Chapter 4: Classification and Prediction

Repeated Questions:

1. Classification Basics:

- **Q.** Differentiate between Classification and Prediction / Explain classification Vs. Prediction. **(SUMMER 2022, 3 Marks OR; WINTER 2023, 4 Marks; WINTER 2025, 3 Marks OR)**

Answer:

Aspect	Classification	Prediction
Output	Predicts categorical class label (discrete)	Predicts continuous numerical value
Model	Builds a classifier (e.g., decision tree, rule-based)	Builds a predictor (e.g., regression model)
Evaluation	Accuracy, confusion matrix, precision/recall	Mean squared error, correlation coefficient
Example	Classify loan applicant as “safe” or “risky”	Predict future salary of an employee

---

- **Q.** Differentiate between supervised learning and unsupervised learning. **(SUMMER 2022, 4 Marks OR; WINTER 2022, 4 Marks OR; WINTER 2023, 3 Marks OR; SUMMER 2024, 3 Marks)**

Answer:

Aspect	Supervised Learning	Unsupervised Learning
Training data	Labeled (input–output pairs)	Unlabeled (only inputs)
Goal	Learn mapping from input to output	Discover hidden structure/patterns
Techniques	Classification, regression	Clustering, association rules
Example	Decision tree, Naïve Bayes	K-means, Apriori

---

2. Decision Tree:

- **Q.** Explain Decision Tree Classification algorithm with the help of example / Explain decision tree algorithm with suitable example. **(WINTER 2022, 7 Marks OR; SUMMER 2023, 4 Marks OR; WINTER 2023, 7 Marks; SUMMER 2024, 7 Marks; WINTER 2025, 7 Marks)**

Answer:

- **Decision Tree:** A flowchart-like tree structure where each internal node denotes a test on an attribute, each branch represents an outcome, and each leaf node holds a class label.

Algorithm (ID3/C4.5 style):

- Select the best attribute using an attribute selection measure (Information Gain or Gain Ratio) to split the data.
- Create a node for that attribute and branches for each possible value.
- Partition the data according to attribute values.
- Repeat recursively for each branch until:
  - All tuples belong to the same class, or
  - No attributes remain, or
  - No tuples remain (assign majority class).
- Prune the tree to remove overfitting (optional).

Example (buys_computer):

Age	Income	Student	Credit_Rating	Buys_Computer
Youth	High	No	Fair	No
Youth	High	No	Excellent	No
Middle	High	No	Fair	Yes
Senior	Medium	No	Fair	Yes
Senior	Low	Yes	Fair	Yes
Senior	Low	Yes	Excellent	No
Middle	Low	Yes	Excellent	Yes
Youth	Medium	No	Fair	No
Youth	Low	Yes	Fair	Yes
Senior	Medium	Yes	Fair	Yes
Youth	Medium	Yes	Excellent	Yes
Middle	Medium	No	Excellent	Yes
Middle	High	Yes	Fair	Yes
Senior	Medium	No	Excellent	No

- Root: Age has highest information gain → split into Youth, Middle, Senior.
- Youth branch: Student? Yes → Yes; No → No.
- Middle branch: All Yes → leaf Yes.
- Senior branch: Credit_Rating? Fair → Yes; Excellent → No.

Final Tree:

```
Age?
├─ Youth → Student?
│          ├─ Yes → Yes
│          └─ No → No
├─ Middle → Yes
└─ Senior → Credit_Rating?
           ├─ Fair → Yes
           └─ Excellent → No
```

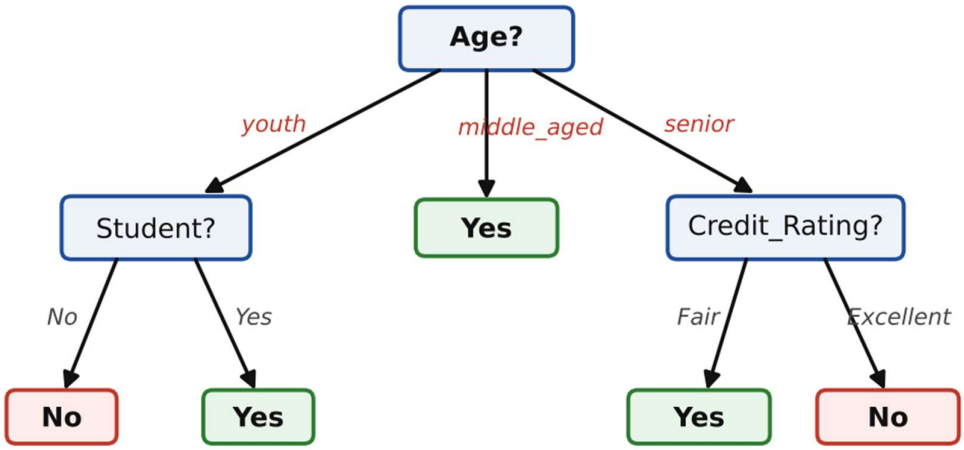


Fig. 4: Decision tree for buys_computer

3. Regression:

- **Q.** Differentiate linear and logistic regression / Briefly explain Linear and Non-linear regression. **(WINTER 2023, 4 Marks OR; SUMMER 2024, 3 Marks OR; WINTER 2025, 4 Marks; SUMMER 2026, 3 Marks)**

Answer:

Aspect	Linear Regression	Logistic Regression
Output	Continuous numeric value	Probability (0 to 1) for binary classification
Equation	$y = \beta_o + \beta_1 x_1 + \cdots + \beta_n x_n$	$p = 1 / (1 + e^{-(\beta_o + \beta_1 x_1 + \ldots)})$
Curve	Straight line	S-shaped (sigmoid) curve
Use case	Predicting salary, temperature	Predicting disease (yes/no), spam detection

- **Non-linear regression:** Fits a non-linear function (e.g., polynomial, exponential) to data where the relationship between variables is not a straight line. Example: $y = ax^2 + bx + c$.

---

4. Neural Network:

- **Q.** Draw the topology of a multilayer, feed-forward Neural Network / Explain ANN. [Diagram required – refer to the original question paper] **(WINTER 2023, 4 Marks; WINTER 2025, 7 Marks OR; SUMMER 2026, 4 Marks)**

Answer:

- **Artificial Neural Network (ANN):** A computational model inspired by biological neural networks, consisting of interconnected nodes (neurons) organized in layers.

Topology of Multilayer Feed-Forward Network:

- **Input Layer:** Receives input attributes (no computation).
- **Hidden Layers:** One or more layers of neurons; each neuron computes a weighted sum of inputs, adds bias, and applies an activation function (e.g., sigmoid, ReLU).
- **Output Layer:** Produces final prediction (class label or value).
- **Feed-forward:** Connections go only from input to output (no cycles).
- **Fully connected:** Each neuron in one layer connects to every neuron in the next layer.

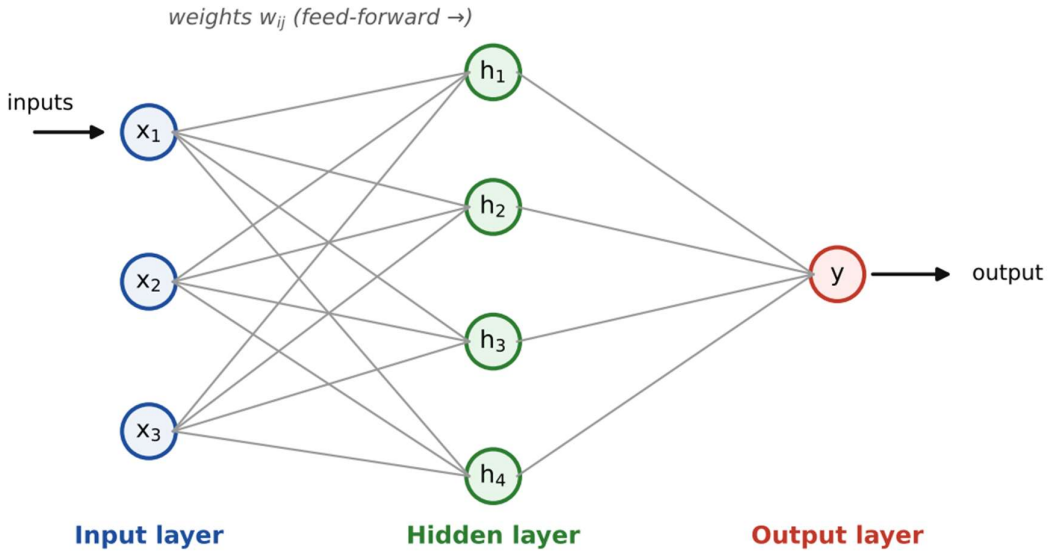


Fig. 5: Multilayer feed-forward neural network (3–4–1)

Learning (Backpropagation):

- Forward pass: Compute output from inputs using current weights.
- Error calculation: Compare predicted output with actual target.
- Backward pass: Propagate error backward and adjust weights using gradient descent to minimize error.

---

5. Accuracy Evaluation:

- **Q.** Explain various methods of evaluating accuracy of classifier / Discuss cross-validation method. **(SUMMER 2023, 3 Marks OR; SUMMER 2024, 4 Marks OR; WINTER 2025, 3 Marks; WINTER 2025, 4 Marks OR)**

Answer:

Methods for Evaluating Classifier Accuracy:

- **Holdout Method:** Split data into training (e.g., 2/3) and test (1/3) sets; train on training set, evaluate on test set.
- **Random Subsampling:** Repeat holdout k times; average accuracy.
- **Cross-Validation (k-fold):** Partition data into k equal folds; train on k−1 folds, test on remaining fold; repeat k times; average accuracy. Stratified cross-validation preserves class distribution in each fold.
- **Bootstrap:** Sampling with replacement; useful for small datasets.
- **Confusion Matrix:** Table of actual vs. predicted classes; compute accuracy, precision, recall, F-measure.
- **Cross-Validation:** Most common is 10-fold cross-validation — data split into 10 folds; each fold used once as test set; final accuracy is average of 10 runs.

Other Important Questions:

1. Classification Theory:

- **Q.** What is classification? Explain classification as a two step process with diagram. [Diagram required – refer to the original question paper] **(SUMMER 2026, 4 Marks OR)**

Answer:

- **Classification:** A data mining function that assigns items to predefined classes based on their attribute values.

Two-Step Process:

- **Learning Step (Training Phase):** Build a classification model (classifier) from a training dataset where class labels are known. The model learns the relationship between attributes and class labels.
- **Classification Step (Testing/ Prediction Phase):** Use the model to predict class labels for new, unseen data. Evaluate the model's accuracy on a test set.

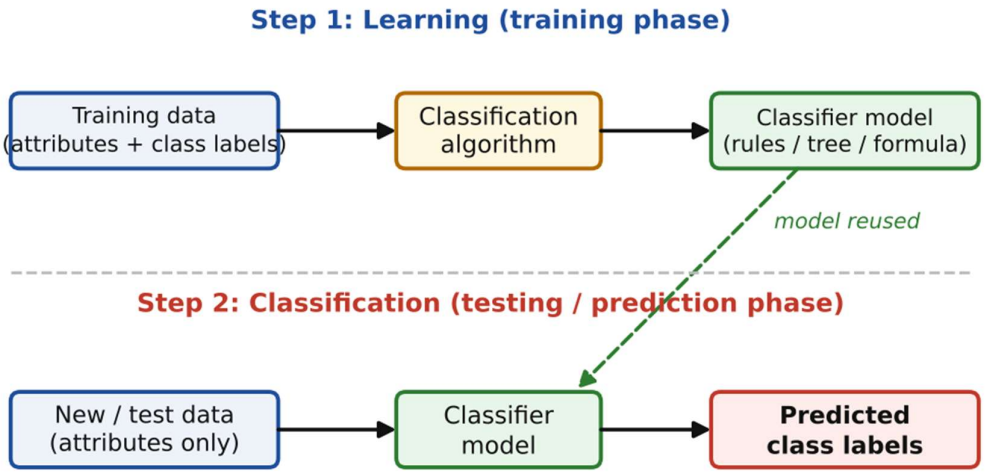


Fig. 6: Classification as a two-step process

- **Q.** Explain the issues regarding the classification and prediction. **(SUMMER 2022, 7 Marks OR)**

Answer:

Issues in Classification and Prediction:

- **Data Preparation:** Handling missing values, noisy data, and irrelevant attributes; feature selection and transformation.
- **Model Selection:** Choosing the right algorithm (decision tree, Bayes, neural network) based on data characteristics and problem type.
- **Overfitting:** Model fits training data too closely and performs poorly on new data; solved by pruning, cross-validation.
- **Underfitting:** Model too simple; fails to capture underlying patterns.
- **Accuracy vs. Interpretability:** Complex models (e.g., neural networks) may be accurate but hard to interpret; simple models (e.g., decision trees) are interpretable but may be less accurate.
- **Class Imbalance:** Unequal distribution of classes; accuracy may be misleading; use precision, recall, F-measure.
- **Scalability:** Algorithms must handle large datasets efficiently.

- **Cost of Errors:** Different misclassification errors may have different costs (e.g., false negative in cancer diagnosis is more costly).

- **Q.** Explain proximity measures used in distance based classifiers. **(WINTER 2025, 4 Marks OR)**

Answer:

- **Proximity Measures:** Quantify similarity or distance between data objects; used in nearest-neighbor classifiers, clustering, etc.
- **Euclidean Distance:** $d(x, y) = \sqrt{(\Sigma_{i=1}^{n} (x_i - y_i)^2)}$ — straight-line distance.
- **Manhattan Distance:** $d(x, y) = \Sigma_{i=1}^{n} |x_i - y_i|$ — city-block distance.
- **Minkowski Distance:** $d(x, y) = (\Sigma_{i=1}^{n} |x_i - y_i|^p)^{(1/p)}$ — generalization of Euclidean (p=2) and Manhattan (p=1).
- **Cosine Similarity:** $\cos(x, y) = (x \cdot y) / (\|x\| \|y\|)$ — measures angle between vectors; used for text data.
- **Jaccard Coefficient:** For binary attributes; ratio of intersection to union.

- **Q.** Write down short note on backpropagation. **(SUMMER 2024, 4 Marks)**

Answer:

- **Backpropagation:** A supervised learning algorithm for training multilayer feed-forward neural networks using gradient descent.

Steps:

- Forward pass: Input is propagated through the network layer by layer to compute the output.
- Error calculation: Compute the error (difference) between predicted output and actual target using a loss function (e.g., squared error).
- Backward pass: Propagate the error backward from output layer to input layer; compute gradient of error with respect to each weight using the chain rule.
- Weight update: Adjust weights using $w_{new} = w_{old} - \eta (\partial E/\partial w)$, where η is learning rate.
- Repeat for multiple epochs until error converges.

- **Q.** What are the hierarchical methods used in classification? **(WINTER 2022, 4 Marks)**

Answer:

- **Hierarchical Methods for Classification:** Build a tree-like hierarchy of classes or use hierarchical decomposition for classification.
- **Decision Tree Induction:** Recursively partition data based on attribute tests; produces a hierarchy of decisions (e.g., ID3, C4.5, CART).
- **Hierarchical Classification:** Classify instances into a hierarchy of categories (e.g., document classification into "Sports → Cricket → Test Match").
- **Rule-Based Hierarchical Classification:** Use hierarchical rules where higher-level rules guide lower-level classification.
- **Nearest Neighbor with Hierarchical Clustering:** Use hierarchical clustering to define class neighborhoods.

- **Q.** Describe in detail about Rule based Classification. **(WINTER 2022, 7 Marks)**

Answer:

- **Rule-Based Classification:** Uses a set of IF-THEN rules for classification. Each rule is of the form: IF condition THEN class_label

Rule Generation:

- **Direct Method:** Extract rules directly from data using algorithms like RIPPER, CN2.
- **Indirect Method:** Extract rules from other models (e.g., decision trees converted to rules).

Rule Evaluation:

- **Coverage:** Fraction of instances satisfying the condition.
- **Accuracy:** Fraction of covered instances that belong to the rule's class.
- **Conflicts:** Multiple rules may match a new instance; resolve by:
    - Size ordering: Use the rule with most attributes (most specific).
    - Rule ordering: Predefine priority order.
    - Voting: Each rule votes for its class; majority wins.

Example Rules:

- IF age = youth AND student = yes THEN buys_computer = yes
- IF age = middle_aged THEN buys_computer = yes
- IF age = senior AND credit_rating = excellent THEN buys_computer = no
- **Advantages:** Interpretable, easy to understand, can handle missing values.
- **Disadvantages:** Rule generation can be complex; may overfit.

---

2. Decision Tree Concepts:

- **Q.** What is attribute selection measure? **(WINTER 2022, 3 Marks OR)**

Answer:

- **Attribute Selection Measure:** A heuristic used to select the best attribute for splitting data at each node of a decision tree. It measures how well an attribute separates the training data into distinct classes.

Common measures:

- **Information Gain:** Reduction in entropy after splitting.
- **Gain Ratio:** Normalized information gain to avoid bias toward multi-valued attributes.
- **Gini Index:** Measures impurity; used in CART.

- **Q.** What do you mean by Tree Pruning? Explain with example. **(WINTER 2023, 3 Marks)**

Answer:

- **Tree Pruning:** The process of removing branches from a decision tree that provide little predictive power, to reduce overfitting and improve generalization.
- **Example:** A decision tree has a branch that splits on an attribute but all resulting leaf nodes belong to the same class. This branch can be pruned (replaced by a leaf).

Types:

- **Pre-pruning:** Stop tree growth early (e.g., if node has fewer than k instances).
- **Post-pruning:** Grow full tree, then remove subtrees that do not improve accuracy on validation data.

- **Q.** Explain the following as attribute selection measure: (i) Information Gain (ii) Gain Ratio **(WINTER 2023, 4 Marks)**

Answer:

- **(i) Information Gain:** Measures the reduction in entropy (impurity) after splitting on an attribute.

$$Gain(A) = Entropy(D) - \Sigma \text{ over } v \in values(A) \text{ of } (|D_v| / |D|) \times Entropy(D_v)$$

where Entropy(D) = $-\Sigma\ p_i \log_2 p_i$. Choose attribute with highest gain.

- **(ii) Gain Ratio:** Normalizes information gain by the split information to penalize attributes with many values.

$$SplitInfo(A) = -\Sigma \text{ over } v \text{ of } (|D_v| / |D|) \times \log_2(|D_v| / |D|)$$

$$GainRatio(A) = Gain(A) / SplitInfo(A)$$

Choose attribute with highest gain ratio; used in C4.5.

---

3. Bayesian Classification:

- **Q.** What are Bayesian Classifiers? **(WINTER 2022, 3 Marks)**

Answer:

- **Bayesian Classifiers:** Probabilistic classifiers based on Bayes' theorem that predict the class with the highest posterior probability given the attribute values.

Key features:

- Based on Bayes' theorem: P(C|X) = P(X|C) P(C) / P(X).
- Naïve Bayes: Assumes conditional independence of attributes given the class.
- Advantages: Simple, fast, works well with high-dimensional data, handles missing values.
- Applications: Text classification, spam filtering, medical diagnosis.

- **Q.** What is Bayesian theorem? **(SUMMER 2022, 4 Marks)**

Answer:

- **Bayes' Theorem:** Describes the probability of a hypothesis based on prior knowledge and evidence.

$$P(H|X) = P(X|H) \cdot P(H) / P(X)$$

Where:

- P(H|X) = posterior probability of hypothesis H given evidence X.
- P(X|H) = likelihood: probability of evidence X given H.
- P(H) = prior probability of H.
- P(X) = marginal probability of evidence X (normalizing constant).
- **Use in classification:** For a tuple X with attributes, compute $P(C_i|X)$ for each class C_i; assign the class with maximum posterior probability.

- **Q.** Demonstrate how Bayesian classification helps in predicting class membership probabilities. **(SUMMER 2022, 7 Marks)**

Answer:

Bayesian Classification for Class Membership Probabilities:

- **Given:** A tuple $X = (x_1, x_2, \ldots, x_n)$ and a set of classes $C_1, C_2, \ldots, C_m$.
- **Goal:** Compute $P(C_i|X)$ for each class — the probability that tuple X belongs to class C_i.
- **Apply Bayes' theorem:**

$$P(C_i|X) = P(X|C_i) \cdot P(C_i) / P(X)$$

- **Naïve assumption:** Attributes are conditionally independent given the class:

$$P(X|C_i) = \Pi \text{ over } k = 1..n \text{ of } P(x_k|C_i)$$

- **Compute probabilities:**
- $P(C_i)$ = relative frequency of class C_i in training data.

- $P(x_k|C_i)$ = relative frequency of attribute value x_k among tuples of class C_i.
- **Predict:** Choose class C_i with maximum $P(C_i|X)$ (since $P(X)$ is same for all classes).
- **Example:** For tuple X = (age=youth, income=medium, student=yes, credit=fair), compute:
- $P(yes|X) \propto P(youth|yes) \cdot P(medium|yes) \cdot P(yes|yes) \cdot P(fair|yes) \cdot P(yes)$
- $P(no|X) \propto$ similarly for class "no".
- Compare and assign class with higher probability.
- **Result:** Bayesian classification provides class membership probabilities, not just a label — useful for risk assessment and decision-making.

- **Q.** Explain Naïve Bayes Classifier with example. **(WINTER 2023, 7 Marks OR)**

Answer:

- **Naïve Bayes Classifier:** A probabilistic classifier based on Bayes' theorem with the assumption that attributes are conditionally independent given the class.

Algorithm:

- Training: For each class C_i, compute prior $P(C_i)$ and conditional probabilities $P(x_k|C_i)$ for each attribute value.
- Prediction: For a new tuple X, compute $P(C_i|X) \propto P(C_i) \Pi P(x_k|C_i)$ for each class.
- Assign the class with maximum probability.
- **Example (buys_computer):** Training data (14 tuples) with classes Yes (9) and No (5).

For tuple X = (age=youth, income=medium, student=yes, credit_rating=fair):

- Prior: P(yes) = 9/14 = 0.643, P(no) = 5/14 = 0.357.

Conditional probabilities (Yes):

- P(youth|yes) = 2/9 = 0.222
- P(medium|yes) = 4/9 = 0.444
- P(yes|yes) = 6/9 = 0.667
- P(fair|yes) = 6/9 = 0.667

Conditional probabilities (No):

- P(youth|no) = 3/5 = 0.600
- P(medium|no) = 2/5 = 0.400
- P(yes|no) = 1/5 = 0.200
- P(fair|no) = 2/5 = 0.400

Compute:

- $P(yes|X) \propto 0.643 \times 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.0282$
- $P(no|X) \propto 0.357 \times 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.00686$
- **Prediction:** Since P(yes|X) > P(no|X), classify as buys_computer = Yes.

---

4. Classification Problems (Numerical):

- **Q.** The following table shows the midterm and final exam grades obtained by students in a database course. Use the method of least squares to find an equation for the prediction of a student's final exam grade based on the student's midterm grade in the course. Predict the final exam grade of a student who received 86 grade in the midterm exam. [Diagram required – refer to the original question paper] **(SUMMER 2023, 4 Marks)**

Answer:

- **Note:** Original data table not provided. Method shown; apply to actual table.

- **Given:** Data pairs (x_i, y_i) — midterm grade x and final grade y.
- **To Find:** Equation y = a + bx and predict y for x = 86.

Formula:

$$b = (n \Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)) / (n \Sigma x_i^2 - (\Sigma x_i)^2), \quad a = (\Sigma y_i - b \Sigma x_i) / n$$

Solution Steps:

- Compute n, Σx_i, Σy_i, $\Sigma x_i y_i$, Σx_i^2.
- Substitute in formulas to find b and a.
- Write equation y = a + bx.
- For x = 86, compute y = a + 86b.
- **Final Answer:** Predicted final grade = (value obtained after calculation).

- **Q.** Apply ID3 classification algorithm on the following data and construct a decision tree. Show all the stepwise calculations clearly. [Diagram required – refer to the original question paper] **(SUMMER 2023, 7 Marks)**

Answer:

- **Note:** Original data table not provided. Method shown.

ID3 Algorithm Steps:

- Compute entropy of the target attribute for the whole dataset:

$$Entropy(S) = -\Sigma\, p_i \log_2 p_i$$

- For each attribute A, compute Information Gain:

$$Gain(S, A) = Entropy(S) - \Sigma \text{ over } v \text{ of } (|S_v| / |S|) \times Entropy(S_v)$$

- Select attribute with maximum gain as root.
- Split data by that attribute; repeat recursively for each branch.
- Stop when:
  - All instances in a branch belong to same class → leaf.
  - No attributes left → majority class leaf.
  - No instances → parent majority class leaf.
- **Example (buys_computer):** (same as Q3)
- Root: Age (highest gain).
- Youth branch: Student? → Yes/No leaves.
- Middle branch: leaf Yes.
- Senior branch: Credit_Rating? → Fair/Excellent leaves.
- **Final Tree:** (same as Q3 diagram description).

- **Q.** Predict class label of the tuple X = (age = youth, income = medium, student = yes, credit_rating = fair) with the help of Naïve Bayesian classification method and the following data. [Diagram required – refer to the original question paper] **(SUMMER 2023, 7 Marks OR)**

Answer:

- **Note:** Original data table not provided. Method shown; apply to actual table.
- **Given:** Tuple X = (age=youth, income=medium, student=yes, credit_rating=fair) and training data.
- **To Find:** Class label (buys_computer = yes or no).

Steps:

- Compute prior probabilities $P(C_i)$ from training data.
- For each class, compute conditional probabilities $P(x_k|C_i)$ for each attribute value in X.

- Compute $P(C_i|X) \propto P(C_i) \Pi P(x_k|C_i)$ for each class.
- Compare probabilities; assign class with maximum.
- **Example Calculation:** (same as Q19)
- $P(yes|X) \propto 0.643 \times 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.0282$
- $P(no|X) \propto 0.357 \times 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.00686$
- **Final Answer:** Class label = Yes (buys_computer = yes).

Chapter 5: Cluster Analysis

Repeated Questions:

1. K-Means and K-Medoids:

- **Q.** Explain K-means algorithm / Discuss the K-means clustering algorithm using examples. **(SUMMER 2022, 7 Marks; WINTER 2023, 7 Marks; WINTER 2025, 7 Marks)**

Answer:

- **K-Means Clustering:** A partition-based clustering algorithm that divides n objects into k clusters by minimizing the sum of squared distances between objects and their cluster centers (centroids).

Algorithm Steps:

- Choose k — the number of clusters to form.
- Initialize k centroids — randomly select k data points as initial cluster centers.
- Assign each data point to the cluster whose centroid is nearest (using Euclidean distance).
- Recompute centroids — for each cluster, calculate the mean of all points assigned to it.
- Repeat Steps 3–4 until:
  - Cluster assignments no longer change, or
  - Maximum iterations reached, or
  - Centroids stabilize (change below threshold).
- Output the final k clusters.

Objective Function (to minimize):

$$J = \Sigma \text{ over } j = 1..k \ \ \Sigma \text{ over } x_i \in C_j \ \ ||x_i - \mu_j||^2$$

where μ_j = centroid of cluster C_j.

- **Example (2D points, k=2):** Points: A(1,1), B(2,1), C(4,3), D(5,4). Initial centroids: A(1,1) and C(4,3).

Iteration 1:

- Distances: B→A = 1, B→C = √8 ≈ 2.83 → B in cluster 1.
- D→A = 5, D→C = √2 ≈ 1.41 → D in cluster 2.
- Cluster 1: {A, B}, centroid = (1.5, 1)
- Cluster 2: {C, D}, centroid = (4.5, 3.5)

Iteration 2:

- A→C1 = 0.5, A→C2 = $\sqrt{(3.5^2+2.5^2)} \approx 4.30$ → A in C1.
- B→C1 = 0.5, B→C2 = $\sqrt{(2.5^2+2.5^2)} \approx 3.54$ → B in C1.
- C→C1 = $\sqrt{(2.5^2+2^2)} \approx 3.20$, C→C2 = $\sqrt{(0.5^2+0.5^2)} \approx 0.71$ → C in C2.
- D→C1 = $\sqrt{(3.5^2+3^2)} \approx 4.61$, D→C2 = $\sqrt{(0.5^2+0.5^2)} \approx 0.71$ → D in C2.
- Assignments unchanged → stop.
- **Final Clusters:** C1 = {A, B}, C2 = {C, D}.

- **Q.** How K-Mean clustering method differs from K-Medoid clustering method? **(WINTER 2022, 4 Marks OR; SUMMER 2023, 4 Marks OR; WINTER 2023, 3 Marks; SUMMER 2024, 4 Marks OR; SUMMER 2026, 4 Marks)**

Answer:

Aspect	K-Means	K-Medoids
Centroid	Mean of points (may not be an actual data point)	Medoid — an actual data point closest to the center
Sensitivity	Sensitive to outliers (mean is affected)	Robust to outliers (medoid is less affected)
Distance	Uses Euclidean distance	Uses any distance metric (Manhattan, etc.)
Complexity	$O(n \cdot k \cdot t)$ — faster	$O(k(n-k)^2)$ per iteration — slower

(PAM is the most common K-Medoid algorithm.)

---

2. Clustering Basics:

- **Q.** Differentiate classification and clustering. **(SUMMER 2023, 3 Marks; WINTER 2025, 4 Marks)**

Answer:

Aspect	Classification	Clustering
Learning type	Supervised (labeled data)	Unsupervised (unlabeled data)
Goal	Assign items to predefined classes	Discover natural groups/clusters
Output	Class labels for new data	Cluster groupings of data
Example	Decision tree, Naïve Bayes	K-means, hierarchical clustering

---

- **Q.** Explain the requirements of Clustering in Data Mining. **(SUMMER 2022, 7 Marks OR; WINTER 2022, 3 Marks; SUMMER 2023, 4 Marks OR; WINTER 2023, 7 Marks OR; SUMMER 2024, 7 Marks OR)**

Answer:

Requirements of Clustering:

- **Scalability:** Algorithms must handle large databases with millions of objects efficiently.
- **Ability to deal with different types of attributes:** Handle numeric, categorical, binary, and mixed data types.
- **Discovery of clusters with arbitrary shape:** Not limited to spherical clusters; must find elongated, ring-shaped, or irregular clusters.
- **Minimal requirements for domain knowledge:** Require few input parameters (e.g., k) and minimal user guidance.
- **Ability to deal with noise and outliers:** Robust — not sensitive to noisy data or isolated points.
- **Insensitivity to order of input records:** Results should not depend on the order in which data is presented.
- **High dimensionality:** Handle datasets with many attributes efficiently.
- **Interpretability and usability:** Results should be understandable and actionable for users.

---

3. Hierarchical Clustering:

- **Q.** Differentiate between partitioning and hierarchical methods for clustering. **(WINTER 2023, 4 Marks OR; SUMMER 2024, 4 Marks)**

Answer:

Aspect	Partitioning Methods	Hierarchical Methods
Approach	Divide data into k clusters in one level	Build a hierarchy (tree) of clusters
Output	Single partition of data	Dendrogram / nested clusters
Input	Requires k (number of clusters)	No need to specify k initially
Examples	K-means, K-medoids, CLARA	Agglomerative (AGNES), Divisive (DIANA)

- **Q.** Explain hierarchical and agglomerative method of clustering. **(SUMMER 2023, 3 Marks OR; WINTER 2025, 7 Marks)**

Answer:

- **Hierarchical Clustering:** Builds a hierarchy of clusters by either merging smaller clusters into larger ones (agglomerative) or splitting larger clusters into smaller ones (divisive). Result is represented as a dendrogram (tree diagram).

Agglomerative Method (Bottom-Up):

- Start: Each data point is its own cluster (n clusters).
- Compute proximity matrix — distances between all pairs of clusters.
- Merge the two closest clusters into one.
- Update the proximity matrix to reflect new cluster distances.
- Repeat Steps 3–4 until all points belong to a single cluster (or until a stopping criterion is met).
- Cut the dendrogram at a desired level to obtain the final clusters.

Linkage Methods (for measuring cluster distance):

- **Single Link (MIN):** Distance between closest pair of points.
- **Complete Link (MAX):** Distance between farthest pair.
- **Average Link:** Average distance between all pairs.
- **Centroid Link:** Distance between cluster centroids.

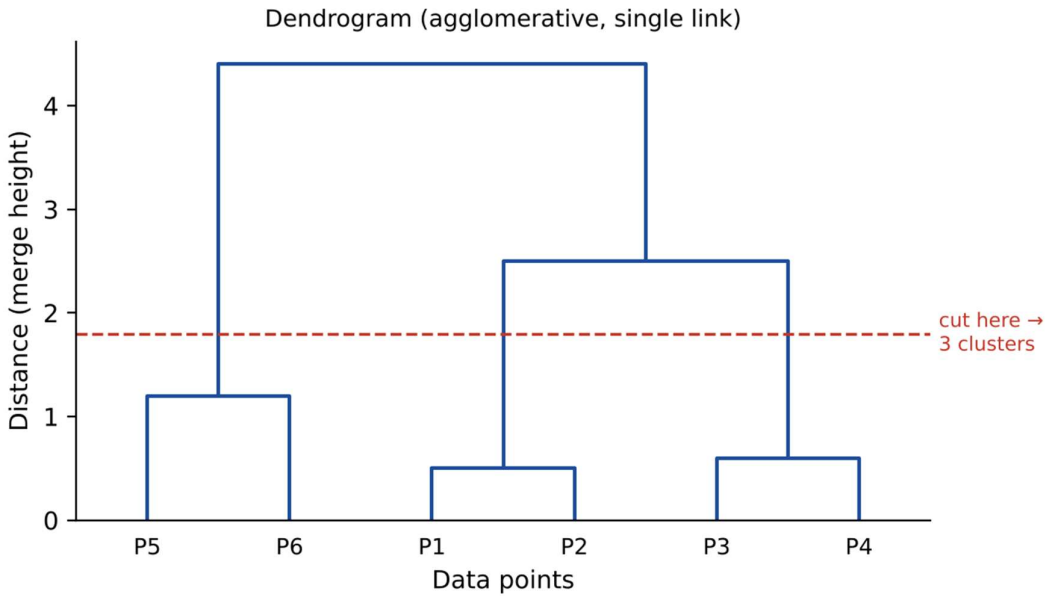


Fig. 7: Dendrogram for agglomerative (bottom-up) clustering

Other Important Questions:

1. Clustering Concepts:

- **Q.** Define "clustering"? Mention any two applications of clustering. **(SUMMER 2026, 3 Marks)**

Answer:

- **Clustering:** The process of grouping a set of data objects into clusters such that objects within a cluster are highly similar to each other and dissimilar to objects in other clusters (unsupervised learning — no predefined class labels).

Two Applications:

- **Market Segmentation:** Group customers with similar buying behavior for targeted marketing.
- **Image Segmentation:** Group pixels with similar color/intensity for object recognition in images.

- **Q.** How to evaluate clustering algorithm? **(WINTER 2025, 3 Marks OR)**

Answer:

Evaluation of Clustering Algorithms:

- **Internal measures:** Use only the data and clustering itself — e.g., Silhouette coefficient (measures how similar an object is to its own cluster vs. others), Davies-Bouldin index, Dunn index.
- **External measures:** Compare clustering results with known ground truth — e.g., Purity, Rand index, Jaccard coefficient, Entropy.
- **Relative measures:** Compare two clustering results to determine which is better (e.g., comparing different values of k).

- **Q.** Discuss data matrix and dissimilarity matrix with respect to clustering. **(SUMMER 2023, 4 Marks)**

Answer:

- **Data Matrix:** An n × p matrix (n objects × p attributes) storing the actual attribute values of data objects. It is the input to clustering algorithms.

```
[ x11   x12   ...   x1p ]
[ x21   x22   ...   x2p ]
[  :     :            :  ]
[ xn1   xn2   ...   xnp ]
```

Each row represents an object; each column an attribute.

- **Dissimilarity Matrix:** An n × n matrix storing the pairwise distances (dissimilarities) between all pairs of objects.

d(i, j) = distance between object i and object j

Properties: d(i, i) = 0, d(i, j) = d(j, i) ≥ 0. Used by clustering algorithms to determine which objects are close.

- **Relationship:** The dissimilarity matrix is computed from the data matrix using a distance function (e.g., Euclidean distance).

- **Q.** What is meant by hierarchical clustering? **(WINTER 2022, 3 Marks OR)**

Answer:

- **Hierarchical Clustering:** A clustering technique that creates a hierarchical decomposition of the data objects, producing a tree-like structure called a dendrogram. It does not require the number of clusters k to be specified in advance — the user can cut the dendrogram at any level to obtain the desired number of clusters. Two approaches: agglomerative (bottom-up merging) and divisive (top-down splitting).

- **Q.** Differentiate Agglomerative and Divisive Hierarchical Clustering? **(WINTER 2022, 4 Marks)**

Answer:

Aspect	Agglomerative	Divisive
Direction	Bottom-up	Top-down
Start	Each point is its own cluster (n clusters)	All points in one cluster (1 cluster)
Operation	Merge two closest clusters at each step	Split a cluster into two at each step
Complexity	$O(n^2)$ or $O(n^2 \log n)$	$O(2^n)$ — expensive
Example	AGNES	DIANA

---

2. Clustering Algorithms:

- **Q.** Explain PAM algorithm. **(WINTER 2025, 7 Marks OR)**

Answer:

- **PAM (Partitioning Around Medoids):** A K-Medoid clustering algorithm that uses actual data points (medoids) as cluster representatives instead of computed means, making it robust to outliers.

Algorithm Steps:

- Select k initial medoids — choose k data points as representative objects.
- Assign each non-medoid object to the cluster of the nearest medoid.
- For each medoid m and each non-medoid object o: (Swap step)
  - Compute the total cost of swapping m with o (change in sum of distances of all objects to their medoids).
  - If the swap reduces total cost, perform the swap (replace medoid m with o).
- Repeat Steps 2–4 until no swap improves the total cost (local optimum reached).
- Output the final k clusters.

Cost Function:

$$Cost = \Sigma \text{ over } j = 1..k \ \ \Sigma \text{ over } x_i \in C_j \ \ d(x_i, m_j)$$

where m_j = medoid of cluster C_j.

Advantages over K-means:

- Robust to noise and outliers (medoid is an actual point).
- Works with any distance metric (not just Euclidean).
- **Disadvantage:** Computationally expensive — $O(k(n-k)^2)$ per iteration.

- **Q.** Explain following clustering algorithm in details: 1) CLARA 2) BIRCH **(SUMMER 2024, 7 Marks)**

Answer:

1) CLARA (Clustering Large Applications):

- An extension of PAM for large datasets.
- Idea: Instead of taking the entire dataset, draw multiple random samples, apply PAM on each sample, and select the best clustering.

Steps:

- Draw s random samples of size 40+2k from the dataset (multiple times).
- Apply PAM to each sample to find k medoids.
- Evaluate the quality of each clustering on the full dataset (compute average distance).
- Keep the clustering with the lowest average distance.
- **Advantage:** Handles large datasets efficiently; more scalable than PAM.
- **Disadvantage:** Results depend on sample quality; may miss some structures.

2) BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies):

- Designed for very large datasets; uses a CF-tree (Clustering Feature tree) to compress data.
- Clustering Feature (CF): (N, LS, SS) — number of points, linear sum, and squared sum.

Steps:

- Build CF-tree: Scan data; insert each point into the CF-tree (a height-balanced tree). If a leaf is full, split it. Optionally rebuild with larger threshold.
- Clustering: Apply a clustering algorithm (e.g., agglomerative) to the leaf nodes of the CF-tree.
- Refinement (optional): Multi-phase — make further passes to improve clusters.
- **Advantages:** Single scan of data; handles noise; memory-efficient; scalable.
- **Limitation:** Handles only numeric data; sensitive to order of input.

---

3. Outlier Analysis:

- **Q.** What is an outlier? **(SUMMER 2022, 3 Marks)**

Answer:

- **Outlier:** A data object that deviates significantly from the rest of the dataset — it appears to be generated by a different mechanism. Outliers are also called anomalies, exceptions, or discordant observations. They may represent errors (noise) or genuinely interesting rare events (e.g., fraud).

- **Q.** Briefly explain outlier analysis. **(WINTER 2025, 3 Marks)**

Answer:

- **Outlier Analysis:** The process of detecting, identifying, and analyzing data objects that deviate significantly from the general behavior of the data.

Approaches:

- **Statistical methods:** Fit a probability model; objects with low probability are outliers.
- **Distance-based methods:** Objects far from most other objects (beyond a threshold) are outliers.
- **Density-based methods:** Objects in low-density regions (e.g., LOF — Local Outlier Factor) are outliers.
- **Clustering-based methods:** Objects not belonging to any cluster are outliers.
- **Applications:** Fraud detection, network intrusion detection, medical anomaly detection.

- **Q.** Discuss outlier analysis as a data mining functionality with the help of an example. **(SUMMER 2023, 3 Marks OR)**

Answer:

- **Outlier Analysis as a Data Mining Functionality:** Discovering data objects that do not comply with the general behavior or model of the data. It is a distinct data mining task (complementary to clustering and classification).
- **Example:** In a credit card transaction database, most transactions follow normal spending patterns (e.g., amounts within a range, locations frequently visited). An outlier analysis system flags a transaction of a very large amount from a foreign country at an unusual time — this deviation may indicate credit card fraud. The outlier is either removed (if noise) or investigated (if a genuine rare event).

---

4. K-Means (Numerical):

- **Q.** Suppose that the data mining task is to cluster the following eight points (with (x, y) representing location) into three clusters: A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1, and C1 as the center of each cluster, respectively. With the help of k-means algorithm calculate, (i) The three cluster centers after the first round execution (ii) The final three clusters **(SUMMER 2023, 7 Marks)**

Answer:

Given:

- Data points: A1(2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9)
- Number of clusters: k = 3
- Initial centers: A1(2, 10), B1(5, 8), C1(1, 2)
- Distance function: Euclidean distance

$$d = \sqrt{((x_2 - x_1)^2 + (y_2 - y_1)^2)}$$

To Find:

- (i) The three cluster centers after the first round execution
- (ii) The final three clusters

Solution — First Round (Round 1):

Compute distance of each point from the three initial centers:

Point	Distance to A1(2,10)	Distance to B1(5,8)	Distance to C1(1,2)	Assigned Cluster
A1(2,10)	0	√(9+4)=3.61	√(1+64)=8.06	C1 (A1)
A2(2,5)	√(0+25)=5.00	√(9+9)=4.24	√(1+9)=3.16	C3 (C1)
A3(8,4)	√(36+36)=8.49	√(9+16)=5.00	√(49+4)=7.28	C2 (B1)
B1(5,8)	√(9+4)=3.61	0	√(16+36)=7.21	C2 (B1)
B2(7,5)	√(25+25)=7.07	√(4+9)=3.61	√(36+9)=6.71	C2 (B1)

Point	Distance to A1(2,10)	Distance to B1(5,8)	Distance to C1(1,2)	Assigned Cluster
B3(6,4)	√(16+36)=7.21	√(1+16)=4.12	√(25+4)=5.39	C2 (B1)
C1(1,2)	√(1+64)=8.06	√(16+36)=7.21	0	C3 (C1)
C2(4,9)	√(4+1)=2.24	√(1+1)=1.41	√(9+49)=7.62	C2 (B1)

Cluster assignments after Round 1:

- Cluster 1 (centered at A1): {A1(2,10)}
- Cluster 2 (centered at B1): {A3(8,4), B1(5,8), B2(7,5), B3(6,4), C2(4,9)}
- Cluster 3 (centered at C1): {A2(2,5), C1(1,2)}

New cluster centers after Round 1 (mean of points in each cluster):

- Center 1: (2, 10)
- Center 2: ((8+7+6+5+4)/5, (4+5+4+8+9)/5) = (30/5, 30/5) = (6, 6)
- Center 3: ((2+1)/2, (5+2)/2) = (1.5, 3.5)
- **Answer (i):** Cluster centers after first round = (2, 10), (6, 6), (1.5, 3.5)

Solution — Second Round (Round 2):

Recompute distances from new centers:

Point	Dist to (2,10)	Dist to (6,6)	Dist to (1.5,3.5)	Assigned
A1(2,10)	0	√(16+16)=5.66	√(0.25+42.25)=6.52	C1
A2(2,5)	√(0+25)=5.00	√(16+1)=4.12	√(0.25+2.25)=1.58	C3
A3(8,4)	√(36+36)=8.49	√(4+4)=2.83	√(42.25+0.25)=6.52	C2
B1(5,8)	√(9+4)=3.61	√(1+4)=2.24	√(12.25+20.25)=5.70	C2
B2(7,5)	√(25+25)=7.07	√(1+1)=1.41	√(30.25+2.25)=5.70	C2
B3(6,4)	√(16+36)=7.21	√(0+4)=2.00	√(20.25+0.25)=4.53	C2
C1(1,2)	√(1+64)=8.06	√(25+16)=6.40	√(0.25+2.25)=1.58	C3
C2(4,9)	√(4+1)=2.24	√(4+9)=3.61	√(6.25+30.25)=6.04	C1

Cluster assignments after Round 2:

- Cluster 1: {A1(2,10), C2(4,9)}
- Cluster 2: {A3(8,4), B1(5,8), B2(7,5), B3(6,4)}
- Cluster 3: {A2(2,5), C1(1,2)}

New centers after Round 2:

- Center 1: ((2+4)/2, (10+9)/2) = (3, 9.5)
- Center 2: ((8+5+7+6)/4, (4+8+5+4)/4) = (6.5, 5.25)
- Center 3: ((2+1)/2, (5+2)/2) = (1.5, 3.5)

Solution — Third Round (Round 3):

Recompute distances from the new centers:

Point	Dist to (3,9.5)	Dist to (6.5,5.25)	Dist to (1.5,3.5)	Assigned
A1(2,10)	√(1+0.25)=1.12	√(20.25+22.56)=6.54	√(0.25+42.25)=6.52	C1
A2(2,5)	√(1+20.25)=4.61	√(20.25+0.06)=4.51	√(0.25+2.25)=1.58	C3
A3(8,4)	√(25+30.25)=7.43	√(2.25+1.56)=1.95	√(42.25+0.25)=6.52	C2
B1(5,8)	√(4+2.25)=2.50	√(2.25+7.56)=3.13	√(12.25+20.25)=5.70	C1
B2(7,5)	√(16+20.25)=6.02	√(0.25+0.06)=0.56	√(30.25+2.25)=5.70	C2
B3(6,4)	√(9+30.25)=6.26	√(0.25+1.56)=1.35	√(20.25+0.25)=4.53	C2
C1(1,2)	√(4+56.25)=7.76	√(30.25+10.56)=6.39	√(0.25+2.25)=1.58	C3
C2(4,9)	√(1+0.25)=1.12	√(6.25+14.06)=4.51	√(6.25+30.25)=6.04	C1

Cluster assignments after Round 3:

- Cluster 1: {A1(2,10), B1(5,8), C2(4,9)}
- Cluster 2: {A3(8,4), B2(7,5), B3(6,4)}
- Cluster 3: {A2(2,5), C1(1,2)}

New centers after Round 3:

- Center 1: ((2+5+4)/3, (10+8+9)/3) = (3.67, 9)
- Center 2: ((8+7+6)/3, (4+5+4)/3) = (7, 4.33)
- Center 3: ((2+1)/2, (5+2)/2) = (1.5, 3.5)
- **Round 4 check:** Recomputing distances from these new centers yields the same assignments — no point changes cluster. Algorithm converges.

Final Answer:

(i) Cluster centers after first round execution:

- Center 1 = (2, 10)
- Center 2 = (6, 6)
- Center 3 = (1.5, 3.5)

(ii) Final three clusters:

- Cluster 1: {A1(2,10), B1(5,8), C2(4,9)}
- Cluster 2: {A3(8,4), B2(7,5), B3(6,4)}
- Cluster 3: {A2(2,5), C1(1,2)}

Chapter 6: Web mining and other data mining

Repeated Questions:

1. Web Mining:

- **Q.** What is web log? Explain web structure mining and web usage mining in detail. **(WINTER 2023, 7 Marks OR; SUMMER 2026, 7 Marks OR)**

Answer:

- **Web Log:** A record maintained by a web server of user access requests. It typically contains IP address, timestamp, requested URL, HTTP method, status code, bytes transferred, referrer, and user-agent. Web logs are the main data source for web usage mining.

Web Structure Mining:

- Discovers knowledge from the hyperlink structure and document structure of the web.
- Intra-page structure mining: Analyzes structure within a page (HTML/DOM tree, tags, headings).
- Inter-page structure mining: Analyzes hyperlinks between pages as a directed graph.
- Uses link analysis algorithms such as PageRank and HITS to find authoritative/hub pages.
- Applications: search-engine ranking, site link optimization, authority discovery, web community detection.

Web Usage Mining:

- Extracts user access patterns from web logs and related data.
- Data sources: server logs, proxy logs, client logs, cookies, registration data.
- Steps: data collection → preprocessing (cleaning, user/session identification) → pattern discovery → pattern analysis.
- Techniques: association rules, clustering, classification, sequential pattern mining.
- Applications: personalization, recommendation, prefetching/caching, site restructuring, fraud detection.

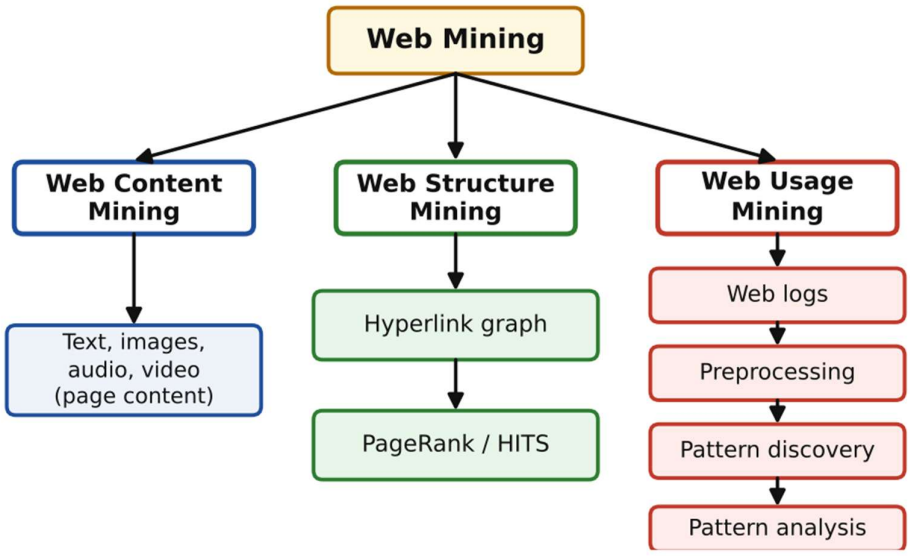


Fig. 8: Web mining taxonomy

- **Q.** Explain web usage mining. **(SUMMER 2022, 4 Marks; WINTER 2022, 7 Marks OR; SUMMER 2023, 7 Marks OR)**

Answer:

- **Web Usage Mining:** The process of extracting interesting user access patterns from web usage data such as server logs, proxy logs, and cookies.
- **Data Collection:** Gather server logs, proxy logs, client-side logs, cookies, and user registration data.
- **Data Preprocessing:** Clean noisy log entries; identify users and sessions; remove robot/spider accesses; path completion.
- **Pattern Discovery:** Apply data mining techniques — association rules (pages accessed together), clustering (similar user groups), classification, and sequential patterns (navigation paths).
- **Pattern Analysis:** Filter and interpret discovered patterns using visualization, OLAP, or knowledge-query mechanisms.
- **Techniques Used:** Association rule mining, clustering, classification, sequential pattern mining, statistical analysis.
- **Applications:** Personalized web pages, recommender systems, improved site design, web caching/prefetching, e-commerce marketing, fraud/intrusion detection.
- **Challenge:** Log data is noisy, distributed, and incomplete due to caching and proxy servers.

- **Q.** Discuss web structure mining. **(SUMMER 2023, 3 Marks; SUMMER 2026, 7 Marks OR)**

Answer:

- **Web Structure Mining:** The process of discovering knowledge from the structural organization of web pages and hyperlinks.
- **Intra-page Structure Mining:** Analyzes the structure inside a single web page — HTML tags, DOM tree, headings, tables, and formatting.
- **Inter-page Structure Mining:** Models the web as a directed graph where pages are nodes and hyperlinks are edges.
- **Link Analysis:** Uses hyperlink topology to infer importance and relationships among pages.
- **PageRank:** Ranks pages by the number and quality of incoming links; a page is important if important pages link to it.
- **HITS Algorithm:** Identifies hubs (pages linking to many authorities) and authorities (pages linked by many hubs).
- **Applications:** Search-engine ranking, finding authoritative pages, web community discovery, improving website navigation.
- **Limitation:** Link structure alone may be affected by spam links and link farms.

(For 3 marks: write points 1, 2, 3, and 6.)

- **Q.** Write a short note: Web content mining. **(SUMMER 2022, 4 Marks OR; WINTER 2022, 7 Marks; SUMMER 2024, 7 Marks OR)**

Answer:

- **Web Content Mining:** The extraction of useful knowledge from the content of web pages — text, images, audio, video, and structured/semi-structured data.
- **Data types:** unstructured text, semi-structured HTML/XML, and multimedia content.
- **Techniques:** information retrieval, text summarization, NLP, named-entity extraction, classification, clustering, and wrapper induction.
- **Structured data mining:** extracts data from tables, templates, and XML/HTML tags.

- **Applications:** search engines, news filtering, opinion mining, product comparison, question answering.

---

2. Temporal and Multimedia Mining:

- **Q.** Explain temporal mining. **(SUMMER 2022, 3 Marks; SUMMER 2024, 3 Marks)**

Answer:

- **Temporal Mining:** Mining patterns from data that has a time dimension — such as time-series, temporal sequences, or timestamped transactions.
- **Temporal data types:** time-series data, sequence data, temporal databases, and timestamped events.
- **Techniques:** trend analysis, similarity search, sequential pattern mining, periodicity detection, and forecasting.
- **Applications:** stock-market prediction, weather forecasting, customer purchase sequences, medical event analysis.

- **Q.** Discuss multimedia mining. **(SUMMER 2022, 3 Marks OR; SUMMER 2023, 4 Marks)**

Answer:

- **Multimedia Mining:** Extraction of hidden knowledge and patterns from multimedia data such as images, audio, video, and text.
- **Data types:** image, audio, video, graphics, and combined multimedia documents.
- **Techniques:** feature extraction (color, texture, shape, audio features), classification, clustering, similarity search, and content-based retrieval.
- **Challenges:** high dimensionality, large size, semantic gap between low-level features and high-level meaning.
- **Applications:** content-based image/video retrieval, surveillance, medical imaging, multimedia recommendation.

---

3. Distributed Data Mining and Hadoop:

- **Q.** Discuss the Applications of Distributed and parallel Data Mining. **(SUMMER 2024, 3 Marks OR; SUMMER 2026, 4 Marks OR)**

Answer:

- **Large-scale data analysis:** Mines massive datasets spread across multiple machines/sites.
- **Privacy-preserving mining:** Mines distributed data without centralizing sensitive records.
- **Speed and scalability:** Parallel processing reduces mining time for big data.
- **Fault tolerance and resource sharing:** Uses distributed computing resources and continues working despite node failures.
- **Applications:** distributed fraud detection, sensor networks, web-scale analytics, cloud data mining.

- **Q.** Discuss the main features of Hadoop Distributed File System / Why Hadoop is important? **(WINTER 2023, 4 Marks OR; SUMMER 2026, 7 Marks OR)**

Answer:

HDFS Main Features:

- **Distributed storage:** Stores large files across many commodity machines.

- **Block-based storage:** Files are split into large blocks (default 128 MB) and stored across DataNodes.
- **Replication:** Each block is replicated (default 3 copies) for fault tolerance.
- **Master–slave architecture:** NameNode manages metadata; DataNodes store actual data blocks.
- **Write-once, read-many:** Optimized for batch processing rather than random updates.
- **High throughput:** Designed for streaming access to large datasets.
- **Scalability:** Easily scales by adding more DataNodes.

Why Hadoop is Important:

- Handles big data on low-cost commodity hardware.
- Provides fault tolerance through replication.
- Supports parallel processing with MapReduce.
- Forms the base of a large big-data ecosystem (Hive, Pig, HBase, Spark).

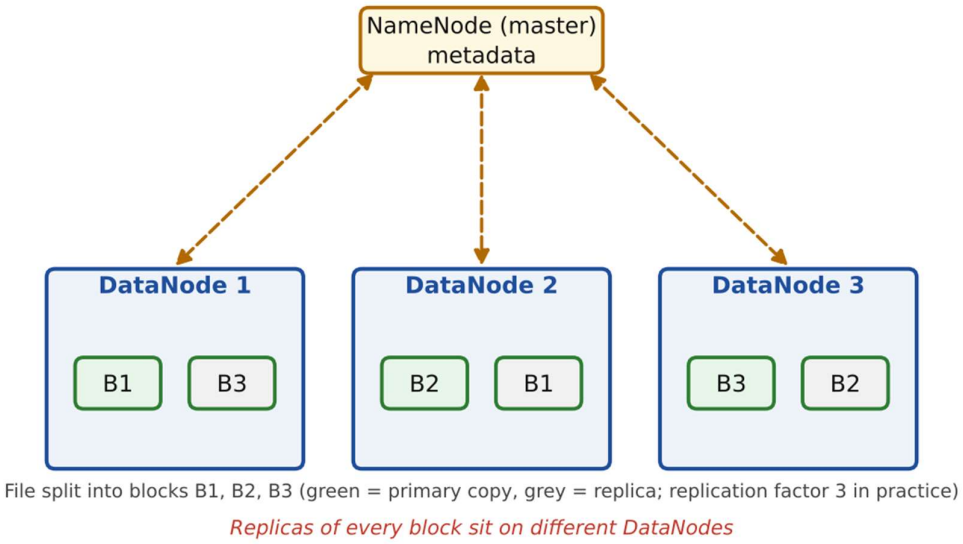


Fig. 9: HDFS master–slave architecture with block replication

---

Other Important Questions:

1. Spatial and Web Mining:

- **Q.** What is spatial data mining? Explain with its application. **(WINTER 2025, 3 Marks)**

Answer:

- **Spatial Data Mining:** Mining knowledge from spatial data such as maps, satellite images, GPS traces, and geographic databases.
- **Spatial data types:** points, lines, polygons, regions, and raster/vector data.
- **Techniques:** spatial clustering, spatial association rules, spatial classification, and spatial outlier detection.
- **Applications:** GIS analysis, traffic prediction, urban planning, location-based services, environmental monitoring.

- **Q.** Discuss spatial Data Mining, Temporal Mining, and Multimedia Mining with their application. **(WINTER 2025, 7 Marks OR)**

Answer:

- **Spatial Data Mining:** Mines patterns from geographic/spatial data — points, lines, polygons, maps, satellite images. Techniques: spatial clustering, spatial association, spatial classification. Applications: GIS, traffic management, urban planning, location-based services.
- **Temporal Mining:** Mines patterns from time-dependent data — time-series, sequences, timestamped events. Techniques: trend analysis, sequential pattern mining, periodicity detection. Applications: stock forecasting, weather prediction, purchase sequence analysis.
- **Multimedia Mining:** Mines patterns from images, audio, video, and text. Techniques: feature extraction, content-based retrieval, classification, clustering. Applications: image/video search, surveillance, medical imaging, multimedia recommendation.
- **Combined use:** Spatial-temporal mining analyzes moving objects and geographic events over time.
- **Common challenge:** High dimensionality, noise, and large data volume.

- **Q.** Discuss application of web mining. **(WINTER 2025, 4 Marks OR)**

Answer:

- **E-commerce personalization:** Recommends products based on browsing and purchase patterns.
- **Search-engine ranking:** Uses link structure and content to rank web pages.
- **Web analytics:** Analyzes visitor behavior to improve site design and marketing.
- **Recommender systems:** Suggests pages, products, or content based on usage patterns.
- **Security and fraud detection:** Detects abnormal access, intrusion, and click fraud.

- **Q.** Explain web mining using example. **(SUMMER 2026, 3 Marks OR)**

Answer:

- **Web Mining:** Application of data mining techniques to web data (content, structure, and usage).
- **Example:** An e-commerce website stores server logs. Web usage mining finds the association rule: “users who view mobile phones also view phone cases” with high support and confidence. The site uses this rule to recommend phone cases on the mobile-phone page. Web structure mining can also use hyperlinks to rank important product pages.

- **Q.** Give difference between text mining and web mining. **(WINTER 2023, 3 Marks OR)**

Answer:

Aspect	Text Mining	Web Mining
Data source	Plain text documents	Web content, hyperlinks, and web logs
Scope	Extracts knowledge from unstructured text	Mines content, structure, and usage of the web
Techniques	NLP, information extraction, summarization	Link analysis, log mining, content extraction
Example	Mining customer reviews for sentiment	Ranking web pages using PageRank

---

2. Hadoop and MapReduce:

- **Q.** Explain Hadoop storage - HDFS. **(SUMMER 2026, 3 Marks)**

Answer:

- **HDFS (Hadoop Distributed File System):** A distributed, fault-tolerant file system designed to store very large files across commodity machines.
- **Architecture:** Master–slave — NameNode stores metadata; DataNodes store data blocks.
- **Block storage:** Files split into large blocks (default 128 MB), each replicated (default 3 times).
- **Fault tolerance:** Replicas on different DataNodes ensure availability.
- **Access model:** Write-once, read-many; optimized for high-throughput batch processing.

- **Q.** Explain MapReduce. Explain any example using MapReduce. **(SUMMER 2026, 7 Marks)**

Answer:

- **MapReduce:** A distributed programming model for processing large datasets in parallel across a cluster.

Phases:

- Input splitting: Input data is split into independent chunks.
- Map phase: Each mapper processes a chunk and emits intermediate key–value pairs.
- Shuffle and sort: Intermediate pairs are grouped by key.
- Reduce phase: Reducers aggregate values for each key and produce final output.
- Output: Final results are written to HDFS.

Example — Word Count:

- Input: “cat dog cat”
- Map: emits (cat,1), (dog,1), (cat,1)
- Shuffle/Sort: groups cat → [1,1], dog → [1]
- Reduce: sums values → (cat,2), (dog,1)
- Output: word frequencies.

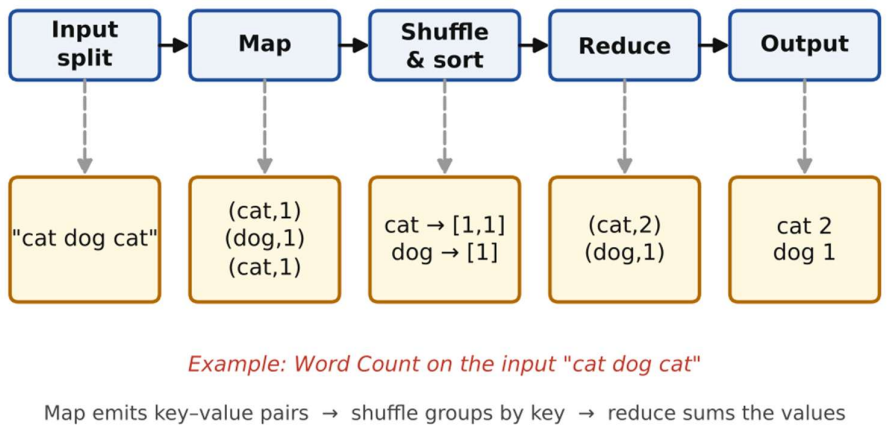


Fig. 10: MapReduce data flow (Word Count example)

****