Naive Bayes Viva Questions and Answers

1. What is the K-Means algorithm?

K-Means is an unsupervised machine learning algorithm used for clustering data into groups (clusters) based on feature similarity. It minimizes the variance within each cluster.

2. How does the K-Means algorithm work?

1. Initialize K cluster centroids randomly.

2. Assign each data point to the nearest centroid.

3. Recalculate centroids as the mean of all points in each cluster.

4. Repeat steps 2 and 3 until centroids no longer change or a maximum number of iterations is reached.

3. What is the objective function in K-Means?

The objective is to minimize the sum of squared distances between data points and their respective cluster centroids, also called the Within-Cluster Sum of Squares (WCSS).

4. How do you choose the value of K?

The value of K can be chosen using:

1. Elbow Method: Plot WCSS for different K values and select the point where the decrease in WCSS slows down.

2. Silhouette Score: Measures how similar a point is to its cluster compared to other clusters.

5. What are the limitations of K-Means?

1. Requires predefining the number of clusters (K).

2. Sensitive to the initial placement of centroids.

3. Struggles with non-spherical or overlapping clusters.

4. Sensitive to outliers.

6. What are the advantages of K-Means?

1. Simple and easy to implement.

2. Scales well with large datasets.

3. Works well with well-separated clusters.

7. What is the difference between hard clustering and soft clustering?

In hard clustering (like K-Means), each data point belongs to exactly one cluster. In soft clustering (like Gaussian Mixture Models), a data point can belong to multiple clusters with probabilities.

8. What is the role of centroids in K-Means?

Centroids represent the center of a cluster. They are recalculated as the mean of all data points assigned to the cluster during each iteration.

9. What are some common distance metrics used in K-Means?

The most common metric is Euclidean distance. Other metrics like Manhattan or Cosine distance can also be used depending on the data.

10. How does K-Means handle categorical data?

K-Means is designed for numerical data. For categorical data, variations like K-Modes or K-Prototypes can be used.

Naive Bayes Viva Questions and Answers

11. How do you evaluate the performance of K-Means?

Common evaluation metrics include:

1. Elbow Method for optimal K.

2. Silhouette Score for cluster quality.

3. Davies-Bouldin Index for compactness and separation.

12. What are some applications of K-Means?

1. Customer segmentation in marketing.

2. Image compression by reducing colors.

3. Document clustering in text analysis.

13. What happens if K-Means converges to a local minimum?

K-Means can converge to a local minimum depending on the initial centroids. Running the algorithm multiple times with different initializations can help find better solutions.

14. What is the difference between K-Means and Hierarchical Clustering?

K-Means is fast and works well for large datasets but requires a predefined K. Hierarchical clustering is more interpretable but computationally expensive for large datasets.

15. How does K-Means handle outliers?

K-Means is sensitive to outliers as they can significantly affect the position of centroids. Preprocessing steps like removing outliers or using robust clustering methods can help.

16. Can K-Means be used for high-dimensional data?

Yes, but performance may degrade due to the 'curse of dimensionality.' Dimensionality reduction techniques like PCA can improve performance.