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

- noise เป็นข้อมูลที่ไม่เข้ากับพวก → Ex. ช่องจังหวัดแต่ไปกรอกชื่อ
- Missing Data ข้อมูลที่ไม่ได้กรอก ไม่ได้เก็บ

- เก็บเอง
- Sensor - เก็บอัตโนมัติ มี noise, missing เหมือนกัน Ex. อยู่ดีๆไฟดับ

- Data Preprocessing: An Overview
- Data Cleaning → Data ที่เก็บมาก็มีหลายแหล่ง
- Data Integration → เอา Data จากหลายแหล่งมารวมกัน
- ลดจำนวน Data ลดด้าน Dimension ลดด้านตั้ง
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

What is Data Preprocessing? — Major Tasks

ขั้นตอนการเตรียมจ้า ↗

- **Data cleaning**
  - จัดการ missing / กำจัด noise / กำจัด Outliers / จัดการ inconsistencies
  - Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
  - ไม่เป็นไปทางเดียวกันหรือ ไม่เป็น Data เดียวกัน / หลุดจาก Mean data ไปสุดโต่ง / 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
  - เพราะข้อมูลที่ใส่เข้ามาเป็นหรือข้อมูลที่ผิด,ถูก
  - Accuracy: correct or wrong, accurate or not
  - ความสมบูรณ์ของข้อมูล
  - Completeness: not recorded, unavailable, …
  - Consistency: some modified but some not, dangling, …
  - ตรงในเวลานั้น
  - Timeliness: timely update? → อัพเดทตามกาลเวลา
  - น่าเชื่อถือมั้ย?
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, เครื่องมือเสียชั่วคราว Ex. ทำงาน ไปละเครื่องร้อน human or computer error, and transmission error
  - ไม่สมบูรณ์ Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data ไม่กรอกข้อมูลมา, กรอกผิด
    - e.g., *Occupation* = " " (missing data)
  - Noisy: containing noise, errors, or outliers
    - e.g., *Salary* = "−10" (an error)
  - Inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age* = "42", *Birthday* = "03/07/2010"
    - 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?

Incomplete (Missing) Data Dataไม่สมบูรณ์

- 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 - เครื่องเสีย
  - Inconsistent with other recorded data and thus deleted - เครื่องทำ Record data ไม่สมบูรณ์
  - Data were not entered due to misunderstanding
  - Certain data may not be considered important at the time of entry
  - Did not register history or changes of the data
- Missing data may need to be inferred

Missing data อาจต้องอาศัย ประมาณค่าแทนได้

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 เอาทิ้งออกหมด
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with กรอก Missing ด้วย ไม่รู้
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean เอาค่า mean ได้จากค่าต่าง ๆ บวกหาร
  - the attribute mean for all samples belonging to the same class: smarter
  - **the most probable value: inference-based such as Bayesian formula or decision tree**