is a modified version of k nearest neighbors. One of the many issues that affect the performance of the kNN algorithm is the choice of the hyperparameter k.
- II. If k is too small, the algorithm would be more sensitive to outliers. If k is too large, then the neighborhood may include too many points from other classes.
- III. Another issue is the approach to combining the class labels. The simplest method is to take the majority vote, but this can be a problem if the nearest neighbors vary widely in their distance and the closest neighbors more reliably indicate the class of the object.
- IV. The intuition behind weighted kNN, is to give more weight to the points which are nearby and less weight to the points which are farther away.
- V. Any function can be used as a kernel function for the weighted knn classifier whose value decreases as the distance increases.

9. What do you mean by the r-near algorithm?

Ans.

- I. R-Nearest neighbors are a modified version of the k-nearest neighbors.
- II. The issue with k-nearest neighbors is the choice of k. With a smaller k, the classifier would be more sensitive to outliers. If the value of k is large, then the classifier would include many points from other classes.

10. What is the disadvantage of the near neighbor approach?

Ans.

- I. **Does not work well with large datasets:** In large datasets, the cost of calculating the distance between the new point and each existing point is huge which degrades the performance of the algorithm.
- II. **Does not work well with high dimensions:** The KNN algorithm doesn't work well with high dimensional data because with a large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.
- III. **Need feature scaling:** We need to do feature scaling (standardization and normalization) before applying KNN algorithm to any dataset. If we don't do so, KNN may generate wrong predictions.
- IV. **Sensitive to noisy data, missing values and outliers:** KNN is sensitive to noise in the dataset. We need to manually calculate missing values and remove outliers.

11. What is the branch and bound algorithm? What is its advantage?

Ans.

Branch and bound is one of the techniques used for problem solving. It is similar to the backtracking since it also uses the state space tree.

It is used for solving optimization problems and minimization problems. If we have given a maximization problem then we can convert it using the Branch and bound technique by simply converting the problem into a maximization problem.

Advantages

1. Branch-and-Bound traverse the tree in any manner, DFS or BFS.
2. Branch-and-Bound involves a bounding function.
3. Branch-and-Bound is used for solving the optimization Problem.
4. In Branch-and-Bound as the optimum solution may be present anywhere in the state space tree, so the tree needs to be searched completely.
5. Useful in solving Knapsack Problem, Travelling Salesman Problem.

13. What is the Bayes theorem?

Ans.

Bayes' Theorem is the basic foundation of probability. It is the determination of the conditional probability of an event.

This *conditional probability is known as a hypothesis*. This hypothesis is calculated through previous evidence or knowledge.

This conditional probability is the probability of the occurrence of an event, given that some other event has already happened.

$$p(H | E) = \frac{p(E | H) p(H)}{p(E)}$$

Where ,

$P(H | E)$ = posterior probability, $P(E | H)$ = Probability of Hypothesis, $P(H)$ = Prior Probability, $P(E)$ = Probability of Evidence

DS QUESTION BANK ANSWERS

14. What is the naïve Bayes approach?

Ans. Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

It is mainly used in *text classification* that includes a high-dimensional training dataset.

Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

$P(A|B)$ is Posterior probability: Probability of hypothesis A on the observed event B.

$P(B|A)$ is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

$P(A)$ is Prior Probability: Probability of hypothesis before observing the evidence.

$P(B)$ is Marginal Probability: Probability of Evidence.

15. Numerical based on Naïve Bayes approach (https://youtu.be/mzPHmNm_NrM)

④ Naïve Bayes

Fruite = { yellow, Sweet, long }

| fruite | Yellow | Sweet | Long | Total |
|--------|--------|-------|------|-------|
| Orange | 350 | 450 | 0 | 650 |
| Banana | 400 | 300 | 350 | 400 |
| Others | 50 | 100 | 50 | 150 |
| Total | 800 | 850 | 400 | 1200 |

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

\downarrow \downarrow (column)
 set of Row of
 features fruite

$$P(\text{Yellow}|\text{orange}) = \frac{\frac{350}{800} \times \frac{800}{1200}}{\frac{650}{1200}}$$

$$= \frac{0.4375 \times 0.6666}{0.5416}$$

$$= \frac{0.2916}{0.5416} = 0.5$$

$$P(\text{Sweet}|\text{orange}) = \frac{\frac{450}{850} \times \frac{850}{1200}}{\frac{650}{1200}}$$

$$= \frac{0.5294 \times 0.7083}{0.5416}$$

$$= \frac{0.3749}{0.5416} = 0.69$$

$$P(\text{Long}|\text{orange}) = 0 \therefore P(\text{fruite}|\text{orange}) = 0.5 \times 0.69 \times 0 = 0$$

DS QUESTION BANK ANSWERS

| | |
|----------|-----|
| PAGE No. | |
| DATE | / / |

$$P(\text{fruite} | \text{Banana}) = 1 \times 0.75 \times 0.87$$

$$= 0.65$$

$$P(\text{fruite} | \text{others}) = 0.33 \times 0.66 \times 0.33$$

$$= 0.072$$

$\therefore \text{fruite} = \text{Banana}$

DS QUESTION BANK ANSWERS

16. What are the salient features of decision trees?

1. Hierarchical or flowchart structured manner which is easy to understand and analyze the data.
2. There are 2 nodes 1st one is decision node which will divided further and the 2nd one is leaf node for the final classification of the data
3. Decision trees are also used for regression
4. The test is performed on the features/ attributes.
5. We can remove impurity using GINI and Entropy.
6. We can remove irrelevant attributes or low importance attributes for our test by using the Pruning method.

17. What are axis parallel, oblique split and non-linear tests?

$$\Delta i(N) = 0.3974$$

13/5/22

Splitting at the nodes :-

- Each decision outcome a node is called as a split.
- Depending on the split rule, we can have different types of splitting.

i) Axis parallel split :

For this type of split we apply axis parallel test.

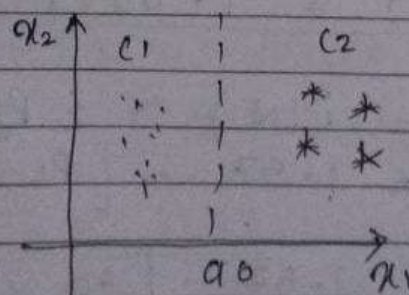
It is of the form :

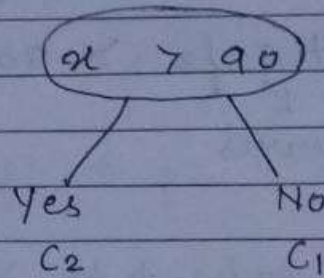
$$x > a_0$$

where, $x \rightarrow$ feature

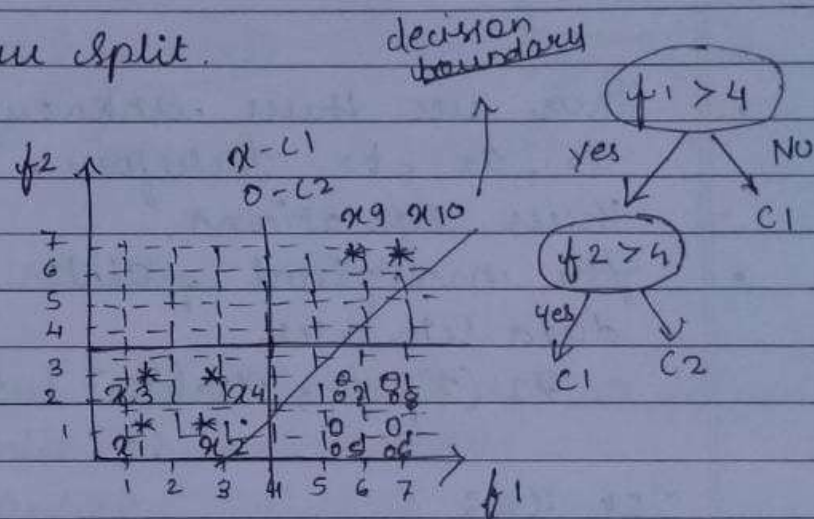
$a_0 \rightarrow$ Threshold value

This involves only one feature





ii) Oblique split.



- For this type of split, we apply test based on linear combination of features.
- The test is of the form $\sum_i a_i x_i > 0$ where
 x_i is a i^{th} feature/col
 a_i is a weight associated with it.
- The decision hyperplane is not parallel to any of the axis.

Hyper plane : more than 2 feature.

(decision hyper plane).

marginal : max to classifier.

Page

Date

$$a_1 f_1 + a_2 f_2 > a_0$$

\downarrow \downarrow \rightarrow Threshold
 no. of features.

$$a_1 f_1 + a_2 f_2 + b_0 > 0$$

- There are three unknown value a_1, a_2, b_0 . Therefore, we require three equations.
- The marginal points from the dataset are $x_2(2,1)$, $x_{10}(7,7)$ & $x_7(6,2)$

For x_2 :

$$a_1 f_1 + a_2 f_2 + b_0 > 0$$

$$2a_1 + a_2 + b_0 > 0 \quad \text{--- ①}$$

For x_{10} :

$$7a_1 + 7a_2 + b_0 > 0 \quad \text{--- ②}$$

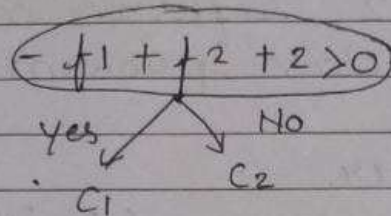
For x_7 :

$$6a_1 + 2a_2 + b_0 > 0 \quad \text{--- ③}$$

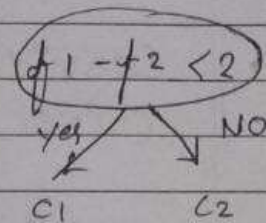
$$a_1 = -1, a_2 = 1, b_0 = 2$$

$$-f_1 + f_2 + 2 > 0$$

$$f_1 - f_2 < 2$$



or



iii) multivariate split

- It uses more than 1 attribute for the split at a node.
- It contains the test of the form $f(x) > 0$ where, $f(x)$ can be any non-linear funⁿ as well.
- Axis parallel & Oblique split are the special cases of multivariate split

classifier performance :

Confusion matrix :

It is a table that categorizes

18. State various impurities in decision trees

Ans.

Entropy: - Entropy can be defined as a measure of the purity of the sub split. Entropy always lies between 0 to 1. The entropy of any split can be calculated by this formula.

$$H(s) = -P_{(+)} \log_2 P_{(+)} - P_{(-)} \log_2 P_{(-)}$$

Here $P_{(+)}/P_{(-)}$ = % of + ve class 1% of - ve class

The algorithm calculates the entropy of each feature after every split and as the splitting continues on, it selects the best feature and starts splitting according to it.

Gini Impurity:

The internal working of Gini impurity is also somewhat similar to the working of entropy in the Decision Tree. In the Decision Tree algorithm, both are used for building the tree by splitting as per the appropriate features but there is quite a difference in the computation of both the methods. Gini Impurity of features after splitting can be calculated by using this formula.

$$GI = 1 - \sum_{i=1}^n (p_i)^2$$

$$GI = 1 - [(P_{(+)})^2 + (P_{(-)})^2]$$

Variance :

Its task is regression

Formula - $\frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$

y_i is label for an instance, N is the number of instances and μ is the mean given by $\frac{1}{N} \sum_{i=1}^N y_i$.

19. Numerical based on impurity calculations(Entropy : <https://youtu.be/knwfCgc9ymU> , Gini : <https://youtu.be/TXp6lr1MSmk>)

DS QUESTION BANK ANSWERS

| | |
|----------|-----|
| PAGE No. | |
| DATE | / / |

* Numerical based on impurity calculations

- ① Entropy
- ② Gini

Given Dataset

| Outlook | Temp | Humidity | Wind | Play |
|----------|------|----------|--------|------|
| Sunny | Hot | High | Weak | No |
| S | H | H | Strong | N |
| Overcast | H | H | W | Yes |
| Rain | Mild | H | W | Y |
| R | cool | Normal | W | Y |
| R | C | N | S | N |
| O | C | N | S | Y |
| S | M | H | W | N |
| S | C | N | W | Y |
| R | M | N | W | Y |
| S | M | N | S | Y |
| O | M | H | S | Y |
| O | H | N | W | Y |
| R | M | H | S | N |

9 Yes

5 No

Total 14

DS QUESTION BANK ANSWERS

| | |
|----------|-----|
| PAGE No. | |
| DATE | / / |

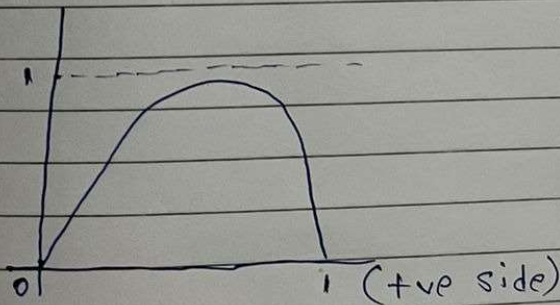
Entropy : It is a measure of disorder of uncertainty in given example dataset

$$\text{Entropy } H(Y) = - \sum_{i=1}^n P_i \log_2 P_i$$

n = number of classes in classification
 \therefore in this example there are two classes Yes and No

P_i = Probability of class i

$$\begin{aligned} H(Y) &= - P_+ \log_2 P_+ - P_- \log_2 P_- \\ &= - \frac{9}{14} \log_2 \frac{9}{14} - \frac{5}{14} \log_2 \frac{5}{14} \\ &= 0.94 \end{aligned}$$



14 Yes, 0 No = 0 Entropy

0 Yes, 14 No = 0 Ent

7 Yes, 7 No = 1 Entropy.

Gini Impurity

Page No.
 Date

Play Tennis

| Outlook | Temperature | Humidity | Wind | Play |
|----------|-------------|----------|--------|------|
| Sunny | Hot | High | Weak | N |
| S | H | H | Strong | N |
| Overcast | H | H | W | Y |
| Rain | Mild | H | W | Y |
| R | cool | Normal | W | Y |
| R | C | N | S | N |
| O | C | N | S | Y |
| S | M | H | W | N |
| S | C | N | W | Y |
| R | M | N | W | Y |
| S | M | N | S | Y |
| O | M | H | S | Y |
| O | H | N | W | Y |
| R | M | H | S | N |

$$G(y) = 1 - \sum_{i=1}^n (p_i)^2$$

$$= 1 - \left(\left(\frac{9}{14} \right)^2 + \left(\frac{5}{14} \right)^2 \right)$$

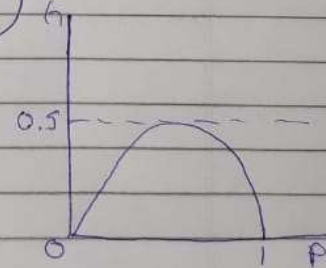
$$= 0.4598$$

$$0^+, 14^- \Rightarrow G = 0$$

$$1^+, 13^- \Rightarrow G = 0.1328$$

$$7^+, 7^- \Rightarrow G = 0.5$$

$$14^+, 0^- \Rightarrow G = 0$$



20. What is decision tree pruning?

Ans. Decision trees are tree data structures that are generated using learning algorithms for the purpose of classification and regression. **Pruning** is a data compression technique in machine learning and search algorithms that reduces the size of decision trees by removing sections of the tree

21. What are the advantages of a decision tree?

- Ans.
1. Simple to understand and interpret
 2. Requires little data preparation
 3. Ability to handle both categorical and numerical data
 4. Robust performance
 5. Well performance with large datasets

22. What do you mean by inter cluster and intra cluster distance?

Ans.

Intercluster Distance

Intercluster distance is the distance between two objects belonging to two different clusters. It is of 5 types –

1. **Single Linkage Distance** : The single linkage distance is the closest distance between two objects belonging to two different clusters defined as

$$\delta_1(S, T) = \min \left\{ d(x, y) \mid x \in S, y \in T \right\}$$

2. **Complete Linkage Distance** : The complete linkage distance is the distance between two most remote objects belonging to two different clusters defined as –

$$\delta_2(S, T) = \max \left\{ d(x, y) \mid x \in S, y \in T \right\}$$

3. **Average Linkage Distance** : The average linkage distance is the average distance between all the objects belonging to two different clusters defined as –

$$\delta_3(S, T) = \frac{1}{|S||T|} \sum_{\substack{x \in S \\ y \in T}} d(x, y)$$

4. **Centroid Linkage Distance** : The centroid linkage distance is the distance between the centers v_s and v_t of two clusters S and T respectively, defined as –

$$\delta_4(S, T) = d(v_s, v_t)$$

where,

$$v_s = \frac{1}{|S|} \sum_{x \in S} x, v_t = \frac{1}{|T|} \sum_{y \in T} y$$

5. **Average Centroid Linkage Distance** : The average centroid linkage distance is the distance between the center of a cluster and all the objects belonging to a different cluster, defined as –

$$\delta_5(S, T) = \frac{1}{|S|+|T|} \left\{ \sum_{x \in S} d(x, v_t) + \sum_{y \in T} d(y, v_s) \right\}$$

Intracuster Distance:

Intracuster distance is the distance between two objects belonging to same cluster. It is of 3 types –

1. **Complete Diameter Distance** : The complete diameter distance is the distance between two most remote objects belonging to the same cluster defined as –

$$\Delta_1(S) = \max\{d(x, y)\}$$

2. **Average Diameter Distance** : The average diameter distance is the average distance between all the objects belonging to the same cluster defined as –

$$\Delta_2(S) = \frac{1}{|S| \cdot (|S| - 1)} \sum_{\substack{x, y \\ x \neq y}} \{d(x, y)\}$$

3. **Centroid Diameter Distance** : The centroid diameter distance is double average distance between all of the objects and the cluster center of s defined as –

$$\Delta_3(S) = 2 \left\{ \frac{\sum_{x \in S} d(x, \bar{v})}{|S|} \right\}$$

where,

$$\bar{v} = \frac{1}{|S|} \sum_{x \in S} x$$

If a clustering algorithm makes clusters so that the Intercluster distance between different clusters is more and Intracluster distance of same cluster is less, then we can tell that it is a good clustering algorithm.

23. What are the various types of clustering approaches?

Ans.

1. Density-Based Clustering

- In this method, the clusters are created based upon the density of the data points which are represented in the data space.

DS QUESTION BANK ANSWERS

- The regions that become dense due to the huge number of data points residing in that region are considered as clusters.
- The data points in the sparse region are considered as noise or outliers.
- The clusters created in these methods can be of arbitrary shape.

2. Hierarchical Clustering

- Hierarchical Clustering groups (Agglomerative or also called Bottom-Up Approach) or divides (Divisive or also called as Top-Down Approach) the clusters based on the distance metrics.
- In Agglomerative clustering, each data point acts as a cluster initially, and then it groups the clusters one by one.
- The clustering of the data points is represented by using a dendrogram

3. Fuzzy Clustering

- In fuzzy clustering, the assignment of the data points in any of the clusters is not decisive. Here, one data point can belong to more than one cluster. It provides the outcome as the probability of the data point belonging to each of the clusters. One of the algorithms used in fuzzy clustering is Fuzzy c-means clustering.
- This algorithm is similar in process to the K-Means clustering and it differs in the parameters that are involved in the computation like fuzzifier and membership values

4. Partitioning Clustering

- This method is one of the most popular choices for analysts to create clusters. In partitioning clustering, the clusters are partitioned based upon the characteristics of the data points.
- We need to specify the number of clusters to be created for this clustering method.

24.What is polythetic clustering?

Ans. Clustering can be done on various criteria. When all features are used at once and clustered objects have no similar property then it is called **Polythetic Clustering**. For example, distance-based clustering.

25. What is monothetic clustering?

DS QUESTION BANK ANSWERS

Ans. Objects clustered using features one by one is called Monothetic Clustering. Such clusters have some properties in common. Examples include clusters of cold-blooded and warm-blooded mammals etc.

The divisive algorithms are considered monothetic if they consider only one feature at a time.

26. What are the advantages and disadvantages of agglomerative clustering?

Ans.

Advantages:

- 1) It is highly versatile and can adapt to clusters of various shapes
- 2) It is good at identifying smaller clusters.
- 3) It is less complex as compared to divisive algorithms.

Disadvantages:

- 1) It is inefficient as compared to divisive algorithms.
- 2) It is computational intensive as it requires calculation and storage of matrix.
- 3) It is a bit less accurate as compared to divisive algorithms as it fails to take the global distribution of data points into picture.

27. What is the k-means clustering approach?

Ans. It is an unsupervised learning clustering algorithm, which groups the unlabeled dataset into different clusters. It classified the data into certain no of assumed 'k' clusters in advance.

It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.

The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

28. What do you mean by holdout, resubstitution, cross-validation and bootstrapping strategies for training and testing?

Ans.

- 1) **The hold-out method** for training the machine learning models is a technique that involves splitting the data into different sets: one set for training, and other sets for validation and testing. The hold-out method is used to check how well a machine learning model will perform on the new data.

2)

29. What do you mean by a confusion matrix?

Ans.

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data.

It is a table that categorizes predictions according to whether they match the actual value in data. One of the table's dimensions indicates the possible categories of predicted values while the other dimension indicates the same for actual values.

DS QUESTION BANK ANSWERS

A confusion matrix can be created for a model predicting any no. of classes.

| n = total predictions | Actual: No | Actual: Yes |
|-----------------------|----------------|----------------|
| Predicted: No | True Negative | False Positive |
| Predicted: Yes | False Negative | True Positive |

The above table has the following cases:

- **True Negative:** Model has given prediction No, and the real or actual value was also No.
- **True Positive:** The model has predicted yes, and the actual value was also true.
- **False Negative:** The model has predicted no, but the actual value was Yes, it is also called as **Type-II error**.
- **False Positive:** The model has predicted Yes, but the actual value was No. It is also called a **Type-I error**.

30. What are various terms like accuracy, precision, recall, sensitivity, specificity, f-score?

1) Accuracy: It represents the number of correctly classified data instances over the total number of data instances.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Accuracy may not be a good measure if the dataset is not balanced.

2) Precision: It refers to the quality of a positive prediction made by the model. Precision refers to the number of true positives divided by the total number of positive predictions. Precision should ideally be 1 (high) for a good classifier.

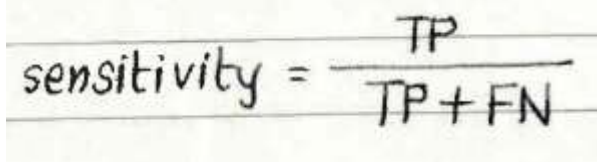
$$Precision = \frac{TP}{TP + FP}$$

DS QUESTION BANK ANSWERS

- 3) Recall: The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

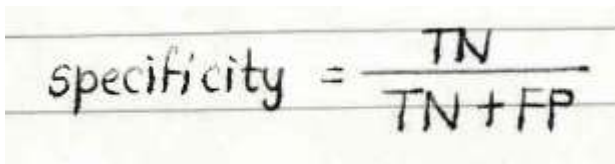
$$Recall = \frac{TP}{TP + FN}$$

- 4) Sensitivity: The sensitivity of a model, also called the positive rate, measures the proportion of positive examples that were correctly classified.



A handwritten formula on lined paper: sensitivity = TP / (TP + FN)

- 5) Specificity: Specificity, also called the true negative rate, measures the proportion of negative examples that were correctly classified.



A handwritten formula on lined paper: specificity = TN / (TN + FP)

- 6) F-score: F1-score is a metric which takes into account both precision and recall.

$$F1 \text{ Score} = 2 * \frac{Precision * Recall}{Precision + Recall}$$