Chapter 3: Data Preprocessing

Dong-Kyu Chae

PI of the Data Intelligence Lab @HYU
Department of Computer Science & Data Science
Hanyang University

Data
Intelligence
LAB

Contents: Major Tasks in Data Preprocessing

- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization

Why Pre-process the Data?

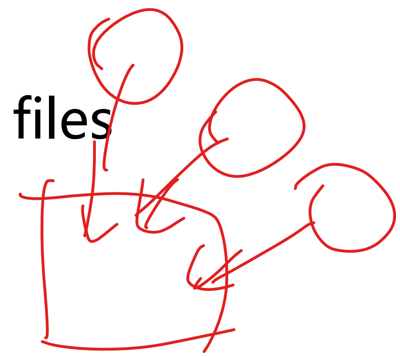
❑ A multidimensional view

- ❑ **Accuracy**: your data mining/machine learning results are not accurate, despite multiple trials...
- ❑ **Completeness**: not recorded, unavailable, ...
- ❑ **Consistency**: some modified but some not, ... y-m-d m/d/y
  Birthday
- ❑ **Timeliness**: timely updated? data t
- ❑ **Believability**: how trustable the data are?
- ❑ **Interpretability**: how easily the data can be understood?

preprocessing이 필요한 이유

Overview

- ❑ **Data cleaning** dirty한거 제거기
  - ❑ Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- ❑ **Data integration** scatter되어있는걸 integrate
  - ❑ Integration of multiple databases, data cubes, or files



- ❑ **Data reduction**
  - ❑ Dimensionality reduction
  - ❑ Numerosity reduction
  - ❑ Data compression

여러개로 data를
줄임

- ❑ **Data transformation and data discretization**
  - ❑ Normalization
  - ❑ Concept hierarchy generation

Data Cleaning

- **Data in the Real World Is Dirty:** Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error 날씨 때문에 video가 terrible
  - incomplete: lacking feature values, lacking certain features of interest, or containing only aggregate data
    - e.g., *Occupation*= " " (missing data) missing value
  - noisy: containing noise, errors, or outliers
    - e.g., *Salary*= "−10" (an error)
  - inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age*= "42", *Birthday*= "03/07/2000" inconsistent
    - In some DBs, *rating* is "1, 2, 3", but some other DBS, "A, B, C"
    - discrepancy between duplicate records

Missing Data

missing value를 ignore 한다.

- **Remove the object:** usually done when class label is missing—not effective when the % of missing values is large

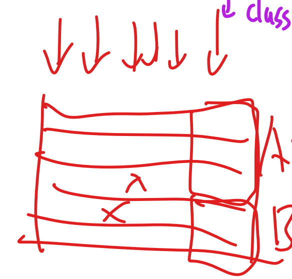
  Class lable이 X이면 전체를 삭제 그러나 다른 feature가 X이면 삭제 X

- **Fill in the missing value manually:** might be accurate, but tedious + infeasible

  cost ↑

- **Fill in it automatically with**

  class lable

  - a simple constant (default value, or "unknown")
  - the feature mean
  - the feature mean for all samples belonging to the same group (e.g., same class, same cluster, etc...)
  - **the inferred value: such as based on some regression or classification model**

missing value를 다른 model로 predict 한 값으로

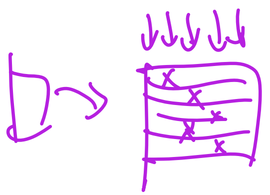
feature가 D에 대한 종속

Noisy & Inconsistent Data

- **Noise :** random error or variance in a measured feature
  - Mainly due to faulty data collection instruments
  - Noisy data is often expressed as an **outlier**
    - Outlier detection -> delete outliers -> find missing values
  - Thus we can apply an **outlier detection** method (will learn it later)

salary = -10,450 90000 1100100 → outlier

noise data를 remove 한다



- **Inconsistent data**
  - *Age*= "42" , but *Birthday*= "03/07/2000"
  - For a duplicate records, one name is "cm " but the other is "inch"

1 group을 정하고 → convert 한다. inch → cm

  - Human inspection will be needed

→ 나는 CEO의 salary를 포함하지 않겠다.

    - Computer performs outlier detection, then human will inspect it

Contents

Data Integration

- **Data integration:**
  - Combines multiple datasets from multiple sources into a coherent store
- **Schema integration: e.g., A.cust-id ≡ B.cust-#**
  - Integrate metadata from different sources

이름이 다르지만
meaning이 같은거

- **Detecting and resolving data value conflicts**
  - For the same real world entity, feature values from different sources are different
  - e.g., cm vs. inch, meter vs. mile

A에서 mile
B에서 km
인 경우 conflict

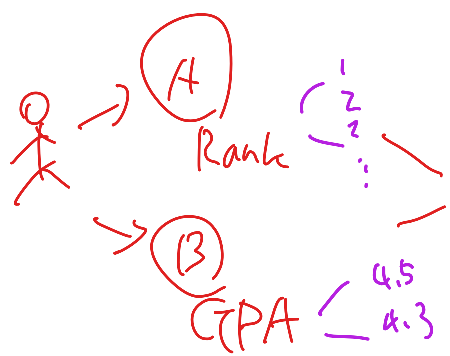
Handling Redundancy in Data Integration

- **Redundant data occur often when integration of multiple databases**
  - Derivable data : One feature may be a "derived" feature in another table, e.g., birthdate vs. age

redundant 둘중하나제거

A B birthday age 둘중하나제거

- **Redundant features can be automatically detected by correlation analysis and covariance analysis**
- **Reducing/avoiding redundancies and inconsistencies improves mining speed and quality**



correlation이 높다. using only 1

Correlation Analysis (Nominal Features)

- **We want to know that "like_science_fiction" and "play_chess" are correlated**

	Play chess (Y)	Not play chess (N)	Sum (row)	
Like science fiction (Y)	250(90)	200(360)	450	1:4
Not like science fiction (N)	50(210)	1000(840)	1050	1:4
Sum(col.)	300	1200	1500	1:4

expected values

'independent 한 경우'

높으면 하나 지움

- **X² (chi-square) test**

$$\chi^2 = \sum_{\text{each cell}} \frac{(Observed - Expected)^2}{Expected}$$

↓ → independent

↑ → correlated

  - The larger the X² value, the more likely the features are corelated
  - The cells that contribute the most to the X² value are those whose actual count is very different from the expected count
  - Expected value is estimated under the independence assumption

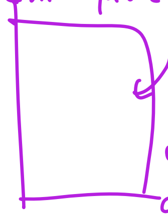
Correlation Analysis (Nominal Features)

	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

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

chi-square table



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

Correlation Analysis (Numeric Features)

- **Correlation coefficient (also called Pearson's correlation coefficient, PCC) among features A and B:**

average

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \overline{A})(b_i - \overline{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\overline{A}\overline{B}}{(n-1)\sigma_A \sigma_B}$$

number of data points

n is the number of data, $\overline{B}$ and $\overline{A}$ are the respective means of A and B, σ_A and σ_B are the respective standard deviation of A and B, and $\Sigma(a_i b_i)$ is the sum of the AB cross-product.

- **If $r_{A,B}$ > 0:** A and B are positively correlated

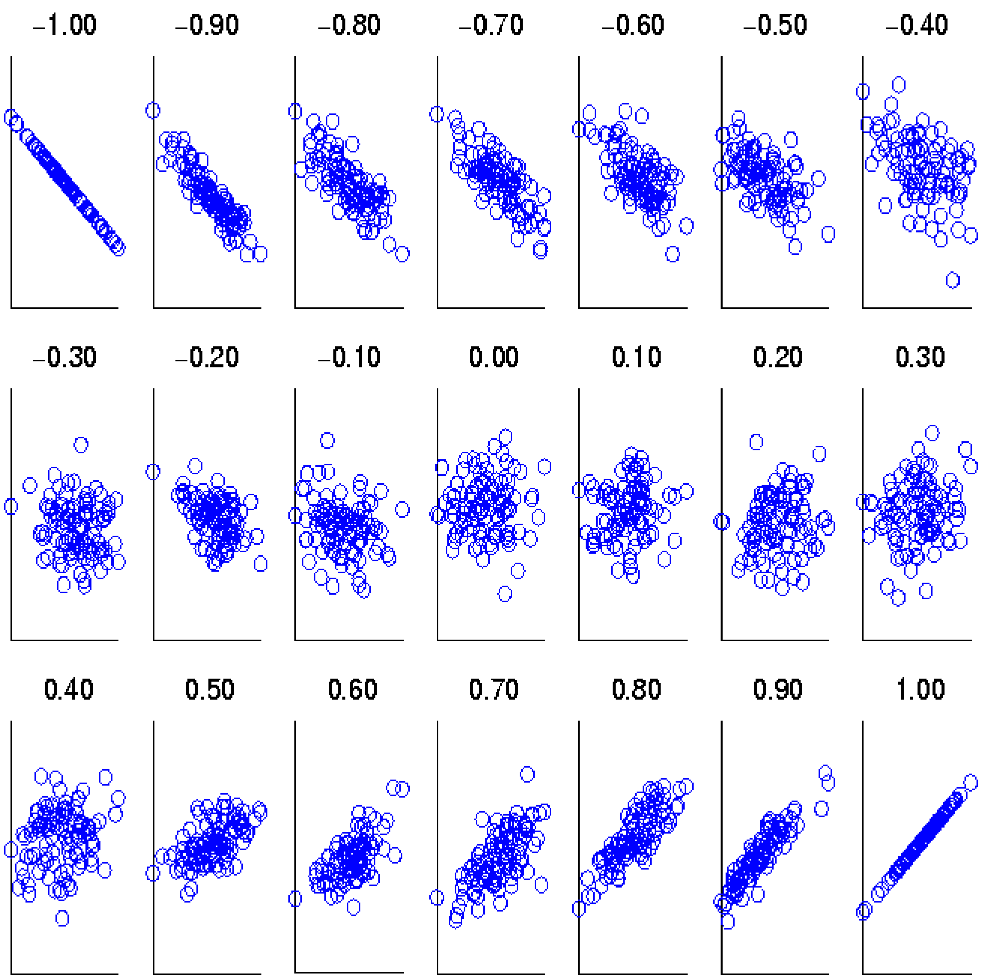
A가 증가하면 b도 tend to be increasing

  - A's values increase as B's). The higher, the stronger correlation
- **$r_{A,B}$ = 0:** independent

completely independent

- **r_{AB} < 0:** negatively correlated

Visually Evaluating Correlation



Scatter plots showing the correlation from –1 to 1.

Correlation(상관관계) does not imply causality(인과관계)
=> “# of hospitals” and “# of car-theft” in a city are correlated. However, both may be causally linked to another feature: **population**

Covariance (Numeric Features)

- **Covariance is similar to correlation**

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

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

where n is the number of data, $\bar{A}$ and $\bar{B}$ are the respective mean or expected values of A and B, σ_A and σ_B are the respective standard deviation of A and B.

- **Positive covariance:** If $Cov_{A,B} > 0$, then A and B both tend to be larger than their expected values.

specific range 가 없다
↓
위에선 normalize 한거
cov

- **Negative covariance:** If $Cov_{A,B} < 0$ then if A is larger than its expected value, B is likely to be smaller than its expected value.

- **Independence:** $Cov_{A,B} = 0$

Covariance: An Example

skip

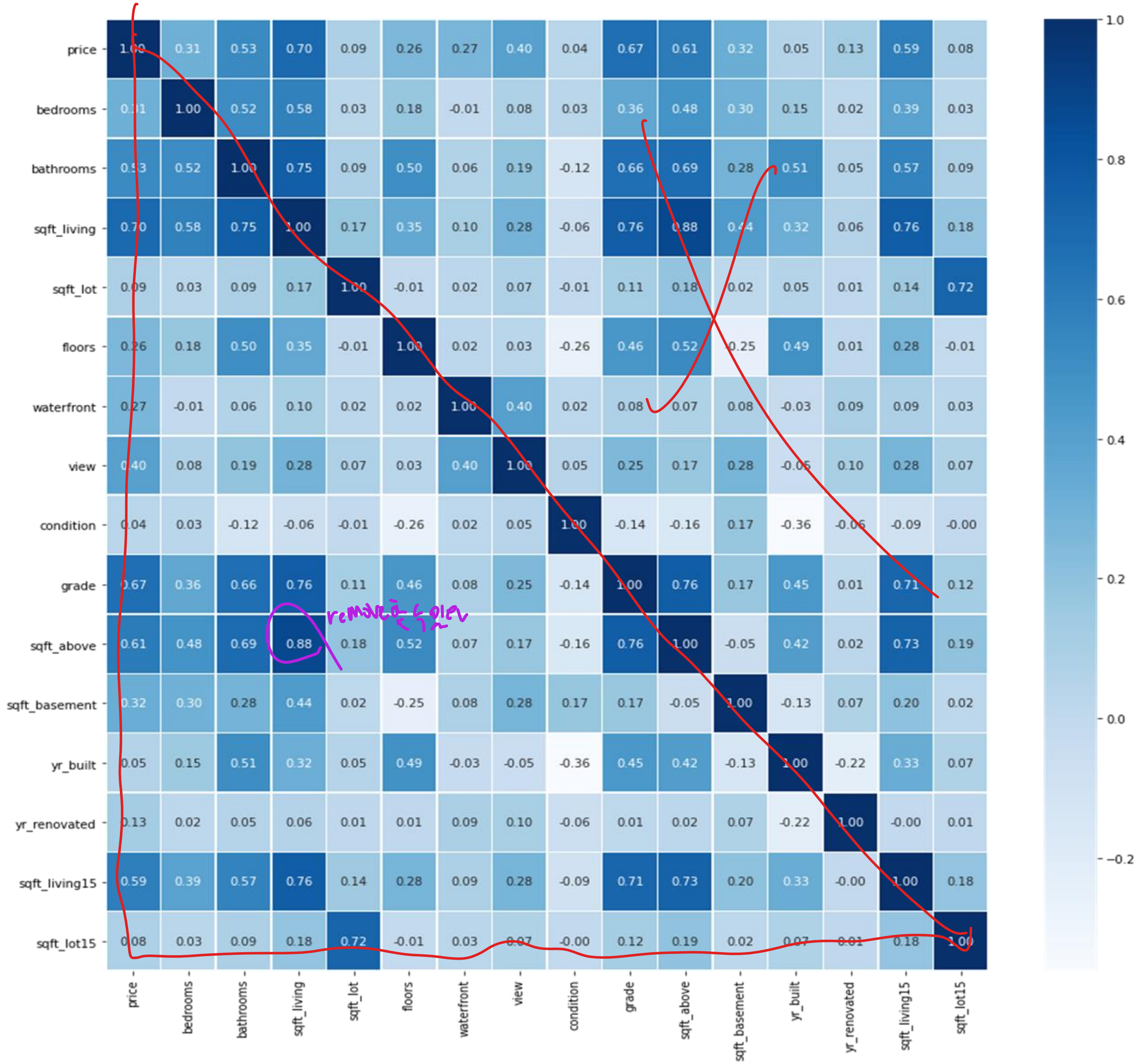
$$Cov(A,B) = E((A-\bar{A})(B-\bar{B})) = \frac{\sum_{i=1}^{n}(a_i-\bar{A})(b_i-\bar{B})}{n}$$

- **It can be simplified in computation as**

$$Cov(A,B) = E(A \cdot B) - \bar{A}\bar{B}$$

- **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.**

Covariance/Correlation Matrix Visualization



highly correlation 이면 둘 중 하나를 선택

Contents

Thank You

Data
Intelligence
Lab