### Study Notes: Data Mining Chapter 2 – Getting to Know Your Data

**Focus Areas: Data Objects, Attribute Types, Statistical Descriptions, Visualization, Similarity Measures**

#### **1. Data Objects and Attribute Types**

* **Data Object**: Represents an entity (e.g., customer, patient) described by attributes.
* **Attribute Types**:
  + **Nominal**: Categories with no order (e.g., hair color, ZIP codes).
    - Use **mode** for central tendency.
  + **Binary**: Two states (0/1).
    - **Symmetric**: Both states equally important (e.g., gender).
    - **Asymmetric**: One state is more important (e.g., medical test results).
  + **Ordinal**: Ordered but differences unknown (e.g., rankings: small, medium, large).
    - Use **median** or **mode**.
  + **Numeric**:
    - **Interval**: Equal intervals, no true zero (e.g., temperature in °C).
    - **Ratio**: True zero (e.g., height, weight).
  + **Discrete vs. Continuous**: Finite vs. infinite possible values.

#### **2. Basic Statistical Descriptions**

* **Central Tendency**:
  + **Mean**: Sensitive to outliers.
  + **Median**: Robust to outliers.
  + **Mode**: Most frequent value.
  + **Midrange**: .
* **Dispersion**:
  + **Range**, **Variance**, **Standard Deviation**, **Quartiles** (, ), **IQR** = .
  + **Boxplot**: Visualizes min, , median, , max, and outliers (1.5 × IQR rule).
* **Distribution**:
  + **Normal Distribution**:
    - within ,
    - within ,
    - within .

#### **3. Data Visualization**

* **Techniques**:
  + **Histograms**: Show frequency distribution.
  + **Scatter Plots**: Identify correlations (positive/negative/uncorrelated).
  + **Boxplots**: Compare distributions across groups.
  + **Parallel Coordinates**: High-dimensional data.
  + **Chernoff Faces**: Multivariate data using facial features.
  + **Tree-Maps**: Hierarchical data as nested rectangles.

#### **4. Similarity and Dissimilarity Measures**

**Key Formulas**:  
1. **Jaccard Coefficient** (binary data):

- : Shared 1s, : Mismatches.

1. **Minkowski Distance** (numeric data):
   * : Manhattan ().
   * : Euclidean ().
   * : Supremum ().
2. **Cosine Similarity** (text/document data):
3. **Z-Score Standardization**:

#### **5. Practice Questions**

1. Compute Jaccard similarity for and .
   * **Answer**: , , → .
2. Calculate Euclidean distance between and .
   * **Answer**:

### How to Use This Text:

1. **Copy** the entire block above.
2. **Paste** into a Markdown editor (e.g., [Typora](https://typora.io/), [Obsidian](https://obsidian.md/)).
3. **Export as PDF** (most editors support this).

If you need a pre-rendered PDF, you can use free tools like [Pandoc](https://pandoc.org/) or [Overleaf](https://www.overleaf.com/) (for LaTeX). Let me know if you need help! 😊

# **Study Notes: Data Mining - Chapter 3**

## **Data Preprocessing**

### **1. Why Preprocess the Data?**

* **Data Quality Measures:**
  + **Accuracy:** Correctness of values.
  + **Completeness:** Availability of all required data.
  + **Consistency:** Consistency across different data sources.
  + **Timeliness:** Data should be up to date.
  + **Believability & Interpretability:** Trustworthiness and ease of understanding.

### **2. Major Data Preprocessing Tasks**

1. **Data Cleaning**
   * Handling **missing values**
     + Ignore the tuple (if missing data is in the class label).
     + Fill with a global constant (e.g., “unknown”).
     + Fill with mean, median, or most probable value (using Bayesian formulas or decision trees).
   * Handling **noisy data** (random errors)
     + **Binning:** Smoothing by bin means, medians, or boundaries.
     + **Regression:** Fit data into regression models.
     + **Clustering:** Detect and remove outliers.
   * Handling **inconsistent data**
     + Use metadata constraints (e.g., age should be positive).
     + Detect duplicate records and resolve conflicts.
2. **Data Integration**
   * **Combining multiple databases, data cubes, or files.**
   * **Entity Identification Problem:** Schema integration (matching attributes from different sources).
   * **Handling Redundancy:** Use correlation analysis to detect redundant attributes.
   * **Handling Data Conflicts:** Standardize measurement units, resolve naming inconsistencies.
3. **Data Reduction**
   * **Dimensionality Reduction:** Reduce the number of attributes.
     + **Principal Component Analysis (PCA)**
     + **Feature Selection**
   * **Numerosity Reduction:** Reduce volume without losing key information.
     + **Regression & Log-Linear Models**
     + **Clustering & Sampling**
   * **Data Cube Aggregation:** Summarizing data at different levels (e.g., quarterly sales vs. yearly sales).
4. **Data Transformation & Discretization**
   * **Normalization (Scaling)**
     + Min-Max Normalization:
     + Z-Score Normalization:
     + Decimal Scaling:
   * **Discretization (Converting continuous data to categorical)**
     + **Binning**
     + **Histogram Analysis**
     + **Clustering**
     + **Decision Tree Analysis**
   * **Concept Hierarchy Generation**
     + Organizing attributes into levels of abstraction (e.g., city → state → country).

### **3. Key Equations**

* **Minkowski Distance** (generalized distance metric)
  + Special cases:
    - Manhattan Distance ((h = 1))
    - Euclidean Distance ((h = 2))
    - Chebyshev Distance ((h ))
* **Cosine Similarity** (used for text and high-dimensional data)
* **Chi-Square Test** (for correlation between nominal attributes)
* where:
  + ( O\_{ij} ) = observed frequency
  + ( E\_{ij} ) = expected frequency
* **Correlation Coefficient (Pearson’s r)**
  + Values range from -1 to 1, where 1 = strong positive correlation, -1 = strong negative correlation.

### **4. Summary**

* **Preprocessing is essential for improving data quality and efficiency.**
* **Common tasks include cleaning, integration, reduction, transformation, and discretization.**
* **Key methods include handling missing/noisy data, normalizing values, detecting outliers, and aggregating data.**
* **Understanding correlation, distance metrics, and similarity measures is critical for feature engineering and data preprocessing.**

Let me know if you need further clarifications! 🚀