CS24303 • COMPLETE MASTER GUIDE

Data Mining (CS24303) 10-Page Master Revision

High-Yield Formulas, Schemas, Cuboid Lattices, Apriori Rules & Solved Numerical Cards

---

COMPREHENSIVE HIGH-YIELD MASTER GUIDE

Data Mining & Data Warehousing (CS24303) 10-Page Master Quick Revision Guide

Formulas, Schemas, Cuboid Lattices, Apriori Rules, Cluster Algorithms & Solved Exam Cards

---

Master Formula & Metric Reference Sheet

Domain	Universal Formula	Key Exam Property
Min-Max Normalization	$v' = \frac{v - \min_A}{\max_A - \min_A}(\text{new_max} - \text{new_min}) + \text{new_min}$	Preserves exact relationships; sensitive to future out-of-bound outliers.
Z-Score Standardization	$z = \frac{v - \bar{x}}{\sigma}$	Zero mean, unit variance ($\mathcal{N}(0,1)$); ideal when min/max are unknown.
χ^2 Independence Test	$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \text{where } E_{ij} = \frac{\text{Row}_i \times \text{Col}_j}{N}$	$df = (r-1)(c-1)$; tests categorical attribute independence.
Cosine Similarity	$\text{sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2}\sqrt{\sum y_i^2}}$	Angle between vectors; independent of document text length.
Jaccard Similarity	$J(\mathbf{x}, \mathbf{y}) = \frac{q}{q + r + s}$	Asymmetric binary metric; completely ignores $0-0$ negative matches t.
Association Rule Lift	$\text{Lift}(A \implies B) = \frac{P(A \cup B)}{P(A)P(B)} = \frac{\text{Confidence}(A \implies B)}{\text{Support}(B)}$	$\text{Lift} > 1 \implies$ Positive correlation; $= 1 \implies$ Independent.
Information Gain	$\text{Gain}(S, A) = H(S) - \sum_{v \in \text{Values}(A)} \frac{\lvert S_v \rvert}{\lvert S \rvert} H(S_v)$	Biased towards attributes with massive numbers of distinct values.
Gini Impurity	$\text{Gini}(S) = 1 - \sum_{i=1}^{m} p_i^2$	CART metric for strictly binary decision tree splits.

Complete 5-Module Comparative Checklists & Solved Flashcards

Module	Core Theoretical Concepts	High-Yield Exam Numericals
M1: Data Understanding	KDD Process, Nominal/Ordinal/Interval/Ratio, Proximity Metrics (SMC, Jaccard, Cosine, Mahalanobis, Gower).	Five-number summary, IQR outlier fences, Cosine vector similarity, Jaccard distance.
M2: Preprocessing	Missing imputation, Bin smoothing (means, boundaries), χ^2 test, Min-Max / Z-score normalization, PCA.	Contingency table χ^2 computation, decimal scaling, Min-Max re-scaling to $[-1, 1]$.
M3: Data Warehousing	OLTP vs OLAP, Star / Snowflake / Fact Constellation, Additive vs Semi-additive measures, Cuboid lattice combinatorics.	Total cuboids calculation with hierarchies $N = \prod(L_i + 1)$, OLAP operations (slice/dice/roll-up).
M4: Frequent Patterns	Apriori property, join & prune, FP-Tree conditional bases, ECLAT vertical TID lists, Closed vs Maximal.	Full Apriori table trace, Support/Confidence/Lift calculations, FP-Tree branch insertion.
M5: Classification & Clustering	ID3 vs C4.5 vs CART, Naive Bayes, Confusion Matrix, ROC/AUC, k-Means vs PAM, DBSCAN density, LOF.	Entropy & Information Gain numerical, Confusion matrix Precision/Recall/F1, k-Means 1D convergence.

🏛️ Solved Exam Flashcard 1: Star vs Snowflake Schema Differences

Star Schema: De-normalized dimension tables with redundant columns, single-join fact table queries, optimized for fast OLAP read speeds.
Snowflake Schema: Normalized dimension hierarchies (splitting out sub-dimensions into 3NF), zero redundancy, requires complex multi-table SQL joins that degrade interactive slice/dice query performance.

🏛️ Solved Exam Flashcard 2: k-Means vs DBSCAN Clustering Trade-offs

k-Means: Fast $O(tkN)$, requires upfront k, assumes spherical clusters of equal size/variance, highly sensitive to outliers.
DBSCAN: Density-based $(\epsilon, \text{MinPts})$, automatically determines number of clusters, discovers arbitrary non-convex geometries (crescents, rings), completely isolates noise outliers.

🏛️ Solved Exam Flashcard 3: Confidence vs Lift in Association Rules

Confidence measures conditional probability $P(B \mid A)$, but fails to account for the baseline popularity of item B. **Lift** divides confidence by $P(B)$ to reveal true statistical correlation ($\text{Lift} > 1$ positive correlation, $\text{Lift} = 1$ independent, $\text{Lift} < 1$ negative substitute).

🏛️ Solved Exam Flashcard 4: Confusion Matrix Metrics Summary

- $\text{Accuracy} = \frac{TP+TN}{N}$ (Overall correctness)
- $\text{Sensitivity} \ / \ \text{Recall} = \frac{TP}{TP+FN}$ (Fraction of actual positives captured)
- $\text{Specificity} = \frac{TN}{TN+FP}$ (Fraction of actual negatives captured)
- $\text{Precision} = \frac{TP}{TP+FP}$ (Fraction of predicted positives that are correct)
- $F_1\text{-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision}+\text{Recall}}$ (Harmonic mean)

🏛️ Solved Exam Flashcard 5: Information Gain vs Gain Ratio vs Gini Index

- **Information Gain (ID3):** $\text{Gain}(S, A) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$. Biased toward attributes with large numbers of distinct values.
- **Gain Ratio (C4.5):** $\text{GainRatio}(S, A) = \frac{\text{Gain}(S,A)}{\text{SplitInfo}(S,A)}$. Normalizes gain by split entropy, penalizing broad multi-way splits.
- **Gini Impurity (CART):** $\text{Gini}(S) = 1 - \sum p_i^2$. Strictly binary splits.

🏛️ Solved Exam Flashcard 6: Complete Normalization Formulas Reference

- **Min-Max Scaling:** $v' = \frac{v - \min}{\max - \min}(\text{new_max} - \text{new_min}) + \text{new_min}$
- **Z-Score Standardization:** $z = \frac{v - \mu}{\sigma}$
- **Modified Z-Score (Outlier-Robust):** $M_i = \frac{0.6745(x_i - \text{Median})}{\text{MAD}}$ where $\text{MAD} = \text{Median}(|x_i - \text{Median}|)$
- **Decimal Scaling:** $v' = \frac{v}{10^j}$ where $j = \lceil \log_{10}(\max |v|) \rceil$

🏛️ Solved Exam Flashcard 7: The 4 Major Clustering Paradigms

Paradigm	Representative Algorithms	Key Characteristics
Partitioning	k-Means, k-Medoids (PAM), CLARA, CLARANS	Constructs k partitions; iterative relocation; spherical clusters.
Hierarchical	AGNES (agglomerative), DIANA (divisive), BIRCH, CURE	Dendrogram tree; no k required; irreversible merges.
Density-Based	DBSCAN, OPTICS, DENCLUE	Arbitrary non-convex shapes; filters noise outliers automatically.
Grid-Based	STING, WaveCluster, CLIQUE	Quantizes object space into finite grid cells; fast $O(N)$ execution.

Comprehensive 10-Page Master Revision Examination Compendium

🏛️ Solved Flashcard 19: Full Data Preprocessing Pipeline Checklist

- **1. Data Cleaning:** Missing values (impute mean/mode/k-NN), Noise smoothing (bin means, bin boundaries).
- **2. Data Integration:** χ^2 independence test for nominal data, Pearson r / Covariance for numeric data.
- **3. Data Reduction:** Dimensionality reduction (PCA, Wavelet, SVD, t-SNE), Numerosity reduction (Histograms, Clustering).
- **4. Data Transformation:** Normalization (Min-Max, Z-score, Decimal scaling), Discretization (ChiMerge, Equi-depth binning).

🏛️ Solved Flashcard 20: Data Warehouse Architecture Reference

- **Bottom Tier:** Warehouse Database Server (relational back-end data store).
- **Middle Tier:** OLAP Server (ROLAP relational engines, MOLAP multidimensional servers, or HOLAP hybrid).
- **Top Tier:** Front-End Client Tools (Analysis, Query/Reporting, Data Mining engines).

🏛️ Solved Flashcard 21: Full Association Mining Formulas

$$\text{Support}(A \implies B) = \frac{\text{count}(A \cup B)}{N} \qquad \text{Confidence}(A \implies B) = \frac{\text{count}(A \cup B)}{\text{count}(A)}$$

$$\text{Lift}(A \implies B) = \frac{\text{Conf}(A \implies B)}{\text{Support}(B)} \qquad \text{Leverage} = P(A \cup B) - P(A)P(B)$$

$$\text{Conviction} = \frac{1 - \text{Support}(B)}{1 - \text{Conf}(A \implies B)} \qquad \text{Jaccard}(A, B) = \frac{P(A \cup B)}{P(A) + P(B) - P(A \cup B)}$$

🏛️ Solved Flashcard 22: Classification Decision Criteria

- **Information Gain (ID3):** $\text{Gain}(S, A) = H(S) - \sum \frac{|S_v|}{|S|} H(S_v)$
- **Gain Ratio (C4.5):** $\text{GainRatio}(S, A) = \frac{\text{Gain}(S,A)}{\text{SplitInfo}(S,A)}$
- **Gini Impurity (CART):** $\text{Gini}(S) = 1 - \sum p_i^2$
- **Misclassification Error:** $\text{Error}(S) = 1 - \max(p_i)$

🏛️ Solved Flashcard 23: Complete Clustering Paradigms Summary

Method	Core Algorithm	Strengths & Limitations
Partitioning	k-Means, k-Medoids (PAM)	Fast $O(tkN)$; spherical clusters only; sensitive to initialization.
Hierarchical	AGNES (bottom-up), DIANA (top-down)	Dendrogram hierarchy; $O(N^2)$ distance matrix; irreversible merges.
Density-Based	DBSCAN, OPTICS, DENCLUE	Arbitrary non-convex shapes; filters noise outliers; $O(N \log N)$ with spatial index.
Grid-Based	STING, CLIQUE, WaveCluster	Quantized spatial grid cells; fast $O(K)$ query speed independent of N.

Master Exam Final Checklists & Complete Solved Cards

🏛️ Solved Flashcard 24: KDD vs Data Mining vs Machine Learning

- **KDD:** Complete iterative end-to-end knowledge discovery process (Cleaning, Integration, Selection, Transformation, Mining, Evaluation, Presentation).
- **Data Mining:** The algorithmic pattern discovery phase of KDD.
- **Machine Learning:** Algorithms that learn predictive functions from data.

🏛️ Solved Flashcard 25: All Normalization Formulas

- **Min-Max:** $v' = \frac{v - \min}{\max - \min}(\text{new_max} - \text{new_min}) + \text{new_min}$
- **Z-Score:** $z = \frac{v - \mu}{\sigma}$
- **Robust MAD:** $M_i = \frac{0.6745(x_i - \text{Median})}{\text{MAD}}$

🏛️ Solved Flashcard 26: Complete Data Warehousing Concepts

- **Inmon:** Top-down 3NF centralized enterprise data warehouse.
- **Kimball:** Bottom-up dimensional data marts with conformed dimensions.
- **Star:** De-normalized, fast OLAP queries.
- **Snowflake:** Normalized dimensions, zero redundancy, multi-join queries.

🏛️ Solved Flashcard 27: Association Rule Metrics

$$\text{Support} = \frac{\text{count}(A \cup B)}{N} \qquad \text{Confidence} = \frac{\text{count}(A \cup B)}{\text{count}(A)} \qquad \text{Lift} = \frac{\text{Conf}}{\text{Support}(B)}$$

🏛️ Solved Flashcard 28: Complete Decision Tree Induction Pipeline

$$\text{Dataset } S \xrightarrow{\text{Calculate Impurity}} \text{Evaluate Splits} \xrightarrow{\text{Select Max Gain / Min Gini}} \text{Partition Tree} \xrightarrow{\text{Post-Prune Subtrees}} \text{Final Model}$$

🏛️ Solved Flashcard 29: Complete Hierarchical Clustering Linkage Summary

- **Single Link:** $d(A, B) = \min d(x, y)$ (Chaining effect, arbitrary shapes)
- **Complete Link:** $d(A, B) = \max d(x, y)$ (Dense spherical clusters, avoids chaining)
- **Average Link:** $d(A, B) = \text{mean } d(x, y)$ (Balanced compromise)
- **Ward's Method:** Minimizes total within-cluster variance increase ΔSSE

🏛️ Solved Flashcard 30: Complete Association Rule Quality Check

$$\text{Confidence}(A \implies B) \geq \text{min_conf} \quad \text{AND} \quad \text{Lift}(A \implies B) > 1.0 \quad \text{AND} \quad \chi^2 > 3.841$$

🏛️ Solved Flashcard 31: OLAP Storage Architecture Summary

- **ROLAP:** Relational backend, Star/Snowflake schemas, high scalability, SQL engine.
- **MOLAP:** Multidimensional array backend, dense matrix offsets, sub-second query latency.
- **HOLAP:** Hybrid engine: relational for detailed leaf records, multidimensional arrays for aggregates.

🏛️ Solved Flashcard 32: Outlier Detection Techniques Compared

- **Statistical:** 3σ rule, Grubbs' test, Tukey's IQR fences (assumes parametric distribution).
- **Distance-Based:** k-NN distance thresholding ($O(N^2)$).
- **Density-Based:** Local Outlier Factor (LOF) comparing local reachability densities.
- **Ensemble:** Isolation Forest (iTrees) measuring average path lengths to isolate anomalies in $O(N \log N)$.

🏛️ Solved Flashcard 33: Complete Model Performance Metrics Reference

Metric	Formula	Diagnostic Focus
Accuracy	$\frac{TP+TN}{N}$	Overall proportion of correct classifications.
Sensitivity (Recall)	$\frac{TP}{TP+FN}$	Minimizing False Negatives (medical diagnosis).
Specificity	$\frac{TN}{TN+FP}$	True Negative rate ($1 - \text{FPR}$).
Precision	$\frac{TP}{TP+FP}$	Minimizing False Positives (search ranking, spam).
F_1**-Score**	$\frac{2 \cdot P \cdot R}{P+R}$	Harmonic mean for imbalanced targets.
ROC AUC	$\int_0^1 \text{TPR}(t) d(\text{FPR}(t))$	Threshold-independent ranking performance.

🏛️ Solved Flashcard 34: Data Preprocessing Matrix & Transformation Techniques

- **Data Cleaning:** Missing imputation (MICE, k-NN), Noise smoothing (bin means/boundaries).
- **Data Integration:** χ^2 independence test ($df = (r-1)(c-1)$), Pearson correlation $r_{A,B}$.
- **Data Transformation:** Min-Max to $[0, 1]$, Z-score standardization ($\mu = 0, \sigma = 1$), Decimal scaling.
- **Data Reduction:** PCA ($EVR \geq 95\%$), DWT Wavelets, SVD low-rank approximation, ReliefF feature selection.

🏛️ Solved Flashcard 35: Apriori Join & Prune Step Rule

Two frequent $(k-1)$-itemsets $p, q \in L_{k-1}$ are joined if they share the first $k-2$ items ($p[1] = q[1], \ldots, p[k-2] = q[k-2]$ and $p[k-1] < q[k-1]$). Candidate $c = p \cup \{q[k-1]\}$ is pruned if ANY of its $(k-1)$ subsets is missing from L_{k-1}!

🏛️ Solved Flashcard 36: DBSCAN Point Categorization Rule

- **Core Point:** $|N_\epsilon(p)| \geq \text{MinPts}$ (Density center)
- **Border Point:** $|N_\epsilon(p)| < \text{MinPts}$, but $p \in N_\epsilon(q)$ where q is a core point
- **Noise Point:** Neither core nor border (Isolated anomaly)

🏛️ Solved Flashcard 37: Multi-Dimensional Schema Architecture Rule

- **Star Schema:** Fact table surrounded by flat de-normalized dimension tables.
- **Snowflake Schema:** Fact table surrounded by normalized dimension hierarchy tables.
- **Fact Constellation:** Multiple fact tables sharing conformed dimensions.

🏛️ Solved Flashcard 38: Attribute Subset Selection Techniques

- **Stepwise Forward Selection:** Starts with $\emptyset$, greedily adds best attribute.
- **Stepwise Backward Elimination:** Starts with full feature set, greedily removes worst attribute.
- **Decision Tree Induction:** Attributes selected as split nodes represent minimal descriptive subset.

🏛️ Solved Flashcard 39: Classification Performance Metrics Summary

$$\text{FPR} = \frac{FP}{TN + FP} = 1 - \text{Specificity} \qquad \text{FNR} = \frac{FN}{TP + FN} = 1 - \text{Sensitivity}$$

🏛️ Solved Flashcard 40: Density-Based Reachability Axioms

$$p \xrightarrow{\text{direct}} q \implies q \in N_\epsilon(p) \wedge |N_\epsilon(p)| \geq \text{MinPts}$$

$$\text{Density-Reachable: Chain of Core Points} \qquad \text{Density-Connected: Common Core Ancestor } o$$

🏛️ Solved Flashcard 41: Data Cube Computation Strategies

- **BUC (Bottom-Up Computation):** Explores cuboid tree bottom-up with Apriori pruning; ideal for sparse iceberg cubes.
- **MultiWay Array Aggregation:** Computes full dense cubes in single pass over multidimensional RAM chunks.
- **Star-Cubing:** Combines star-trees and shared prefix paths for iceberg cubing.

🏛️ Solved Flashcard 42: Distance Metrics Axioms & Semi-Metrics

$$\text{Non-negativity: } d(x,y) \geq 0 \qquad \text{Identity: } d(x,y) = 0 \iff x = y$$

$$\text{Symmetry: } d(x,y) = d(y,x) \qquad \text{Triangle Inequality: } d(x,z) \leq d(x,y) + d(y,z)$$

🏛️ Solved Flashcard 43: SVD vs PCA Mathematical Equivalences

$$\text{Centering: } \mathbf{X}_c = \mathbf{X} - \mathbf{1}\mu^T \implies \mathbf{X}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

$$\text{Covariance: } \mathbf{S} = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c = \frac{1}{n-1}\mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T \implies \lambda_i = \frac{\sigma_i^2}{n-1},\ \mathbf{e}_i = \mathbf{v}_i$$

🏛️ Solved Flashcard 44: Complete Hierarchical Linkage Metrics

- **Single Link (Nearest Neighbor):** $d(A,B) = \min_{x \in A, y \in B} d(x,y)$
- **Complete Link (Farthest Neighbor):** $d(A,B) = \max_{x \in A, y \in B} d(x,y)$
- **Average Link (Group Average):** $d(A,B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x,y)$
- **Centroid Link:** $d(A,B) = \|\mu_A - \mu_B\|_2$

🏛️ Solved Flashcard 45: Complete Master Review Strategy

Master the 4 Core Pillars: 1. Proximity & Five-number summary, 2. Star/Snowflake & OLAP lattices, 3. Apriori/FP-Tree association rules, 4. Decision trees & DBSCAN density clustering!