CS313: Data Mining and Application

Data Pre-Processing

Vo Nguyen Le Duy

University of Information Technology, VNUHCM
RIKEN, Japan

Fall 2025

Agenda

- Data Pre-Processing: An Overview
- Statistical Descriptions of Data
- Data Cleaning
- Data Integration
- Data Transformation and Discretization
- Data Reduction
- Summary

Data Pre-Processing: An Overview

- Real data is noisy, incomplete and inconsistent
- Low-quality data $\Rightarrow$ low-quality mining results
- Example: please find issues in the following table

Student ID	Year	Mid-term Score	Final Score
101	2024	9	8.5
102	2024	12	8
103	2024		9.5
104	2024	7.5	
101	2024	8	8.5

Data Pre-Processing: An Overview

- Pre-process data

  $\Rightarrow$ Improve the quality of the data

  $\Rightarrow$ Improve the quality of mining results

  $\Rightarrow$ Improve the efficiency and ease of the mining process

- Data quality:
  - Accuracy: no errors, or values that deviate from the expected
  - Completeness: no missing data
  - Consistency
  - Timeliness
  - Believability: how much the data are trusted by users
  - Interpretability: how easy the data are understood

Inaccurate, Incomplete, and Inconsistent Data

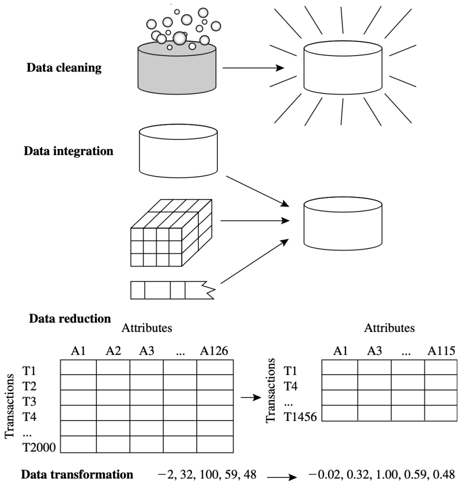
Position	Experience	Salary
CEO	10+	15,000
CTO	10+	10,000
Tech Lead	5 ~ 10	5,000
Data Scientist	1 ~ 3	12,000
Data Engineer	1 ~ 3	NULL
Product Manager	NULL	3,000
Data Scientist	1 ~ 3	2,800

- Reasons
  - Data collection devices may be defective
  - Users purposely submit incorrect data values
  - Errors in data transmission
  - Lost information
  - Human errors
- Other examples?

Timeliness, Believability, and Interpretability

- Timeliness
  - Sales records
  - Flight status
  - Student scores
- Believability
  - Health monitoring app
  - Scientific results
  - Clinical trial data
- Interpretability
  - Accounting codes

Major Tasks in Data Preprocessing



Han et al. (2022)

Statistical Descriptions of Data

- Measuring the central tendency
  - Mean: Let $x_1, x_2, ..., x_N$ be the set of N observations of a random variable X, the mean of this set of values is
$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N}$$
  - Weighted average: If each value x_i is associated with a weight w_i, $\forall i \in \{1, 2, ..., N\}$,
$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + ... + w_N x_N}{w_1 + w_2 + ... + w_N}$$
  - Median:
$$median = \begin{cases} x_{(N+1)/2} & \text{if } N \text{ is odd} \\ \frac{x_{N/2} + x_{N/2+1}}{2} & \text{if } N \text{ is even} \end{cases}$$
  - Mode:
$$mode = the\ value\ that\ occurs\ most\ frequently\ in\ the\ set$$

- Measuring the dispersion of data
  - Range: the difference between the largest and smallest values
  - Quartiles:
    - 1^{st} quartile (Q1): the 25^{th} percentile
    - 2^{nd} quartile (Q2): the 50^{th} percentile
    - 3^{rd} quartile (Q3): the 75^{th} percentile
  - Interquartile range:

$$IQR = Q3 \text{ - } Q1$$

  - Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Example: find Q1, Q2, Q3, and IQR of the following set

$$\mathcal{S} = \{2,\ 2,\ 4,\ 5,\ 5,\ 6,\ 8,\ 9,\ 10,\ 12,\ 14\}$$

Statistical Descriptions of Data

- Measuring the dispersion of data
  - Range: the difference between the largest and smallest values
  - Quartiles:
    - 1^{st} quartile (Q1): the 25^{th} percentile
    - 2^{nd} quartile (Q2): the 50^{th} percentile
    - 3^{rd} quartile (Q3): the 75^{th} percentile
  - Interquartile range:

$$IQR = Q3 \text{ - } Q1$$

  - Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

- Example: find Q1, Q2, Q3, and IQR of the following set

$$\mathcal{S} = \{2,\ 2,\ 4,\ 5,\ 5,\ 6,\ 8,\ 9,\ 10,\ 12,\ 14\}$$

$$\mathcal{S} = \{2,\ 2,\ \underbrace{4}_{Q1},\ 5,\ 5,\ \underbrace{6}_{Q2},\ 8,\ 9,\ \underbrace{10}_{Q3},\ 12,\ 14\}$$

Data Cleaning

- Fill in missing values
- Smooth out noise
- Correct inconsistencies

Fill in Missing Values

- Ignore the tuple
  - Deleting all tuples with missing data
- Fill in the missing value manually
  - This may be the most reliable way. However, it is time and effort consuming
- Fill in automatically
  - Use a global constant
  - Use the mean or median of the attribute value
  - Use the mean or median of the attribute value **within the same class**
  - Use the most probable value
    - Predict the missing values by regression
- Note: a missing value may not imply an error in the data

Smooth Out Noise

- Binning
  - Data values are sorted and distributed into a number of bins
  - Smoothing: smoothing by bin means, smoothing by bin medians, smoothing by bin boundaries

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Han et al. (2022)

- Regression and outlier analysis

Correct Inconsistencies

- Correct inconsistent data manually
- Correct inconsistent data automatically

Data Integration

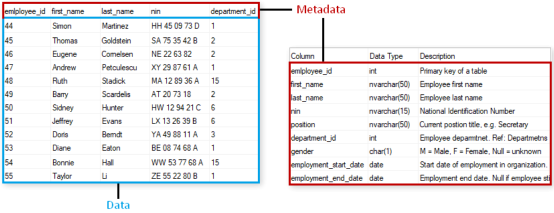
- Data integration: the merging of data from multiple data stores
    - Entity identification problem
    - Redundancy and correlation analysis
    - Tuple duplication
    - Data value conflict
- Careful integration

  $\Rightarrow$ Reduce and avoid redundancies and inconsistencies

  $\Rightarrow$ Improve the accuracy and speed of the data mining process

Entity Identification Problem

- How can equivalent entities from multiple data sources be matched up?
- Examples
  - *customer_id* in one database (D1) and *cust_number* in another database (D2)
  - *R&D* in D1 and *Research and Development* in D2
  - *Male* and *Female* in D1 and *M* and *F* in D2
- Metadata is helpful



emlployee_id	first_name	last_name	nin	department_id
44	Simon	Martinez	HH 45 09 73 D	1
45	Thomas	Goldstein	SA 75 35 42 B	2
46	Eugene	Comelsen	NE 22 63 82	2
47	Andrew	Petculescu	XY 29 87 61 A	1
48	Ruth	Stadick	MA 12 89 36 A	15
49	Barry	Scardelis	AT 20 73 18	2
50	Sidney	Hunter	HW 12 94 21 C	6
51	Jeffrey	Evans	LX 13 26 39 B	6
52	Doris	Berndt	YA 49 88 11 A	3
53	Diane	Eaton	BE 08 74 68 A	1
54	Bonnie	Hall	WW 53 77 68 A	15
55	Taylor	Li	ZE 55 22 80 B	1

Column	Data Type	Description
emlployee_id	int	Primary key of a table
first_name	nvarchar(50)	Employee first name
last_name	nvarchar(50)	Employee last name
nin	nvarchar(15)	National Identification Number
position	nvarchar(50)	Current postion title, e.g. Secretary
department_id	int	Employee deparmtnet. Ref: Departmetns
gender	char(1)	M = Male, F = Female, Null = unknown
employment_start_date	date	Start date of employment in organization.
employment_end_date	date	Employment end date. Null if employee sti

https://dataedo.com/kb/data-glossary/what-is-metadata

Redundancy and Correlation Analysis

- An attribute may be redundant if it can be "derived" from another attribute or set of attributes
- Some redundancies can be detected by correlation analysis
  - Given two attributes A and B, correlation analysis measures how strongly A implies B
  - For numeric data, we can use the *correlation coefficient* and *covariance*
  - For nominal data, we use the χ^2 (*chi-square*) test

- Consider two numeric attributes A and B, and a set of n observations $\{(a_1, b_1), ..., (a_n, b_n)\}$,

$$Cov(A, B) = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{n},$$

where $\bar{a} = \frac{\sum_{i=1}^{n} a_i}{n}$ and $\bar{b} = \frac{\sum_{i=1}^{n} b_i}{n}$

- The correlation coefficient is then calculated by

$$r_{A,B} = \frac{Cov(A, B)}{\sigma_A \sigma_B},$$

where σ_A and σ_B are the standard deviations of A and B, respectively

  - $r_{A,B} \in [-1, 1]$
  - If $r_{A,B} > 0$, then A and B are positively correlated
  - If $r_{A,B} = 0$, then A and B are independent
  - If $r_{A,B} < 0$, then A and B are negatively correlated

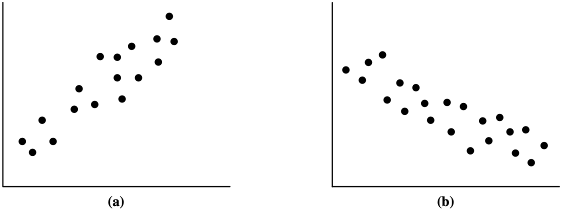
Example

Time point	*AllElectronics*	*HighTech*
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

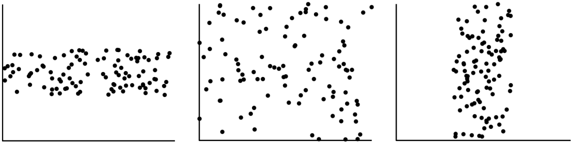
Han et al. (2022)

- Calculate *Cov(AllElectronics, HighTech)*

Covariance and Correlation Coefficient for Numerical Data



Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.



Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.

Han et al. (2022)

χ^2 Test for Nominal Data

- Analyze the correlation between two discrete attributes A and B
  - A has c distinct values: $a_1, a_2, ..., a_c$
  - B has r distinct values: $b_1, b_2, ..., b_r$
  - Hypothesis test:

$$\mathrm{H}_0 : A \text{ and } B \text{ are independent} \quad \text{vs.} \quad \mathrm{H}_1 : A \text{ and } B \text{ are correlated}$$

  - The χ^2 value is computed as

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

    where o_{ij} is the *observed frequency* of the *joint* event $(A = a_i, B = b_j)$ and e_{ij} is the *expected frequency* of $(A = a_i, B = b_j)$, i.e.,

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}$$

  - The test is based on a significance level α, with $(r - 1) \times (c - 1)$ degrees of freedom

Example

- Are *gender* and *preferred_reading* correlated?

	male	*female*	*Total*
fiction	250	200	450
non_fiction	50	1000	1050
Total	300	1200	1500

Han et al. (2022)

Example (cont.)

- Check Chi-square distribution table

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322

www.statology.org

- For 1 degree of freedom, the χ^2 value needed to reject the H_0 at $\alpha = 0.001$ is 10.828
- Since our computed value is above 10.828, we can reject H_0 and conclude that the two attributes are (strongly) correlated

Tuple Duplication

- Duplication should also be detected at the tuple level
- The use of denormalized tables is another source of data redundancy

Data Value Conflict Detection and Resolution

- For a given real-world entity, attribute values may vary across different sources due to differences in representation, scaling, or encoding
  - Representation: "2022/12/14" vs. "14/12/2022"
  - Scaling: GPA $[0, 4]$ vs GPA $[0, 10]$
  - Encoding: "pass" and "fail" vs. 1 and 0

Data Transformation and Discretization

- Data transformation is the process in which the data are transformed into form appropriate for mining
  - Smoothing: binning, regression, and clustering
    $\Rightarrow$ remove noise from the data
  - Attribute construction: new attributes are constructed to help the mining process
  - Aggregation: summary or aggregation operations are applied to the data (e.g., average, count, max, min, sum)
    - The daily sales data is aggregated to compute monthly and annual total amounts.
  - Normalization: the attribute data are scaled to fall within a smaller range
  - Discretization: raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0 - 10, 11 - 20, etc.) or conceptual labels (e.g., youth, adult, senior)
  - Concept hierarchy generation

Normalization

- Min-max normalization
  - $v_A \in [min_A, max_A]$
  - Min-max normalization maps v_A to v'_A in the range $[new\_min_A, new\_max_A]$ by computing

$$v'_A = \frac{v_A - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- Example: suppose *income* $\in [500,\ 5,000]$, please perform min-max normalization to maps a value of 1000 for *income* to the range $[0.0, 1.0]$

Normalization

- z-score normalization
  - Let $\bar{v}_A$ and σ_A are the mean and standard deviation of attribute A, respectively. Then,

$$v'_A = \frac{v_A - \bar{x}_A}{\sigma_A}$$

  - Example: suppose that the mean and variance of the values for the attribute income are $2,750$ and 2500, please perform z-score normalization for a value of $2,900$

Discretization

- Binning
- Cluster analysis
- Decision tree
- Correlation analyses
- Concept hierarchy generation

Data Reduction

- Complex data analysis and mining on huge amounts of data can take a long time
  $\Rightarrow$ Making the analysis impractical or infeasible
- Data reduction: reduce the representation of the data set, yet closely maintains the integrity of the original data
  - Principal components analysis
  - Attribute subset selection
  - Regression
  - Histograms
  - Clustering
  - Sampling

References

- Jiawei Han, Jian Pei, and Hanghang Tong. *Data mining: concepts and techniques*. Morgan kaufmann, 2022.
- David J. Hand, Heikki Mannila and Padhraic Smyth. *Principles of Data Mining*. The MIT Press, 2001.

Q & A

duyvnl@uit.edu.vn