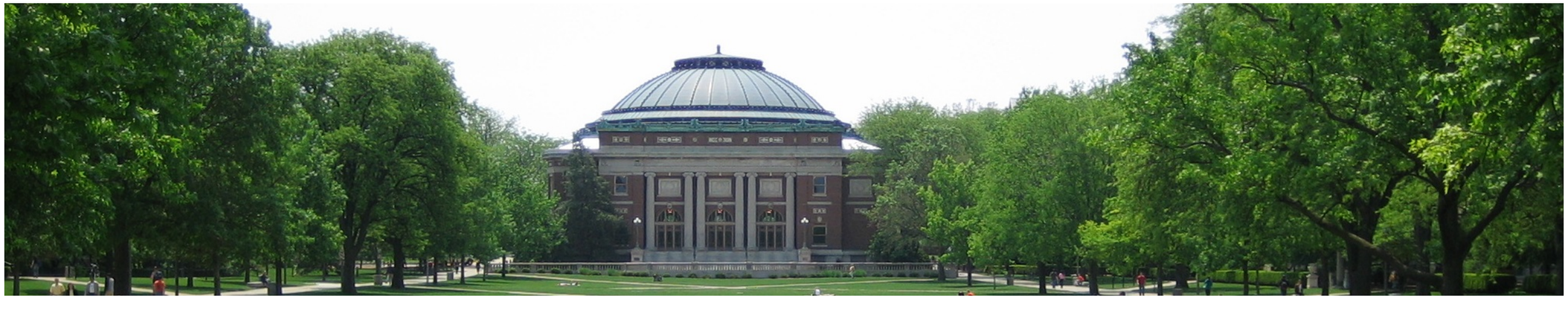

CS 412 Intro. to Data Mining

Chapter 3. Data Preprocessing

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

การเตรียมข้อมูล การประมวลผลข้อมูล

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning
- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

What is Data Preprocessing? — Major Tasks

- **Data cleaning** *ขั้นตอนสำหรับการตัดข้อมูลที่ไม่เกี่ยวข้องออกไป*
  - Handle missing data, smooth noisy data, identify or remove outliers, and resolve inconsistencies
- **Data integration** *รวม Data เข้าด้วยกัน / รวมข้อมูลหลายๆที่มารวมไว้ที่เดียวกัน*
  - Integration of multiple databases, data cubes, or files
- **Data reduction** *การลดมิติของข้อมูล*
  - Dimensionality reduction *เลือก Data ที่สำคัญ*
  - Numerosity reduction
  - Data compression
- **Data transformation and data discretization** *แบ่งข้อมูลในขั้นตอนการคัดเลือก ให้เหมาะสำหรับขั้นตอนการทำเหมืองข้อมูล*
  - Normalization
  - Concept hierarchy generation

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, … กรอกข้อมูลไม่ครบถ้วน
  - Consistency: some modified but some not, dangling, … ทำ Normalization ไม่ดีข้อมูลจะไปอยู่ในหลายที่
  - Timeliness: timely update? Data ที่มีการกรอกช่วงหนึ่ง และในเวลาต่อมาจะทำให้มีการเปลี่ยนแปลง
  - Believability: how trustable the data are correct?
  - Interpretability: how easily the data can be understood?

Chapter 3: Data Preprocessing

- Data Preprocessing: An Overview
- Data Cleaning

การทำความสะอาดของข้อมูล

- Data Integration
- Data Reduction and Transformation
- Dimensionality Reduction
- Summary

Data Cleaning การทำความสะอาดของข้อมูล

- Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, and transmission error ข้อมูลที่ผิด
  - Data ที่ ไม่สมบูรณ์ 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) ตัวอย่างเช่น หรือเสียงที่ติดมา
  - Inconsistent: containing discrepancies in codes or names, e.g.,
    - *Age* = “42”, *Birthday* = “03/07/2010” อายุกับวันเกิดไม่ตรงกัน
    - 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?

ข้อมูลที่ไม่สมบูรณ์ ไม่ชัดเจนข้อมูล

Incomplete (Missing) Data

- Data is not always available ข้อมูลไม่พร้อมใช้งาน
  - 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 อุปกรณ์ทำงานผิดปกติ
  - 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

วิธีจัดการกับข้อมูลที่ขาดหายไป

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 (เติม) + infeasible (ลงแรง)?
- Fill in it automatically with
  - a global constant : e.g., “unknown”, a new class?!
  - the attribute mean เอาค่ากลางมาแทน missing
  - the attribute mean for all samples belonging to the same class: smarter แทนด้วย mean ที่อยู่ในคลาสเดียวกัน
  - **the most probable value: inference-based such as Bayesian formula or decision tree**

Noisy Data ข้อมูลที่มีเสียง / ข้อมูลรบกวน

- **Noise:** random error or variance in a measured variable
- **Incorrect attribute values** may be due to
  - Faulty data collection instruments ข้อมูลผิดพลาด
  - Data entry problems การป้อนข้อมูล
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming convention
- **Other data problems** ปัญหาของข้อมูล
  - Duplicate records ข้อมูลซ้ำ
  - Incomplete data ข้อมูลไม่สมบูรณ์
  - Inconsistent data ข้อมูลไม่สอดคล้องกัน

How to Handle Noisy Data?

- Binning ค่าของข้อมูลที่เรียงลำดับเป็นไปอย่างราบรื่น โดยพิจาณาจากข้อมูลที่อยู่เคียงๆ
  - First sort data and partition into (equal-frequency) bins
  - Then one can **smooth by bin means, smooth by bin median, smooth by bin boundaries**, etc.
- Regression สามารถปรับข้อมูลให้เรียบได้ โดยปรับข้อมูลให้เหมาะกับฟังก์ชัน
  - Smooth by fitting the data into regression functions
- Clustering
  - Detect and remove outliers
- Semi-supervised: Combined computer and human inspection
  - Detect suspicious values and check by human (e.g., deal with possible outliers)