ML-5

1. **Clustering** groups similar data points (unsupervised learning).
2. **Supervised vs Unsupervised Clustering**:
   - *Supervised*: Uses labeled data (rare).
   - *Unsupervised*: No labels (common, e.g., K-means).
3. **Applications**: Customer segmentation, image compression, anomaly detection.
4. **K-means Algorithm**:
   1. Choose *K* clusters.
   2. Assign points to nearest centroid.
   3. Update centroids.
   4. Repeat until convergence.
5. **K-means Pros/Cons**:
   - *Pros*: Fast, scalable.
   - *Cons*: Sensitive to initial centroids, assumes spherical clusters.
6. **Hierarchical Clustering**: Builds a tree of clusters (agglomerative/divisive).
7. **Linkage Criteria**:
   - *Single*: Min distance between clusters.
   - *Complete*: Max distance.
   - *Average*: Mean distance.
8. **DBSCAN**: Density-based clustering (core, border, noise points).
9. **DBSCAN Parameters**: *eps* (radius), *min_samples* (density threshold).
10. **Clustering Evaluation**: Silhouette score, Davies-Bouldin index.
11. **Silhouette Score**: Measures cohesion/separation. Range: [-1, 1]. Higher = better.
12. **High-Dim Challenges**: Curse of dimensionality → sparse data.
13. **Density-Based Clustering**: Finds arbitrary-shaped clusters (e.g., DBSCAN).
14. **GMM vs K-means**: GMM uses probability distributions; K-means uses hard assignments.
15. **Limitations**: Assumes cluster shape (K-means), struggles with varying densities.
16. **Spectral Clustering**: Uses graph theory for non-convex clusters.
17. **Affinity Propagation**: Automates cluster count via message passing.
18. **Categorical Variables**: Use distance metrics (e.g., Hamming) or encoding.

19. **Elbow Method**: Plot WCSS vs *K*; choose *K* at "elbow".

20. **Emerging Trends**: Deep clustering, subspace clustering.

21. **Anomaly Detection**: Identifies rare events (e.g., fraud).

22. **Anomaly Types**:

- *Point*: Single outlier.
- *Contextual*: Abnormal in context (e.g., temp spike in winter).
- *Collective*: Unusual sequence.

23. **Supervised vs Unsupervised**:

- *Supervised*: Needs labeled anomalies.
- *Unsupervised*: Assumes anomalies are rare.

24. **Isolation Forest**: Isolates anomalies using random splits (shorter paths = anomaly).

25. **One-Class SVM**: Learns "normal" boundary; outliers fall outside.

26. **High-Dim Challenges**: Distance metrics become meaningless.

27. **Novelty Detection**: Identifies new/unseen anomalies.

28. **Applications**: Fraud detection, network intrusion.

29. **LOF**: Compares local density to neighbors (low density = outlier).

30. **Evaluation**: Precision-recall, F1-score (if labels available).

31. **Feature Engineering**: Normalize, reduce dimensions (PCA).

32. **Limitations**: Assumes anomalies are rare; sensitive to noise.

33. **Ensemble Methods**: Combine multiple detectors (e.g., Isolation Forest + LOF).

34. **Autoencoder-Based**: Reconstructs normal data poorly for anomalies.

35. **Imbalanced Data**: Use anomaly score thresholds or resampling.

36. **Semi-Supervised**: Uses few labeled anomalies + unlabeled data.

37. **Trade-offs**: Lower false positives → more false negatives (and vice versa).

38. **Interpretation**: Analyze anomaly scores/feature contributions.

39. **Research Challenges**: Explainability, adaptive thresholds.

40. **Contextual Anomalies**: Depend on context (e.g., time, location).

41. **Time Series Analysis**: Studies temporal data (trend, seasonality).

42. **Univariate vs Multivariate**:

- *Univariate*: Single metric over time.
- *Multivariate*: Multiple interdependent metrics.

43. **Decomposition**: Splits series into trend, seasonality, residuals.

44. **Components**:

- *Trend*: Long-term direction.
- *Seasonality*: Periodic patterns.
- *Residuals*: Random noise.

45. **Stationarity**: Mean/variance constant over time (required for ARIMA).

46. **Stationarity Tests**: ADF (Augmented Dickey-Fuller) test.

47. **ARIMA**: Models non-seasonal data with *p* (AR), *d* (I), *q* (MA) terms.

48. **ARIMA Parameters**:

- *p*: Autoregressive lags.
- *d*: Differencing order.
- *q*: Moving average terms.

49. **SARIMA**: Adds seasonal terms (*P*, *D*, *Q*, *m*).

50. **Lag Order Selection**: Use ACF/PACF plots or grid search.

51. **Differencing**: Makes series stationary by subtracting past values.

52. **Box-Jenkins**: Methodology for ARIMA model selection (identify, estimate, diagnose).

53. **ACF/PACF Plots**:

- *ACF*: Total correlation at lag *k*.
- *PACF*: Direct correlation at lag *k*.

54. **Missing Values**: Interpolate or impute (e.g., forward fill).

55. **Exponential Smoothing**: Weighted average of past observations.

56. **Holt-Winters**: Handles trend + seasonality (additive/multiplicative).

57. **Long-Term Forecasting Challenges**: Uncertainty accumulation, regime shifts.

58. **Seasonality**: Regular, time-based patterns (e.g., monthly sales peaks).

59. **Evaluation Metrics**: MAE, RMSE, MAPE.

60. **Advanced Techniques**: Prophet, LSTM neural networks.