CS 412 Intro. to Data Mining

การจัดการ Data ก่อนที่จะไปประมวลผล

Chapter 3. Data Preprocessing

Jiawei Han, Computer Science, Univ. Illinois at Urbana-Champaign, 2017

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview

เก็บเอง

sensor - เก็บอัตโนมัติ

มี noise, missing เหมือนกัน

missing → ข้อมูลที่ไม่ได้กรอก

noise → เป็นข้อมูลที่ไม่เข้ากับพวก

- Data Cleaning Data ที่เก็บมามันหลายแหล่ง
- Data Integration เอา Data จากหลายแหล่งมารวมกัน

ลดจำนวน. Data ลดด้าน Dimension ลดด้านตัวแปร

- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

What is Data Preprocessing? — Major Tasks

ขั้นตอนการทำ ↗

- **Data cleaning**
  - จัดการ missing   กำจัด noisy   กำจัด outliers   จัดการ inconsistencies
  - Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
  - ออกห่างจาก mean data ไปไกลหน่อย
  - ข้อมูลที่ผิดไปเลย
- **Data integration**
  - Integration of multiple databases, data cubes, or files
- **Data reduction** ลดจำนวนข้อมูล
  - Dimensionality reduction
  - Numerosity reduction
  - Data compression
- เปลี่ยนแปลงชนิดของข้อมูล
- **Data transformation and data discretization**
  - Normalization
  - Concept hierarchy generation

Why Preprocess the Data? — Data Quality Issues

- Measures for data quality: A multidimensional view
  - เพราะข้อมูลที่ใส่เข้ามาเป็นทั้งข้อมูลที่ถูก / ผิด
    Accuracy: correct or wrong, accurate or not
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - กรอกในเวลานั้น
    Timeliness: timely update? → อัพเดทตามกาลเวลา
  - น่าเชื่อถือหรือไม่?
    Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Data Cleaning

เครื่องมือเสียชั่วคราว

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error

มีต.จำนปีน

ไม่สมบูรณ์

- Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data

  ไม่กรอกข้อมูล

  - e.g., *Occupation* = " " (missing data)

ทำการคลีน เพราะ ข้อมูล ผิด

- Noisy: containing noise, errors, or outliers
  - e.g., *Salary* = "−10" (an error) กรอกผิด

ข้อมูลไม่สามัคคีกัน

- Inconsistent: 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 (e.g., *disguised missing* data)
  - Jan. 1 as everyone's birthday?

Data ไม่สมบูรณ์

Incomplete (Missing) Data

- 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 เกิดจากการไม่กรอกข้อมูล จึงเกิด Missing
  - Equipment malfunction
  - ไม่สอดคล้อง
    Inconsistent with other recorded data and thus deleted
  - Data were not entered due to misunderstanding เข้าใจผิด
  - Certain data may not be considered important at the time of entry
  - Did not register history or changes of the data ข้อมูลเปลี่ยนแปลง
- Missing data may need to be inferred

→ บางอันอาจจะประมาณค่ามันได้

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 Data record ไหนมี Missing เอาก็ลบทิ้งออก
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with ไม่รู้
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean เอาค่า Mean มาแทนค่า Missing
  - the attribute mean for all samples belonging to the same class: smarter แทน Mean ที่อยู่ใน class เดียวกัน
  - **the most probable value: inference-based such as Bayesian formula or decision tree**