


Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview 
- Data Quality
- Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Major Tasks in Data Preprocessing

- **Data cleaning**
 - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration**
 - Integration of multiple databases, data cubes, or files
- **Data reduction**
 - Dimensionality reduction
 - Numerosity reduction
 - Data compression
- **Data transformation and data discretization**
 - Normalization
 - Concept hierarchy generation

Data Quality: Why Preprocess the Data?

- Measures for data quality: A multidimensional view
 - Accuracy: correct or wrong, accurate or not ความถูกต้อง
 - Completeness: not recorded, unavailable, ... ความสมบูรณ์
 - Consistency: some modified but some not, dangling, ... ครอบคลุมไม่ซ้ำกัน
 - Timeliness: timely update? normalization ยาก
 - Believability: how trustable the data are correct?
 - Interpretability: how easily the data can be understood?

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
 - Data Quality
 - Major Tasks in Data Preprocessing
- Data Cleaning
- Data Integration
- Data Reduction
- Data Transformation and Data Discretization
- Summary

Data Cleaning

ทำจนสะอาดจนดู

ข้อมูลสกปรก → ดิบ / ฝุ่น

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error
- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
ไม่สมบูรณ์
- e.g., *Occupation*=" " (missing data)
- noisy: containing noise, errors, or outliers
สกปรก / ผิด
- e.g., *Salary*="−10" (an error)
- inconsistent: containing discrepancies in codes or names, e.g., ไม่เหมือนกัน
- *Age*="42", *Birthday*="03/07/2010" ขงขงยังไม่ครบถ้วน
- Was rating "1, 2, 3", now rating "A, B, C"
- discrepancy between duplicate records
- Intentional (e.g., *disguised missing data*)
- Jan. 1 as everyone's birthday? default

Incomplete (Missing) Data

ค่าที่หายไป

เกิดจากคนไม่กรอก หรือ ไม่รู้ ไม่สนใจ

- 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 เครื่องพัง
- inconsistent with other recorded data and thus deleted เครื่องที่ record ไม่เหมือนกัน
- data not entered due to misunderstanding
- certain data may not be considered important at the time of entry
- not register history or changes of the data
- Missing data may need to be inferred

หาค่าที่หายไป / ที่คนกรอกไม่ได้
การ infer Missing หมายถึง คิดเอาที่ควรจะเป็น

How to Handle Missing Data?

- Ignore the tuple: usually done when class label is missing (when doing classification)—not effective when the % of missing values per attribute varies considerably *ไม่เชิง ๆ ไม่เหมาะสม*
- Fill in the missing value manually: tedious + infeasible? *ไม่เชิง ๆ ไม่เหมาะสม*
- Fill in it automatically with
 - a global constant : e.g., “unknown”, a new class?! *อาจทำได้*
 - the attribute mean *หาค่าเฉลี่ย*
 - the attribute mean for all samples belonging to the same class: smarter *หาค่าเฉลี่ยของแต่ละคลาส*
- the most probable value: inference-based such as Bayesian formula or decision tree