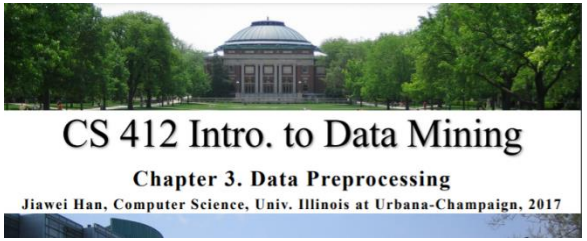
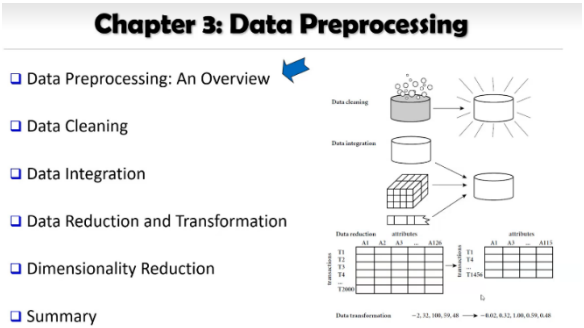
นางสาวอุมาพร คำภิชัย รหัสนักศึกษา 623020547-0
คณะวิทยาศาสตร์ สาขาสถิติ หลักสูตรสารสนเทศสถิติ ชั้นปีที่ 3

สรุป Chapter 3



Data Preprocessing คือ การจัดการข้อมูลก่อนไปประมวลผล

Pre แปลว่า ก่อน



- ขั้นตอนก็จะอยู่ฝั่งซ้าย รูปร่างหน้าตามันก็อยู่ฝั่งขวา

- เริ่มมาก็จะทำการ Cleaning Data

(นอยด์ คือ ข้อมูลที่ไม่จริงข้อมูลที่กรอกผิด ไม่เข้ากับพวก) (Missing คือ ข้อมูลที่ไม่ได้กรอก ไม่ได้เก็บ หายไปพอดี)

- ขั้นที่ 2 Data Integration คือ กรรวมข้อมูลมาจากหลายแหล่ง เช่น รวมเป็นตาราง เป็นต้น

- Data Reduction (การลดข้อมูลแนวนอน) **เป็นเทคนิคโบราณ and Transformation

Transformation (แปลงข้อมูลยังไงให้ประมวลผลได้) -> แนวนอนเป็น Data point แนวตั้งเป็น ฟิวเจอร์

- Dimensionality (การลดข้อมูลแนวตั้ง)

What is Data Preprocessing? — Major Tasks

- **Data cleaning**
  - Handle missing data, 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

*** อธิบายจากด้านบน***

Why Preprocess the Data? — Data Quality Issues

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

- ทำไมถึงต้องทำ Data Preprocess เพราะว่า ข้อมูลที่ใส่เข้ามามีทั้งข้อมูลที่ผิดและข้อมูลที่ถูกเป็นข้อมูลที่ไม่ค่อยแม่นยำ มาจากหลายแหล่ง คนกรอกบ้าง เครื่องบ้าง

Data Cleaning

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error
  - Incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
    - 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?

- Data Cleaning มีความจำเป็นเพราะว่าข้อมูลในโลกนี้มีความสกปรกและไม่สะอาด ไม่ถูก สาเหตุมาจากเครื่องมือเสียอาจจะทำให้ข้อมูลเพี้ยนไป หรือพิมพ์ข้อมูลผิด กรอกข้อมูลมาแล้วผิดพลาดจากคอมพิวเตอร์แทน

- ข้อมูลไม่สมบูรณ์ อย่างเช่น ไม่กรอกข้อมูลมา กรอกผิด เช่นเงินเดือนกรอกเป็นติดลบ

- ข้อมูลไม่สอดคล้องกัน เช่น วันเดือนปีเกิด กับ อายุ ไม่ตรงกัน

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
  - Equipment malfunction
  - Inconsistent with other recorded data and thus deleted
  - 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

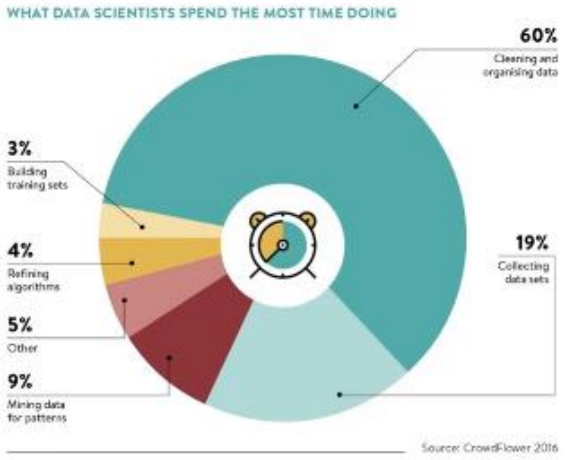
- ข้อมูลไม่สมบูรณ์ เกิดจาก คนไม่ได้กรอก ไม่มีข้อมูลนั้น

- Data Missing เกิดจากการที่เราสามารถประมาณค่าให้มันได้

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
- Fill in the missing value manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean
  - 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**

Time spending for different data tasks



ตัวเลือกกับการจัดการ Missing

- ง่ายที่สุดคือการลบ Missing ออก

- สิ่งที่เกิดขึ้น ถ้าเรคคอดไหนมี มิสสิ่ง จาก 1 ล้านเหลือ 2 พัน อาจจะเป็นตัวเลือกที่ไม่ค่อยดีเท่าไหร่ แต่ก็สามารถทำได้

- นั่งดูเองว่าข้อมูลเรคคอดนี้เป็นยังไงบ้าง เราจะเติมค่าไรให้มันหรือจะลบออก

- กรอกมิชชิ่ง สร้างคลาสใหม่ขึ้นมา เขียนด้วยคำว่า ‘’ไม่รู้’’ หรือจะเอาค่าเฉลี่ยมากรอกแทน