ML Assignment -4

1. What is clustering in machine learning?
Clustering is the process of grouping similar data points into clusters based on certain characteristics.

2. Explain the difference between supervised and unsupervised clustering.
Supervised clustering uses labeled data, while unsupervised clustering works with unlabeled data.

3. What are the key applications of clustering algorithms?
Applications include customer segmentation, image segmentation, anomaly detection, and pattern recognition.

4. Describe the K-means clustering algorithm.
K-means partitions data into K clusters by minimizing within-cluster variance.

5. What are the main advantages and disadvantages of K-means clustering?
Advantages: Simple, fast. Disadvantages: Sensitive to initial centroids, requires specifying K.

6. How does hierarchical clustering work?
Hierarchical clustering creates a tree of clusters, merging or splitting them based on distance.

7. What are the different linkage criteria used in hierarchical clustering?
Linkage criteria include single, complete, average, and Ward linkage.

8. Explain the concept of DBSCAN clustering.
DBSCAN clusters based on density, grouping together dense areas and marking sparse areas as outliers.

9. What are the parameters involved in DBSCAN clustering?
Key parameters: Epsilon (maximum distance), MinPts (minimum points to form a cluster).

10. Describe the process of evaluating clustering algorithms.
Evaluation is done using metrics like silhouette score, Davies-Bouldin index, or external validation measures.

11. What is the silhouette score, and how is it calculated?
Silhouette score measures the quality of clustering by evaluating both cohesion and separation.

12. Discuss the challenges of clustering high-dimensional data.
Challenges include curse of dimensionality, distance metric issues, and high computational cost.

13. Explain the concept of density-based clustering.
Density-based clustering groups points in high-density regions and marks low-density points as outliers.

14. How does Gaussian Mixture Model (GMM) clustering differ from K-means?
GMM models data as a mixture of Gaussian distributions, allowing for soft clustering, unlike K-means.

15. What are the limitations of traditional clustering algorithms?
Limitations include sensitivity to initial conditions, difficulty with non-convex shapes, and scalability issues.

16. Discuss the applications of spectral clustering.
Spectral clustering is used in image segmentation, graph clustering, and dimensionality reduction.

17. Explain the concept of affinity propagation.
Affinity propagation identifies clusters based on message passing between data points, without requiring K.

18. How do you handle categorical variables in clustering?
Categorical variables can be handled using methods like one-hot encoding or similarity measures (e.g., Jaccard index).

19. Describe the elbow method for determining the optimal number of clusters.
The elbow method plots the cost function against the number of clusters, selecting the point where the curve bends.

20. What are some emerging trends in clustering research?
Emerging trends include deep clustering, clustering with noisy data, and scalable algorithms for big data.

21. What is anomaly detection, and why is it important?
Anomaly detection identifies unusual patterns that do not conform to expected behavior, useful for fraud detection, etc.

22. Discuss the types of anomalies encountered in anomaly detection.
Types include point anomalies, contextual anomalies, and collective anomalies.

23. Explain the difference between supervised and unsupervised anomaly detection techniques.
Supervised techniques require labeled data, while unsupervised techniques work with unlabeled data.

24. Describe the Isolation Forest algorithm for anomaly detection.
Isolation Forest isolates anomalies by randomly selecting features and splitting data into smaller subsets.

25. How does One-Class SVM work in anomaly detection?
One-Class SVM separates normal data from anomalies by finding a decision boundary in feature space.

26. Discuss the challenges of anomaly detection in high-dimensional data.
Challenges include the curse of dimensionality, overfitting, and difficulty in defining "normal" behavior.

27. Explain the concept of novelty detection.
Novelty detection identifies previously unseen patterns or data points as anomalies in a model.

28. What are some real-world applications of anomaly detection?
Applications include fraud detection, network security, and equipment failure prediction.

29. Describe the Local Outlier Factor (LOF) algorithm.
LOF measures the local density deviation of data points to detect outliers in a dataset.

30. How do you evaluate the performance of an anomaly detection model?
Performance is evaluated using metrics like precision, recall, F1 score, or ROC curves.

31. Discuss the role of feature engineering in anomaly detection.
Feature engineering helps highlight key characteristics of data that differentiate normal from anomalous behavior.

32. What are the limitations of traditional anomaly detection methods?
Limitations include difficulty handling large datasets, high-dimensionality, and label scarcity.

33. Explain the concept of ensemble methods in anomaly detection.
Ensemble methods combine multiple anomaly detection models to improve accuracy and robustness.

34. How does autoencoder-based anomaly detection work?
Autoencoders reconstruct input data; anomalies are detected based on reconstruction error.

35. What are some approaches for handling imbalanced data in anomaly detection?
Approaches include resampling techniques (over-sampling/under-sampling) and anomaly-aware algorithms.

36. Describe the concept of semi-supervised anomaly detection.
Semi-supervised anomaly detection uses a small labeled dataset of normal data to identify anomalies in the unlabeled data.

37. Discuss the trade-offs between false positives and false negatives in anomaly detection.
Reducing false positives may increase false negatives, and vice versa, affecting model reliability.

38. How do you interpret the results of an anomaly detection model?
Results are interpreted based on how well the model distinguishes between normal and anomalous data.

39. What are some open research challenges in anomaly detection?
Challenges include scalability, high-dimensional data handling, and interpretability of results.

40. Explain the concept of contextual anomaly detection.
Contextual anomaly detection detects outliers based on the context or surrounding data points.

41. What is time series analysis, and what are its key components?
Time series analysis involves analyzing data points indexed in time order, with components like trend, seasonality, and noise.

42. Discuss the difference between univariate and multivariate time series analysis.
Univariate involves a single time-dependent variable, while multivariate involves multiple variables.

43. Describe the process of time series decomposition.
Time series decomposition breaks a series into trend, seasonality, and residual components.

44. What are the main components of a time series decomposition?
Components include trend, seasonal variation, and residual (noise).

45. Explain the concept of stationarity in time series data.
Stationarity means statistical properties of the time series do not change over time.

46. How do you test for stationarity in a time series?
Tests like Augmented Dickey-Fuller (ADF) or KPSS test can check for stationarity.

47. Discuss the autoregressive integrated moving average (ARIMA) model.
ARIMA models time series by combining autoregression, differencing, and moving averages.

48. What are the parameters of the ARIMA model?
ARIMA parameters are (p, d, q): p (autoregressive), d (differencing), and q (moving average).

49. Describe the seasonal autoregressive integrated moving average (SARIMA) model.
SARIMA extends ARIMA by adding seasonal components for modeling seasonal time series data.

50. How do you choose the appropriate lag order in an ARIMA model?
Lag order is selected using ACF/PACF plots or criteria like AIC/BIC.

51. Explain the concept of differencing in time series analysis.
Differencing removes trends by subtracting previous values from current values to make a series stationary.

52. What is the Box-Jenkins methodology?
Box-Jenkins methodology is a systematic approach for identifying, modeling, and forecasting time series data using ARIMA.

53. Discuss the role of ACF and PACF plots in identifying ARIMA parameters.
ACF and PACF help determine the order of AR and MA components in ARIMA.

54. How do you handle missing values in time series data?
Missing values can be imputed using interpolation, forward filling, or model-based imputation.

55. Describe the concept of exponential smoothing.
Exponential smoothing gives more weight to recent observations for forecasting future values.

56. What is the Holt-Winters method, and when is it used?
Holt-Winters is a time series forecasting method for handling both trend and seasonality.

57. Discuss the challenges of forecasting long-term trends in time series data.
Challenges include high uncertainty, changes in external factors, and model overfitting.

58. Explain the concept of seasonality in time series analysis.
Seasonality refers to periodic fluctuations in time series data that occur at regular intervals.

59. How do you evaluate the performance of a time series forecasting model?
Performance is evaluated using metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or RMSE.

60. What are some advanced techniques for time series forecasting?
Advanced techniques include machine learning models like LSTM, Prophet, and hybrid models combining traditional and machine learning methods.