1. What is your definition of clustering? What are a few clustering algorithms you might think of?

2. What are some of the most popular clustering algorithm applications?

3. When using K-Means, describe two strategies for selecting the appropriate number of clusters.

4. What is mark propagation and how does it work? Why would you do it, and how would you do it?

5. Provide two examples of clustering algorithms that can handle large datasets. And two that look for high-density areas?

6. Can you think of a scenario in which constructive learning will be advantageous? How can you go about putting it into action?

7. How do you tell the difference between anomaly and novelty detection?

8. What is a Gaussian mixture, and how does it work? What are some of the things you can do about it?

9. When using a Gaussian mixture model, can you name two techniques for determining the correct number of clusters?

Answer:

1. Clustering is a type of unsupervised machine learning where the goal is to group similar data points together in order to discover underlying patterns or structures in the data. Some popular clustering algorithms include K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models.
2. Clustering algorithms are used in a variety of applications such as image segmentation, customer segmentation, anomaly detection, and document clustering, among others.
3. Two common strategies for selecting the appropriate number of clusters in K-Means are the "elbow method" and the "silhouette score". The elbow method involves plotting the within-cluster sum of squared distances (WCSS) against the number of clusters and selecting the number of clusters where the curve begins to level off, resembling an elbow. The silhouette score measures the distance between a data point and its assigned cluster compared to other clusters, and the goal is to maximize the average silhouette score across all data points.
4. Mark propagation is a type of clustering algorithm that involves assigning a "label" to each data point and then propagating those labels to neighboring data points based on a set of rules or criteria. The goal is to group together data points that share similar labels. Mark propagation can be useful when the number of clusters is unknown or when the data does not conform to traditional clustering assumptions, such as being normally distributed or having spherical clusters.
5. Two examples of clustering algorithms that can handle large datasets are K-Means and DBSCAN, which use different methods to efficiently cluster large amounts of data. Two clustering algorithms that look for high-density areas are Mean Shift and OPTICS, which use density-based clustering to identify regions of high density and group data points within those regions.
6. Constructive learning is a type of machine learning where the model builds upon previously learned concepts and adapts to new information over time. A scenario where constructive learning may be advantageous is in online learning, where data is constantly streaming and the model needs to adapt in real-time. To put constructive learning into action, the model needs to be designed to handle incremental updates and incorporate new data while maintaining previous knowledge.
7. Anomaly detection involves identifying data points that deviate significantly from the majority of the data, while novelty detection involves identifying data points that are significantly different from previously seen data. The main difference between the two is that anomaly detection is concerned with identifying rare or unusual events within the existing data, while novelty detection is focused on detecting previously unseen patterns or data.
8. A Gaussian mixture is a type of probabilistic model that involves assuming that the data is generated from a mixture of several Gaussian distributions, each with its own mean and variance. The parameters of the model are learned using an iterative optimization algorithm, such as Expectation-Maximization (EM), which estimates the distribution parameters and assigns data points to the most likely component distribution. Some things you can do with Gaussian mixtures include estimating the number of clusters in the data, generating new data samples, and identifying data points that are outliers or anomalies.
9. Two techniques for determining the correct number of clusters in a Gaussian mixture model are the "BIC score" and the "AIC score". The Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) are statistical measures that penalize models with more parameters and help determine the optimal number of components in the Gaussian mixture model.