Viva Prep Question Bank

1. Write the formula for Normalization
   depends on type:
   Min-max : x' = (x - min) / (max - min)
   Z-score : x' = (X - μ) / σ
   Decimal Scaling : x' = X / 10^d d $\rightarrow$ the smallest integer such that the largest absolute value of any data point is less than 1
2. What are nominal, ordinal and interval variables?
   Nominal variables are categorical variables that do not have an inherent order, such as colors or types of animals. Ordinal variables have a natural order, such as grades (A, B, C, etc.) or star ratings (1 star, 2 stars, 3 stars, etc.). Interval variables are numeric variables with a consistent scale, such as temperature in Celsius or time in seconds.
3. What are outliers? How are they different from noise?
   Outliers are data points that are significantly different from other data points in a dataset. They can be caused by measurement errors or rare events. Noise, on the other hand, is random variation in data that does not necessarily have a clear cause or pattern.
4. What is the difference between Classification and regression?
   Classification is a supervised learning technique used to predict categorical outcomes, while regression is a supervised learning technique used to predict continuous numerical outcomes.
5. What are the 3 ways of dealing with null values?
   The three ways of dealing with null values are
   deletion : removing any rows or columns with null values from the dataset, but this can lead to a loss of valuable information.
   imputation : filling in the missing values with estimated values, such as the mean or median of the remaining data
   prediction : using ML algorithms that can handle missing data. eg : decision trees, random forests etc.
6. What are the 3 types of subset selection problems?
   The three types of subset selection problems are filter methods, wrapper methods, and embedded methods.
7. What is feature weighting?
   Feature weighting is the process of assigning importance scores to the features in a dataset, based on their relevance to the target variable. It is often used in feature selection and feature extraction techniques.
8. What is feature creation?
   Feature creation is the process of generating new features from existing features in a dataset, to improve the performance of machine learning models. It can involve techniques such as feature scaling, polynomial features, or feature engineering.

9. What is a boxplot? What is the 5 number summary?
   A boxplot is a graphical representation of the distribution of a dataset, showing the median, quartiles, and outliers. The 5 number summary consists of the minimum, the maximum, the first quartile (Q1), the median, and the third quartile (Q3) of the data.
10. Give the formula for Minkowski Distance, Euclidean Distance
    Minkowski distance is a generalization of other distance measures, including Euclidean distance and Manhattan distance. The formula for Minkowski distance is D(x,y) = (|x1-y1|^p + |x2-y2|^p + ... + |xn-yn|^p)^(1/p), where p is a tuning parameter that determines the order of the distance metric. When p = 2, the Minkowski distance is equivalent to the Euclidean distance, which is simply the square root of the sum of the squared differences between corresponding elements in two vectors.
11. What are the 3 properties for metrics?
    The three properties for metrics are positivity, identity, and symmetry. Positivity requires that the distance between two points is always non-negative. Identity requires that the distance between a point and itself is always zero. Symmetry requires that the distance between two points is the same regardless of the order in which they are considered.
12. Write the formula for Gini Index.
    The formula for Gini Index is: Gini(p) = 1 - Σ (pi^2), where p is the probability of each class.
13. Write the formula for Information Gain.
    The formula for Information Gain is: IG(T, X) = H(T) - H(T|X), where T is the dataset, X is the feature, H(T) is the entropy of T, and H(T|X) is the conditional entropy of T given X.
14. Write the formula for Entropy.
    The formula for entropy is: H(S) = - Σ (pi * log2(pi)), where pi is the probability of each class.
15. Why are decision trees prone to overfitting? What is the solution?
    Decision trees are prone to overfitting because they can create complex trees that fit the training data perfectly but do not generalize well to new data. The solution is to prune the tree by removing branches that do not improve the performance on the validation set.
16. What is k-fold cross validation?
    K-fold cross validation is a technique for estimating the performance of a machine learning model by dividing the dataset into k equal-sized folds, training the model on k-1 folds, and testing it on the remaining fold. This process is repeated k times, with each fold being used as the test set exactly once. The results are averaged to give an estimate of the model's performance.
17. What is Leave One Out Cross Validation?
    Leave One Out Cross Validation (LOOCV) is a technique for estimating the performance of a machine learning model by training the model on all but one sample and testing it on the remaining sample. This process is repeated for each

sample in the dataset, with each sample being used as the test set exactly once. The results are averaged to give an estimate of the model's performance.

18. Merits of K-Fold over LOOCV and vice versa.
    K-Fold Cross Validation is less computationally expensive than LOOCV since it trains the model on k-1 folds rather than all but one sample. However, LOOCV has a lower bias than K-Fold since it uses more samples for training and testing. The choice between the two depends on the size of the dataset and the computational resources available.

19. What is the training set and test set?
    The training set is a subset of the dataset used to train a machine learning model. The test set is a subset of the dataset used to evaluate the performance of the model after it has been trained.

20. Give formula for Recall, Precision and F1 Score?
    Recall = TP / (TP + FN)
    Precision = TP / (TP + FP)
    F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
    where TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives.

21. Draw the confusion matrix
    A confusion matrix is a table used to evaluate the performance of a classification model by comparing the predicted and actual values. It has four quadrants: True Positive, False Positive, False Negative, and True Negative. The rows represent the actual values and the columns represent the predicted values.

22. What is coverage and accuracy in Rule based classifiers?
    Coverage is the percentage of instances in the dataset that are covered by the rules in the rule set. Accuracy is the percentage of instances in the dataset that are correctly classified by the rule set.

23. What is a mutually exclusive ruleset?
    A mutually exclusive ruleset is a set of rules where each instance in the dataset is covered by at most one rule.

24. What is an exhaustive ruleset?
    An exhaustive ruleset is a set of rules where each instance in the dataset is covered by at least one rule.

25. What do we do when a ruleset is not exhaustive?
    When a ruleset is not exhaustive, we can either add more rules to the set or use a different classification algorithm that covers the remaining instances.

26. What are the two ordering schemes?
    The two ordering schemes used in association rule mining are: itemset ordering and rule ordering.

27. Difference between Eager and Lazy Learners
    Eager learners build a model from the training data before being presented with new, unseen data. Examples of eager learners include decision trees and Naive Bayes. Lazy learners, on the other hand, defer the processing of data until a query

is made, making them faster to train but slower to make predictions. Examples of lazy learners include k-nearest neighbors and case-based reasoning.

28. What are K nearest neighbors?
    K nearest neighbors is a non-parametric classification algorithm that assigns a class label to an instance based on the class labels of its k nearest neighbors in the training data.
29. What is the weighted K Nearest neighbor?
    Weighted K Nearest neighbor is a variant of the KNN algorithm where the contribution of each neighbor to the classification decision is weighted by its distance from the test instance.
30. Can the classification of a point change according to the value of k?
    - Yes, the classification of a point can change according to the value of k in k-nearest neighbor algorithm. A smaller value of k can lead to overfitting and a larger value of k can lead to underfitting, which can result in different classifications for the same point.
31. List merits and demerits of K Nearest neighbor
    - Merits:
        - Simple to implement
        - Non-parametric, so no assumptions about the underlying data distribution
        - Can handle multi-class classification problems
        - Can work well with both numerical and categorical data
    - Demerits:
        - Computationally expensive for large datasets
        - Sensitive to irrelevant features and noisy data
        - Choosing the optimal value of k can be challenging
        - Does not work well with high-dimensional data
32. Give the formula for Naive Bayes Algorithm
    - The Naive Bayes Algorithm calculates the probability of a given class (C) given a set of features (X) using the Bayes theorem:
        - P(C|X) = (P(X|C) * P(C)) / P(X)
        - where P(C|X) is the posterior probability of class C given the features X
        - P(X|C) is the likelihood of the features X given class C
        - P(C) is the prior probability of class C
        - P(X) is the probability of the features X
33. What is the use of association Mining?
    - Association mining is used to find interesting patterns and relationships between variables in large datasets. It is commonly used in market basket

analysis to identify items that are frequently purchased together. It can also be used in medical diagnosis to find the co-occurrence of symptoms and diseases.

34. Give the formula for support, confidence and lift?
    - Support:
        - The support of an itemset is the proportion of transactions in the dataset that contain the itemset.
    - Formula: Support(X) = (Number of transactions containing X) / (Total number of transactions)
    - Confidence:
        - The confidence of a rule measures the proportion of transactions containing the antecedent that also contain the consequent.
        - Formula: Confidence(X → Y) = Support(X ∪ Y) / Support(X)
    - Lift:
        - The lift of a rule measures the degree of association between the antecedent and consequent, taking into account the background occurrence of both items.
        - Formula: Lift(X → Y) = (Support(X ∪ Y) / N) / (Support(X) / N * Support(Y) / N) = Support(X ∪ Y) / (Support(X) * Support(Y))
        - where N is the total number of transactions in the dataset.
35. State the apriori Principle
    - The Apriori principle states that
      if an itemset is frequent ⇒ all its subsets must also be frequent
      Also if itemset infrequent ⇒ all its supersets must also also infrequent
36. What is Fk-1 * Fk-1 Principle. What is the merging principle?
    - The Fk-1 * Fk-1 principle is a property of frequent itemsets that states that any (k-1)-itemset that is frequent must be a subset of at least one k-itemset that is frequent. This principle is used in the Apriori algorithm to efficiently generate frequent itemsets.
    - The merging principle is used in the generation of candidate itemsets in the Apriori algorithm. It states that if
37. What is Fk-1 * F1 Principle?

    The Fk-1 * F1 principle is a property of frequent itemset mining algorithms that states that any subset of a frequent itemset must also be a frequent itemset.
38. What is average transaction width?

    Average transaction width is the average number of items in a transaction in a dataset used for association rule mining.
39. What is Clustering?

    Clustering is a technique used in data mining and machine learning to group together similar data points based on some similarity measure. The goal of

clustering is to find natural groupings in the data without any prior knowledge of the groups.

40. What is the difference between hierarchical clustering and partial clustering?

    Hierarchical clustering is a type of clustering algorithm that builds a hierarchy of clusters by recursively dividing the data into smaller clusters based on a similarity measure. Partial clustering is a technique where only a subset of the data is clustered.

41. What is complete vs partial clustering?

    Complete clustering is a technique where all data points are assigned to a cluster. Partial clustering is a technique where only a subset of the data is assigned to a cluster.

42. Exclusive Vs Overlapping vs Fuzzy Clustering

    Exclusive clustering is a type of clustering where each data point belongs to only one cluster. Overlapping clustering is a type of clustering where each data point can belong to multiple clusters. Fuzzy clustering is a type of clustering where each data point is assigned a probability of belonging to each cluster.

43. Explain the types of clusters (Well Separated, Prototype Based, Density Based, Graph Based, Conceptual Clustering).

    - Well-separated clusters are clusters that are clearly separated from each other.
    - Prototype-based clustering is a type of clustering where each cluster is represented by a prototype, which is a data point that is considered to be representative of the cluster.
    - Density-based clustering is a type of clustering where clusters are defined based on areas of high density in the data.
    - Graph-based clustering is a type of clustering where clusters are defined based on the connectivity of the data points.
    - Conceptual clustering is a type of clustering where clusters are defined based on a set of predefined concepts or rules.

44. What is the k-means algorithm? What all proximity measures can be used to calculate the distance between points?

    The k-means algorithm is a clustering algorithm that partitions the data into k clusters based on a distance measure. The algorithm works by randomly selecting k initial cluster centers and then iteratively assigning data points to the nearest cluster center and updating the cluster centers based on the new data points assigned to the cluster. The algorithm continues to iterate until convergence is achieved.

    Various proximity measures can be used to calculate the distance between points, including Euclidean distance, Manhattan distance, and cosine similarity.

45. What are the limitations of random initialization of centroids in K Means?

    Random initialization of centroids in K Means can lead to suboptimal solutions, as the initial centroids may be chosen in a way that does not reflect the true structure

of the data. Additionally, the algorithm may get stuck in local optima if the initial centroids are not chosen carefully.

46. How to address empty clusters in K Means?

    Empty clusters can be addressed in K Means by reinitializing the empty cluster with a new centroid, either by selecting a new random point from the data or by using a more sophisticated method to select the new centroid.

47. What is SSE?

    SSE stands for Sum of Squared Errors and is a measure of the distance between each data point and its assigned cluster center in K Means clustering. The objective of K Means is to minimize SSE.

48. How can you reduce SSE in K Means?

    SSE (Sum of Squared Errors) is a measure of the total distance between the data points and their respective centroids in K Means clustering. To reduce SSE in K Means, we can follow these strategies:

    1. Increase the number of clusters: With more clusters, the centroids will be closer to the data points and SSE will reduce.
    2. Choose better initial centroids: Random initialization of centroids can sometimes result in high SSE. To reduce this, we can use methods like K-Means++ initialization or hierarchical clustering to choose better initial centroids.
    3. Run K Means multiple times: Running K Means multiple times with different initial centroids and choosing the one with the lowest SSE can help reduce the SSE.
    4. Use different distance measures: Instead of Euclidean distance, other distance measures like Manhattan distance or cosine distance can be used to measure the distance between data points and centroids. This can sometimes result in lower SSE.
    5. Remove outliers: Outliers can have a significant impact on SSE. Removing them before running K Means can help reduce SSE.

49. In what cases can K Means not handle clusters?
    K Means algorithm fails to handle non-linearly separable data. It also assumes that clusters are isotropic and have similar sizes. Additionally, it requires a pre-defined number of clusters as input, which may not always be known.

50. What are dendrograms?
    Dendrograms are a graphical representation of the hierarchy of clusters in hierarchical clustering. They show how the clusters are merged or divided at each step, and the distance between them.

51. What is agglomerative vs Divisive clustering?
    Agglomerative clustering is a bottom-up approach where individual points or small clusters are gradually merged together to form larger clusters. Divisive clustering, on the other hand, is a top-down approach where a single large cluster is divided into smaller clusters.

52. What are the three types of proximity links in Hierarchical Clustering?
    The three types of proximity links in hierarchical clustering are Single Linkage, Complete Linkage, and Average Linkage. Single Linkage considers the minimum distance between any two points in the clusters being merged, Complete Linkage considers the maximum distance, and Average Linkage considers the average distance.

53. Can the cluster of a point change according to the type of linkage used?
    Yes, the cluster of a point can change according to the type of linkage used in hierarchical clustering. This is because different linkage methods calculate the distance between clusters in different ways, which can affect the final clustering.

54. What are the issues in hierarchical clustering?
    Some issues with hierarchical clustering include: it can be computationally expensive, the choice of linkage method can affect the results, and it is sensitive to noise and outliers.

55. What is DB Scan algorithm?
    DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that groups together points that are close to each other based on a density threshold. It can find clusters of arbitrary shapes and sizes, and can also identify noise points.

56. What are core points, border points and noise points?
    In DBSCAN, core points are points that have a minimum number of neighboring points within a given distance. Border points have fewer neighbors than the minimum threshold, but are within the distance of a core point. Noise points have no neighbors within the minimum distance.

57. How can you select the value of Eps and Minpoint?
    The values of Eps and Minpoints can be selected by visual inspection of the data or by using a heuristic approach. A common method is to plot the distance to the kth nearest neighbor for each point and look for a knee point where the distance sharply increases. The value of k can be estimated using the number of dimensions in the data.

58. What are the advantages and disadvantages of DBScan?
    Advantages of DBSCAN include its ability to handle non-linearly separable clusters, its ability to detect noise points, and its ability to find clusters of arbitrary shapes and sizes. Disadvantages include its sensitivity to the choice of parameters, its dependence on the density of the data, and its computational complexity.