CS24303 • MODULE 5

Module 5: Classification & Cluster Analysis

Topics 37 to 46 • ID3/C4.5/CART Trees, Naive Bayes, Confusion Matrix, k-Means, PAM & DBSCAN

Module 5 Table of Contents (Topics 37 to 46)

- **Topic 37:** Classification vs. Prediction Overview
- **Topic 38:** Decision Tree Induction (ID3, C4.5, CART)
- **Topic 39:** Naive Bayes & Bayesian Belief Networks
- **Topic 40:** Instance-Based Learning (k-NN) & SVM
- **Topic 41:** Model Evaluation (Confusion Matrix, ROC, AUC)
- **Topic 42:** k-Means & k-Medoids (PAM) Partitioning
- **Topic 43:** Hierarchical Clustering (AGNES, DIANA, BIRCH)
- **Topic 44:** Density-Based Clustering (DBSCAN, OPTICS)
- **Topic 45:** Outlier Detection (Statistical, LOF, Isolation Forest)
- **Topic 46:** 15 Solved University Examination Bank

Topic 37 & 38: Decision Tree Induction (Entropy, Gain Ratio, Gini)

Classification predicts discrete categorical class labels, whereas **Numeric Prediction** models continuous valued functions. Decision tree induction builds a flowchart-like tree structure where each internal node denotes an attribute test, each branch denotes a test outcome, and each leaf node holds a class label.

Algorithm	Splitting Criterion	Attribute Types	Pruning & Missing Values
ID3 (Quinlan 1986)	**Information Gain:** $\text{Gain}(S, A) = H(S) - \sum \frac{\lvert S_v \rvert}{\lvert S \rvert} H(S_v)$	Categorical only (No numeric support).	No pruning (prone to severe overfitting); cannot handle missing values.
C4.5 (Quinlan 1993)	**Gain Ratio:** $\text{GainRatio}(S, A) = \frac{\text{Gain}(S,A)}{\text{SplitInfo}(S,A)}$	Categorical & Continuous (dynamic threshold splitting).	Post-pruning via Pessimistic Error Pruning (PEP); native fractional missing value imputation.
CART (Breiman 1984)	**Gini Impurity:** $\text{Gini}(S) = 1 - \sum_{i=1}^{m} p_i^2 \implies \text{Strictly Binary Trees}$	Categorical & Continuous.	Cost-Complexity Pruning with cross-validation; surrogate splits for missing data.

🏛️ Step-by-Step Solved Problem: ID3 vs. CART Splitting Calculations

A dataset S has 14 instances: 9 "Yes" and 5 "No". Attribute **Wind** has values $\{\text{Weak: } [6 \text{ Yes}, 2 \text{ No}], \text{Strong: } [3 \text{ Yes}, 3 \text{ No}]\}$.

1. Base Entropy $H(S)$ and Information Gain (ID3):

$$H(S) = -\left(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14}\right) = -[0.6429(-0.6374) + 0.3571(-1.4854)] = 0.4098 + 0.5305 = \mathbf{0.9403} \text{ bits}$$

$$H(S_{\text{Weak}}) = -\left(\frac{6}{8}\log_2\frac{6}{8} + \frac{2}{8}\log_2\frac{2}{8}\right) = -[0.75(-0.415) + 0.25(-2.0)] = 0.3113 + 0.5000 = \mathbf{0.8113} \text{ bits}$$

$$H(S_{\text{Strong}}) = -\left(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6}\right) = \mathbf{1.0000} \text{ bits}$$

$$\text{Gain}(\mathbf{S}, \text{Wind}) = \mathbf{0.9403} - \left[\frac{\mathbf{8}}{\mathbf{14}}(\mathbf{0.8113}) + \frac{\mathbf{6}}{\mathbf{14}}(\mathbf{1.0000})\right] = \mathbf{0.9403} - [\mathbf{0.4636} + \mathbf{0.4286}] = \mathbf{0.0481} \text{ bits}$$

2. Gini Impurity (CART):

$$\text{Gini}(S) = 1 - \left[\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right] = 1 - [0.4133 + 0.1276] = 1 - 0.5409 = \mathbf{0.4591}$$

$$\text{Gini}(S_{\text{Weak}}) = 1 - \left[\left(\frac{6}{8}\right)^2 + \left(\frac{2}{8}\right)^2\right] = 1 - [0.5625 + 0.0625] = \mathbf{0.3750}$$

$$\text{Gini}(S_{\text{Strong}}) = 1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] = 1 - [0.25 + 0.25] = \mathbf{0.5000}$$

$$\mathbf{\Delta}\text{Gini}(\mathbf{S}, \text{Wind}) = \mathbf{0.4591} - \left[\frac{\mathbf{8}}{\mathbf{14}}(\mathbf{0.3750}) + \frac{\mathbf{6}}{\mathbf{14}}(\mathbf{0.5000})\right] = \mathbf{0.4591} - [\mathbf{0.2143} + \mathbf{0.2143}] = \mathbf{0.0305}$$

Topic 41: Model Evaluation Metrics (Confusion Matrix, ROC, AUC)

Evaluation Metric	Mathematical Formulation	Key Focus
Accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$	Overall correctness; misleading in imbalanced datasets!
Sensitivity / Recall (TPR)	$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{\text{Positives Correctly Identified}}{\text{Actual Positives}}$	Critical in disease diagnosis (minimizing False Negatives).
Specificity (TNR)	$\text{Specificity} = \frac{TN}{TN + FP} = \frac{\text{Negatives Correctly Identified}}{\text{Actual Negatives}}$	True negative rate (FPR = $1 -$ Specificity).
Precision (PPV)	$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{True Positives}}{\text{Predicted Positives}}$	Critical in fraud detection & search ranking.
F_1-Score	$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$	Harmonic mean balancing precision and recall.

Topic 42 to 45: Clustering Paradigms & Density-Based DBSCAN

Clustering Paradigm	Core Algorithm	Key Strengths & Cluster Shapes
Partitioning Methods	**k-Means:** Minimizes Sum of Squared Errors (SSE); **k-Medoids (PAM):** Uses actual medoid points (robust to outliers).	Fast ($O(tkN)$); restricted strictly to spherical, convex clusters of equal density.
Hierarchical Methods	**AGNES (Agglomerative):** Bottom-up tree; **DIANA (Divisive):** Top-down; **BIRCH:** CF-Tree for massive datasets.	Produces dendrogram hierarchy; no need to specify k upfront; irreversible merges $O(N^2)$.
Density-Based Methods	**DBSCAN:** Clusters are continuous dense regions separated by low-density noise regions (ϵ, MinPts).	Discovers **arbitrary non-convex shapes**; automatically detects and filters noise outliers!

🏛️ The DBSCAN Density Formalism

- **ϵ-Neighborhood:** $N_\epsilon(p) = \{q \in D \mid \text{dist}(p, q) \leq \epsilon\}$.
- **Core Point:** Point p is a Core Point if $|N_\epsilon(p)| \geq \text{MinPts}$.
- **Directly Density-Reachable:** Point q is directly density-reachable from p if $q \in N_\epsilon(p)$ and p is a Core Point.
- **Density-Reachable:** Point q is density-reachable from p if there exists a chain of core points $p_1, \ldots, p_k$ where $p_1 = p, p_k = q$.
- **Density-Connected:** Points p and q are density-connected if there exists a common core point o such that both p and q are density-reachable from o.
- **Noise Outlier:** Any point that is neither a core point nor density-reachable from any core point.

Topic 46: Master University Examination Solved Question Bank

Q1. Trace k-Means Clustering on 1D points: $\{2, 4, 10, 12, 3, 20, 30, 11, 25\}$ for $k = 2$ with initial seeds $m_1 = 2, m_2 = 4$. (10 Marks)

Iteration 1:
- Distance to $m_1 = 2$ vs $m_2 = 4$: Points $\{2, 3\}$ assigned to C_1; Points $\{4, 10, 12, 20, 30, 11, 25\}$ assigned to C_2.
- Update Means: $m_1 = \frac{2+3}{2} = \mathbf{2.5}$; $m_2 = \frac{4+10+12+20+30+11+25}{7} = \frac{112}{7} = \mathbf{16.0}$.

Iteration 2:
- Points closer to 2.5: $\{2, 3, 4\} \implies m_1 = \frac{2+3+4}{3} = \mathbf{3.0}$.
- Points closer to 16.0: $\{10, 11, 12, 20, 25, 30\} \implies m_2 = \frac{10+11+12+20+25+30}{6} = \frac{108}{6} = \mathbf{18.0}$.

Iteration 3:
- Points closer to 3.0: $\{2, 3, 4, 10\}$ ... mid-point $\frac{3+18}{2} = 10.5 \implies$ Point 10 joins $C_1 \implies C_1 = \{2, 3, 4, 10\}$ $(m_1 = 4.75)$; $C_2 = \{11, 12, 20, 25, 30\}$ $(m_2 = 19.6)$.

Iteration 4: Midpoint $\frac{4.75+19.6}{2} = 12.175 \implies$ Point $11, 12$ join $C_1 \implies C_1 = \{2, 3, 4, 10, 11, 12\}$ $(m_1 = 7.0)$; $C_2 = \{20, 25, 30\}$ $(m_2 = 25.0)$.

Iteration 5: Midpoint $\frac{7.0+25.0}{2} = 16.0$. Assignments remain unchanged. **Converged!**

Q2. Explain the Local Outlier Factor (LOF) Density-Based Outlier Detection Algorithm. (8 Marks)

Unlike global distance methods, **LOF (Breunig et al. 2000)** detects outliers by comparing the local density of an object with the densities of its k-nearest neighbors:
1. **k-Distance of p:** Distance $d(p, o)$ to its k-th nearest neighbor.
2. **Reachability Distance:** $\text{reach-dist}_k(p, o) = \max(k\text{-distance}(o), d(p, o))$.
3. **Local Reachability Density (lrd):** $\text{lrd}_k(p) = \frac{|N_k(p)|}{\sum_{o \in N_k(p)} \text{reach-dist}_k(p,o)}$.
4. **LOF Score:** $\text{LOF}_k(p) = \frac{\sum_{o \in N_k(p)} \frac{\text{lrd}(o)}{\text{lrd}(p)}}{|N_k(p)|}$.
- $\text{LOF} \approx 1$: Point is inside a homogeneous cluster.
- $\text{LOF} \gg 1$: Point has significantly lower density than its neighbors $\implies$ Local Outlier!

Topic 46.1: Cluster Validity Indices & Support Vector Machine Math

The Silhouette Coefficient for Clustering Validation: For each data point i: - Let $a(i)$ = average distance from i to all other points in the **same cluster** (Compactness). - Let $b(i)$ = minimum average distance from i to all points in any **other cluster** (Separation).

$$\mathbf{s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \qquad s(i) \in [-1, +1]}$$

- $s(i) \approx +1$: Point is exceptionally well clustered (compact and well-separated). - $s(i) \approx 0$: Point lies exactly on the decision boundary between two clusters. - $s(i) < 0$: Point is assigned to the wrong cluster!

🏛️ **Step-by-Step Solved Problem: Silhouette Score Calculation on 2 Clusters**

Consider 4 points: $C_1 = \{P_1(0,0), P_2(1,0)\}, C_2 = \{P_3(5,0), P_4(6,0)\}$. Calculate Silhouette score for P_1:

- Intra-cluster distance: $a(P_1) = d(P_1, P_2) = 1.0$.
- Inter-cluster distance to C_2: $b(P_1) = \frac{d(P_1,P_3) + d(P_1,P_4)}{2} = \frac{5.0+6.0}{2} = \mathbf{5.5}$.

$$\mathbf{s(P_1) = \frac{b(P_1) - a(P_1)}{\max(a(P_1), b(P_1))} = \frac{5.5 - 1.0}{\max(1.0, 5.5)} = \frac{4.5}{5.5} = +0.8182}$$

Result: High positive Silhouette score $(+0.818)$ indicates near-optimal cluster assignment!

Q3. Compare Agglomerative Hierarchical Linkage Criteria: Single Link, Complete Link, Average Link, and Ward's Method. (8 Marks)

Linkage Criterion	Distance Formulation	Cluster Shape & Properties
Single Link (MIN)	$d(C_i, C_j) = \min_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})$	Can handle non-elliptical shapes; highly sensitive to noise and **chaining effect** (elongated stringy clusters).
Complete Link (MAX)	$d(C_i, C_j) = \max_{\mathbf{x} \in C_i, \mathbf{y} \in C_j} d(\mathbf{x}, \mathbf{y})$	Produces compact, spherical clusters of equal diameter; avoids chaining; sensitive to outliers.
Average Link (UPGMA)	$d(C_i, C_j) = \frac{1}{\lvert C_i \rvert \lvert C_j \rvert} \sum_{\mathbf{x}} \sum_{\mathbf{y}} d(\mathbf{x}, \mathbf{y})$	Robust balance between single and complete link; resistant to noise.
Ward's Method	Merges pair that minimizes increase in total within-cluster Sum of Squared Errors (ΔSSE).	Produces exceptionally balanced, dense spherical clusters (most popular in business analytics).

Topic 46.2: Advanced Ensemble Learning & Density-Based OPTICS Algorithm

Ensemble Method	Core Algorithmic Paradigm	Variance vs. Bias Reduction
Bagging (Bootstrap Aggregating)	Trains B independent base models (e.g. deep unpruned Decision Trees) on bootstrap sampled subsets (N with replacement). Averages predictions (regression) or takes majority vote (classification).	**Dramatically reduces Variance** without increasing bias (e.g. Random Forests).
Boosting (AdaBoost / GBM / XGBoost)	Trains a sequence of weak learners iteratively. Each new learner is trained on the residual errors / weighted mistakes of previous learners.	**Dramatically reduces Bias** and variance; prone to overfitting on heavy noise.
Stacking (Stacked Generalization)	Trains multiple heterogeneous base models (SVM, Random Forest, k-NN); uses a meta-classifier (e.g. Logistic Regression) to learn optimal combination weights.	Maximizes generalization performance by combining diverse hypothesis spaces.

🏛️ Step-by-Step Solved Problem: The OPTICS (Ordering Points To Identify Clustering Structure) Algorithm

Unlike DBSCAN (which requires a single rigid global radius ϵ), **OPTICS (Ankerst et al. 1999)** creates an ordered list of points storing two dynamic distance attributes:

- **Core-Distance of p:** The smallest distance ϵ' such that p is a core point with respect to MinPts:

$$\text{core-dist}_{\epsilon,\text{MinPts}}(\mathbf{p}) = \begin{cases} \text{UNDEFINED} & \text{if } |N_\epsilon(p)| < \text{MinPts} \\ \text{Distance to MinPts-th nearest neighbor} & \text{otherwise} \end{cases}$$

- **Reachability-Distance of p from o:**

$$\text{reach-dist}_{\epsilon,\text{MinPts}}(\mathbf{p}, \mathbf{o}) = \begin{cases} \text{UNDEFINED} & \text{if } |N_\epsilon(o)| < \text{MinPts} \\ \max(\text{core-dist}(o), \text{dist}(o,p)) & \text{otherwise} \end{cases}$$

Reachability Plot: Plotting reachability distances in the OPTICS ordering produces valleys corresponding to dense clusters and peaks corresponding to noise dividers, revealing hierarchical multi-density cluster structures across all ϵ simultaneously!

Topic 46.3: Master University Exam Problem Bank (Part II)

🏛️ Solved Problem 5: Naive Bayes Classification with Laplace Smoothing Calculation

A training dataset has 10 samples: 6 Class C_1 and 4 Class C_2. An attribute **Color** has values $\{\text{Red}, \text{Green}, \text{Blue}\}$. In class C_1, $\text{count}(\text{Red}) = 0, \text{count}(\text{Green}) = 4, \text{count}(\text{Blue}) = 2$.

1. Zero-Frequency Problem (Without Smoothing):

$$P(\text{Red} \mid C_1) = \frac{0}{6} = 0 \implies \text{Total posterior } P(C_1 \mid \text{Red}, \dots) = 0 \quad (\text{Wipes out all evidence!})$$

2. Laplace (Add-1) Smoothing ($k = 3$ attribute values):

$$\mathbf{P(\text{Red} \mid C_1) = \frac{\text{count}(\text{Red}) + 1}{|C_1| + k} = \frac{0+1}{6+3} = \frac{1}{9} = 0.1111}$$

$$\mathbf{P(\text{Green} \mid C_1) = \frac{4+1}{6+3} = \frac{5}{9} = 0.5556} \qquad \mathbf{P(\text{Blue} \mid C_1) = \frac{2+1}{6+3} = \frac{3}{9} = 0.3333}$$

$$\mathbf{\text{Sum of Smoothed Probabilities: } \frac{1}{9} + \frac{5}{9} + \frac{3}{9} = \frac{9}{9} = 1.0000 \quad (\text{Strictly Valid!})}$$

Q4. Compare k-Means Clustering with Gaussian Mixture Models (GMM) and Expectation-Maximization (EM). (8 Marks)

• **k-Means:** Hard clustering (each point belongs strictly to exactly one cluster with probability 1 or 0), assumes spherical clusters of equal variance, minimizes Euclidean SSE.
• **GMM with EM: Soft (Probabilistic) Clustering** where each data point has a posterior membership probability $\gamma_{ik} \in [0, 1]$ across all K Gaussian components ($\sum_k \gamma_{ik} = 1$). GMM learns arbitrary ellipsoidal covariance matrices $\mathbf{\Sigma}_k$ with varying cluster orientations, densities, and scales!

Q5. Explain the Davies-Bouldin Index (DBI) for Evaluating Clustering Quality. (8 Marks)

The **Davies-Bouldin Index (DBI)** measures clustering quality based on the ratio of within-cluster scatter (S_i) to between-cluster separation (R_{ij}):

$$\mathrm{DBI} = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{S_i + S_j}{M_{ij}}\right)$$

Where S_i is the average distance from points in cluster i to their centroid c_i, and $M_{ij} = \|c_i - c_j\|$ is the distance between centroids. **A lower DBI value indicates better clustering** (tighter compactness and wider separation)!

Topic 46.4: Master University Exam Problem Bank (Part III)

🏛️ **Solved Problem 6: Complete Confusion Matrix & ROC Curve Coordinates Calculation**

A binary classifier produces probabilities for 5 test samples: $(S_1 : 0.9, +), (S_2 : 0.8, +), (S_3 : 0.6, -), (S_4 : 0.4, +), (S_5 : 0.1, -)$ $(P = 3, N = 2)$:

Threshold T	Predictions $(\hat{Y})$	TP	FP	$\mathrm{TPR} = \frac{TP}{3}$	$\mathrm{FPR} = \frac{FP}{2}$	ROC Point (x, y)
> 0.9	$[-\,-\,-\,-\,-]$	0	0	0.00	0.00	$(0, 0)$
0.85	$[+\,-\,-\,-\,-]$	1	0	0.33	0.00	$(0, 0.33)$
0.70	$[+\,+\,-\,-\,-]$	2	0	0.67	0.00	$(0, 0.67)$
0.50	$[+\,+\,+\,-\,-]$	2	1	0.67	0.50	$(0.50, 0.67)$
0.30	$[+\,+\,+\,+\,-]$	3	1	1.00	0.50	$(0.50, 1.00)$
< 0.1	$[+\,+\,+\,+\,+]$	3	2	1.00	1.00	$(1.00, 1.00)$

Area Under the Curve (AUC) by Trapezoidal Rule:

$$\mathrm{AUC} = (0.50 - 0)(0.67) + (1.00 - 0.50)\left(\frac{1.00 + 0.67}{2}\right) = 0.335 + 0.50(0.835) = 0.335 + 0.4175 = \mathbf{0.7525}$$

Q6. Explain BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) Algorithm and Clustering Feature (CF) Trees. (8 Marks)

BIRCH (Zhang, Ramakrishnan, Livny 1996) clusters massive datasets in a single database scan using **Clustering Feature (CF) Vectors**:

$$\mathbf{CF} = \langle \mathbf{N}, \mathbf{LS}, \mathbf{SS} \rangle = \langle \mathbf{N}, \sum_{i=1}^{N}\mathbf{x_i}, \sum_{i=1}^{N}\mathbf{x_i^2} \rangle$$

• **Additivity Theorem:** $CF_1 + CF_2 = \langle N_1 + N_2, \mathbf{LS}_1 + \mathbf{LS}_2, SS_1 + SS_2 \rangle$.
• **CF-Tree:** A height-balanced B-tree storing CF vectors. Inserts points incrementally into leaf nodes based on radius thresholds without holding raw data points in memory! Runs in linear time $O(N)$!

Topic 46.5: Master University Exam Problem Bank (Part IV)

🏛️ Solved Problem 7: Complete Support Vector Machine (SVM) Margin & Hyperplane Equations

Given 3 support vectors in 2D space: $\mathbf{x}_1 = (1, 1)^T, y_1 = -1; \mathbf{x}_2 = (2, 0)^T, y_2 = -1; \mathbf{x}_3 = (2, 3)^T, y_3 = +1$:

- Normal weight vector: $\mathbf{w} = (0, 1.5)^T$.
- Bias intercept: $b = -3.5$.
- **Optimal Separating Hyperplane:** $\mathbf{w}^T\mathbf{x} + b = 0 \implies 0(x_1) + 1.5(x_2) - 3.5 = 0 \implies \mathbf{x_2 = 2.333}$.
- **Margin Width (M):**

$$\mathbf{M = \frac{2}{\|w\|} = \frac{2}{\sqrt{0^2 + 1.5^2}} = \frac{2}{1.5} = 1.3333 \text{ units}}$$

- **Classification Decision Rule:**

$$f(\mathbf{x}) = \text{sign}(1.5x_2 - 3.5) = \begin{cases} +1 & \text{if } x_2 \geq 2.333 \\ -1 & \text{if } x_2 < 2.333 \end{cases}$$

Q7. Explain the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) Time Complexity and Spatial Indexing. (8 Marks)

• **Naive Complexity:** Without spatial indexing, finding the ϵ-neighborhood for all N points requires computing $N \times N$ pairwise distance comparisons $\implies O(N^2)$ time complexity.
• **Spatial Indexing Acceleration:** Using R^*-Trees, k-d Trees, or Ball Trees to index spatial coordinates reduces ϵ-range query lookups from $O(N)$ to $O(\log N)$, slashing overall DBSCAN execution complexity to $\mathbf{O(N \log N)}$!

Q8. Detail the Isolation Forest Algorithm for Fast Outlier Detection. (8 Marks)

Isolation Forest (Liu et al. 2008) detects anomalies by explicitly isolating outliers rather than profiling normal data points:
1. Randomly build an ensemble of t Isolation Trees (iTrees) by recursively choosing a random feature and selecting a random split value between $\min$ and $\max$.
2. **Core Axiom:** Anomalies require far fewer random splits to isolate than normal cluster points, resulting in drastically shorter tree path lengths $h(x)$!
3. **Anomaly Score:** $s(x, n) = 2^{-\frac{\mathbb{E}[h(x)]}{c(n)}}$. If $s \to 1$, instance is definitively an outlier. Runs in blazing fast linear time $O(t \cdot \psi \log \psi)$!

Topic 46.6: Complete Solved Laboratory Problem Bank on Clustering & Classification

🏛️ Solved Numerical 8: k-Medoids Partitioning Around Medoids (PAM) Execution Step

Given 5 objects with pairwise L_1 distances: $P_1(1, 1), P_2(2, 2), P_3(3, 3), P_4(8, 8), P_5(9, 9)$. Initial medoids $M_1 = P_2(2, 2), M_2 = P_4(8, 8)$:

- Initial Total Cost = $d(P_1, P_2) + d(P_3, P_2) + d(P_5, P_4) = (|1 - 2| + |1 - 2|) + (|3 - 2| + |3 - 2|) + (|9 - 8| + |9 - 8|) = 2 + 2 + 2 = \mathbf{6.0}$.
- Swap P_2 with non-medoid candidate $P_3(3, 3)$:
- New Cost = $d(P_1, P_3) + d(P_2, P_3) + d(P_5, P_4) = 4 + 2 + 2 = \mathbf{8.0}$.
- $\Delta\text{Cost} = 8.0 - 6.0 = +2.0 > 0 \implies$ Reject swap! Keep $\mathbf{P_2}$ as medoid!

🏛️ Solved Problem 13: Cost-Complexity Tree Pruning ($Minimal \alpha$ Subtree Selection)

In CART decision trees, the cost-complexity measure for subtree T with parameter $\alpha \geq 0$ is:

$$\mathbf{R_\alpha(T) = R(T) + \alpha|T|}$$

Where $R(T)$ is the misclassification error rate and $|T|$ is the number of leaf nodes. By calculating $\alpha = \frac{R(t) - R(T_t)}{|T_t| - 1}$ for each internal node t, CART prunes nodes that produce the smallest error increase per leaf saved!

🏛️ Solved Problem 14: Non-Linear Kernel Support Vector Machine (RBF Gaussian Kernel)

The Radial Basis Function (RBF) kernel is formulated as:

$$\mathbf{K(x, y) = \exp\left(-\gamma \|x - y\|^2\right) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)}$$

Maps input points into an *infinite-dimensional Hilbert space*, allowing linear separation of any arbitrary non-linear dataset!

🏛️ Solved Numerical 12: Grid-Based Spatial Clustering with STING Algorithm

STING (Statistical Information Grid - Wang et al. 1997) divides spatial area into hierarchical rectangular cells at multiple resolution levels. Stores statistical parameters (count n, mean m, standard deviation s, min, max, distribution type) in each grid cell. Queries execute in blazing fast $O(K)$ time (where K is number of grid cells at bottom layer), completely independent of total dataset size N!

🏛️ Solved Problem 15: Decision Tree Pruning via Pessimistic Error Estimation (PEP)

In C4.5, a subtree T with L leaves and E training errors has pessimistic error estimate:

$$\mathbf{E_{pessimistic}(T) = E + \frac{L}{2} + 1.96\sqrt{\frac{(E + L/2)(N - E - L/2)}{N}}}$$

If the pessimistic error of collapsing the subtree into a single leaf node is lower than keeping the complex subtree, C4.5 prunes the subtree, improving generalization on unseen test data!

🏛️ Solved Problem 16: Complete Naive Bayes Full Likelihood Table Proof

Given prior probabilities $P(\text{Play} = \text{Yes}) = \frac{9}{14}$, $P(\text{Play} = \text{No}) = \frac{5}{14}$, and evidence $X = (\text{Outlook} = \text{Sunny}, \text{Temp} = \text{Cool}, \text{Humidity} = \text{High}, \text{Wind} = \text{Strong})$:

$$P(X \mid \text{Yes}) = \left(\frac{2}{9}\right)\left(\frac{3}{9}\right)\left(\frac{3}{9}\right)\left(\frac{3}{9}\right) = \frac{2 \times 3 \times 3 \times 3}{6561} = \frac{54}{6561} \approx \mathbf{0.00823}$$

$$P(X \mid \text{No}) = \left(\frac{3}{5}\right)\left(\frac{1}{5}\right)\left(\frac{4}{5}\right)\left(\frac{3}{5}\right) = \frac{3 \times 1 \times 4 \times 3}{625} = \frac{36}{625} = \mathbf{0.05760}$$

$$\mathbf{P(X, \text{Yes}) = 0.00823 \times \frac{9}{14} = 0.00529} \qquad \mathbf{P(X, \text{No}) = 0.05760 \times \frac{5}{14} = 0.02057}$$

$$\text{Posterior: } \mathbf{P(\text{No} \mid X)} = \frac{\mathbf{0.02057}}{\mathbf{0.00529 + 0.02057}} = \frac{\mathbf{0.02057}}{\mathbf{0.02586}} = \mathbf{79.54\%} \implies \text{Predict: NO PLAY!}$$

🏛️ Solved Problem 17: Multi-Class Support Vector Machine (One-vs-Rest vs One-vs-One)

• **One-vs-Rest (OvR):** Trains K binary classifiers separating class k from all other $K-1$ classes combined. Fast (K models), but sensitive to class imbalance.
• **One-vs-One (OvO):** Trains $\frac{K(K-1)}{2}$ binary classifiers separating all pairwise combinations. Slower training, but each classifier is trained on balanced subsets!