Data Mining: Data Preparation

Lecture Notes for Chapter 3

Data Preparation

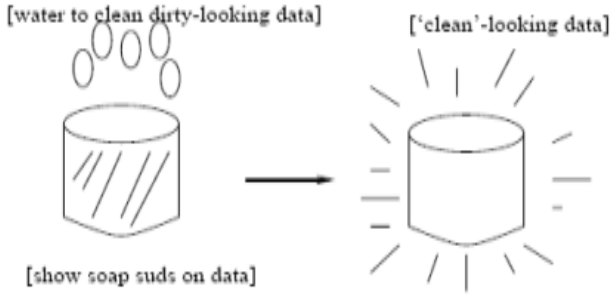
Outline

Unit 3: Data Preparation [6 Hrs.]

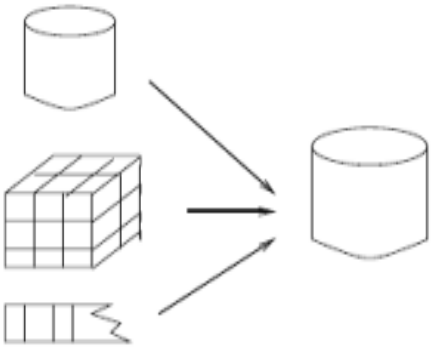
- Data cleaning: Missing values, noisy data, inconsistent data
- Data integration
- Data transformation
- Data reduction
- Discretization and generating concept hierarchies

Forms of data preprocessing

Data Cleaning

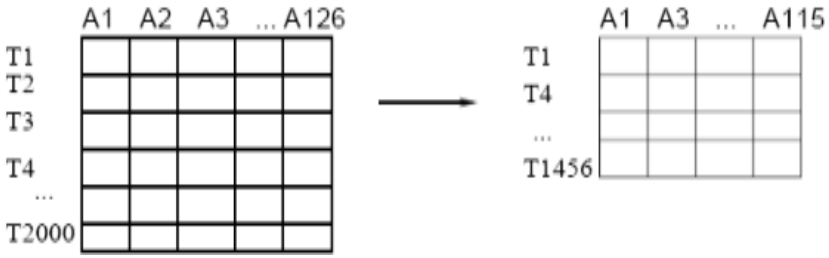


Data Integration



Data Transformation -2, 32, 100, 59, 48 → -0.02, 0.32, 1.00, 0.59, 0.48

Data Reduction



Data Cleaning

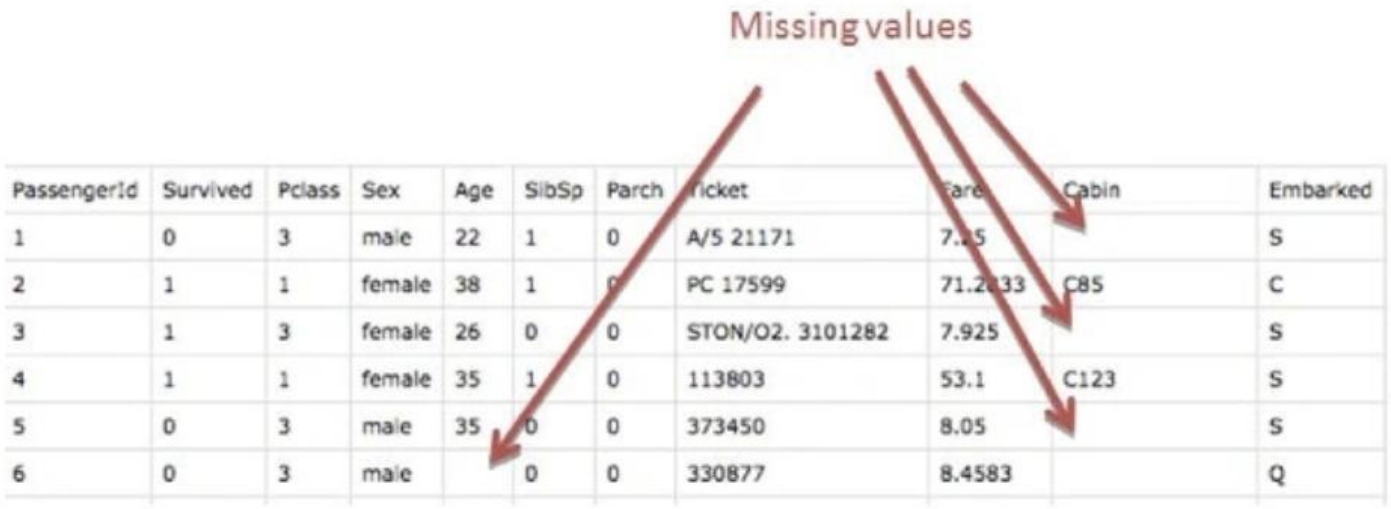
- Removing problems from raw data.

Main Tasks:

- **Handling Missing Values:**
  ➔ Fill, delete, or predict missing entries.
- **Removing Noisy Data:**
  ➔ Smooth data (e.g., using averages) or remove outliers.
- **Fixing Inconsistent Data:**
  ➔ Correct contradictions (e.g., "Male" and "M" treated differently).

What is a Missing Value?

- Missing data is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset.
- Below is a sample of the missing data from the Titanic dataset.
- You can see the columns ‘Age’ and ‘Cabin’ have some missing values.



PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/5 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

How is a Missing Value Represented in a Dataset?

- **NaN (Not a Number): Sometimes,** missing values are represented as NaN.
  - This is the default for libraries like Pandas in Python.

- **NULL or None:** In databases and some programming languages, missing values are often represented as NULL or None.
  - For instance, in SQL databases, a missing value is typically recorded as NULL.

- **Empty Strings:** Sometimes, missing values are denoted by empty strings (""). This is common in text-based data or CSV files where a field might be left blank.

...

- **Special Indicators:** Datasets might use specific indicators like -999, 9999, or other unlikely values to signify missing data.
  - This is often seen in older datasets or specific industries where such conventions were established.

- **Blanks or Spaces:** In some cases, particularly in fixed-width text files, missing values might be represented by spaces or blank fields.

Why is Data Missing From the Dataset?

Types of Missing Values

Figure 1 - Different Types of Missing Values in Datasets

...

1. Missing Completely At Random (MCAR)

- Occurs when the probability of data being missing is uniform across all observations.
- No relationship between missingness and observed or unobserved data.
- The missing data is purely random with no pattern.
- **Example**: In a survey about library books, some overdue book values are missing due to human error in recording.

2. Missing At Random (MAR)

- The probability of missing data depends only on observed data, not on the missing data itself.
- There is a pattern, but it can be explained by variables for which you have complete information.
- **Example**: In a survey, ‘Age’ values might be missing for those who did not disclose their ‘Gender’. Missingness of ‘Age’ depends on ‘Gender’.

...

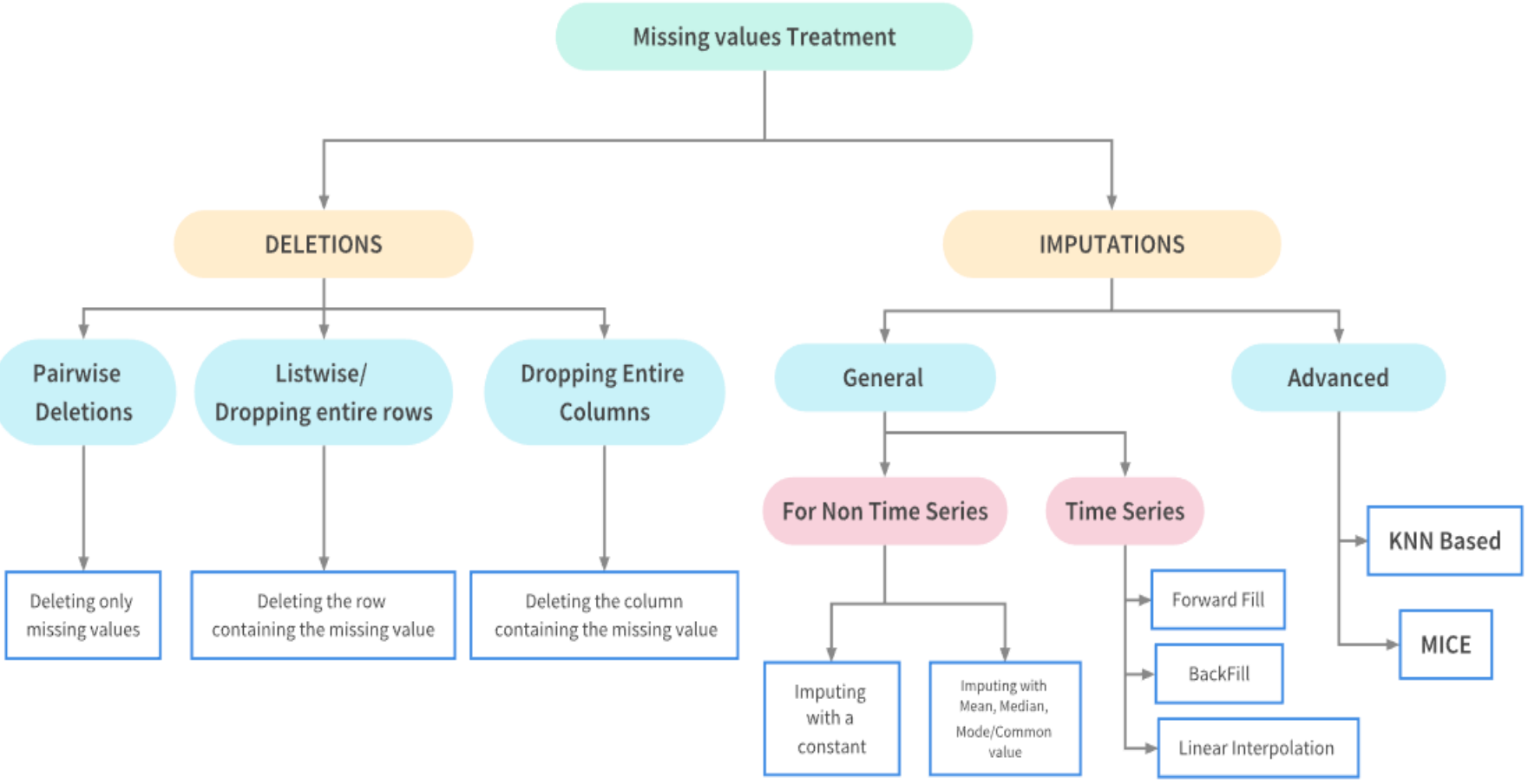
3. Missing Not At Random (MNAR)

- The missingness is related to the unobserved (missing) data itself.
- There is a specific pattern, but it cannot be explained by observed variables.
- **Example**: In a survey about library books, people with more overdue books are less likely to respond.
  - Missing number of overdue books depends on the number itself.

Importance of Understanding Missing Data Types

- Choosing the right strategy to handle missing values depends on the type (MCAR, MAR, MNAR).
- Essential for maintaining the integrity of statistical analyses.
- Techniques like handling missing values, filling missing values, and missing value imputation are crucial.

Treating missing values



Handling Missing Values

1. Deletion

- **Row Deletion**: Remove rows containing missing values.
- **Column Deletion**: Remove columns with too many missing values.
- **Advantages**: Simple and quick.
- **Disadvantages**: Risk of losing too much data, which can weaken your analysis and conclusions.

...

2. Imputation

Replacing missing values with estimated ones. Common imputation techniques include:

- **Mean/Median/Mode Imputation**
  - Replace missing entries with the column's mean, median, or mode.
  - **Advantage**: Fast and simple.
  - **Disadvantage**: Can introduce bias if missingness is not random.

...

- **K-Nearest Neighbors (KNN) Imputation**
  - Find the closest data points (neighbors) based on existing features.
  - Use their values to estimate missing data.
  - **Advantage**: Good when missing values are scattered and data is large.
  - **Disadvantage**: Computationally intensive for very large datasets.

...

- **Model-based Imputation**
  - Build a predictive model (like regression, decision trees, etc.) to estimate missing values based on other features.
  - **Advantage**: Powerful and can be highly accurate.
  - **Disadvantage**: Requires expertise and can be computationally expensive.

How to Impute Missing Values for Categorical Features?

There are two ways to impute missing values for categorical features as follows:

1. Impute the Most Frequent Value

- Use ‘SimpleImputer’
  - and as this is a non-numeric column, we can’t use mean or median, but we can use the most frequent value and constant.

2. Impute the Value “Missing”

- We can impute the value “missing,” which treats it as a separate category.

Noisy and Inconsistent Data

Noisy Data = Random errors or variations.

- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - technology limitation
  - inconsistency in naming convention
- Other data problems which requires data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

...

- Binning method:
  - first sort data and partition into (equi-depth) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human
- Regression
  - smooth by fitting the data into regression functions

Binning

- smooths a sorted data value by consulting the neighborhood", or values around it.
  - The sorted values are distributed into a number of 'buckets', or bins.
  - Because binning methods consult the neighborhood of values, they perform local smoothing.

▪▪▪

- In this technique,
- 1. The data for first sorted
- 2. Then the sorted list partitioned into equi-depth of bins.
- 3. Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
  - a. Smoothing by bin means: Each value in the bin is replaced by the mean value of the bin.
  - b. Smoothing by bin medians: Each value in the bin is replaced by the bin median.
  - c. Smoothing by boundaries: The min and max values of a bin are identified as the bin boundaries.
    - Each bin value is replaced by the closest boundary value.

...

- Example: Binning Methods for Data Smoothing

o Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

o Partition into (equi-depth) bins(equi depth of 3 since each bin contains three values):

- Bin 1: 4, 8, 9, 15
- Bin 2: 21, 21, 24, 25
- Bin 3: 26, 28, 29, 34

...

- Smoothing by bin means:

- Bin 1: 9, 9, 9, 9

- Bin 2: 23, 23, 23, 23

- Bin 3: 29, 29, 29, 29

- Smoothing by bin boundaries:

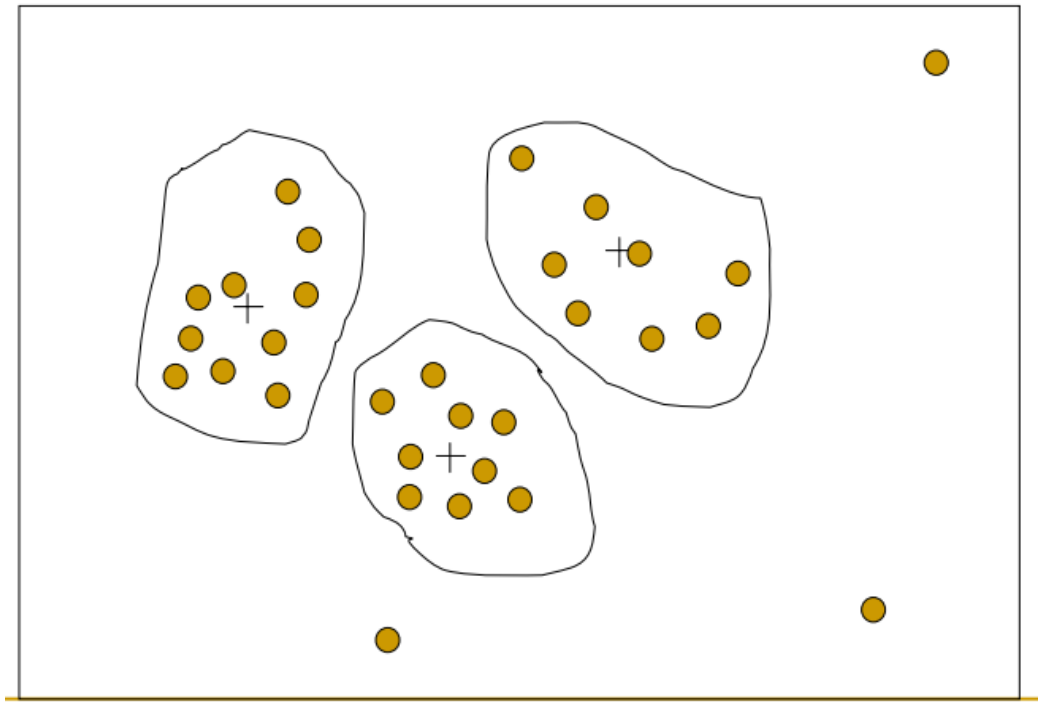
- Bin 1: 4, 4, 4, 15

- Bin 2: 21, 21, 25, 25

- Bin 3: 26, 26, 26, 34

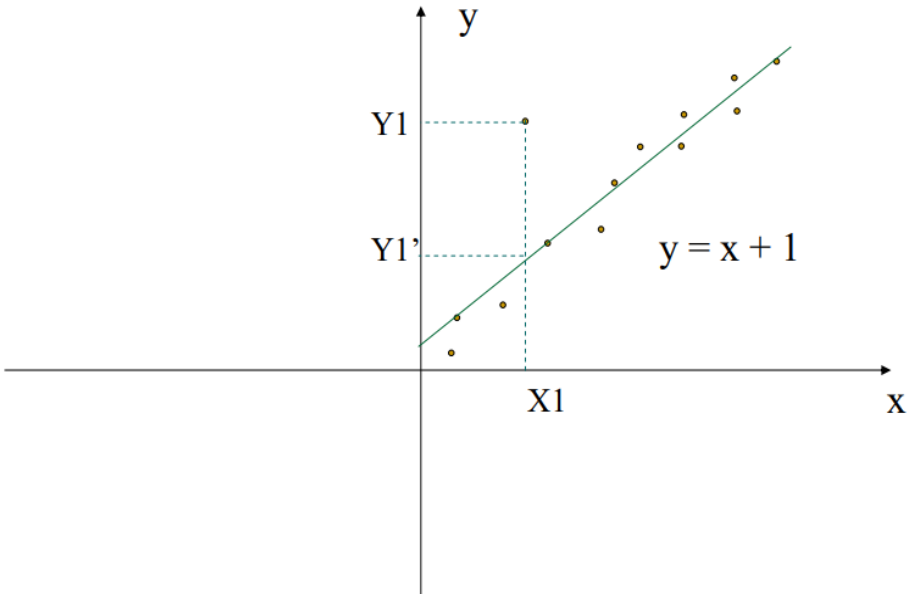
Cluster analysis

- Outliers in the data may be detected by clustering, where similar values are organized into groups, or ‘clusters’.
- Values that fall outside of the set of clusters may be considered outliers.



Regression Analysis

- smooth by fitting the data into regression functions.
- Linear regression involves finding the best of line to fit two variables, so that one variable can be used to predict the other.



...

- Using regression to find a mathematical equation to fit the data helps smooth out the noise.
- **Field overloading:** is a kind of source of errors that typically occurs when developers compress new attribute definitions into unused portions of already defined attributes.
- **Unique rule** is a rule says that each value of the given attribute must be different from all other values of that attribute.
- **Consecutive rule** is a rule says that there can be no missing values between the lowest and highest values of the attribute and that all values must also be unique.
- **Null rule** specifies the use of blanks, question marks, special characters or other strings that may indicate the null condition and how such values should be handled.

Data Integration

- It combines data from multiple sources into a coherent store.
- There are number of issues to consider during data integration
- Issues:
  - **Schema integration**: refers integration of metadata from different sources.
  - **Entity identification problem**: Identifying entity in one data source similar to entity in another table.
    - For example, customer_id in one **db** and customer_no in another **db** refer to the same entity.

…

- **Detecting and resolving data value conflicts:** Attribute values from different sources can be different due to different representations, different scales. E.g. metric vs. British units
- **Redundancy:** is another issue while performing data integration. Redundancy can occur due to the following reasons:
  - **Object identification:** The same attribute may have different names in different database.
  - **Derived Data:** one attribute may be derived from another attribute.

Handling redundant data in data integration

1. **Correlation analysis**

   a. **For numeric data**

   Some redundancy can be identified by correlation analysis.

   The correlation between two variables A and B can be measured by

$$r_{A,B} = \frac{\Sigma(A - \overline{A})(B - \overline{B})}{(n-1)\sigma_A \sigma_B}$$

$\overline{A}$, $\overline{B}$ are respective mean values of A and B

σ_A, σ_B are respective standard deviation of A and B

n is the number of tubles

...

- The result of the equation is > 0, then A and B are positively correlated, which means the value of A increases as the values of B increases. The higher value may indicate redundancy that may be removed.
- The result of the equation is = 0, then A and B are independent and there is no correlation between them.
- If the resulting value is < 0, then A and B are negatively correlated where the values of one attribute increase as the value of one attribute decrease which means each attribute may discourages each other.
- also called **Pearson's product moment coefficient**

For categorical data

- X^2 (chi-square) test

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

- The larger the X^2 value, the more likely the variables are related
- The cells that contribute the most to the X^2 value are those whose actual count is very different from the expected count
- Correlation does not imply causality
  - # of hospitals and # of car-theft in a city are correlated
  - Both are causally linked to the third variable: population

Example:

	Play chess	Not play chess	Sum (row)
Like science fiction	250(90)	200(360)	450
Not like science fiction	50(210)	1000(840)	1050
Sum(col.)	300	1200	1500

- X^2 (chi-square) calculation (numbers in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840} = 507.93$$

- It shows that like_science_fiction and play_chess are correlated in the group

Data transformation

- Normalization:
  - Scaling attribute values to fall within a specified range.
    - Example: to transform V in [min, max] to V' in [0,1], apply V'=(V-Min)/(Max-Min)
  - Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers):
    - V'=(V-Mean)/StDev
- Aggregation: moving up in the concept hierarchy on numeric attributes.
- Generalization: moving up in the concept hierarchy on nominal attributes.
- Attribute construction: replacing or adding new attributes inferred by existing attributes.

...

- Data transformation can involve the following:
- **Smoothing:** which works to remove noise from the data.
- **Aggregation:** where summary or aggregation operations are applied to the data.
  - For example, the daily sales data may be aggregated so as to compute weekly and annual total scores.

...

- **Generalization of the data:** where low-level or “primitive” (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher-level concepts, like city or country.

- **Normalization:** where the attribute data are scaled so as to fall within a small specified range, such as −1.0 to 1.0, or 0.0 to 1.0.

- **Attribute construction (feature construction):** this is where new attributes are constructed and added from the given set of attributes to help the mining process.

Normalization

- In which data are scaled to fall within a small, specified range, useful for classification algorithms involving neural networks, distance measurements such as nearest neighbor classification and clustering.
- There are 3 methods for data normalization. They are:

1) min-max normalization
2) z-score normalization
3) normalization by decimal scaling

Data Reduction

1. Reducing the number of attributes
   - Data cube aggregation: applying roll-up, slice or dice operations.
   - Removing irrelevant attributes: attribute selection (filtering and wrapper methods), searching the attribute space (see Lecture 5: Attribute-oriented analysis).
   - Principle component analysis (numeric attributes only): searching for a lower dimensional space that can best represent the data..

...

2. Reducing the number of attribute values
   - Binning (histograms): reducing the number of attributes by grouping them into intervals (bins).
   - Clustering: grouping values in clusters.
   - Aggregation or generalization

3. Reducing the number of tuples
   - Sampling

Discretization and generating concept hierarchies

- refers to the organization of data into a tree-like structure,
  - where each level of the hierarchy represents a concept that is more general than the level below it.
- This hierarchical organization of data allows for more efficient and effective data analysis, as well as the ability to drill down to more specific levels of detail when needed.
- used to organize and classify data in a way that makes it more understandable and easier to analyze.

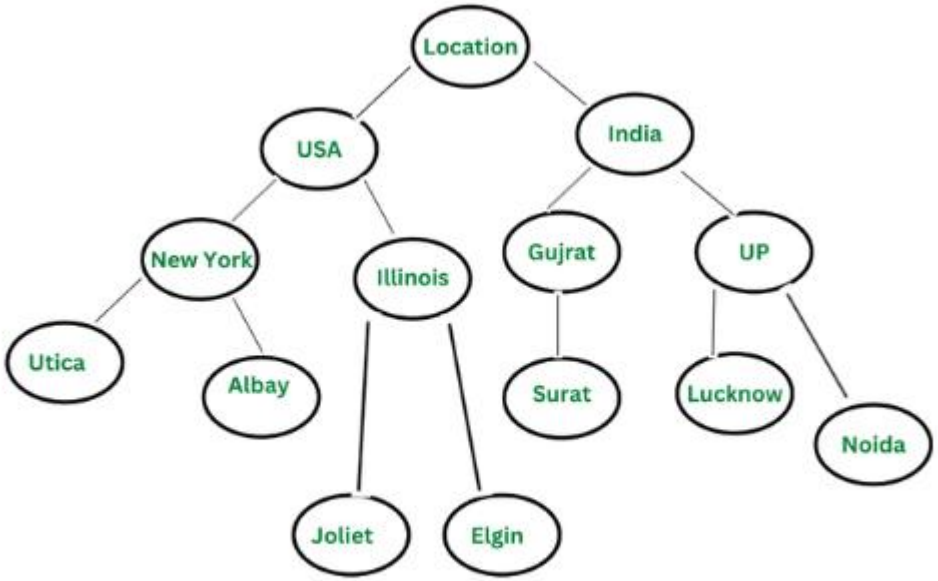
...

- The main idea behind the concept of hierarchy is that the same data can have different levels of granularity or levels of detail and that by organizing the data in a hierarchical fashion, it is easier to understand and perform analysis.

Concept hierarchy for the dimension location, where the user can easily retrieve the data.

In order to evaluate it easily the data is represented in a tree-like structure.

The top of the tree consists of the main dimension location and further splits into various sub-nodes. The root node is located, and it further splits into two nodes countries ie. USA and India...



Concept Hierarchy for Dimension Location

Types of Concept Hierarchies

1. Schema Hierarchy

- Organizes database schema in a logical and meaningful structure.
- Groups similar objects: tables, attributes, and relationships.
- Helps in integrating data from multiple sources into a unified view.
- Commonly used in data warehousing environments.

2. Set-Grouping Hierarchy

- Based on **set theory**: each set is defined by its membership in other sets.
- Useful for:
  - Data cleaning
  - Data preprocessing
  - Data integration
- Helps identify and remove:
  - Outliers
  - Noise
  - Inconsistencies

3. Operation-Derived Hierarchy

- Data is organized by applying **transformations or operations**.
- Hierarchy proceeds **top-down**:
  - Higher levels = more abstract/general views.
- Common in:
  - Clustering
  - Dimensionality reduction
- Operations include:
  - Aggregation
  - Normalization
  - Other statistical transformations

...

4. Rule-Based Hierarchy

- Built by applying **rules or conditions** to data.
- Supports tasks like:
  - Classification
  - Decision-making
  - Data exploration
- Assigns **class labels or decisions** based on attribute characteristics.
- Helps discover **patterns and relationships** in data.

Need for Concept Hierarchy in Data Mining

- **Improved Data Analysis**
  - Simplifies and organizes complex data.
  - Facilitates pattern and trend identification.
  - Uncovers hidden or unexpected insights.
- **Enhanced Data Visualization and Exploration**
  - Organizes data in a **tree-like** structure.
  - Enables users to **drill down** into specific details.
  - Useful for interactive **dashboards** and **reports**.

...

- **Improved Algorithm Performance**
  - Hierarchical structuring improves data accessibility.
  - Leads to **faster** and **more accurate** data mining outcomes.
- **Supports Data Cleaning and Preprocessing**
  - Helps **detect outliers** and **remove noise**.
  - Improves data quality before mining begins.
- **Captures Domain Knowledge**
  - Represents **domain expertise** in a structured form.
  - Helps in better understanding of **problem domains**.

Applications of Concept Hierarchy

- **Data Warehousing**
  - Integrates multiple sources into a unified structure.
  - Improves reporting and analysis consistency.
- **Business Intelligence**
  - Supports decision-making via trend and pattern discovery.
  - Example: Customer behavior analysis for product development.
- **Online Retail**
  - Organizes products into categories and subcategories.
  - Enhances **user navigation** and product discovery.

...

- **Healthcare**
  - Groups patients by diagnosis, treatment, or outcomes.
  - Identifies patterns to **improve treatments**.
- **Natural Language Processing (NLP)**
  - Organizes unstructured text into themes or topics.
  - Aids in **information extraction** from large corpora.
- **Fraud Detection**
  - Analyzes financial data in hierarchical structures.
  - Helps detect **fraud patterns** or suspicious activity.

...

Three types of attributes:

- Nominal — values from an unordered set, e.g., color, profession
- Ordinal — values from an ordered set, e.g., military or academic rank
- Continuous — real numbers, e.g., integer or real numbers

...

- There are five methods for numeric concept hierarchy generation.
- These include:

1. binning,
2. histogram analysis,
3. clustering analysis,
4. entropy-based Discretization, and
5. data segmentation by “natural partitioning".

...

- An information-based measure called “entropy" can be used to recursively partition the values of a numeric attribute A, resulting in a hierarchical Discretization.

  - Given a set of samples S, if S is partitioned into two intervals S_1 and S_2 using boundary T, the information gain after partitioning is

$$I(S,T) = \frac{|S_1|}{|S|} Entropy(S_1) + \frac{|S_2|}{|S|} Entropy(S_2)$$

  - Entropy is calculated based on class distribution of the samples in the set. Given m classes, the entropy of S_1 is

$$Entropy(S_1) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

    where p_i is the probability of class i in S_1
  - The boundary that minimizes the entropy function over all possible boundaries is selected as a binary discretization
  - The process is recursively applied to partitions obtained until some stopping criterion is met
  - Such a boundary may reduce data size and improve classification accuracy

Segmentation by Natural Partitioning

The Problem:

We have profit numbers from different company branches in 1997.

These profits range **from big losses (-$351,976) to big profits ($4,700,896.50)**.

We want to **group these numbers into levels or categories** (called a **concept hierarchy**) so that we can:

- Better understand the data
- Summarize and analyze it
- Use it in data mining

- A simply 3-4-5 rule can be used to segment numeric data into relatively uniform, "natural" intervals.
  - If an interval covers 3, 6, 7 or 9 distinct values at the most significant digit, partition the range into 3 equi-width intervals
  - If it covers 2, 4, or 8 distinct values at the most significant digit, partition the range into 4 intervals
  - If it covers 1, 5, or 10 distinct values at the most significant digit, partition the range into 5 intervals

Solution steps

Step 1: Identify Range

We find:

- **MIN** profit = **–$351,976** (worst branch)
- **MAX** profit = **$4,700,896.50** (best branch)

Also, we ignore the most extreme 5% on both ends (outliers) and focus on:

- **LOW** = –$159,876
- **HIGH** = $1,838,761
- This range covers the **middle 90%** of data.

Step 2: Round the Range to Clean Numbers

We round LOW down and HIGH up to **nearest million**:

- LOW → **–$1,000,000**
- HIGH → **$2,000,000**
- We now have a rounded, clean range: **–1M to 2M**

Step 3: Apply the 3-4-5 Rule

The **3-4-5 rule** says:

- If the range spans **3 values**, divide into **3 parts** (or 4 or 5 if more detail is needed)

Here:

- From –1M to 2M = span of 3 million
- So, we create **3 equal intervals**:
  - (–$1,000,000 to $0]
  - ($0 to $1,000,000]
  - ($1,000,000 to $2,000,000]
- This is the **top level of the hierarchy**.

Step 4: Adjust the Edges

We check if the original MIN and MAX fit inside these 3 intervals.

- MIN = –$351,976 → **Doesn't go down to –$1,000,000**, so shrink first interval
  - Round down to **–$400,000**
  - New first interval: **(–$400,000 to $0]**
- MAX = $4.7M → **Exceeds $2,000,000**, so we **add another interval**
  - Round up to $5,000,000
  - Add a new interval: **($2,000,000 to $5,000,000]**

Now the **top level** has **4 intervals**:

1. (–$400,000 to $0]
2. ($0 to $1,000,000]
3. ($1,000,000 to $2,000,000]
4. ($2,000,000 to $5,000,000]

Step 5: Break Each Interval Down (Lower Levels)

We now create **smaller sub-intervals** within each top-level category:

- (–$400,000 to $0] → divide into **4 parts**:
  - (–400K to –300K], (–300K to –200K], (–200K to –100K], (–100K to 0]
- ($0 to $1,000,000] → divide into **5 parts**:
  - ($0 to $200K], $200K to $400K], etc.
- ($1M to $2M] → also 5 parts:
  - $1M–$1.2M], $1.2M–$1.4M], ..., $1.8M–$2M]
- ($2M to $5M] → 3 parts:
  - $2M–$3M], $3M–$4M], $4M–$5M]

Thank you for your time and attention. ☺