TÀI LIỆU THỰC HÀNH
CHUẨN HÓA DỮ LIỆU

I. Giới thiệu

- Chuẩn hóa dữ liệu (normalization) là đưa miền giá trị của thuộc tính về miền giá trị nhỏ, thông thường là [-1, 1] hoặc [0, 1].
- Chuẩn hóa dữ liệu có lợi cho một số thuật toán phân lớp như mạng nơron nhân tạo, cho phép tăng tốc độ huấn luyện.
- Chuẩn hóa dữ liệu cũng giúp ngăn ngừa sự lấn át của các thuộc tính có miền giá trị lớn với những thuộc tính có miền giá trị nhỏ khi tính khoảng cách giữa các mẫu.
- Một số phương pháp chuẩn hóa thông dụng:
  - Min-max
  - Z-score
  - Decimal scaling (tính tỷ lệ)

II. Một vài phương pháp chuẩn hóa

1. Chuẩn hóa Min-max

Gọi min_A và max_A lần lượt là giá trị nhỏ nhất và lớn nhất của thuộc tính A.

Giá trị sau khi chuẩn hóa min-max cho giá trị v của thuộc tính A vào miền giá trị mới $[new\_min_A, new\_max_A]$ là:

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

Ví dụ: Cho giá trị nhỏ nhất và lớn nhất của thuộc tính "Thu nhập" là \$12,000 và \$98,000. Yêu cầu chuẩn hóa min-max các giá trị thuộc tính "Thu nhập" của từng mẫu về miền [0, 1].

Giả sử, ta có một mẫu có giá trị thu nhập là \$73,600 ⇒ giá trị của nó sau khi chuẩn hóa sẽ là:

$$\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1 - 0) + 0 = 0.716$$

2. *Chuẩn hóa Z-score*

Các giá trị của thuộc tính sẽ được chuẩn hóa dựa trên giá trị trung bình và độ lệch chuẩn của thuộc tính đó.

Một giá trị v của thuộc tính A được chuẩn hóa theo Z-score sẽ được tính bằng công thức:

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

Trong đó,

- $\bar{A}$: là giá trị trung bình của thuộc tính. Tính bằng cách lấy trung bình cộng các giá trị thuộc tính A ở các mẫu trong tập dữ liệu.
- σ_A: độ lệch chuẩn của thuộc tính, tính theo công thức:

$$\sigma_A = \sqrt[2]{\frac{1}{N}\sum_{i=1}^{N}(v_i - \bar{A})^2}$$

Ví dụ: Cho tập dữ liệu, trong đó thuộc tính "Thu nhập" có giá trị trung bình là \$54,000 và độ lệch chuẩn là \$16,000.

Giả sử, ta có một mẫu có giá trị thu nhập là \$73,600 ⇒ giá trị của nó sau khi chuẩn hóa sẽ là:

$$\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$$

3. *Chuẩn hóa thang thập phân*

Việc chuẩn hóa thực chất là dời dấu phẩy động sang trái, dời bao nhiêu chữ số là tùy thuộc vào giá trị tuyệt đối lớn nhất của thuộc tính.

Công thức chuẩn hóa là:

$$v' = \frac{v}{10^j}$$

Với j là số chữ số muốn di chuyển.

Ví dụ: Một tập dữ liệu có thuộc tính có giá trị nhỏ nhất là *-986* và lớn nhất là *917*. Như vậy, giá trị tuyệt đối lớn nhất là *-986* ⇒ dịch chuyển sang trái 3 chữ số (*j=3*).

Khi đó, giá trị *321*, sẽ được chuẩn hóa thành *0.321*.