Chương 4

Khai phá dữ liệu

Nội dung

1. Tiền xử lý dữ liệu.
2. Phương pháp khai phá bằng luật kết hợp.
3. Phương pháp cây quyết định.
4. Các phương pháp phân cụm.
5. Các phương pháp khai phá dữ liệu phức tạp.

Tiền xử lý dữ liệu

- Dữ liệu phát sinh trong quá trình tác nghiệp gọi là dữ liệu thô (raw/original data),
- Dữ liệu thô:
  - ✓ Từ các nguồn file/cơ sở dữ liệu (database),
  - ✓ Không hoàn chỉnh: thiếu thuộc tính, giá trị cần.
  - ✓ Chứa giá trị nhiễu: có lỗi hoặc có giá trị lệch,
  - ✓ Không nhất quán.
- Để có thể khai phá các khía cạnh khác của chúng cần phải biến đổi về dạng thích hợp,

Tiền xử lý dữ liệu

- Chất lượng dữ liệu
  - ✓ Tính chính xác (accuracy): giá trị được ghi nhận đúng với giá trị thực,
  - ✓ Tính hiện hành (currency/timeliness): giá trị được ghi nhận không bị lỗi thời.
  - ✓ Tính toàn vẹn (completeness): tất cả các giá trị dành cho một biến/thuộc tính đều được ghi nhận.
  - ✓ Tính nhất quán (consistency): tất cả giá trị dữ liệu đều được biểu diễn như nhau trong tất cả các trường hợp.

Tiền xử lý dữ liệu

- Các kỹ thuật tiền xử lý:
- Tích hợp dữ liệu (Data integration):
  - ✓ Làm tăng lượng thông tin.
  - ✓ Tuy nhiên có thể làm dư thừa và không nhất quán.
- Làm sạch dữ liệu (Data cleaning):
  - ✓ Bổ sung giá trị thiếu,
  - ✓ Loại dữ liệu nhiễu,
  - ✓ Loại giá trị lệch,
  - ✓ Nhất quá hóa dữ liệu.

Tiền xử lý dữ liệu

- Các kỹ thuật tiền xử lý (tt):
- Chuyển dạng dữ liệu (Data transformation):
  - Chuẩn hóa (normalization),
  - Gộp nhóm (aggregation).
- Rút gọn dữ liệu (Data reduction):
  - Giảm số chiều,
  - Giảm biểu diễn số lớn,
  - Lựa chọn tập thuộc tính,
  - …

Tiền xử lý dữ liệu

- Tóm tắt – mô tả về dữ liệu:
  - Xác định các thuộc tính (properties) tiêu biểu của dữ liệu về xu hướng chính (central tendency) và sự phân tán (dispersion) của dữ liệu.
  - Làm nổi bật các giá trị dữ liệu nên được xem như nhiễu (noise) hoặc phần tử biên (outliers), cung cấp cái nhìn tổng quan về dữ liệu.

Tiền xử lý dữ liệu

- Các yếu tố cần quan tâm khi nghiên cứu khai phá dữ liệu:
  - Xu hướng tập trung (central tendency): đặc trưng bởi các đại lượng thống kê: trung bình (Mean), trung vị (Median), mode, khoảng trung bình (midrange), …
  - Sự phân ly (dispersion): tứ nhân vị (quartile), khoảng tứ phân vị (interquartile range), phương sai (variance), độ lệch chuẩn (standard deviation)

Tiền xử lý dữ liệu

- Công thức tính của các độ đo xu hướng chính của dữ liệu:
  - Mean: $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}$
  - Weighted arithmetic mean: $\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}$
  - Median: $Median = \begin{cases} x_{\lceil N/2 \rceil} \\ (x_{N/2} + x_{N/2+1})/2 \end{cases}$

Tiền xử lý dữ liệu

- Công thức tính của các độ đo xu hướng chính của dữ liệu (tt):
  - Mode: giá trị xuất hiện thường xuyên nhất trong tập dữ liệu
  - Midrange: Giá trị trung bình của các trị lớn nhất và nhỏ nhất trong tập dữ liệu.

Tiền xử lý dữ liệu

- Công thức tính của các độ đo về sự phân tán của dữ liệu (tt):
  - Quartiles (tứ phân vị):
    - The first quartile: Q1 = 25 * (n+1) / 100,
    - The second quartile: Q2 = 50 * (n+1) / 100,
    - The third quartile: Q3 = 75 * (n+1) / 100.
  - Interquartile Range (IQR) = Q3 – Q1
    - Outliers (trị biên): trên Q3/dưới Q1 = 1.5*IQR
  - Variance: (phương sai)

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N}\left[\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right]$$

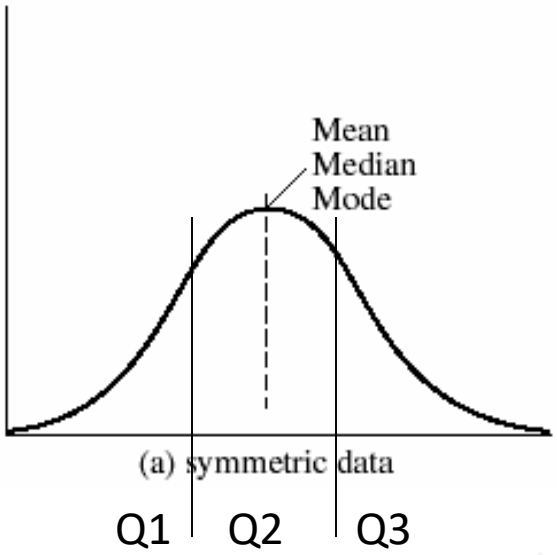
Tiền xử lý dữ liệu

- Công thức tính của các độ đo về sự phân tán của dữ liệu (tt):
  - Tính quartiles:
    - ✓ Sắp xếp các số theo thứ tự tăng dần,
    - ✓ Cắt dãy số thành 4 phàn bằng nhau,
    - ✓ Tứ phân vị là các giá trị tại vị trí cắt
  - Ví dụ: Cho dãy số 5, 8, 4, 4, 6, 3, 8
    - ➢ Sắp xếp: 3, 4, 4, 5, 6, 8, 8
      - ⇒ Q1 = 4; Q2 = 5; Q3 = 8

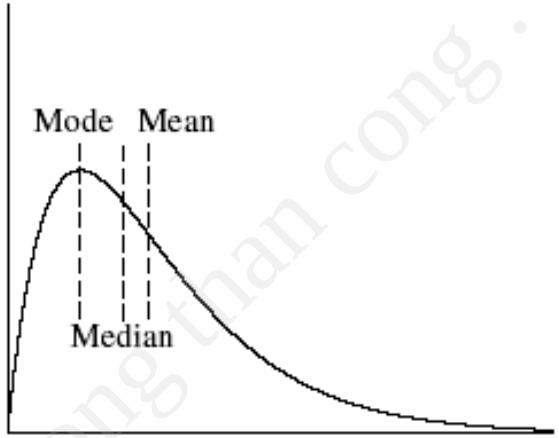
Nếu vị trí cắt ở giữa 2 số thì tứ phân vị là giá trị trung bình của 2 số đó.

Tiền xử lý dữ liệu

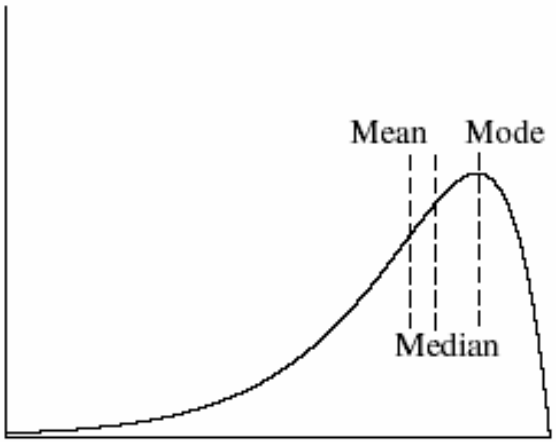
- Tóm tắt mô tả về dữ liệu:



(a) symmetric data



(b) positively skewed data



(c) negatively skewed data

- (a): Dữ liệu cân đối
- (b): Dữ liệu lệch dương
- (c): Dữ liệu lệch âm
- Minimum < Q1 < Median < Q3 < Maximum

Tiền xử lý dữ liệu

- Tóm tắt mô tả về dữ liệu:
  - Độ lệch chuẩn (Standard deviation):

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i \,.$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2}\,,$$

Tiền xử lý dữ liệu

- Làm sạch dữ liệu:
  - Xử lý dữ liệu bị thiếu (missing data),
  - Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data),
  - Xử lý dữ liệu không nhất quán (inconsistent data)

Tiền xử lý dữ liệu

- Làm sạch dữ liệu (tt):
  - Xử lý dữ liệu bị thiếu (missing data):
    - ✓ Định nghĩa của dữ liệu bị thiếu
      - Dữ liệu không có sẵn khi cần được sử dụng
    - ✓ Nguyên nhân gây ra dữ liệu bị thiếu
      - Khách quan (không tồn tại lúc được nhập liệu, sự cố, …)
      - Chủ quan (tác nhân con người)

Tiền xử lý dữ liệu

- Làm sạch dữ liệu (tt):
  - Xử lý dữ liệu bị thiếu (missing data):
    - Giải pháp cho dữ liệu bị thiếu
      - Bỏ qua
      - Xử lý tay (không tự động, bán tự động),
      - Dùng giá trị thay thế (tự động): hằng số toàn cục, trị phổ biến nhất, trung bình toàn cục, trung bình cục bộ, trị dự đoán, …
      - Ngăn chặn dữ liệu bị thiếu: thiết kế tốt CSDL và các thủ tục nhập liệu (các ràng buộc dữ liệu).

Tiền xử lý dữ liệu

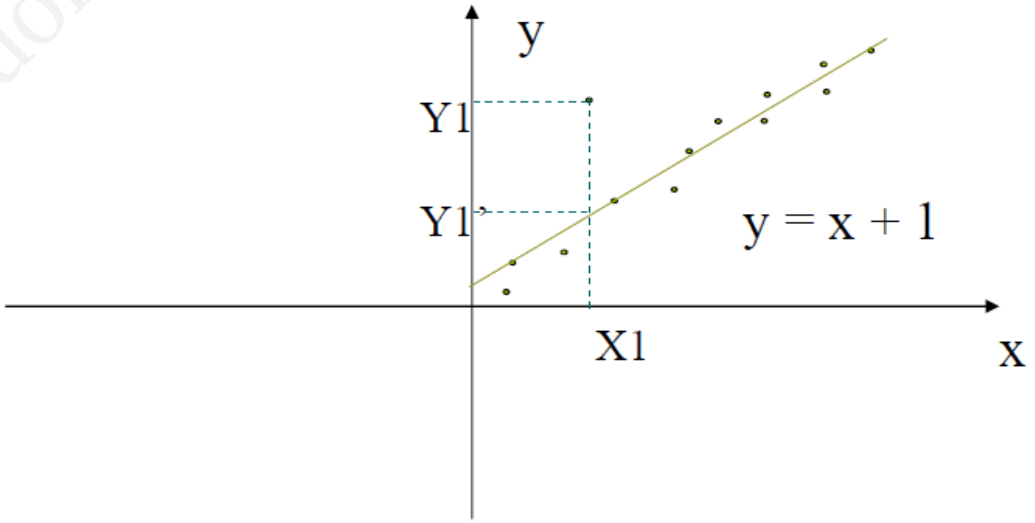
- Làm sạch dữ liệu (tt):
  - Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data):
    - ✓ Outliers: những dữ liệu (đối tượng) không tuân theo đặc tính/hành vi chung của tập dữ liệu (đối tượng).
    - ✓ Noisy data: outliers bị loại bỏ (rejected/discarded outliers) như là những trường hợp ngoại lệ (exceptions).

Tiền xử lý dữ liệu

- Làm sạch dữ liệu (tt):
  - Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data):
    - ✓ Giải pháp nhận diện phần tử biên
      - Dựa trên phân bố thống kê (statistical distribution - based)
      - Dựa trên khoảng cách (distance-based)
      - Dựa trên mật độ (density-based)
      - Dựa trên độ lệch (deviation-based)

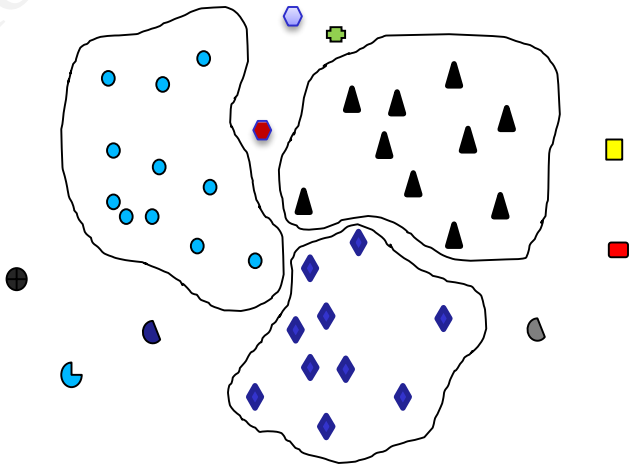
Tiền xử lý dữ liệu

- Làm sạch dữ liệu (tt):
  - Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data):
    - ✓ Giải pháp giảm thiểu nhiễu
      - Hồi quy (regression)



Tiền xử lý dữ liệu

- Làm sạch dữ liệu (tt):
  - Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data):
    - ✓ Giải pháp giảm thiểu nhiễu
      - Phân tích cụm (cluster analysis)



Tiền xử lý dữ liệu

- Làm sạch dữ liệu (tt):
  - Nhận diện phần tử biên (outliers) và giảm thiểu nhiễu (noisy data):
    - Giải pháp xử lý dữ liệu không nhất quán (inconsistent)
      - Tận dụng siêu dữ liệu, ràng buộc dữ liệu, sự kiểm tra của nhà phân tích dữ liệu cho việc nhận diện.
      - Điều chỉnh dữ liệu không nhất quán bằng tay.
      - Biến đổi, chuẩn hóa dữ liệu tự động.

Tiền xử lý dữ liệu

2. Biến đổi dữ liệu: Tạo tính tương thích giữa dữ liệu của nhiều nguồn khác nhau.
   - ✓ Làm mịn: loại bỏ trường hợp nhiễu.
   - ✓ Tổng hợp: Rút gọn dữ liệu và tạo khối dữ liệu cho việc phân tích.
   - ✓ Khái quát hóa: Chuyển dữ liệu mức thấp sang mức cao.
   - ✓ Chuẩn hóa: Chuyển khoảng giá trị rộng thành khoảng giá trị nhỏ hơn ([10..1.000] -> [0.0..1.0])
   - ✓ Xác định thêm thuộc tính.

Tiền xử lý dữ liệu

2. Biến đổi dữ liệu:
   - Một số phương pháp biến đổi:
     - ✓ Min-Max:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_B) + new\_min_A$$

- min_A, max_A: giá trị lớn nhất và nhỏ nhất của thuộc tính A
- New_min_A, new_max_A: miền giá trị mới.

Tiền xử lý dữ liệu

2. Biến đổi dữ liệu:
   - Một số phương pháp biến đổi:
     - ✓ Z-score:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

       - Ā: giá trị trung bình của thuộc tính A,
       - σ_A: độ lệch chuẩn.
     - ✓ Thay đổi tỷ lệ.
     - ✓ Lựa chọn tập thuộc tính con