**INT423**

**SAMPLE QP**

**Machine Learning Clustering Algorithms-I:**

**Introduction to Unsupervised Learning**

1. **What is unsupervised learning?**
   * A) Learning with labeled data
   * B) Learning without labeled data
   * C) Learning with both labeled and unlabeled data
   * D) Learning through trial and error
   * **Answer:** B
2. **Which of the following is NOT a type of unsupervised learning?**
   * A) Clustering
   * B) Regression
   * C) Dimensionality reduction
   * D) Association
   * **Answer:** B

**What is Clustering?**

1. **Clustering is primarily used to:**
   * A) Predict future data points
   * B) Group similar data points
   * C) Label data
   * D) Solve classification problems
   * **Answer:** B
2. **Which of the following is an application of clustering?**
   * A) Image compression
   * B) Market segmentation
   * C) Spam detection
   * D) Optical character recognition
   * **Answer:** B

**K-Means Intuition and Algorithm**

1. **K-means clustering partitions data into how many clusters?**
   * A) Predefined number of clusters (K)
   * B) Any number of clusters
   * C) One cluster
   * D) Two clusters
   * **Answer:** A
2. **What is the objective of K-means clustering?**
   * A) To maximize the distance between clusters
   * B) To minimize the distance between points within the same cluster
   * C) To classify data
   * D) To predict continuous values
   * **Answer:** B
3. **Which type of distance metric is commonly used in K-means clustering?**
   * A) Manhattan distance
   * B) Euclidean distance
   * C) Cosine similarity
   * D) Pearson correlation
   * **Answer:** B

**Optimization Objective and Initializing K-means**

1. **The optimization objective of K-means clustering is to:**
   * A) Minimize within-cluster variance
   * B) Maximize within-cluster variance
   * C) Minimize between-cluster variance
   * D) Maximize the number of clusters
   * **Answer:** A
2. **Which method can be used to initialize K-means clustering?**
   * A) Random initialization
   * B) K-means++
   * C) Lloyd’s algorithm
   * D) All of the above
   * **Answer:** D

**Choosing the Number of Clusters**

1. **How is the number of clusters (K) determined in K-means?**
   * A) It is a hyperparameter set before running the algorithm
   * B) It is determined automatically
   * C) It is the output of the algorithm
   * D) It is always set to 2
   * **Answer:** A
2. **Which method is commonly used to determine the optimal number of clusters?**
   * A) Silhouette method
   * B) Elbow method
   * C) Cross-validation
   * D) Gradient descent
   * **Answer:** B

**Hard vs Soft Clustering**

1. **In hard clustering, each data point:**
   * A) Belongs to one and only one cluster
   * B) Can belong to multiple clusters
   * C) Is not assigned to any cluster
   * D) Is ignored in the clustering process
   * **Answer:** A
2. **In soft clustering, each data point:**
   * A) Can belong to multiple clusters with a probability
   * B) Is assigned to only one cluster
   * C) Is not assigned to any cluster
   * D) Is always misclassified
   * **Answer:** A

**Machine Learning Clustering Algorithms-II:**

**Finding Unusual Events and Gaussian (Normal) Distribution**

1. **Which distribution is most commonly used in anomaly detection?**
   * A) Gaussian (Normal) distribution
   * B) Uniform distribution
   * C) Exponential distribution
   * D) Poisson distribution
   * **Answer:** A
2. **Anomalies in data typically represent:**
   * A) Rare events that deviate from the norm
   * B) Frequent and common occurrences
   * C) Average data points
   * D) Normal clusters
   * **Answer:** A

**Anomaly Detection Algorithms**

1. **Anomaly detection is useful in which of the following applications?**
   * A) Credit card fraud detection
   * B) Market segmentation
   * C) Handwriting recognition
   * D) Optical character recognition
   * **Answer:** A
2. **In anomaly detection, the assumption is that most data points:**
   * A) Belong to the normal distribution
   * B) Are anomalies
   * C) Are outliers
   * D) Belong to multiple clusters
   * **Answer:** A

**Anomaly Detection vs Supervised Learning**

1. **The key difference between anomaly detection and supervised learning is:**
   * A) Anomaly detection does not require labeled data
   * B) Anomaly detection requires labeled data
   * C) Supervised learning works only for numerical data
   * D) Supervised learning works without labels
   * **Answer:** A
2. **Anomaly detection is considered a type of:**
   * A) Unsupervised learning
   * B) Supervised learning
   * C) Reinforcement learning
   * D) Classification
   * **Answer:** A

**Principal Component Analysis (PCA)**

1. **Principal Component Analysis (PCA) is primarily used for:**
   * A) Dimensionality reduction
   * B) Classification
   * C) Regression
   * D) Clustering
   * **Answer:** A
2. **PCA works by transforming data into:**
   * A) A new set of variables called principal components
   * B) The original set of variables
   * C) Predictions
   * D) Labels
   * **Answer:** A

**Agglomerative Clustering**

1. **Agglomerative clustering is a type of:**
   * A) Hierarchical clustering
   * B) Flat clustering
   * C) K-means clustering
   * D) Density-based clustering
   * **Answer:** A
2. **In agglomerative clustering, clusters are:**
   * A) Merged in a bottom-up approach
   * B) Split in a top-down approach
   * C) Formed using centroids
   * D) Created based on density
   * **Answer:** A

**DBSCAN Clustering**

1. **What does DBSCAN stand for?**
   * A) Density-Based Spatial Clustering of Applications with Noise
   * B) Density-Based Sample Clustering of Anomalous Nodes
   * C) Decision-Based Spatial Clustering Algorithm
   * D) Dynamic Boundary Spatial Clustering Analysis
   * **Answer:** A
2. **Which of the following is true about DBSCAN?**
   * A) It can identify clusters of varying shapes
   * B) It always forms spherical clusters
   * C) It is sensitive to the number of clusters
   * D) It requires labeled data
   * **Answer:** A

**Clustering Metrics:**

**Silhouette Score**

1. **The silhouette score measures:**
   * A) How similar an object is to its own cluster compared to other clusters
   * B) The number of clusters
   * C) The variance within a cluster
   * D) The distance between centroids
   * **Answer:** A

**Davies-Bouldin Index**

1. **The Davies-Bouldin Index is used to:**
   * A) Evaluate clustering quality by comparing the ratio of within-cluster scatter to between-cluster separation
   * B) Count the number of clusters
   * C) Optimize the clustering algorithm
   * D) Measure dimensionality reduction
   * **Answer:** A

**Dunn Index**

1. **A higher Dunn Index value indicates:**
   * A) Better clustering performance
   * B) Poor clustering performance
   * C) More clusters
   * D) Fewer clusters
   * **Answer:** A

**Adjusted Rand Index (ARI)**

1. **The Adjusted Rand Index (ARI) is a measure of:**
   * A) The similarity between two data clusterings
   * B) The distance between two centroids
   * C) The number of clusters
   * D) Data transformation
   * **Answer:** A

**Normalized Mutual Information (NMI)**

1. **Normalized Mutual Information (NMI) measures the:**
   * A) Quality of clustering by considering mutual information between true labels and predicted clusters
   * B) Variance between data points
   * C) Sum of squared errors
   * D) Distance between clusters
   * **Answer:** A

**Homogeneity, Completeness, and V-measure**

1. **A clustering result is considered homogeneous if:**
   * A) All clusters contain only data points that are members of a single class
   * B) Each cluster contains multiple classes
   * C) Data points are spread equally across clusters
   * D) There is an equal number of data points in each cluster
   * **Answer:** A
2. **Completeness in clustering refers to:**
   * A) All data points of a given class are assigned to the same cluster
   * B) The number of clusters is optimal
   * C) The dataset is complete
   * D) The clustering algorithm has been fully executed
   * **Answer:** A

**Fowlkes-Mallows Index**

1. **The Fowlkes-Mallows Index is used to evaluate:**
   * A) Clustering similarity
   * B) Model accuracy
   * C) Data preprocessing
   * D) Anomaly detection
   * **Answer:** A

**Adjusted Mutual Information (AMI)**

1. **Adjusted Mutual Information (AMI) is designed to:**
   * A) Correct bias in mutual information for varying cluster sizes
   * B) Measure distance between clusters
   * C) Adjust the silhouette score
   * D) Correct the elbow method
   * **Answer:** A

UNIT-1

**Introduction to Unsupervised Learning**

1. **What is the primary goal of unsupervised learning?**
   * A) To classify data into known categories
   * B) To find patterns or structures within data without explicit labels
   * C) To predict outcomes based on input data
   * D) To reinforce learning from feedback
   * **Answer:** B
2. **Which of the following is NOT a characteristic of unsupervised learning?**
   * A) No labeled output
   * B) Focus on feature extraction
   * C) Use of target variables
   * D) Discovery of hidden patterns
   * **Answer:** C
3. **Unsupervised learning can be used for which of the following tasks?**
   * A) Classification
   * B) Clustering
   * C) Regression
   * D) All of the above
   * **Answer:** B

**What is Clustering?**

1. **Clustering is defined as the task of:**
   * A) Predicting a target variable
   * B) Grouping similar data points together
   * C) Reducing dimensionality
   * D) Collecting data from various sources
   * **Answer:** B
2. **In clustering, the quality of the grouping is often evaluated using:**
   * A) Regression metrics
   * B) Classification accuracy
   * C) Clustering metrics
   * D) None of the above
   * **Answer:** C

**K-Means Intuition**

1. **The K-means algorithm is primarily used for:**
   * A) Classification tasks
   * B) Clustering tasks
   * C) Regression tasks
   * D) Dimensionality reduction
   * **Answer:** B
2. **In K-means, "K" refers to:**
   * A) The total number of data points
   * B) The number of clusters to form
   * C) The number of iterations
   * D) The dimension of the data
   * **Answer:** B
3. **The intuition behind K-means is to:**
   * A) Minimize the distance between clusters
   * B) Maximize the distance between clusters
   * C) Minimize the distance between data points within the same cluster
   * D) Maximize the distance between data points within the same cluster
   * **Answer:** C

**K-Means Algorithm**

1. **Which of the following is the first step in the K-means algorithm?**
   * A) Assigning data points to clusters
   * B) Calculating the distance between points
   * C) Initializing cluster centroids
   * D) Updating cluster centroids
   * **Answer:** C
2. **In K-means, after assigning points to clusters, the next step is to:**
   * A) Re-initialize centroids randomly
   * B) Calculate the mean of the points in each cluster
   * C) Evaluate the quality of clustering
   * D) Visualize the clusters
   * **Answer:** B
3. **What criterion is used to assign data points to clusters in K-means?**
   * A) Random selection
   * B) Minimum distance to the centroid
   * C) Maximum similarity
   * D) Maximum distance from the centroid
   * **Answer:** B

**Optimization Objective**

1. **The objective function in K-means aims to minimize:**
   * A) Total within-cluster sum of squares
   * B) Between-cluster variance
   * C) Total number of clusters
   * D) Distance between clusters
   * **Answer:** A
2. **Which of the following best describes the within-cluster sum of squares?**
   * A) The sum of distances between data points and their respective cluster centroids
   * B) The sum of distances between cluster centroids
   * C) The total distance of all data points from the origin
   * D) The average distance of all points in the dataset
   * **Answer:** A

**Initializing K-Means**

1. **One common method for initializing cluster centroids is called:**
   * A) Random initialization
   * B) K-means++
   * C) Hierarchical initialization
   * D) Mean initialization
   * **Answer:** B
2. **Why is the initialization of centroids important in K-means?**
   * A) It affects the convergence speed
   * B) It determines the final clusters
   * C) It can lead to different clustering results
   * D) All of the above
   * **Answer:** D

**Choosing the Number of Clusters**

1. **Which method is commonly used to determine the optimal number of clusters in K-means?**
   * A) The silhouette method
   * B) The elbow method
   * C) The gap statistic
   * D) The average linkage method
   * **Answer:** B
2. **In the elbow method, the "elbow" point represents:**
   * A) The number of clusters with the lowest variance
   * B) The optimal number of clusters
   * C) The maximum number of clusters
   * D) The minimum total within-cluster sum of squares
   * **Answer:** B

**Hard versus Soft Clustering**

1. **Hard clustering assigns each data point to:**
   * A) Multiple clusters
   * B) No cluster
   * C) One specific cluster
   * D) A set of probabilities for each cluster
   * **Answer:** C
2. **Soft clustering allows for data points to:**
   * A) Belong to only one cluster
   * B) Have probabilities of belonging to multiple clusters
   * C) Be assigned to a single cluster
   * D) None of the above
   * **Answer:** B

**Using the Elbow Method to Find the Optimal Number of Clusters**

1. **The elbow method involves plotting:**
   * A) The total number of clusters versus silhouette score
   * B) The total within-cluster sum of squares versus the number of clusters
   * C) The distance between centroids versus data points
   * D) The average distance between clusters
   * **Answer:** B
2. **In the elbow method plot, a steep decrease in the total within-cluster sum of squares indicates:**
   * A) More clusters are needed
   * B) An optimal number of clusters has been reached
   * C) The model is overfitting
   * D) Clusters are too close together
   * **Answer:** B

**Additional Questions**

1. **K-means clustering is sensitive to:**
   * A) Initialization of centroids
   * B) Outliers
   * C) The choice of K
   * D) All of the above
   * **Answer:** D
2. **If K-means is run multiple times with different initial centroids, the results may vary due to:**
   * A) The algorithm being deterministic
   * B) Random initialization
   * C) Fixed cluster assignments
   * D) None of the above
   * **Answer:** B
3. **In a scenario where K=3 is used, which of the following could happen?**
   * A) Data points may be assigned to 4 clusters
   * B) All data points may be assigned to the same cluster
   * C) Some clusters may be empty
   * D) None of the above
   * **Answer:** C
4. **K-means clustering assumes that clusters are:**
   * A) Of equal size
   * B) Spherical in shape
   * C) Linearly separable
   * D) Irregular in shape
   * **Answer:** B
5. **A potential drawback of K-means is:**
   * A) It is computationally intensive
   * B) It cannot handle large datasets
   * C) It requires prior knowledge of the number of clusters
   * D) It works only on categorical data
   * **Answer:** C
6. **Which of the following is NOT a common application of K-means clustering?**
   * A) Image segmentation
   * B) Document clustering
   * C) Predictive modeling
   * D) Market segmentation
   * **Answer:** C
7. **The choice of K in K-means can significantly affect:**
   * A) The running time of the algorithm
   * B) The quality of clustering
   * C) The initialization method
   * D) The scaling of data
   * **Answer:** B
8. **What happens to the within-cluster sum of squares as the number of clusters increases?**
   * A) It increases
   * B) It decreases
   * C) It remains constant
   * D) It oscillates
   * **Answer:** B
9. **In soft clustering, what does a high probability value indicate for a data point?**
   * A) Strong association with a cluster
   * B) Weak association with a cluster
   * C) No association with any cluster
   * D) Random assignment
   * **Answer:** A

**Final Set of Questions**

1. **Which of the following clustering algorithms is often compared with K-means?**
   * A) Decision Trees
   * B) DBSCAN
   * C) Linear Regression
   * D) Support Vector Machines
   * **Answer:** B
2. **When using the elbow method, what does it mean if there is no clear "elbow"?**
   * A) The dataset is too small
   * B) The data may not have inherent clusters
   * C) The clustering is not successful
   * D) None of the above
   * **Answer:** B
3. **In K-means, the distance metric commonly used to determine the proximity of points to centroids is:**
   * A) Manhattan distance
   * B) Cosine similarity
   * C) Euclidean distance
   * D) Jaccard index
   * **Answer:** C
4. **Which technique can be used to scale data before applying K-means clustering?**
   * A) Normalization
   * B) Standardization
   * C) Both A and B
   * D) Data augmentation
   * **Answer:** C
5. **K-means is most effective when the clusters are:**
   * A) Highly overlapped
   * B) Very distinct and well-separated
   * C) Irregularly shaped
   * D) Non-spherical
   * **Answer:** B
6. **A common technique to improve the results of K-means is to use:**
   * A) Dimensionality reduction
   * B) Random feature selection
   * C) Data shuffling
   * D) Data augmentation
   * **Answer:** A
7. **K-means can handle both categorical and numerical data. (True/False)**
   * A) True
   * B) False
   * **Answer:** B
8. **What is a common use case for K-means clustering in marketing?**
   * A) Predicting sales forecasts
   * B) Identifying customer segments
   * C) Analyzing financial statements
   * D) Creating product recommendations
   * **Answer:** B
9. **The choice of distance metric can significantly impact the clustering results. (True/False)**
   * A) True
   * B) False
   * **Answer:** A
10. **Which of the following is NOT a limitation of K-means?**
    * A) Sensitivity to outliers
    * B) Assumes spherical clusters
    * C) Works with both continuous and categorical variables
    * D) Requires specifying the number of clusters in advance
    * **Answer:** C
11. **Which clustering method is better suited for detecting noise in data?**
    * A) K-means
    * B) Hierarchical clustering
    * C) DBSCAN
    * D) Gaussian Mixture Models
    * **Answer:** C
12. **The K-means algorithm will always converge to the global minimum. (True/False)**
    * A) True
    * B) False
    * **Answer:** B
13. **Which of the following is a way to visualize clustering results?**
    * A) Box plots
    * B) Heatmaps
    * C) Scatter plots
    * D) All of the above
    * **Answer:** C
14. **The silhouette score is used to evaluate the quality of clustering by measuring:**
    * A) Cluster density
    * B) Cohesion and separation
    * C) Number of clusters
    * D) Distance to centroids
    * **Answer:** B
15. **In K-means, what does a high silhouette score indicate?**
    * A) Poor clustering
    * B) Good clustering
    * C) Ambiguous clustering
    * D) Insufficient data
    * **Answer:** B
16. **Which of the following methods can be used to validate clustering results?**
    * A) Cross-validation
    * B) Bootstrap sampling
    * C) Internal validation metrics
    * D) All of the above
    * **Answer:** C
17. **In the context of clustering, what does PCA stand for?**
    * A) Principal Component Analysis
    * B) Principal Cluster Algorithm
    * C) Pattern Clustering Algorithm
    * D) None of the above
    * **Answer:** A
18. **When performing K-means, what is the role of centroids?**
    * A) They represent the maximum data points in a cluster
    * B) They represent the mean position of all the points in a cluster
    * C) They are fixed points that do not change
    * D) They are used to calculate the silhouette score
    * **Answer:** B
19. **What is one of the main advantages of K-means clustering?**
    * A) It is easy to interpret
    * B) It works with non-linear relationships
    * C) It requires no assumptions about the data distribution
    * D) It handles mixed types of data effectively
    * **Answer:** A
20. **Which of the following is a potential application of clustering in healthcare?**
    * A) Predicting patient outcomes
    * B) Grouping similar patients for treatment recommendations
    * C) Identifying new diseases
    * D) None of the above
    * **Answer:** B

UNIT-2

**Finding Unusual Events**

1. **What is the primary goal of anomaly detection?**
   * A) To classify data points into known categories
   * B) To identify outliers or unusual events in data
   * C) To predict future data points
   * D) To visualize data distributions
   * **Answer:** B
2. **Which of the following can be considered an anomaly?**
   * A) A data point that significantly deviates from the expected pattern
   * B) A typical data point within a cluster
   * C) A cluster containing many data points
   * D) None of the above
   * **Answer:** A

**Gaussian (Normal) Distribution**

1. **The Gaussian distribution is characterized by which of the following?**
   * A) Skewness and kurtosis
   * B) Mean, median, and mode being equal
   * C) Bell-shaped curve
   * D) All of the above
   * **Answer:** D
2. **In anomaly detection, data points that fall outside of the Gaussian distribution are typically classified as:**
   * A) Normal
   * B) Outliers
   * C) Clustered
   * D) Weighted
   * **Answer:** B
3. **Which property of the Gaussian distribution is often used in anomaly detection?**
   * A) Mean and variance
   * B) Skewness
   * C) Kurtosis
   * D) Range
   * **Answer:** A

**Anomaly Detection Algorithm**

1. **Which of the following algorithms is commonly used for anomaly detection?**
   * A) K-means
   * B) Decision Trees
   * C) Isolation Forest
   * D) Linear Regression
   * **Answer:** C
2. **The main concept behind the Isolation Forest algorithm is:**
   * A) Clustering data points
   * B) Separating anomalies through random partitioning
   * C) Predicting outcomes based on input features
   * D) Linear classification of data
   * **Answer:** B
3. **One common approach to detect anomalies is to use:**
   * A) Distance metrics
   * B) Probabilistic models
   * C) Density estimation
   * D) All of the above
   * **Answer:** D

**Anomaly Detection vs. Supervised Learning**

1. **Which of the following is a key difference between anomaly detection and supervised learning?**
   * A) Anomaly detection uses labeled data
   * B) Supervised learning focuses on classification tasks
   * C) Anomaly detection requires a large dataset
   * D) None of the above
   * **Answer:** B
2. **In supervised learning, the model is trained on:**
   * A) Unlabeled data
   * B) Labeled data
   * C) Both labeled and unlabeled data
   * D) Data with missing values
   * **Answer:** B
3. **Which use case is more suited for anomaly detection?**
   * A) Predicting sales for a retail store
   * B) Fraud detection in financial transactions
   * C) Image classification
   * D) Weather forecasting
   * **Answer:** B

**Choosing What Features to Use (PCA)**

1. **What does PCA stand for?**
   * A) Principal Component Analysis
   * B) Predictive Cluster Algorithm
   * C) Probability Classification Analysis
   * D) None of the above
   * **Answer:** A
2. **PCA is primarily used for:**
   * A) Classification
   * B) Regression
   * C) Dimensionality reduction
   * D) Clustering
   * **Answer:** C
3. **Which of the following is a benefit of using PCA?**
   * A) Increases the dimensionality of data
   * B) Reduces computational complexity
   * C) Improves the interpretability of results
   * D) Both B and C
   * **Answer:** D
4. **When applying PCA, the new features are called:**
   * A) Data points
   * B) Eigenvalues
   * C) Principal components
   * D) Clusters
   * **Answer:** C

**Organizing Clusters as a Hierarchical Tree**

1. **Hierarchical clustering results in a:**
   * A) Linear model
   * B) Tree-like structure (dendrogram)
   * C) Grid of clusters
   * D) Set of centroids
   * **Answer:** B
2. **Which of the following methods is NOT a type of hierarchical clustering?**
   * A) Divisive clustering
   * B) Agglomerative clustering
   * C) K-means clustering
   * D) Both A and B
   * **Answer:** C
3. **The main advantage of hierarchical clustering is:**
   * A) It requires fewer data points
   * B) It provides a clear representation of data structure
   * C) It guarantees optimal clusters
   * D) It operates in linear time
   * **Answer:** B

**Agglomerative Clustering**

1. **In agglomerative clustering, clusters are merged based on:**
   * A) Distance between data points
   * B) Mean values of clusters
   * C) Hierarchical order
   * D) None of the above
   * **Answer:** A
2. **The first step in agglomerative clustering involves:**
   * A) Merging all data points into a single cluster
   * B) Treating each data point as an individual cluster
   * C) Calculating the distances between clusters
   * D) None of the above
   * **Answer:** B
3. **Agglomerative clustering can be sensitive to:**
   * A) Noise in the data
   * B) Outliers
   * C) The choice of distance metric
   * D) All of the above
   * **Answer:** D

**DBSCAN Clustering**

1. **What does DBSCAN stand for?**
   * A) Density-Based Spatial Clustering of Applications with Noise
   * B) Dynamic Based Clustering Algorithm
   * C) Data-Based Segmentation and Clustering Analysis
   * D) Density-Based Segmentation and Clustering Algorithm
   * **Answer:** A
2. **DBSCAN identifies clusters based on:**
   * A) Spherical shape of data
   * B) Density of data points
   * C) Distance from centroids
   * D) Hierarchical relationships
   * **Answer:** B
3. **In DBSCAN, what is an "epsilon" (ε) parameter used for?**
   * A) To define the minimum number of points in a cluster
   * B) To specify the maximum distance for neighboring points
   * C) To determine the size of clusters
   * D) None of the above
   * **Answer:** B
4. **What type of points are considered "core points" in DBSCAN?**
   * A) Points with no neighboring points
   * B) Points with at least a minimum number of neighboring points within epsilon
   * C) Points that are farthest from the centroid
   * D) Randomly selected points
   * **Answer:** B
5. **Which of the following is a limitation of DBSCAN?**
   * A) It can identify arbitrary-shaped clusters
   * B) It struggles with varying density clusters
   * C) It requires a large amount of labeled data
   * D) It works well with outliers
   * **Answer:** B

**Additional Questions**

1. **Which of the following techniques is NOT commonly used in anomaly detection?**
   * A) Time series analysis
   * B) Supervised learning
   * C) Clustering methods
   * D) Rule-based systems
   * **Answer:** B
2. **In hierarchical clustering, the "cut" in the dendrogram determines:**
   * A) The number of clusters to form
   * B) The distance between clusters
   * C) The dimensions of the data
   * D) The merging order of clusters
   * **Answer:** A
3. **Which distance metric is commonly used in agglomerative clustering?**
   * A) Manhattan distance
   * B) Euclidean distance
   * C) Cosine similarity
   * D) Both A and B
   * **Answer:** D
4. **Which clustering algorithm is particularly effective for large datasets?**
   * A) K-means
   * B) Hierarchical clustering
   * C) DBSCAN
   * D) All of the above
   * **Answer:** A
5. **Which of the following best describes the concept of "outlier" in data?**
   * A) A data point that belongs to multiple clusters
   * B) A data point that is significantly different from the rest of the data
   * C) A data point that is part of a dense cluster
   * D) A data point that is uniformly distributed
   * **Answer:** B
6. **In PCA, the principal components are derived from:**
   * A) Eigenvalues and eigenvectors of the covariance matrix
   * B) Randomly generated features
   * C) The original dataset directly
   * D) Labeled data only
   * **Answer:** A
7. **When using DBSCAN, what is an "outlier"?**
   * A) A core point with too few neighbors
   * B) A point that is not part of any cluster
   * C) A point within the epsilon neighborhood
   * D) A point that belongs to a small cluster
   * **Answer:** B
8. **What is the advantage of using density-based clustering algorithms like DBSCAN?**
   * A) They are less sensitive to noise
   * B) They require prior knowledge of the number of clusters
   * C) They work only with spherical clusters
   * D) They are easy to implement
   * **Answer:** A
9. **Which approach is suitable for detecting anomalies in time-series data?**
   * A) K-means clustering
   * B) Statistical process control
   * C) PCA
   * D) DBSCAN
   * **Answer:** B
10. **What is the main objective of hierarchical clustering?**
    * A) To create a fixed number of clusters
    * B) To form a hierarchy of clusters that reflects the data's structure
    * C) To visualize the distribution of data
    * D) To categorize data based on labels
    * **Answer:** B
11. **Which of the following is a method for evaluating clustering results?**
    * A) Adjusted Rand Index
    * B) Mean Squared Error
    * C) Cross-validation
    * D) A/B Testing
    * **Answer:** A
12. **What is the purpose of scaling data before applying PCA?**
    * A) To reduce the number of features
    * B) To normalize the data and avoid bias towards larger features
    * C) To eliminate outliers
    * D) To convert categorical data to numerical
    * **Answer:** B
13. **In agglomerative clustering, which linkage criteria can be used?**
    * A) Single linkage
    * B) Complete linkage
    * C) Average linkage
    * D) All of the above
    * **Answer:** D
14. **Which feature of DBSCAN allows it to find clusters of arbitrary shapes?**
    * A) It does not require prior knowledge of the number of clusters
    * B) It uses a fixed distance metric
    * C) It focuses only on core points
    * D) It requires normalization of the data
    * **Answer:** A
15. **Which of the following statements about PCA is false?**
    * A) PCA reduces the dimensionality of data.
    * B) PCA retains most of the variance in the dataset.
    * C) PCA can only be applied to labeled data.
    * D) PCA identifies the directions of maximum variance.
    * **Answer:** C
16. **Which type of clustering approach is NOT a part of hierarchical clustering?**
    * A) Divisive
    * B) Agglomerative
    * C) K-medoids
    * D) Both A and B
    * **Answer:** C
17. **Which of the following can be used to preprocess data before clustering?**
    * A) Normalization
    * B) Standardization
    * C) Feature selection
    * D) All of the above
    * **Answer:** D
18. **DBSCAN can be less effective when clusters have:**
    * A) Very distinct densities
    * B) Arbitrary shapes
    * C) A large number of outliers
    * D) None of the above
    * **Answer:** C
19. **What do we call a point that has no neighboring points within the epsilon distance in DBSCAN?**
    * A) Core point
    * B) Border point
    * C) Noise point
    * D) Cluster point
    * **Answer:** C
20. **Which of the following is a main advantage of agglomerative clustering?**
    * A) It can only form a fixed number of clusters
    * B) It is less affected by outliers
    * C) It provides a clear visualization of the data structure
    * D) It requires less computational power
    * **Answer:** C
21. **In which scenario is PCA most beneficial?**
    * A) When all features are of equal importance
    * B) When the dataset has many features that contribute little variance
    * C) When dealing with time series data
    * D) For classification tasks with labeled data
    * **Answer:** B
22. **In anomaly detection, what type of features are typically preferred?**
    * A) Highly correlated features
    * B) Irrelevant features
    * C) Features that provide clear distinctions between normal and anomalous data
    * D) Features with low variance
    * **Answer:** C
23. **Which of the following describes the "linkage" in hierarchical clustering?**
    * A) The method used to merge clusters
    * B) The distance between clusters
    * C) The final number of clusters
    * D) The initial data points
    * **Answer:** A
24. **Which of the following is true about density-based clustering?**
    * A) It assumes clusters are spherical
    * B) It can find clusters of arbitrary shapes
    * C) It always requires prior knowledge of cluster numbers
    * D) It is not affected by outliers
    * **Answer:** B

UNIT-3

**Silhouette Score**

1. **What does the Silhouette Score measure?**
   * A) The distance between clusters
   * B) The similarity of a data point to its own cluster compared to other clusters
   * C) The overall performance of the clustering algorithm
   * D) The computational efficiency of clustering
   * **Answer:** B
2. **A Silhouette Score close to +1 indicates:**
   * A) Poor clustering
   * B) Average clustering
   * C) Well-clustered data points
   * D) Overlapping clusters
   * **Answer:** C
3. **If the Silhouette Score is negative, it suggests that:**
   * A) The data points are clustered correctly
   * B) Data points may have been assigned to the wrong clusters
   * C) The clustering algorithm is highly effective
   * D) All clusters are of equal size
   * **Answer:** B

**Davies-Bouldin Index**

1. **What does the Davies-Bouldin Index assess?**
   * A) The overall size of clusters
   * B) The similarity between clusters and the dissimilarity within clusters
   * C) The computational efficiency of clustering algorithms
   * D) The number of clusters
   * **Answer:** B
2. **A lower Davies-Bouldin Index indicates:**
   * A) Poor clustering quality
   * B) Better clustering quality
   * C) More clusters than necessary
   * D) Clusters that overlap significantly
   * **Answer:** B
3. **The Davies-Bouldin Index is calculated based on:**
   * A) Intra-cluster distances and inter-cluster distances
   * B) Only inter-cluster distances
   * C) Only intra-cluster distances
   * D) The number of features in the dataset
   * **Answer:** A

**Dunn Index**

1. **What does the Dunn Index measure?**
   * A) The variance within clusters
   * B) The ratio of the minimum inter-cluster distance to the maximum intra-cluster distance
   * C) The size of clusters
   * D) The computational complexity of clustering algorithms
   * **Answer:** B
2. **A higher Dunn Index indicates:**
   * A) Poorly defined clusters
   * B) Better clustering quality
   * C) Clusters that are too close together
   * D) An inadequate number of clusters
   * **Answer:** B
3. **The Dunn Index can be particularly useful when:**
   * A) Evaluating clusters with a high degree of overlap
   * B) Comparing clustering solutions with different numbers of clusters
   * C) Analyzing datasets with a single cluster
   * D) Measuring only the density of clusters
   * **Answer:** B

**Adjusted Rand Index (ARI)**

1. **The Adjusted Rand Index (ARI) compares:**
   * A) The number of features in the dataset
   * B) The predicted clusters with the true labels
   * C) The size of the clusters
   * D) The efficiency of clustering algorithms
   * **Answer:** B
2. **What value does the ARI range between?**
   * A) 0 and 1
   * B) -1 and 1
   * C) -1 and 0
   * D) 0 and infinity
   * **Answer:** B
3. **An ARI score of 1 indicates:**
   * A) Random clustering
   * B) Perfect agreement between predicted and true clusters
   * C) Poor clustering performance
   * D) Average clustering quality
   * **Answer:** B

**Normalized Mutual Information (NMI)**

1. **What does Normalized Mutual Information (NMI) measure?**
   * A) The distance between clusters
   * B) The mutual information between predicted and true labels, normalized by maximum possible mutual information
   * C) The size of the clusters
   * D) The computational efficiency of clustering algorithms
   * **Answer:** B
2. **NMI values range from:**
   * A) -1 to 1
   * B) 0 to 1
   * C) -1 to 0
   * D) 0 to infinity
   * **Answer:** B
3. **A higher NMI indicates:**
   * A) Poor agreement between predicted and true labels
   * B) Good agreement between predicted and true labels
   * C) Random clustering
   * D) Overlapping clusters
   * **Answer:** B

**Homogeneity, Completeness, and V-measure**

1. **Homogeneity in clustering means that:**
   * A) All clusters have the same number of points
   * B) All points in a cluster belong to the same class
   * C) Clusters are of equal size
   * D) Clusters are distinct from each other
   * **Answer:** B
2. **Completeness in clustering means that:**
   * A) All points of a class belong to the same cluster
   * B) Clusters are well-separated
   * C) All clusters are of equal size
   * D) Points in a cluster belong to different classes
   * **Answer:** A
3. **The V-measure is a trade-off between:**
   * A) Homogeneity and completeness
   * B) Silhouette score and Davies-Bouldin Index
   * C) Dunn Index and ARI
   * D) NMI and Fowlkes-Mallows Index
   * **Answer:** A

**Fowlkes-Mallows Index**

1. **The Fowlkes-Mallows Index measures:**
   * A) The size of clusters
   * B) The similarity between predicted and true clusterings
   * C) The computational efficiency of clustering algorithms
   * D) The variance within clusters
   * **Answer:** B
2. **A Fowlkes-Mallows Index value of 1 indicates:**
   * A) Poor clustering
   * B) Perfect similarity between the predicted and true clusterings
   * C) Random clustering
   * D) Clusters of equal size
   * **Answer:** B
3. **Which of the following metrics is NOT used to evaluate clustering quality?**
   * A) Silhouette Score
   * B) Adjusted Rand Index
   * C) Mean Squared Error
   * D) Dunn Index
   * **Answer:** C

**Adjusted Mutual Information (AMI)**

1. **What does Adjusted Mutual Information (AMI) account for?**
   * A) Random chance agreement between predicted and true labels
   * B) The distance between clusters
   * C) The size of the clusters
   * D) The computational complexity of clustering algorithms
   * **Answer:** A
2. **The AMI is a normalized version of which metric?**
   * A) Silhouette Score
   * B) Mutual Information
   * C) Dunn Index
   * D) Davies-Bouldin Index
   * **Answer:** B
3. **An AMI value of 0 indicates:**
   * A) Perfect agreement between clustering results
   * B) Random clustering with no correlation to true labels
   * C) Overlapping clusters
   * D) Poor clustering quality
   * **Answer:** B

**Additional Questions**

1. **Which metric is best for evaluating the internal consistency of clusters?**
   * A) Silhouette Score
   * B) Adjusted Rand Index
   * C) Dunn Index
   * D) All of the above
   * **Answer:** A
2. **Which of the following is a common characteristic of clustering evaluation metrics?**
   * A) They all require true labels for evaluation
   * B) They help to identify optimal clustering configurations
   * C) They provide a measure of clustering speed
   * D) They all assume a Gaussian distribution of data
   * **Answer:** B
3. **Which metric is not sensitive to the number of clusters formed?**
   * A) Silhouette Score
   * B) Davies-Bouldin Index
   * C) Dunn Index
   * D) Fowlkes-Mallows Index
   * **Answer:** C
4. **What is a disadvantage of the Silhouette Score?**
   * A) It is not scalable to large datasets
   * B) It does not account for cluster shape
   * C) It can only evaluate two clusters
   * D) Both A and B
   * **Answer:** D
5. **Which of the following metrics can provide insights into cluster separation?**
   * A) Homogeneity
   * B) Completeness
   * C) V-measure
   * D) All of the above
   * **Answer:** D
6. **The main advantage of using NMI is:**
   * A) It is computationally inexpensive
   * B) It is scale-invariant
   * C) It does not require ground truth labels
   * D) It is easy to interpret
   * **Answer:** B
7. **If two clustering solutions have the same AMI score, it indicates that:**
   * A) They perform equally in terms of clustering quality
   * B) They are significantly different
   * C) One is better than the other
   * D) They have random similarities
   * **Answer:** A
8. **Which clustering evaluation metric can indicate the presence of outliers?**
   * A) Silhouette Score
   * B) Davies-Bouldin Index
   * C) Dunn Index
   * D) All of the above
   * **Answer:** D
9. **In terms of clustering metrics, what does "completeness" signify?**
   * A) All data points in a cluster belong to one class
   * B) All points of a class are assigned to the same cluster
   * C) There is a high level of overlap among clusters
   * D) The clusters have equal sizes
   * **Answer:** B
10. **A clustering metric that is particularly useful for imbalanced classes is:**
    * A) Silhouette Score
    * B) Fowlkes-Mallows Index
    * C) Adjusted Rand Index
    * D) Homogeneity
    * **Answer:** C
11. **Which of the following metrics is NOT based on the concept of distance?**
    * A) Dunn Index
    * B) Davies-Bouldin Index
    * C) Fowlkes-Mallows Index
    * D) Silhouette Score
    * **Answer:** C
12. **Which metric would be most useful when the number of clusters is not known?**
    * A) Silhouette Score
    * B) Davies-Bouldin Index
    * C) Adjusted Rand Index
    * D) Dunn Index
    * **Answer:** A
13. **What is the effect of increasing the number of clusters on the Silhouette Score?**
    * A) It always increases the score
    * B) It always decreases the score
    * C) It may increase or decrease the score depending on the dataset
    * D) It has no effect on the score
    * **Answer:** C
14. **Which of the following metrics evaluates clustering without needing ground truth labels?**
    * A) Silhouette Score
    * B) Adjusted Rand Index
    * C) Normalized Mutual Information
    * D) Fowlkes-Mallows Index
    * **Answer:** A
15. **Which clustering evaluation metric is more robust to random noise?**
    * A) Dunn Index
    * B) Fowlkes-Mallows Index
    * C) Silhouette Score
    * D) Adjusted Rand Index
    * **Answer:** A
16. **The Davies-Bouldin Index works on the principle of:**
    * A) Overlap between clusters
    * B) Ratio of intra-cluster distance to inter-cluster distance
    * C) Count of clusters
    * D) Proximity of points within clusters
    * **Answer:** B
17. **A clustering method that combines both homogeneity and completeness is:**
    * A) V-measure
    * B) Dunn Index
    * C) Adjusted Rand Index
    * D) Fowlkes-Mallows Index
    * **Answer:** A
18. **What is a common drawback of using ARI?**
    * A) It is computationally expensive
    * B) It requires a large amount of labeled data
    * C) It may not capture the nuances of clustering performance
    * D) It is not intuitive to interpret
    * **Answer:** C
19. **Which of the following metrics emphasizes how clusters can be derived from a classification perspective?**
    * A) Homogeneity
    * B) Silhouette Score
    * C) Dunn Index
    * D) Davies-Bouldin Index
    * **Answer:** A
20. **For what type of dataset is the Dunn Index most effective?**
    * A) Datasets with clear and distinct clusters
    * B) Datasets with overlapping clusters
    * C) High-dimensional datasets
    * D) Datasets with noise
    * **Answer:** A
21. **Which of the following metrics would best evaluate clustering performance when clusters are expected to have irregular shapes?**
    * A) Dunn Index
    * B) Silhouette Score
    * C) Davies-Bouldin Index
    * D) Fowlkes-Mallows Index
    * **Answer:** B
22. **What is the primary limitation of using the Silhouette Score?**
    * A) It cannot be used with large datasets
    * B) It is sensitive to the choice of distance metric
    * C) It does not provide a clear threshold for clustering
    * D) It does not account for cluster shape
    * **Answer:** B
23. **Which metric is specifically designed to measure the extent of agreement between two different clustering assignments?**
    * A) Dunn Index
    * B) Adjusted Rand Index
    * C) Silhouette Score
    * D) Normalized Mutual Information
    * **Answer:** B
24. **Which metric can indicate both clustering quality and class membership?**
    * A) Normalized Mutual Information
    * B) Dunn Index
    * C) Silhouette Score
    * D) Davies-Bouldin Index
    * **Answer:** A
25. **In clustering evaluation, which metric is particularly sensitive to the number of clusters?**
    * A) Silhouette Score
    * B) Dunn Index
    * C) Adjusted Rand Index
    * D) Fowlkes-Mallows Index
    * **Answer:** A
26. **Which evaluation metric is primarily based on the concept of mutual information?**
    * A) Normalized Mutual Information
    * B) Adjusted Rand Index
    * C) Fowlkes-Mallows Index
    * D) Both A and B
    * **Answer:** D