Data Mining

Marc Reyes
Department of Software Technology
May 2025

Overview & Objectives

- Introduction to data mining and the KDD process.
- Examination of core techniques:
  - Classification
  - Clustering
  - Association Rule Mining (Apriori & FP-Growth)
  - Regression
  - Anomaly Detection
- Discussion of evaluation metrics, software tools, and ethical considerations.

Learning Objectives

- Define data mining and its significance in extracting actionable insights.
- Master the underlying mathematics behind advanced data mining techniques.
- Analyze and implement real-world computer science applications using rigorous evaluation metrics.

What is Data Mining?

Definition:

- The process of discovering hidden patterns, trends, and relationships in large datasets using mathematical models and algorithms.

Key Elements:

- **Data:** The raw material (structured or unstructured).
- **Algorithms:** Mathematical models used to extract insights.
- **Domain Knowledge:** Expertise required for proper interpretation.

Data Mining vs. Related Fields

Data Mining vs. Machine Learning:

- Focuses on pattern discovery using measures like entropy, information gain, and the Gini index:

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

Data Mining vs. Data Analytics:

- Data Analytics is hypothesis-driven, while data mining is exploratory.

The KDD Process: Data Collection

- **Data Collection:**
  - Gather data from diverse sources such as databases, sensors, public datasets, or web scraping.

The KDD Process: Data Preprocessing

- **Data Preprocessing:**
  - Clean and transform data.
  - Example: Normalization

$$x' = \frac{x - \mu}{\sigma}$$

The KDD Process: Data Exploration

- **Data Exploration:**
  - Perform statistical analysis and visualization to understand data distributions and identify trends.

The KDD Process: Data Mining

- **Data Mining:**
  - Apply advanced algorithms to extract patterns from the processed data.

The KDD Process: Evaluation & Deployment

- **Evaluation:**
  - Assess models using quantitative metrics.
- **Deployment:**
  - Integrate insights into real-world decision-making.

Data Collection & Preprocessing: Data Collection

- **Sources:**
  - Databases, sensors, public datasets, web scraping.

Data Collection & Preprocessing: Cleaning

- **Cleaning Techniques:**
  - Remove noise and handle missing values.
  - Example: Z-score

$$z = \frac{x - \mu}{\sigma}$$

Data Collection & Preprocessing: Transformation

- **Transformation Techniques:**
  - Normalize data using min-max scaling:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Exploratory Data Analysis (EDA)

Purpose:

- Understand distributions, identify trends, and detect outliers.

Techniques:

- Statistical summaries (Mean, Variance, Standard Deviation).
- Visualizations (Histograms, Scatter Plots, Box Plots).
- Correlation Coefficient:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Classification Techniques: Overview

- **Objective:**
  - Assign data points to predefined categories.
- **Common Methods:**
  - Decision Trees, Logistic Regression, k-Nearest Neighbors.

Classification: Decision Trees (Entropy)

- **Entropy Formula:**

$$Entropy(S) = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Classification: Decision Trees (Gini Index)

- **Gini Index Formula:**

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

Classification: Logistic Regression

- **Logistic Regression Model:**

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n)}}$$

Classification: k-Nearest Neighbors

- **Euclidean Distance:**

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Clustering Techniques: Overview

- **Objective:**
  - Group similar data points without predefined labels.
- **Common Method:**
  - k-Means Clustering.

Clustering: k-Means (Centroid Calculation)

- **Centroid Calculation:**

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i$$

Clustering: k-Means (Euclidean Distance)

- **Euclidean Distance:**

$$d(x_i, \mu_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - \mu_{jk})^2}$$

Association Rule Mining: Overview

- **Objective:**
  - Discover relationships among items (market basket analysis).

Association Rule Mining: Support

- **Support Formula:**

$$Support(X) = \frac{\text{Number of transactions containing } X}{N}$$

Association Rule Mining: Confidence

- **Confidence Formula:**

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

Association Rule Mining: Lift

- **Lift Formula:**

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

Association Rule Mining: Additional Measures

- **Leverage:**

$$Leverage = Support(X \cup Y) - Support(X) \times Support(Y)$$

- **Conviction:**

$$Conviction = \frac{1 - Support(Y)}{1 - Confidence(X \rightarrow Y)}$$

Apriori Algorithm: Candidate Generation

- **Candidate Generation Formula:**

$$C_k = \{X \cup Y \mid X, Y \in L_{k-1},\ |X \cap Y| = k - 2\}$$

Apriori Algorithm: Pruning

- **Pruning Condition:**

$$L_k = \{X \in C_k \mid \text{all } (k-1)\text{-subsets of } X \text{ are in } L_{k-1}\}$$

FP-Growth Algorithm: FP-Tree & Conditional Pattern Base

- **Conditional Pattern Base:**

$$\{(\beta, \text{count}) \mid \beta \text{ is a prefix path in the FP-tree for a given prefix } \alpha\}$$

FP-Growth Algorithm: Frequent Pattern Extraction

- **Frequent Pattern Extraction:**

$$\text{FPGrowth}(T_\alpha) = \{\alpha \cup \beta \mid \beta \in \text{FPGrowth}(T_\alpha)\}$$

Regression Techniques: Linear Regression

- **Linear Regression Model:**

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon$$

Regression: Evaluation Metrics

- **Residual Sum of Squares (RSS):**

$$RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

- **Coefficient of Determination (R^2):**

$$R^2 = 1 - \frac{RSS}{TSS}, \quad TSS = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

Anomaly Detection: Mahalanobis Distance

- **Mahalanobis Distance:**

$$D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$$

Evaluation Metrics: Classification

- **Precision:**

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:**

$$Recall = \frac{TP}{TP + FN}$$

- **F1 Score:**

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Evaluation Metrics: Clustering & Statistical Tests

- **Silhouette Score:**

$$s = \frac{b - a}{\max(a, b)}$$

- **Chi-Square Test:**

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Tools & Software for Data Mining

Popular Platforms:

- Python (scikit-learn, Pandas, NumPy)
- R (caret, dplyr, ggplot2)
- Weka (GUI-based tool)
- RapidMiner & Orange (Visual interfaces)

Real-World Applications of Data Mining

Industries & Applications:

- Marketing: Customer segmentation and recommendation systems.
- Finance: Fraud detection and risk analysis.
- Healthcare: Diagnostic support and personalized treatment planning.
- E-commerce: Product recommendations and sentiment analysis.

Challenges and Ethical Considerations

Challenges:

- Handling high-dimensional data (the curse of dimensionality).
- Scalability and computational efficiency (e.g., $O(n \cdot k \cdot t)$ for k-means).
- Data quality issues such as noise and missing values.

Ethical Considerations:

- Privacy, bias, and transparency.

Case Study: Market Basket Analysis Using Association Rule Mining

Project Example:

- Market Basket Analysis in a Large Supermarket Chain

Dataset:

- Instacart Online Grocery Shopping Dataset
  - Over 3 million orders, 200,000+ unique products, 200,000+ customers.
  - Key columns: order_id, user_id, product_id, add_to_cart_order, reordered, order_dow, order_hour_of_day, etc.

Case Study: Data Collection & Preprocessing

Phases:

1. **Data Collection:**
    - Obtain the extensive Instacart dataset covering millions of transactions.
2. **Preprocessing:**
    - Clean the dataset by removing duplicates and standardizing product codes.

Case Study: Mining Techniques - Apriori

Mining Using Apriori:

- Generate candidate itemsets:

$$C_k = \{X \cup Y \mid X, Y \in L_{k-1},\ |X \cap Y| = k - 2\}$$

- Prune candidates by ensuring every $(k-1)$-subset is frequent.

Case Study: Mining Techniques - FP-Growth

Mining Using FP-Growth:

- Build an FP-tree to represent the dataset compactly.
- Extract the conditional pattern base:

$$\{(\beta, \text{count}) \mid \beta \text{ is a prefix path in the FP-tree for a given prefix } \alpha\}$$

- Recursively extract frequent patterns:

$$\text{FPGrowth}(T_\alpha) = \{\alpha \cup \beta \mid \beta \in \text{FPGrowth}(T_\alpha)\}$$

Case Study: Evaluation & Deployment

Evaluation:

- Compute Support, Confidence, Lift, Leverage, and Conviction for each rule.
- Identify high-value rules (e.g., $\text{Support}(\text{Bread} \rightarrow \text{Butter}) = 0.15$, $\text{Confidence} = 0.50, \text{Lift} = 2.5$).

Deployment:

- Use the mined rules to optimize store layouts, personalize promotions, and improve inventory management.

Summary & Key Takeaways

Recap:

- Definition and significance of data mining.
- The complete KDD process: data collection, preprocessing, exploration, mining, evaluation, and deployment.
- Core techniques: classification, clustering, and advanced association rule mining (with detailed math for Apriori and FP-Growth), regression, and anomaly detection.
- Advanced mathematical concepts: Gini index, Information Gain Ratio, Mahalanobis distance, Chi-Square test, candidate generation formulas, and conditional pattern base extraction.

Questions & Discussion

Discussion Prompts:

- Which mathematical technique do you find most applicable to real-world computer science problems?
- What challenges do you foresee when working with large-scale datasets like Instacart's?
- How can we balance technical innovation with ethical considerations in data mining?

Thank You

Marc Reyes

Questions & Comments Welcome