GATE

DPP 01

Data Science and Artificial Intelligence

Machine Learning

DPP: 1

Unsupervised Learning & ANN

Q1 Which of the following statements is true about a Random Forest Classifier?
(i) Increasing the number of trees will generally decrease model bias.
(ii) Decreasing the depth of trees will generally increase model bias.
(iii) Increasing the number of trees will increase model variance.
(iv) Decreasing the depth of trees will generally decrease model variance.
(A) i and ii (B) i and iv
(C) ii and iii (D) ii and iv

Q2 Which of the following statements is true about a Random Forest Regressor?
(i) Increasing the minimum samples per leaf will generally decrease model variance.
(ii) Increasing the minimum samples per leaf will generally increase model bias.
(iii) Decreasing the number of trees will generally increase model bias.
(iv) Increasing the number of features considered for splitting at each node will generally decrease model variance.
(A) i and ii
(B) i and iii
(C) ii and iv
(D) iii and iv

Q3 In a Random Forest Classifier, what effect does increasing the max_features parameter have on the model's performance?
(i) Increases model bias.
(ii) Decreases model bias.
(iii) Increases model variance.
(iv) Decreases model variance.
(A) i and iii
(B) ii and iv
(C) i and iv
(D) ii and iii

Q4 In a Random Forest Regressor, what is the impact of increasing the n_estimators parameter?
(i) Steadily increases computational cost.
(ii) Significantly increases model variance after a certain point.
(iii) Generally decreases model variance up to a certain point.
(iv) Has a significant impact on reducing model bias.
(A) i and iii
(B) ii and iv
(C) i and iv
(D) iii and iv

Q5 What happens when you increase the min_samples_split parame-ter in a Random Forest Classifier?
(i) Increases model bias.
(ii) Decreases model bias.

GATE

(iii) Increases model variance.
(iv) Decreases model variance.
(A) i and iii
(B) ii and iv
(C) i and iv
(D) ii and iii

Q6 In a Random Forest Regressor, what is the effect of using a larger bootstrap sample size
(i) Increases model bias.
(ii) Decreases model bias.
(iii) Increases model variance.
(iv) Decreases model variance.
(A) i and iii (B) ii and iv
(C) i and iv (D) ii and iii

Q7 What is the impact of reducing the max_depth of trees in a Random Forest Classifier?
(i) Increases model bias.
(ii) Decreases model bias.
(iii) Increases model variance.
(iv) Decreases model variance.
(A) i and iv
(B) ii and iii
(C) i and iii
(D) ii and iv

Q8 What happens in a Random Forest Regressor when the number of features considered at each split (max_features) is set to its maximum?
(i) Increases model bias.
(ii) Decreases model bias.
(iii) Increases model variance.
(iv) Decreases model variance.
(A) ii and iii
(B) i and iv
(C) ii and iv
(D) i and iii

Q9 In a Random Forest Classifier, what is the effect of increasing the n_estimators parameter beyond an optimal point?
(i) Steadily increases computational cost.
(ii) Significantly increases model variance.
(iii) Generally decreases model variance.
(iv) Has a significant impact on reducing model bias.
(A) i and ii
(B) i and iii
(C) i and iv
(D) ii and iii

Q10 What best describes the K-Nearest Neighbors algorithm?
(i) It is a type of neural network.
(ii) It is a lazy learning algorithm.
(iii) It requires feature scaling for optimal performance.
(iv) It is primarily used for regression tasks.
(A) ii and iii (B) i and iv
(C) ii and iv (D) i and iii

Q11 In the K-Nearest Neighbors algorithm, what is the effect of choosing a very small value for K?
(i) Increases bias.
(ii) Decreases bias.
(iii) Increases variance.
(iv) Decreases variance.
(A) i and iv
(B) ii and iii
(C) ii and iv
(D) i and iii

Q12 What factors significantly affect the performance of a K-Nearest Neighbors model?
(i) Number of features in the dataset.
(ii) The distance metric used.
(iii) The learning rate.
(iv) The number of layers in the model.
(A) i and ii
(B) iii and iv

GATE

(C) i, ii, and iii
(D) ii, iii, and iv

Q13 How does K-Nearest Neighbors typically perform on highly imbal-anced datasets?
(i) Performs well regardless of class distribution.
(ii) Can be biased towards the majority class.
(iii) Requires a large number of K to perform well.
(iv) Is unaffected by the choice of distance metric.
(A) i and iv
(B) ii and iii
(C) ii and iv
(D) i and iii

Q14 Which of the following is true when using K-Nearest Neighbors in complex datasets with many overlapping classes?
(i) KNN tends to perform very well.
(ii) KNN often struggles due to its simplicity.
(iii) The model's performance improves significantly with feature scaling.
(iv) The model's complexity increases with more data.
(A) i and iii
(B) ii and iv
(C) ii and iii
(D) i and iv

Q15 What is a key feature of the Support Vector Machine (SVM) algo-rithm?
(i) It is primarily a probabilistic model.
(ii) It focuses on maximizing the margin between classes.
(iii) It is best suited for large datasets with millions of samples.
(iv) It relies on kernel functions to transform data into higher dimensions.
(A) ii and iv
(B) i and iii
(C) ii and iii
(D) i and iv

Q16 How does the choice of kernel in an SVM affect its performance?
(i) A linear kernel is always the best choice for high-dimensional data.
(ii) Non-linear kernels can capture complex relationships in the data.
(iii) The RBF kernel always overfits the data.
(iv) Kernel choice does not affect model complexity.
(A) ii
(B) i and iii
(C) ii and iii
(D) iv

Q17 In SVM, what parameter adjustments can help prevent overfitting?
(i) Increasing the value of C.
(ii) Decreasing the value of C.
(iii) Using a higher-degree polynomial kernel.
(iv) Using a simpler kernel, like linear.
(A) ii and iv
(B) i and iii
(C) ii and iii
(D) i and iv

Q18 How important is feature scaling for SVM performance?
(i) Feature scaling is not necessary for SVM.
(ii) Feature scaling can significantly impact the performance of SVM.
(iii) Feature scaling only affects linear SVM.
(iv) Feature scaling affects the convergence speed but not the accuracy.
(A) ii

(B) i and iii
(C) iv
(D) i

Q19 How does SVM handle multiclass classification problems?
(i) Directly handles multiclass classification.
(ii) Uses one-vs-one or one-vs-all strategies.
(iii) Not suitable for multiclass classification.
(iv) Requires transformation of data into binary classes.
(A) i and iii
(B) ii
(C) ii and iv
(D) i and iv

Q20 What is a challenge when using SVM with imbalanced datasets?
(i) SVM inherently balances class weights.
(ii) SVM may bias the hyperplane towards the majority class.
(iii) Imbalanced datasets do not affect SVM performance.
(iv) SVM requires undersampling or oversampling techniques.
(A) ii
(B) i and iii
(C) iv
(D) ii and iv

Q21 Which statement is true about the K-Means clustering algorithm?
(i) It requires the number of clusters to be specified in advance.
(ii) It automatically determines the optimal number of clusters.
(iii) It is a hierarchical clustering method.
(iv) It is a density-based clustering method.
(A) i (B) ii
(C) iii (D) iv

Q22 How does the initialization of centroids affect the K-Means algo-rithm?
(i) It has no effect on the final clusters.
(ii) It can affect the convergence speed of the algorithm.
(iii) It can lead to different final clustering results.
(iv) It only affects the algorithm's performance on large datasets.
(A) ii and iii
(B) i
(C) iii and iv
(D) ii and iv

Q23 Why is feature scaling important in K-Means clustering?
(i) Feature scaling is not necessary for K-Means.
(ii) It ensures that all features contribute equally to the distance calculations.
(iii) It changes the shape of clusters.
(iv) It decreases computational complexity.
(A) ii
(B) i and iii
(C) ii and iv
(D) iii

Q24 What is a limitation of the K-Means clustering algorithm?
(i) It works well with non-spherical shapes.
(ii) It requires the number of clusters to be known beforehand.
(iii) It can handle different sizes of clusters efficiently.
(iv) It is suitable for clustering data with large variances in density.
(A) ii (B) i and ii
(C) iii and iv (D) i

Q25

GATE

How does the presence of outliers affect K-Means clustering?
(i) Outliers do not affect K-Means.
(ii) Outliers can significantly distort the shape of clusters.
(iii) K-Means automatically removes outliers.
(iv) Outliers impact the convergence speed of K-Means.
(A) ii
(B) i and iii
(C) iii and iv
(D) ii and iv

Q26 What metric is commonly used to evaluate the quality of clusters formed by K-Means?
(i) Silhouette coefficient.
(ii) Precision and recall.
(iii) Root mean square error.
(iv) F1 score.
(A) i (B) ii and iv
(C) i and iii (D) iii

Q27 What indicates that the K-Means algorithm has converged?
(i) All centroids remain stationary.
(ii) The sum of within-cluster variances is minimized.
(iii) All points are assigned to the nearest cluster.
(iv) The algorithm has completed a fixed number of iterations.
(A) i and ii
(B) ii and iii
(C) i, ii, and iii
(D) iv

Q28 Which statement correctly describes a characteristic of hierarchical clustering?
(i) It requires the number of clusters to be specified in advance.
(ii) It creates a dendrogram to represent the data.
(iii) It is primarily used for large datasets.
(iv) It is faster than K-Means clustering.
(A) ii
(B) i and iii
(C) i and iv
(D) iii and iv

Q29 What are the two main types of hierarchical clustering?
(i) Divisive and Agglomerative.
(ii) K-Means and DBSCAN.
(iii) Agglomerative and K-Means.
(iv) Divisive and Spectral.
(A) i (B) ii
(C) iii (D) iv

Q30 Which distance metric is not commonly used in hierarchical cluster-ing?
(i) Euclidean distance.
(ii) Manhattan distance.
(iii) Cosine similarity.
(iv) Jaccard index.
(A) iii (B) iv
(C) ii (D) i

Q31 What is an advantage of hierarchical clustering over K-Means clus-tering?
(i) It is more computationally efficient.
(ii) It does not require the number of clusters to be specified.
(iii) It always produces spherical clusters.
(iv) It is better suited for very large datasets.
(A) ii
(B) i and iii
(C) iii and iv
(D) i and iv

Q32 What is a limitation of hierarchical clustering?

GATE

(i) It can't handle noisy data.
(ii) It's sensitive to outliers.
(iii) It can be used only with numerical data.
(iv) It's not suitable for any dimensionality reduction techniques.
(A) ii (B) i
(C) iii (D) iv

Q33 In DBSCAN, what defines a core point?
(i) A point that has a minimum number of other points within a specified radius.
(ii) Any point that is not an outlier.
(iii) A point at the center of a cluster.
(iv) A point that is equidistant from all border points of a cluster.
(A) i (B) ii
(C) iii (D) iv

Q34 Which parameters are crucial in determining the behavior of the DBSCAN algorithm?
(i) The number of clusters.
(ii) The (epsilon) radius.
(iii) The minimum number of points required to form a dense region (minPts).
(iv) The maximum distance between two points.
(A) ii and iii
(B) i and iv
(C) ii, iii, and iv
(D) All of the above

Q35 How does DBSCAN handle noise in the dataset?
(i) It classifies noisy points as separate clusters.
(ii) It identifies and excludes noisy points from clusters.
(iii) It uses noise points to connect different clusters.
(iv) It ignores noise points during the clustering process.
(A) ii (B) i
(C) iii (D) iv

Q36 What types of cluster shapes can DBSCAN effectively identify?
(i) Only spherical clusters.
(ii) Clusters of any shape.
(iii) Only elongated clusters.
(iv) Only clusters that are evenly distributed.
(A) ii (B) i
(C) iii (D) iv

Q37 How does DBSCAN perform in high-dimensional spaces?
(i) Its performance improves with an increase in dimensionality.
(ii) It struggles due to the curse of dimensionality.
(iii) It remains unaffected by the dimensionality of the space.
(iv) It requires dimensionality reduction before application.
(A) ii (B) i
(C) iii (D) iv

Q38 What is a challenge when using DBSCAN on datasets with varying densities?
(i) DBSCAN cannot handle datasets with varying densities.
(ii) Choosing a single value of may not be suitable for all regions in the dataset.
(iii) The algorithm becomes computationally more intensive.
(iv) It automatically adjusts for different densities.
(A) ii (B) i
(C) iii (D) iv

Q39 What factor affects the scalability of the DBSCAN algorithm?

GATE

(i) The number of clusters in the dataset.
(ii) The choice of distance metric.
(iii) The computational complexity of finding neighbors.
(iv) The need to specify the number of clusters.
(A) iii (B) ii
(C) i and iv (D) ii and iii

Q40 In PCA, what is the significance of the eigenvalues obtained from the covariance matrix?
(i) They determine the rotation of the principal components.
(ii) They indicate the amount of variance captured by each principal compo-nent.
(iii) They define the number of principal components to be selected.
(iv) They are used for scaling the original features.
(A) ii
(B) i and iii
(C) ii and iii
(D) iv

Q41 How does PCA handle correlated features in a dataset?
(i) It increases the correlation between features.
(ii) It eliminates all correlations between features.
(iii) It combines correlated features into principal components.
(iv) It assigns correlated features to separate components.
(A) iii (B) ii
(C) i (D) iv

Q42 What is an effective use of PCA in the context of data visualization?
(i) To increase the dimensionality for more detailed visualizations.
(ii) To reduce high-dimensional data to two or three dimensions for plotting.
(iii) To visualize the covariance matrix.
(iv) To separate correlated features in a scatter plot.
(A) ii (B) i
(C) iii (D) iv

Q43 What is a limitation of using PCA for dimensionality reduction?
(i) It can lead to information loss.
(ii) It increases the computational complexity.
(iii) It always increases the model's accuracy.
(iv) It cannot be used with non-linear data.
(A) i (B) ii
(C) iii (D) iv

Q44 Why is it important to standardize data before applying PCA?
(i) PCA only works with standardized data.
(ii) Standardization ensures that features with larger scales do not dominate.
(iii) It enhances the non-linear relationships in the data.
(iv) Standardization is not necessary for PCA.
(A) ii (B) i
(C) iii (D) iv

Q45 What role does eigen decomposition play in PCA?
(i) It is used to calculate the covariance matrix.
(ii) It helps in identifying the principal components.
(iii) It standardizes the dataset.
(iv) It is used to increase the dimensionality of the dataset.
(A) ii (B) i

GATE

(C) iii (D) iv

Q46 How do outliers affect the results of PCA?

(i) PCA is robust to outliers.

(ii) Outliers can significantly influence the direction of the principal compo-nents.

(iii) PCA automatically removes outliers.

(iv) Outliers enhance the performance of PCA.

(A) ii (B) i

(C) iii (D) iv

GATE

Answer Key

Q1	(D)
Q2	(A)
Q3	(D)
Q4	(A)
Q5	(C)
Q6	(D)
Q7	(A)
Q8	(A)
Q9	(B)
Q10	(A)
Q11	(B)
Q12	(A)
Q13	(B)
Q14	(C)
Q15	(A)
Q16	(A)
Q17	(A)
Q18	(A)
Q19	(B)
Q20	(D)
Q21	(A)
Q22	(A)
Q23	(A)
Q24	(A)
Q25	(A)
Q26	(A)
Q27	(C)
Q28	(A)
Q29	(A)
Q30	(B)
Q31	(A)
Q32	(A)
Q33	(A)
Q34	(A)
Q35	(A)
Q36	(A)
Q37	(A)
Q38	(A)
Q39	(A)
Q40	(A)
Q41	(A)
Q42	(A)
Q43	(A)
Q44	(A)
Q45	(A)
Q46	(A)

GATE

Hints & Solutions

Q1 Text Solution:

- Increasing the number of trees (i): Adding more trees to a Random Forest generally helps in averaging out predictions, leading to a reduction in variance. It doesn't significantly affect bias because the decision boundaries determined by individual trees don't change; instead, they are averaged over more estimators. Therefore, statement i is incorrect.
- Decreasing the depth of trees (ii): Reducing the depth of the trees in a Random Forest makes individual trees simpler and less capable of capturing complex patterns in the data, which generally increases model bias. So, statement ii is correct.
- Increasing the number of trees (iii): More trees in a Random Forest typically decrease model variance, as the averaging effect of multiple trees stabilizes the predictions. So, statement iii is incorrect.
- Decreasing the depth of trees (iv): Shallower trees in a Random Forest are less complex and less likely to overfit the training data, leading to a decrease in model variance. Therefore, statement iv is correct.

Q2 Text Solution:

- Increasing minimum samples per leaf (i): This action makes the model more conservative, as each leaf will represent a larger portion of the dataset. This reduces the model's complexity and variance, as it becomes less sensitive to fluctuations in the training data. So, statement i is correct.
- Increasing minimum samples per leaf (ii): By requiring more sam-ples per leaf, the model becomes less flexible in fitting the data, potentially missing out on capturing some nuances, thereby increasing bias. There-fore, statement ii is correct.
- Decreasing the number of trees (iii): While decreasing the number of trees can increase variance due to less averaging of predictions, it doesn't significantly affect bias. The bias is more a factor of the individual tree structure rather than the number of trees. So, statement iii is incorrect.
- Increasing the number of features for splitting (iv): Increasing the number of features considered for splitting at each node can potentially make each tree in the forest more complex and potentially more prone to overfitting, which could increase variance rather than decrease it. There-fore, statement iv is incorrect.

Q3 Text Solution:

- **Increasing max_features (ii):** This action allows each tree in the Ran-dom Forest to consider a larger number of features at each split. This can potentially improve the model's ability to fit the training data more accurately, hence decreasing bias. So, ii is correct.
- **Increasing max_features (iii):** While this may decrease bias, it also tends to increase the variance because trees may become more similar to each other, losing the benefit of averaging out errors. So, iii is also correct.

Q4 Text Solution:

- **Increasing n_estimators (i):** Adding more trees to the Random Forest increases the computational cost linearly, as each tree adds to the amount of computation required. So, i is

correct.

- **Increasing n_estimators (iii):** Generally, adding more trees to a Ran-dom Forest reduces variance by averaging out errors across a larger number of estimators. This benefit, however, plateaus after a certain number of trees. So, iii is correct.
- **Increasing n_estimators (ii and iv):** Adding more trees does not sig-nificantly increase model variance beyond a certain point (ii is incorrect), and while it helps reduce variance, it does not have a significant impact on model bias (iv is incorrect).

These questions help in understanding the behavior of Random Forest algo-rithms in machine learning, particularly how different parameters can influence the model's bias, variance, and overall performance.

Q5 Text Solution:

Explanation:

- **Increasing min_samples_split (i):** This makes the model more conser-vative, as it requires more samples to make a split in a tree. It tends to simplify the model, potentially increasing bias.
- **Increasing min_samples_split (iv):** Simplifying the model by increas-ing min_samples_split generally leads to a decrease in variance, as the model becomes less sensitive to fluctuations in the training data.

Q6 Text Solution:

- **Larger bootstrap sample size (ii):** This allows each tree to be trained on a more comprehensive dataset, potentially capturing more patterns and reducing bias.
- **Larger bootstrap sample size (iii):** However, it can also lead to an increase in variance as individual trees may become more similar to each other, reducing the benefit of averaging out errors.

Q7 Text Solution:

- **Reducing max_depth (i):** This action limits the depth of the trees, mak-ing them less complex and potentially increasing bias as the model may not capture all the nuances in the data.
- **Reducing max_depth (iv):** Simpler trees are less likely to overfit the data, which usually results in decreased variance.

Q8 Text Solution:

Explanation:

- **Max max_features (ii):** This allows each split in the trees to consider more features, which can help in capturing more information and reducing bias.
- **Max max_features (iii):** However, it can also increase variance as trees may become more complex and overfit the training data.

Q9 Text Solution:

Explanation:

- **Increasing n_estimators beyond optimal (i):** Adding more trees to the Random Forest increases the computational cost, as each tree adds to the total computation required.
- **Increasing n_estimators beyond optimal (iii):** Generally, adding more trees to a Random Forest reduces variance by averaging out errors across more estimators, but this benefit plateaus after a certain number of trees.

These questions cover various aspects of Random Forests, highlighting how dif-ferent parameters can influence the model's performance, bias, and variance.

Q10 Text Solution:

- **KNN is a lazy learning algorithm (ii):** KNN is considered a lazy learner because it doesn't learn a discriminative function from the training data but memorizes the dataset.
- **KNN requires feature scaling (iii):** Since KNN relies on the distance between feature vectors, feature scaling is important for its performance, as it ensures that all features contribute equally to the distance calcula-tions.

Q11 Text Solution:

Explanation:

- **Small K (ii):** A smaller K value means the algorithm considers fewer neighbors, which can make the model more sensitive to local patterns, thus reducing bias.
- **Small K (iii):** However, this also makes the model more sensitive to noise in the data, potentially increasing variance.

Q12 Text Solution:

Explanation:

- **Number of features (i):** The performance of KNN is heavily influenced by the number of features (the curse of dimensionality). More features can make distance less meaningful.
- **Distance metric used (ii):** The choice of distance metric (e.g., Eu-clidean, Manhattan) significantly affects the model's performance.

Q13 Text Solution:

Explanation:

- **Biased towards the majority class (ii):** KNN can be biased towards the majority class in imbalanced datasets, as the neighbors of a given point are more likely to belong to the majority class.
- **Requires a large number of K (iii):** Adjusting the value of K can sometimes help in balancing the influence of the majority class, but it's not a guaranteed solution.

Q14 Text Solution:

Explanation:

- **Struggles due to simplicity (ii):** KNN may struggle in complex datasets with many overlapping classes due to its simple decision mechanism based on nearest neighbors.
- **Improves with feature scaling (iii):** While feature scaling improves KNN's performance, it may not be sufficient to address the challenges posed by complex datasets with overlapping classes.

These questions cover key concepts of the K-Nearest Neighbors algorithm, in-cluding its characteristics, parameter tuning, performance factors, and behavior in different types of datasets.

Q15 Text Solution:

Explanation:

- **Maximizing margin (ii):** SVM works by finding the hyperplane that maximizes the margin between different classes.
- **Kernel functions (iv):** SVM uses kernel functions to transform data into higher dimensions to find an optimal separating hyperplane in cases where data is not linearly separable in the original space.

Q16 Text Solution:

Explanation:

Non-linear kernels (ii): Non-linear kernels like RBF, polynomial, etc., allow SVM to capture complex relationships in the data by mapping input features into higher-dimensional spaces.

Q17 Text Solution:

Explanation:

- **Decreasing C (ii):** A lower value of the C parameter can help by mak-ing the decision boundary smoother and less sensitive to individual data points.
- **Simpler kernel (iv):** Using a simpler kernel reduces the risk of creating an overly complex model that overfits the data.

Q18 Text Solution:

Explanation:

Feature scaling impact (ii): Feature scaling is critical for SVM as it uses distance measures. Features on larger scales can dominate the decision boundary, so scaling ensures that all features contribute equally.

Q19 Text Solution:

Explanation:

One-vs-one or one-vs-all (ii): SVM is inherently binary but can handle multiclass problems using strategies like one-vs-one (comparing each pair of classes) or one-vs-all (comparing one class against all others).

Q20 Text Solution:

Explanation:

- **Bias towards majority class (ii):** In imbalanced datasets, SVM may bias the decision boundary towards the majority class.
- **Requires balancing techniques (iv):** Techniques like undersampling, oversampling, or adjusting class weights are often necessary to address this imbalance.

Q21 Text Solution:

Number of clusters (i): K-Means requires the user to specify the num-ber of clusters (K) in advance. It does not automatically determine the number of clusters.

Q22 Text Solution:

- **Convergence speed (ii):** Different initializations can lead to faster or slower convergence to the final clusters.
- **Final clustering results (iii):** Depending on the initial centroids, K-Means might converge to different local optima, affecting the final cluster-ing outcome.

Q23 Text Solution:

Explanation:

Equal contribution to distance calculations (ii): Feature scaling is crucial in K-Means because it ensures that all features are on a sim-ilar scale, thus contributing equally to the distance calculations used in forming clusters.

Q24 Text Solution:

Explanation:

Requires number of clusters (ii): One of the key limitations of K-Means is that it requires the number of clusters (K) to be specified in advance.

Q25 Text Solution:

Explanation:

Distort cluster shape (ii): Outliers can greatly affect the K-Means algorithm since it uses the mean of points in a cluster, and outliers can skew this mean significantly.

Q26 Text Solution:

Explanation:

Silhouette coefficient (i): The silhouette coefficient is a common metric for assessing the quality of clusters in K-Means. It measures how

similar an object is to its own cluster compared to other clusters.

Q27 Text Solution:

Explanation:

- **Stationary centroids (i):** Convergence in K-Means is typically indicated when the centroids no longer move significantly.
- **Minimized variance (ii):** Another sign of convergence is when the sum of the squared distances of samples to their closest cluster center is mini-mized.
- **Assignment to nearest cluster (iii):** Points are assigned to the nearest cluster based on the current centroids, and if this assignment does not change between iterations, it indicates convergence.

These questions cover various aspects of the K-Means algorithm, including its initialization, performance, limitations, and evaluation metrics, providing a com-prehensive understanding of its operation and usage.

Q28 Text Solution:

Explanation:

Dendrogram creation (ii): Hierarchical clustering creates a dendro-gram, which is a tree-like diagram that records the sequences of merges or splits.

Q29 Text Solution:

Explanation:

Divisive and Agglomerative (i): The two main types of hierarchical clustering are divisive (top-down approach) and agglomerative (bottom-up approach).

Q30 Text Solution:

Explanation:

Jaccard index (iv): While the Jaccard index is a popular metric for com-paring the similarity and diversity of sample sets, it is less commonly used in hierarchical clustering compared to Euclidean, Manhattan, or Cosine similarity.

Q31 Text Solution:

Explanation:

No need to specify number of clusters (ii): Unlike K-Means, hier-archical clustering doesn't require you to specify the number of clusters beforehand; the dendrogram can be used to determine the number of clus-ters by cutting it at the appropriate level.

Q32 Text Solution:

Explanation:

Sensitivity to outliers (ii): Hierarchical clustering is sensitive to out-liers because outliers can distort the distances between clusters, leading to potentially erroneous merges or splits.

These questions and answers aim to provide a comprehensive understanding of hierarchical clustering, covering its characteristics, types, distance metrics, advantages, and limitations.

Q33 Text Solution:

Explanation:

Minimum number within radius (i): A core point in DBSCAN is defined as a point that has at least a minimum number of other points (defined by minPts) within a specified radius ([redacted]).

Q34 Text Solution:

Explanation:

Epsilon radius (ii) and minPts (iii): The behavior of DBSCAN is pri-marily determined by the [redacted] radius, which defines the neighborhood around

a point, and the minimum number of points (minPts) required to form a dense region.

Q35 Text Solution:

Explanation:

- **Identifies and excludes noise (ii):** DBSCAN identifies noise points points that do not meet the criteria for a core or border point – and excludes them from clusters.

Q36 Text Solution:

Explanation:

Any shape clusters (ii): One of the strengths of DBSCAN is its ability to identify clusters of arbitrary shapes, as it relies on density estimation.

Q37 Text Solution:

Explanation:

Struggles with high dimensionality (ii): Like many clustering al-gorithms, DBSCAN's performance tends to degrade in high-dimensional spaces due to the curse of dimensionality.

Q38 Text Solution:

Explanation:

Single value issue (ii): When dealing with datasets that have regions of varying densities, choosing a single ε value can be challenging as it might be too large for some areas and too small for others.

Q39 Text Solution:

Explanation:

Computational complexity of finding neighbors (iii): The scala-bility of DBSCAN is primarily affected by the computational complexity involved in finding the neighbors of points within the radius, which can be significant, especially for large datasets.

These questions delve into various aspects of DBSCAN, from its basic concepts and handling of noise to its performance in different scenarios, highlighting the complexities and challenges associated with this clustering algorithm.

Q40 Text Solution:

Explanation:

Variance indication (ii): Eigenvalues obtained from the covariance ma-trix in PCA indicate the amount of variance captured by each principal component. Higher eigenvalues correspond to components with more cap-tured variance.

Q41 Text Solution:

Explanation:

Combining correlated features (iii): PCA combines correlated fea-tures into principal components, thereby reducing the dimensionality and retaining most of the variability in the data.

Q42 Text Solution:

Explanation:

Reducing dimensions for plotting (ii): PCA is often used to reduce high-dimensional data to two or three dimensions, making it possible to visualize complex datasets in a 2D or 3D scatter plot.

Q43 Text Solution:

Explanation:

Information loss (i): A key limitation of PCA is that it can lead to information loss, especially if a significant number of principal components are discarded.

Q44 **Text Solution:**

Explanation:

Preventing dominance of larger scales (ii): Standardizing data be-fore applying PCA is important to ensure that features with larger scales do not dominate the principal components, which are sensitive to the scale of the features.

Q45 **Text Solution:**

Explanation:

Identifying principal components (ii): Eigen decomposition of the covariance matrix is a key step in PCA, used to identify the principal components, which are the eigenvectors corresponding to the largest eigen-values.

Q46 **Text Solution:**

Explanation:

Influence on principal components (ii): Outliers can have a signifi-cant impact on PCA because they can substantially alter the covariance structure of the data, influencing the direction of the principal compo-nents.

These questions cover a range of topics related to PCA, from its core concepts and applications to its limitations and the impact of data characteristics on its performance.