4-5 주차. 데이터 전처리

과목명: 데이터사이언스

AI융합학부 박건우

Contents

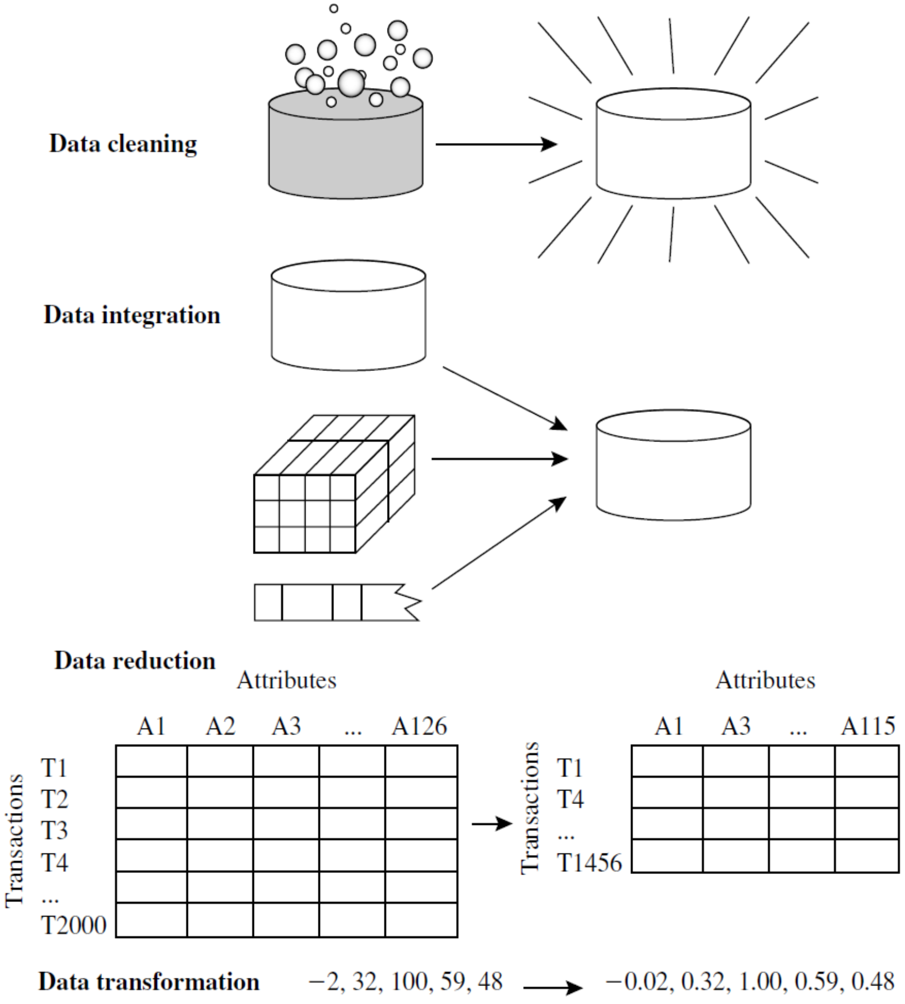
Why Preprocess the Data?

- In many cases, data quality is too low to be used for being mined
- **We aim to improve data quality through preprocessing steps**

- A multidimensional view for data quality
  - **Accuracy (정확도)**: correct or wrong, accurate or not
  - **Completeness (완전성)**: not recorded, unavailable, …
  - **Consistency (일관성)**: some modified but some not, dangling, …
  - **Timeliness (시가적절성)**: timely update?
  - **Believability (신뢰성)**: how trustable the data are correct?
  - **Interpretability (해석가능성)**: how easily the data can be understood?
- Data quality depends on the intended use of the data!

Major Tasks in Data Preprocessing

- **Data cleaning (데이터 정제)**: Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration (데이터 통합)**: Integration of multiple databases or files
- **Data reduction (데이터 축소)**: Dimensionality reduction, Numerosity reduction
- **Data transformation** (데이터 변환) and **discretization** (이산화)



Contents

Data Cleaning (데이터 정제)

- Data in the Real World Is Dirty: Lots of potentially incorrect data
  - incomplete (불완전): lacking attribute values, lacking certain attributes of interest, or containing only aggregate data. e.g., Occupation=" " (missing data)
  - noisy (잡음 있는): containing noise, errors, or outliers. e.g., Salary="−10" (an error)
  - inconsistent (일관성 없는): containing discrepancies in codes or names, e.g.,
    - Age="42", Birthday="03/07/2010"
    - Was rating "1, 2, 3", now rating "A, B, C"
    - discrepancy between duplicate records
  - intentional (의도적인) (e.g., disguised missing data)
    - Jan. 1 as everyone's birthday?

Incomplete (Missing) Data - 결측치

- Data is not always available. e.g., many tuples have no recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - data not entered due to misunderstanding
- In some cases, a missing value may not imply an error in the data!
  - Driver's license number for a credit card application
  - Some applicants may not have a license number
- Missing data may need to be inferred for further analysis

How to Handle Missing Data?

- **Ignore the tuple**: usually done when class label is missing — not effective when the % of missing values per attribute varies considerably
- **Fill in the missing value manually**: tedious + infeasible?
- **Fill in it automatically with**
  - a global constant: e.g., “unknown” or $-\infty$
  - the attribute mean
  - the attribute mean for all samples belonging to the same class
  - the most probable value: by regression, Bayesian methods or decision tree induction

Noisy Data

- **Noise (잡음)**: a random error or variance in a measured variable
- **Incorrect attribute values** may be due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
- **Other data problems** which require data cleaning
  - duplicate records
  - incomplete data
  - inconsistent data

How to Handle Noisy Data?

- **Binning**
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries,
  - will be covered later in the slides
- **Regression**
  - smooth by fitting the data into regression functions
- **Clustering**
  - detect and remove outliers
- **Combined computer and human inspection (human-in-the-loop)**
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data Cleaning as a Process

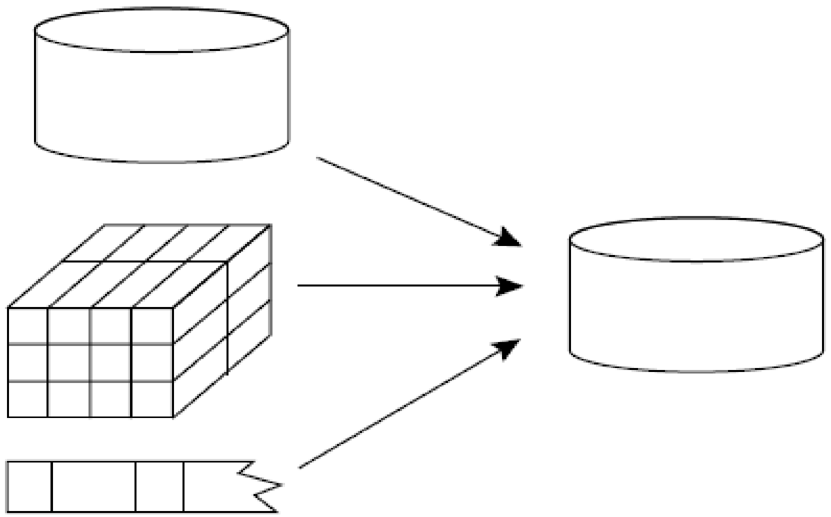
- **How can we proceed with discrepancy detection?** There is no single answer.
- As a starting point, use knowledge about data (metadata: data about data). e.g., domain, range, dependency, distribution
  - What are the data type and domain of each attribute? What are the acceptable values for each attribute?
  - Are the data skewed or symmetric? Do all values fall within the expected range?
- You may write your own scripts to find noise, outliers, unusual values that need investigation.

Contents

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Integration (데이터 통합)

- Combines data from multiple sources into a coherent store
- The same attribute or object may have different names in different databases
- e.g., How can a data scientist be sure that *customer_id* in database A and *cust_number* in database B are the same?
- **Redundancy detection (중복성 탐지)**
  - *Derivable data*: One attribute may be a "derived" attribute in another table,
    e.g., annual revenue, monthly revenue on average
  - Redundancy can be detected by correlation analysis and covariance analysis



χ^2 (Chi-square) Test for Nominal Data

Step 1. Contingency table (분할표) 구성

Sex:	Democrats	Republicans
Male	23	31
Female	28	22

Step 2. Chi-square statistic 계산

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

- o_{ij}: the observed frequency (실제 수치) of the join event (A_i, B_j)
- e_{ij}: the expected frequency of (A_i, B_j), computed as $\frac{count(A=a_i)\times count(B=b_j)}{n}$
- n: the number of data tuples, r: number of rows, c: number of columns

- Interpretation
  - The larger the χ^2 value, the more likely the variables are related
  - The χ^2 statistic tests the hypothesis that A and B are independent, based on a significance level with $(r-1)\times(c-1)$ degrees of freedom

Chi-Square Calculation: An 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

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086

- χ^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

Correlation Coefficient for Numeric Data

- Correlation coefficient (also called **Pearson's correlation coefficient**)

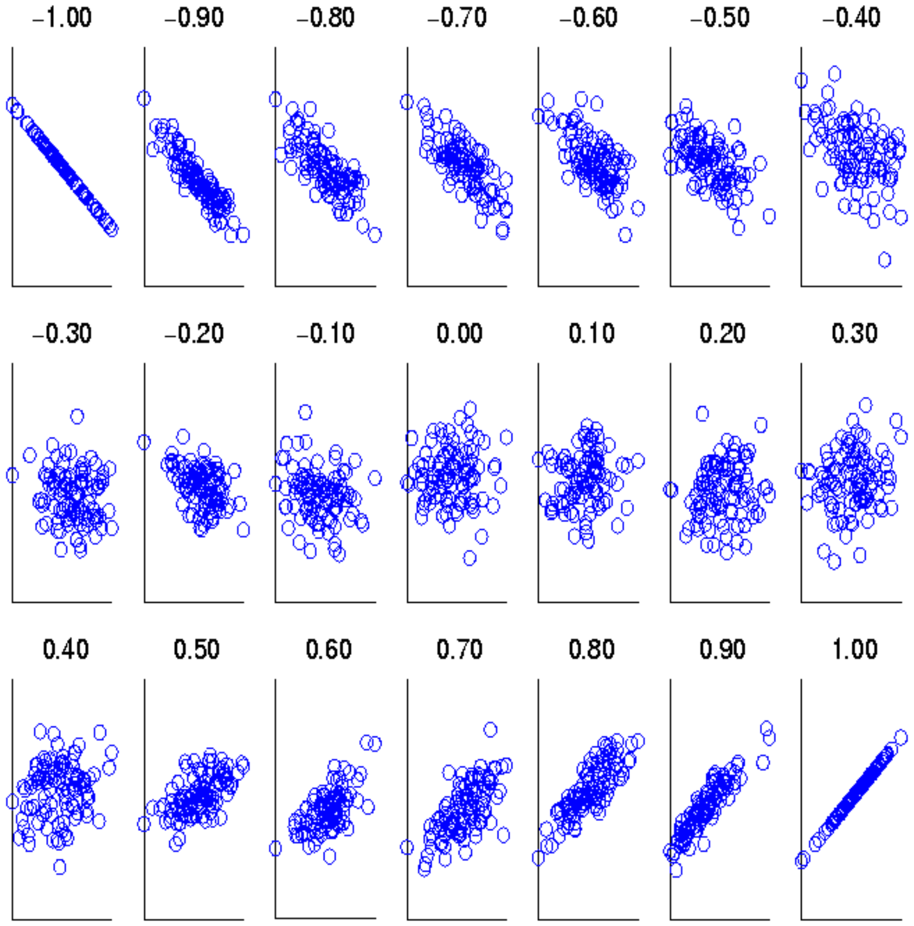
$$r_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^{n}[(x_\mathrm{i}-\mu_X)(y_\mathrm{i}-\mu_Y)/n]}{\sigma_X \sigma_Y}$$

  where X and Y are the respective random variables for two attributes and n is the number of tuples in a dataset.

- If $0 < r_{X,Y} < 1$, X and Y are positively correlated (X's values increase as Y's). The higher, the stronger correlation.
- $r_{X,Y} = 0$; independent; $-1 < r_{X,Y} < 0$: negatively correlated

Visually Evaluating Correlation Coefficient

- Using scatter plots



Scatter plots showing the similarity from –1 to 1.

Covariance (공분산) for Numeric Data

- Covariance is similar to correlation

$$Cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum_{i=1}^{n}[(x_\mathrm{i}-\mu_X)(y_\mathrm{i}-\mu_Y)]}{N}$$

$$r_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y} = \frac{\sum_{i=1}^{N}[(x_\mathrm{i}-\mu_X)(y_\mathrm{i}-\mu_Y)/N]}{\sigma_X\sigma_Y} = \frac{Cov(X,Y)}{\sigma_X\sigma_Y}$$

- **Positive covariance** ($Cov(X,Y) > 0$):
  X and Y both tend to be larger than their expected values.
- **Negative covariance** ($Cov(X,Y) < 0$):
  If X is larger than its expected value, Y is likely to be smaller than its expected value.
- **Independence**: $Cov(X,Y) = 0$ but the converse is not true:
  - Some pairs of random variables may have a covariance of 0 but are not independent.
  - Only under some additional assumptions (e.g., the data follow multivariate normal distributions), a covariance of 0 imply independence

Covariance: An Example

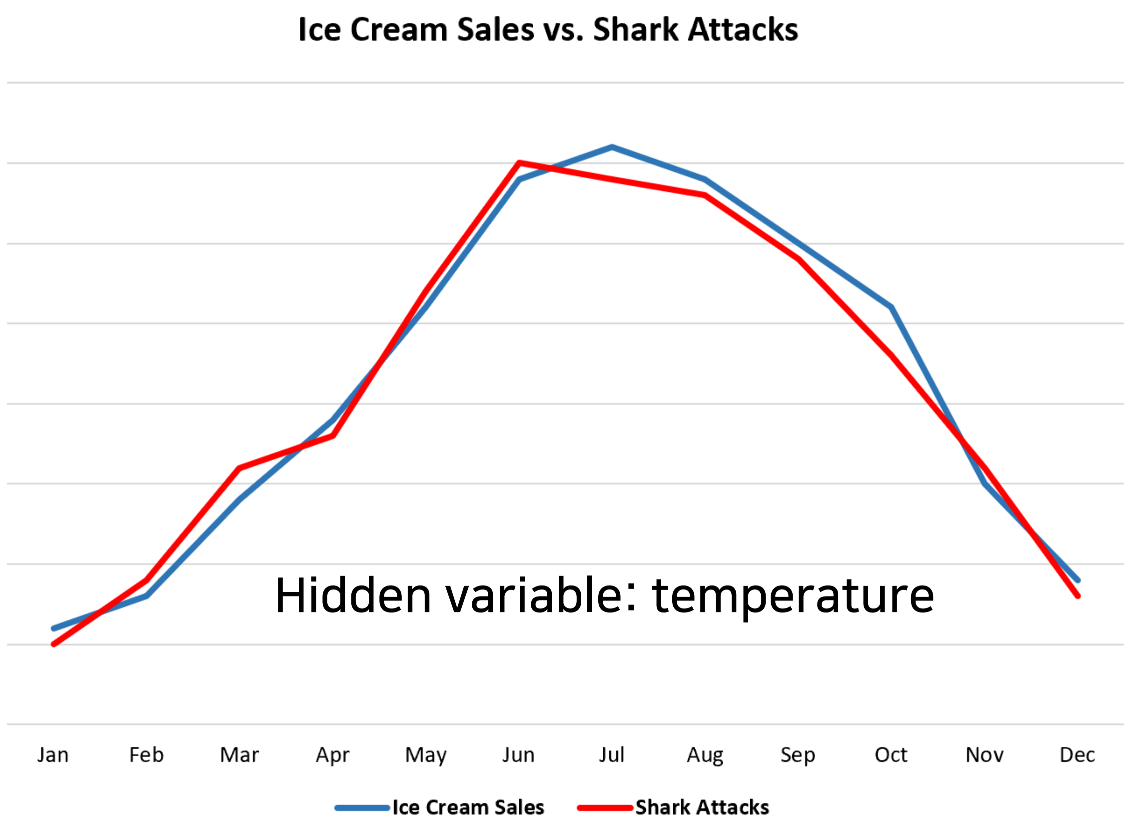
$$Cov(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \frac{\sum_{i=1}^{n}[(x_\mathrm{i}-\mu_X)(y_\mathrm{i}-\mu_Y)]}{N}$$

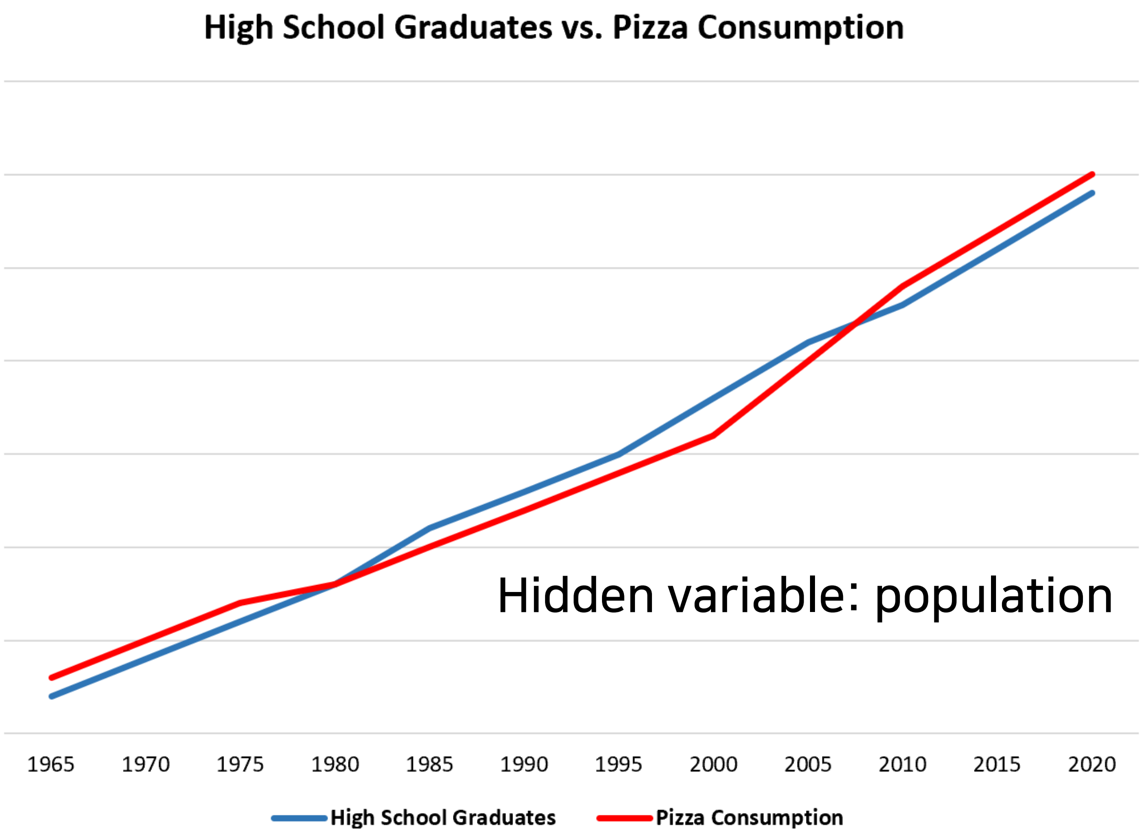
- It can be simplified in computation as

$$Cov(X,Y) = \mathbb{E}\,[X \cdot Y] - \mu_X\mu_Y$$

- Suppose two stocks A and B have the following values in one week:
  (2, 5), (3, 8), (5, 10), (4, 11), (6, 14).
- Question:
  If the stocks are affected by the same industry trends, will their prices rise or fall together?
  - E(A) = (2 + 3 + 5 + 4 + 6)/ 5 = 20/5 = 4
  - E(B) = (5 + 8 + 10 + 11 + 14) /5 = 48/5 = 9.6
  - Cov(A,B) = (2×5+3×8+5×10+4×11+6×14)/5 – 4 × 9.6 = 4
- Thus, A and B rise together since $Cov(A,B) > 0$.

Correlation does not imply causality





Be cautious in arguing causality.

In your analysis, a correlation without causation could be more plausible than the above.

Source: statology