Data Mining Overview

1. **Data Mining and its purpose**
    - data mining is the process of extracting useful patterns, trends, and insights from large datasets using statistical, machine learning, and database techniques.
    - purpose:
        - identify hidden patterns in data
        - predict future trends
        - support decision-making
        - detect anomalies
        - optimize business operations
2. **Supervised, Unsupervised, and Semi-Supervised Learning**
    - **supervised learning:** labeled data is used to train models (e.g., classification, regression). example: spam email detection.
    - **unsupervised learning:** unlabeled data is used to find hidden patterns (e.g., clustering, anomaly detection). example: customer segmentation.
    - **semi-supervised learning:** a combination of labeled and unlabeled data is used to improve learning. example: speech recognition.
3. **How Data Warehouse differs from Transaction Database**
    - **transaction database (OLTP):** handles real-time transactional operations (e.g., banking transactions, order processing).
    - **data warehouse (OLAP):** stores historical data for analytical processing and reporting (e.g., sales trends, customer behavior analysis).
4. **Key functionalities of Data Mining**
    - **data characterization** (summarizing general features of a dataset).
    - **data discrimination** (comparing datasets to identify differences).
    - **association analysis** (finding relationships between variables).
    - **classification and prediction** (categorizing and forecasting data).
    - **clustering** (grouping similar objects).
    - **outlier detection** (finding anomalies).

5. **Concept/Class Description, Data Characterization, Data Discrimination, Impurities**
    - **concept/class description:** summarizing data characteristics for a target class.
    - **data characterization:** extracting key features from a dataset, e.g., summarizing customer demographics.
    - **data discrimination:** contrasting datasets to highlight key differences, e.g., comparing high-spending vs. low-spending customers.
    - **impurities:** inconsistencies or noise in data that affect model accuracy, e.g., incorrect or missing values.
6. **Association**
    - finding relationships between variables in a dataset.
    - example: in market basket analysis, if customers buy bread, they are likely to buy butter.
    - rule format: {bread} → {butter} (confidence: 80%, support: 50%).
7. **Correlation (Example)**
    - measures the strength and direction of relationships between variables.
    - example: correlation between temperature and ice cream sales (positive correlation).
    - formula: Pearson’s correlation coefficient (r).
8. **Classification, Prediction, Clustering, Outlier**
    - **classification:** categorizing data into predefined classes (e.g., spam vs. non-spam email).
    - **prediction:** forecasting future values based on patterns (e.g., predicting stock prices).
    - **clustering:** grouping similar data points without predefined labels (e.g., customer segmentation).
    - **outlier detection:** identifying anomalies that differ significantly from the dataset (e.g., fraudulent transactions).
9. **Explain Discrimination and Characterization in Data Mining with Explanation**
    - **characterization:** describes common properties of a dataset (e.g., summarizing customer spending patterns).

- **discrimination:** highlights differences between groups (e.g., comparing high-income vs. low-income customer purchases).

10. **How does Data Mining contribute to Knowledge Inference?**

- data mining helps infer knowledge by identifying patterns, relationships, and trends from raw data.
- it automates knowledge discovery using algorithms like decision trees, clustering, and neural networks.
- example: analyzing past sales data to infer customer preferences and predict future demand.

11. **Differentiate between Operational and Decision Support System**

- **operational system (OLTP):** used for real-time transaction processing.
    - stores current data, supports day-to-day business operations.
    - example: banking system recording customer deposits.
- **decision support system (OLAP):** used for data analysis and decision-making.
    - stores historical data, supports strategic planning.
    - example: analyzing sales trends for business expansion.

12. **OLTP vs. OLAP**

- **OLTP (Online Transaction Processing):**
    - deals with real-time transactions.
    - normalized databases to avoid redundancy.
    - fast insert/update/delete operations.
    - example: banking transactions.
- **OLAP (Online Analytical Processing):**
    - deals with complex queries and historical data.
    - denormalized databases for faster retrieval.
    - supports aggregation, multi-dimensional analysis.
    - example: sales forecasting.

13. **Design a Data Warehouse Source for a Car Company**

- **data sources:**
  - manufacturing database (vehicle production data).
  - sales database (customer purchases).
  - service records (maintenance and repair history).
  - supplier database (parts inventory).
- **data warehouse design:**
  - fact table: sales transactions (car model, price, date).
  - dimension tables: customer info, dealer info, car features.

14. **Interesting Factors in Data Mining**

- support, confidence in association rule mining.
- correlation between variables.
- classification accuracy.
- cluster compactness.
- outlier significance.
- temporal trends in time-series data.

15. **Evaluating Interesting Factors in Data Mining**

- **support & confidence:** higher values indicate stronger association rules.
- **statistical significance:** chi-square test for correlation.
- **accuracy & precision:** performance metrics for classification models.
- **entropy reduction:** decision trees measure attribute importance.
- **lift ratio:** evaluates the effectiveness of association rules.

16. **Types of Relationships in Association Rule Mining and Correlation**

- **association rule mining:** identifies frequent itemsets and strong rules.
- **correlation analysis:** measures linear dependency between variables.
- **difference:** association rule mining finds patterns, while correlation quantifies strength.

17. **Difference between Structured, Semi-Structured, and Unstructured Data**

- **structured data:** stored in a relational database (e.g., customer table in SQL).
- **semi-structured data:** lacks strict schema but has some organization (e.g., JSON, XML).
- **unstructured data:** lacks predefined format (e.g., images, videos, emails).

18. **Steps in Building a Classification Model using Decision Tree**

- select training dataset.
- choose the target attribute.
- compute attribute selection measures (e.g., entropy, Gini index).
- split data recursively based on best attribute.
- grow tree until stopping criteria met.
- prune the tree to improve generalization.
- evaluate model performance using test data.

19. **Steps in KDD (Knowledge Discovery in Databases) Process**

- data selection: extract relevant data.
- data preprocessing: clean and transform data.
- data transformation: normalize, reduce dimensions.
- data mining: apply algorithms to discover patterns.
- pattern evaluation: validate discovered knowledge.
- knowledge presentation: visualize findings.

20. **Data Mining Architecture**

- **data sources:** databases, data warehouses, text files.
- **data preprocessing:** cleaning, integration, transformation.
- **data mining engine:** core algorithms for extracting patterns.
- **pattern evaluation:** statistical methods for validation.
- **user interface:** visualization tools for decision-makers.

21. **Classification of Data Mining Systems**

- **Based on the type of data:**
  - relational databases (SQL-based).
  - transactional databases (business transactions).
  - data warehouses (historical data storage).
  - text & web mining (unstructured data like documents).
- **Based on mining approach:**
  - machine learning-based (neural networks, decision trees).
  - statistical-based (regression, clustering).
- **Based on application domain:**
  - business intelligence (fraud detection, marketing).
  - scientific research (bioinformatics, climate analysis).

22. **Classification of Database Mining**

- **Transactional Database Mining:** extracts patterns from transactional data.
- **Relational Database Mining:** works on structured tabular data.
- **Object-Oriented Database Mining:** extracts patterns from object-based models.
- **Multimedia Database Mining:** deals with images, videos, and audio analysis.
- **Spatial Database Mining:** extracts patterns from location-based data.
- **Temporal Database Mining:** deals with time-series data.

23. **Classification Based on Types of Knowledge**

- **Association rules:** identifies relationships between data items (e.g., "Customers who buy bread often buy butter").
- **Classification rules:** predicts categorical labels (e.g., "Loan applicants are classified as high or low risk").
- **Clusters:** groups similar objects without predefined labels.
- **Sequential patterns:** identifies trends over time (e.g., "Customers who buy a phone tend to buy accessories within a week").

- **Deviation detection:** finds anomalies (e.g., fraud detection).

24. **Classification Based on Applications Created**

- **Business Analytics:** customer segmentation, sales prediction.
- **Healthcare:** disease prediction, drug discovery.
- **Cybersecurity:** anomaly detection, fraud detection.
- **Finance:** risk assessment, stock market analysis.
- **Retail:** recommendation systems, customer purchase analysis.

25. **Classification Based on Techniques Utilized**

- **Classification and Regression:** Decision Trees, SVM, Naïve Bayes.
- **Clustering:** K-Means, DBSCAN, Hierarchical Clustering.
- **Association Rule Mining:** Apriori, FP-Growth.
- **Outlier Detection:** Isolation Forest, LOF.
- **Neural Networks:** Deep learning, CNNs for image analysis.

26. **Data Mining Task Primitives**

- **Set of tasks that define a data mining process:**
  - data characterization (summarizing data features).
  - data discrimination (contrasting different datasets).
  - association analysis (finding co-occurring patterns).
  - classification and prediction (categorizing data).
  - clustering (grouping similar data points).
  - outlier analysis (identifying anomalies).

27. **EDT (Exploratory Data Analysis Techniques)**

- **Data Visualization:** histograms, scatter plots, box plots.
- **Summary Statistics:** mean, median, standard deviation.
- **Correlation Analysis:** finding relationships between attributes.
- **Data Cleaning:** handling missing and duplicate values.

- **Feature Selection:** identifying important attributes.

28. **Major Issues in Data Mining**

- **Data Quality Issues:** noisy, incomplete, or inconsistent data.
- **Scalability Issues:** large datasets require efficient algorithms.
- **Privacy Concerns:** sensitive data should be protected.
- **Data Integration Challenges:** merging data from different sources.
- **Performance Bottlenecks:** computational cost of mining algorithms.

29. **Performance Issues in Data Mining**

- **High Dimensionality:** large feature spaces slow down processing.
- **Data Preprocessing Overhead:** cleaning and transformation require resources.
- **Complexity of Algorithms:** inefficient algorithms struggle with big data.
- **Scalability Constraints:** limited hardware can slow down mining tasks.
- **Real-Time Processing Needs:** streaming data requires fast updates.

30. **Types of Databases Used in Data Mining**

- **Relational Databases:** SQL-based structured data.
- **Transactional Databases:** records business operations.
- **Data Warehouses:** integrates historical data for analysis.
- **Multimedia Databases:** stores images, videos, and audio.
- **Spatial Databases:** stores location-based data (GIS).
- **NoSQL Databases:** handles semi-structured or unstructured data (MongoDB, Cassandra).

31. **Compare and Contrast Supervised and Unsupervised Learning**

- **Supervised Learning:**
  - Requires labeled data (each input has a known output).
  - Used for classification (e.g., spam detection) and regression (e.g., stock price prediction).

  - Example: training an email spam filter with labeled emails.
- **Unsupervised Learning:**
  - Works with unlabeled data (no predefined output).
  - Used for clustering (e.g., customer segmentation) and association (e.g., market basket analysis).
  - Example: grouping customers based on purchase behavior.

32. **What Do You Mean by Association Rule?**

- **Definition:** Association rules find relationships between items in large datasets.
- **Example:**
  - "If a customer buys milk, they are likely to buy bread."
  - Rule format: `{Milk} → {Bread}` (confidence: 80%, support: 50%).
- **Metrics:**
  - **Support:** Frequency of the itemset appearing in the dataset.
  - **Confidence:** Probability of buying Y given X.
  - **Lift:** Strength of the rule compared to random chance.

33. **Indicate the Importance of Data Preprocessing**

- **Improves Data Quality:** Removes noise, handles missing values, and corrects inconsistencies.
- **Enhances Model Performance:** Clean data improves accuracy and efficiency of algorithms.
- **Reduces Overfitting:** Normalization and transformation help generalize models.
- **Facilitates Integration:** Standardizes data from multiple sources.

34. **Partition Techniques and Binning**

- **Partitioning:**
  - Divides dataset into smaller subsets for easier processing.
  - Example: Training (70%) and Testing (30%) data split.
- **Binning:**

- Groups continuous values into discrete intervals.
- Example: Age bins: `[0-18]`, `[19-35]`, `[36-60]`, `[60+]`.
- **Types:**
  - **Equal-width binning:** Divides range into equal intervals.
  - **Equal-frequency binning:** Divides data so each bin has equal data points.

35. **How to Handle Missing Data?**

- **Deletion Methods:**
  - Remove rows or columns with too many missing values.
- **Imputation Methods:**
  - Fill missing values with mean, median, or mode.
  - Use regression models to estimate missing values.
- **Prediction-Based Methods:**
  - Use machine learning models to predict missing values based on available data.

36. **Normalization (Min-Max, Z-score, Decimal Scaling)**

- **Min-Max Normalization:**
  - Formula: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$
  - Scales values between 0 and 1.
- **Z-score Normalization:**
  - Formula: $X' = \frac{X - \mu}{\sigma}$
  - Centers data around mean 0 with standard deviation 1.
- **Decimal Scaling:**
  - Formula: $X' = \frac{X}{10^j}$, where j is the smallest integer making values <1.

37. **Issues While Doing Data Integration and Interpretation**

- **Schema Mismatch:** Different column names and formats across datasets.
- **Data Redundancy:** Duplicate records leading to inconsistencies.

- **Conflicting Data:** Different sources may have contradictory values.
- **Data Transformation Issues:** Merging different formats (e.g., categorical vs. numerical).

38. **What is Data Correlation? Mechanisms and Methods of Correlation**

- **Definition:** Measures how two variables are related.
- **Mechanisms:**
  - **Positive Correlation:** Both variables increase together.
  - **Negative Correlation:** One increases, the other decreases.
  - **No Correlation:** No relationship between variables.
- **Methods:**
  - Pearson's correlation coefficient (linear relationship).
  - Spearman's rank correlation (non-linear relationships).
  - Chi-square test (categorical data association).

39. **Major Tasks of Data Preprocessing**

- **Data Cleaning:** Handling missing, noisy, and inconsistent data.
- **Data Integration:** Combining data from multiple sources.
- **Data Transformation:** Normalization, aggregation, and feature selection.
- **Data Reduction:** Dimensionality reduction and sampling.
- **Data Discretization:** Converting continuous values into categories.

40. **Noisy Data and How to Handle It**

- **Definition:** Data with errors, inconsistencies, or irrelevant variations.
- **Methods to Handle Noisy Data:**
  - **Binning:** Group similar values together to smooth out noise.
  - **Regression Analysis:** Fit a function to reduce randomness.
  - **Clustering:** Identify and remove outliers.
  - **Moving Averages:** Smooth out fluctuations in time-series data.

41. **Why Should Data Be Normalized?**

- **Ensures Fair Comparisons:** Data with different scales can bias models.
- **Improves Model Performance:** Normalized data speeds up training and improves accuracy.
- **Enhances Convergence in Machine Learning:** Gradient descent algorithms converge faster.
- **Reduces Impact of Outliers:** Helps prevent large values from dominating computations.

42. **Confusion Matrix**

- **Definition:** A matrix used to evaluate classification models.
- **Structure:**
  - True Positives (TP): Correctly classified as positive.
  - False Positives (FP): Incorrectly classified as positive.
  - True Negatives (TN): Correctly classified as negative.
  - False Negatives (FN): Incorrectly classified as negative.
- **Key Metrics Derived:**
  - Accuracy = $\frac{TP+TN}{TP+FP+TN+FN}$
  - Precision = $\frac{TP}{TP+FP}$
  - Recall = $\frac{TP}{TP+FN}$
  - F1 Score = $\frac{2\times Precision\times Recall}{Precision+Recall}$

43. **Chi-Square Test: Steps and Example**

- **Definition:** Tests the independence of categorical variables.
- **Steps:**
  1. Create an **observed frequency** table.
  2. Calculate **expected frequencies** assuming independence.
  3. Use the formula:

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

4. Compare with the chi-square distribution table.

- **Example:** Testing if gender and shopping preference are independent.

44. **Mode, 5-Number Summary (Q1, Q3, Median, Min, Max)**

- **Mode:** Most frequently occurring value in a dataset.
- **Five-Number Summary:**
  - Minimum: Smallest value.
  - First Quartile (Q1): 25th percentile.
  - Median (Q2): 50th percentile.
  - Third Quartile (Q3): 75th percentile.
  - Maximum: Largest value.

45. **Applying Different Distance Metrics on a Dataset**

- **Euclidean Distance:**
  - Formula: $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
  - Measures straight-line distance.
- **Manhattan Distance:**
  - Formula: $d = |x_2 - x_1| + |y_2 - y_1|$
  - Measures block-wise movement.
- **Cosine Similarity:**
  - Measures angle between vectors, useful for text mining.
- **Hamming Distance:**
  - Measures differences in categorical data.

46. **Confidence and Support in Association Rule Mining**

- **Support:**
  - Measures how frequently an itemset appears in the dataset.
  - Formula:

$$Support(X) = \frac{count(X)}{TotalTransactions}$$

- **Confidence:**
  - Measures how often items in Y appear when X is present.
  - Formula:

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)}$$

- **Why Confidence and Support Matter?**
  - Support filters out infrequent patterns.
  - Confidence ensures rules are reliable.

47. **Frequent Pattern Mining**

- **Definition:** Identifying patterns that appear frequently in a dataset.
- **Example:**
  - "Milk & Bread" appear together in 60% of transactions.
- **Algorithms Used:**
  - Apriori Algorithm (uses support to find frequent itemsets).
  - FP-Growth Algorithm (uses tree structures to mine patterns).

48. **Difference Between Apriori and FP-Growth Algorithm**

- **Apriori:**
  - Uses an iterative approach with candidate generation.
  - Slower due to multiple database scans.
- **FP-Growth:**
  - Uses a compact data structure (FP-tree).
  - Faster as it avoids candidate generation.

49. **Advantages and Drawbacks of FP-Growth Algorithm**

- **Advantages:**
  - Efficient for large datasets.

  - Avoids multiple database scans.
- **Drawbacks:**
  - High memory usage due to FP-tree storage.
  - Difficult to implement compared to Apriori.

50. **Vertical Data Problem/Pattern in Frequent Pattern Mining**

- **Definition:**
  - Instead of transactions, data is stored as item occurrences across transactions.
- **Example:**
  - Instead of `{T1: (Milk, Bread)}`, store `{Milk: (T1, T2), Bread: (T1, T3)}`.
- **Benefits:**
  - Efficient pattern counting.
  - Suitable for dense datasets.

51. **Spatial Data Mining**

- **Definition:** Extracting patterns from spatial data such as maps, satellite images, and GPS data.
- **Applications:**
  - Urban planning (e.g., traffic congestion patterns).
  - Agriculture (e.g., soil quality analysis).
  - Disaster management (e.g., flood-prone areas detection).
- **Techniques:**
  - Spatial clustering (e.g., identifying high-crime areas).
  - Spatial classification (e.g., land use categorization).

52. **Given a Transaction Dataset, Find the Largest Frequent Itemset**

- **Steps:**
  1. Set a minimum support threshold.
  2. Identify frequent 1-itemsets (items that appear frequently).

3. Use Apriori or FP-Growth to generate larger itemsets.
4. Find the largest itemset meeting the support threshold.

- **Example:**
  - Transactions: `{Milk, Bread, Butter}`, `{Milk, Bread}`, `{Milk, Butter}`, `{Bread, Butter}`
  - If min support = 50%, `{Milk, Bread}` is the largest frequent itemset.

53. **Advantages and Drawbacks of Apriori Algorithm**

- **Advantages:**
  - Simple and easy to understand.
  - Works well for small datasets.
- **Drawbacks:**
  - Computationally expensive for large datasets due to multiple database scans.
  - Generates a large number of candidate itemsets.

54. **Various Applications of Frequent Pattern Mining**

- **Market Basket Analysis:** Identifying product purchase patterns.
- **Medical Diagnosis:** Finding correlations between symptoms and diseases.
- **Web Usage Mining:** Understanding user behavior on websites.
- **Fraud Detection:** Detecting unusual transaction patterns.

55. **Frequent Itemset and Frequent Subtree**

- **Frequent Itemset:** A set of items that appear together frequently in a dataset.
  - Example: `{Milk, Bread, Butter}` in market basket analysis.
- **Frequent Subtree:** A repeating tree structure in hierarchical data.
  - Example: Similar file structures in a computer filesystem.

56. **Association Rule Problems**

- **Challenges:**

  - Setting the right minimum support and confidence.
  - Handling large datasets efficiently.
  - Interpreting and selecting useful rules from many generated rules.

57. **Apriori Algorithm for Finding Frequent Itemsets (Steps)**

- **Steps:**
  1. Set minimum support threshold.
  2. Identify frequent 1-itemsets.
  3. Generate candidate 2-itemsets and count their occurrences.
  4. Prune itemsets below the support threshold.
  5. Repeat for larger itemsets until no more frequent itemsets are found.

58. **Example of Apriori Algorithm**

- **Transactions:**
  - `{Milk, Bread, Butter}`
  - `{Milk, Bread}`
  - `{Milk, Butter}`
  - `{Bread, Butter}`
- **Step 1:** Find frequent 1-itemsets: `{Milk}`, `{Bread}`, `{Butter}`
- **Step 2:** Generate and filter frequent 2-itemsets: `{Milk, Bread}`, `{Milk, Butter}`, `{Bread, Butter}`
- **Step 3:** Generate 3-itemset `{Milk, Bread, Butter}` if support is high.

59. **Compare and Contrast Between Different Types of Association Mining**

- **Single-Dimensional vs. Multi-Dimensional:**
  - Single: Uses only one attribute (e.g., items bought together).
  - Multi: Includes multiple attributes (e.g., time of purchase + items).
- **Quantitative vs. Qualitative:**
  - Quantitative: Uses numeric values (e.g., age group and spending habits).

    - Qualitative: Uses categorical values (e.g., product categories).

60. **Data Reduction Techniques**

- **Dimensionality Reduction:** Removes irrelevant features (e.g., PCA, LDA).
- **Data Compression:** Stores data efficiently (e.g., wavelet transforms).
- **Numerosity Reduction:** Uses approximations instead of full data (e.g., clustering).
- **Aggregation:** Combines data for summarization (e.g., monthly sales reports).