Unit – 4

🌟 What is Classification

- **Classification** is predicting **categorical (class) labels** for new data using a model trained on past labeled data.
- It's a **predictive data mining technique** (not descriptive).
- Example: Predicting bank loan applications as **safe or risky**.
- Goal: Learn patterns from training data and use them to classify unseen data.
- Difference:
    - **Classification** → predicts class labels (yes/no, safe/risky)
    - **Regression** → predicts continuous values (salary, price)

⚙️ Steps in Classification

1. **Training Phase (Learning)**
    - Train model on a dataset with known class labels.
2. **Testing Phase (Classification)**
    - Use the model to classify new data (unknown labels).

🌳 Decision Tree Induction

- **Decision tree** is a flowchart-like structure.
- **Internal nodes** = attributes
  Branches = attribute tests
  Leaves = class labels

Algorithm (Top-down, greedy):

1. Start with all training data at root.
2. Choose the best attribute (using **information gain / gain ratio / gini index**).
3. Split data based on attribute values.
4. Repeat for each subset until:
    - All tuples in a node belong to the same class
    - Or no attributes left → assign majority class

Famous algorithms:

- ID3 (uses information gain)
- C4.5 (improved ID3)
- CART (binary trees using Gini index)

📌 Attribute Selection Measures

Used to choose the **best splitting attribute**.

1. Information Gain (ID3)

- Measures reduction in entropy (randomness)
- Higher gain = better attribute

2. Gain Ratio (C4.5)

- Solves bias of information gain towards many-valued attributes
- GainRatio = InfoGain / SplitInfo

3. Gini Index (CART)

- Measures impurity
- Lower gini = purer node

✂️ Tree Pruning

- Removes branches caused by noise/outliers to avoid **overfitting**.
- **Prepruning** – stop tree early if gain is low.
- **Postpruning** – grow full tree, then remove weak branches.
- **Cost complexity pruning (CART)** – compares error vs number of leaves.

📊 Bayesian Classification

- Based on **Bayes Theorem** and **conditional probability**.
- Calculates probability that tuple belongs to a class given its attributes.

Bayes Formula:

P(H|X)=P(X|H)P(H)P(X)P(H∣X)=\frac{P(X∣H)P(H)}{P(X)}P(H|X)=P(X)P(X|H)P(H)

- **Naive Bayes** assumes attributes are independent.

- Steps:
    1. Calculate prior probability of each class.
    2. Calculate conditional probability of attributes given class.
    3. Compute posterior for each class and choose highest.

⚖️ Rule-Based Classification

- Uses **IF-THEN rules** for classification.
- Example:
  IF age = youth AND student = yes THEN buys_computer = yes
- **Antecedent (IF)** = conditions, **Consequent (THEN)** = predicted class

Measures

- **Coverage** = % of records that satisfy rule
- **Accuracy** = % of correctly classified among covered records

Conflict resolution strategies

- **Size ordering** – more specific rule wins
- **Rule ordering** – order by accuracy/priority
- **Default rule** – used if no rule matches

📝 Rule Extraction from Decision Tree

- Each root-to-leaf path becomes one rule.
- Rules are **mutually exclusive** (no overlap) and **exhaustive** (cover all cases).

📋 Sequential Covering Algorithm

- Directly learns rules from training data.
- Learn one rule → remove covered tuples → repeat
- Greedy, general-to-specific search
- Examples: AQ, CN2, RIPPER

📈 Model Evaluation and Selection

After building a model, evaluate its accuracy.

Confusion Matrix Terms:

- **TP**: Correctly predicted positive
- **TN**: Correctly predicted negative
- **FP**: Incorrectly predicted positive
- **FN**: Incorrectly predicted negative

Metrics

- **Accuracy** = (TP+TN)/(TP+TN+FP+FN)
- **Error rate** = 1 – Accuracy
- **Precision** = TP/(TP+FP)
- **Recall** = TP/(TP+FN)
- **F1-score** = 2 × (Precision×Recall)/(Precision+Recall)

📦 Evaluation Methods

- **Holdout** – Split data into train/test once
- **Random sampling** – Repeat holdout many times
- **k-Fold Cross Validation** – Split into k parts, train k times, each part once as test
- **Bootstrap (.632)** – Sample with replacement (~63% train, 37% test)

📊 ROC Curve

- Graph of **True Positive Rate (TPR)** vs **False Positive Rate (FPR)**
- Area under ROC curve (AUC) measures model accuracy
- Higher AUC = better

⚡ Ensemble Methods to Improve Accuracy

Combine many models to improve prediction:

- **Bagging**
    - Train many models on bootstrap samples
    - Final prediction by majority vote

- **Boosting**
  - Train models sequentially, each focusing on errors of previous
  - Final prediction by weighted vote
  - Risk: overfitting
- **Random Forest**
  - Many decision trees, each using random subset of attributes
  - Final vote = majority of trees