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

>>>>Clustering is an unsupervised machine learning technique used to group similar data points together based on their inherent patterns or similarities.

Clustering algorithms include K-Means, Hierarchical Clustering, DBSCAN.

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

>>>Customer Segmentation: Grouping customers with similar behaviors for targeted marketing.

Image Segmentation: Partitioning an image into meaningful regions.

Anomaly Detection: Identifying unusual patterns in data.

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

>>>>Elbow Method: Plot the sum of squared distances (inertia) against the number of clusters and look for an "elbow" point where the rate of decrease slows down.

Silhouette Score: Calculate the average silhouette score for different cluster numbers; choose the number with the highest score.

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

>>>>Mark Propagation is a semi-supervised learning technique used for data labeling.

It works by propagating labels from a few labeled instances to unlabeled ones based on similarity.

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

>>>Mini-Batch K-Means: An optimized version of K-Means that uses random subsets of data (mini-batches) for clustering large datasets.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise): Efficiently identifies dense regions in large datasets.

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

>>>>In a scenario with limited labeled data, constructive learning can be advantageous.

You can start with a small labeled dataset, use it to train an initial model, and then actively query the most informative data points for labeling.

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

>>>>Anomaly Detection: Identifies data points that are significantly different from the norm within a dataset, often based on historical data.

Novelty Detection: Detects data points that deviate from the training distribution, potentially indicating new, unseen patterns or outliers.

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

>>>>>GMM is a probabilistic model representing data as a mixture of Gaussian distributions.

It works by estimating the parameters (mean, covariance, and weight) of Gaussian components.

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

>>>>>BIC (Bayesian Information Criterion): A model selection criterion that balances model fit and complexity; choose the model with the lowest BIC.

AIC (Akaike Information Criterion): Similar to BIC but with a different penalty for model complexity.