Khoa Khoa học
và Kỹ thuật Thông tin

BÁO CÁO BÀI TẬP VỀ NHÀ BUỔI 6

Môn học: Thu thập và tiền xử lý dữ liệu - DS103.N21

Tên: Nguyễn Thị Huyền Trang

MSSV: 21520488

Câu 1:

Ta có bộ dữ liệu như bảng sau:

O-Ring Failure	Temperature
Y	53
Y	56
N	57
N	63
N	66
N	67
N	67
N	67
N	68
N	69
Y	70

- Độ lợi thông tin thuộc tính đích: O-Ring Failure

$$Y: 3 => P(Y) = \frac{3}{11}$$

$$N: 8 => P(N) = \frac{8}{11}$$

$$E(O-Ring\ Failure) = E(3,8) = -\frac{3}{11} \cdot log_2\left(\frac{3}{11}\right) - \frac{8}{11} \cdot log_2\left(\frac{8}{11}\right) = 0.85$$

- Khoảng chia: (số phần tử = 11)

$$(\leq 60, > 60)\ và\ (\leq 65, > 65)$$

-Xét khoảng chia $(\leq 60, > 60)$:

Entropy:

$$E(\leq 60) = E(2,1) = -\frac{2}{3} \cdot log_2\left(\frac{2}{3}\right) - \frac{1}{3} \cdot log_2\left(\frac{1}{3}\right) = 0.906$$

$$E(> 60) = E(1,7) = -\frac{1}{8} \cdot log_2\left(\frac{1}{8}\right) - \frac{7}{8} \cdot log_2\left(\frac{7}{8}\right) = 0.54$$

$$E\big(O-Ring\ Failure, (\leq 60, > 60)\big) = \frac{3}{11}.0{,}906 + \frac{8}{11} \cdot 0{,}54 = 0.63$$

$$Information\_Gain(O\text{-}Ring\ Failure, (\leq 60, > 60)) = 0.85 - 0.63 = 0.22$$

-Xét khoảng chia $(\leq 65, > 65)$:

Entropy:

$$E(\leq 65) = E(2,2) = -\frac{2}{4} \cdot log_2\left(\frac{2}{4}\right) - \frac{2}{4} \cdot log_2\left(\frac{2}{4}\right) = 1$$

$$E(> 65) = E(1,6) = -\frac{1}{7} \cdot log_2\left(\frac{1}{7}\right) - \frac{6}{7} \cdot log_2\left(\frac{6}{7}\right) = 0.591$$

$$E\big(O - Ring\ Failure, (\leq 65, > 65)\big) = \frac{4}{11}.1 + \frac{7}{11} \cdot 0{,}591 = 0.73$$

Information_Gain(O-Ring Failure, $(\leq 60, > 60))$ = 0.85 - 0.73 = 0.12

Vậy ta chọn khoảng chia $(\leq \mathbf{60}, > \mathbf{60})$.

Câu 2: (vấn đề 2)

Cho 2 tập hợp sau, là tập điểm số của DS 2 lớp học với thang đo 100

DS1: (70, 70, 10, 70, 30, 30, 45, 10, 10, 70)

DS2: (73, 12, 8, 80, 80, 44, 73, 100, 91, 73)

Biến đổi thành 2 DS1, DS2 bằng phương pháp: min-max, z-score về thang đo [0,10]

Chuẩn hoá bộ dữ liệu DS1 và DS2 theo phương pháp min-max:

Áp dụng công thức:

$$v_i' = \frac{v_i - minA}{max\ A - minA} \cdot (new\_max\ A - new\_minA) + new\_minA$$

*DS1:

min(DS1) = 10

max(DS1) = 70

new_min(DS1) = 0

new_max(DS1) = 10

⇨ **DS11: (10.00, 10.00, 0.00, 10.00, 3.33, 3.33, 5.83, 0.00, 0.00, 10.00)**

*DS2:

min(DS2) = 8

max(DS2) = 100

new_min(DS2) = 0

new_max(DS2) = 10

⇨ **DS21: (7.07, 0.43, 0.00, 7.83, 7.83, 3.91, 7.07, 10.00, 9.02, 7.07)**

Chuẩn hoá bộ dữ liệu DS1 và DS2 theo phương pháp z-score:

Áp dụng công thức:

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

*DS1:

$\overline{DS1} = 41.5$

$\sigma_{DS1} = 25.5$

⇨ **DS12: (1.12, 1.12, -1.24, 1.12, -0.45, -0.45, 0.14, -1.24, -1.24, 1.12)**

*DS2:

$\overline{DS2} = 63.4$

$\sigma_{DS2} = 30.02$

⇨ **DS22: (0.32, -1.71, -1.84, 0.55, 0.55, -0.65, 0.32, 1.22, 0.92, 0.32)**

Độ lệch chuẩn:

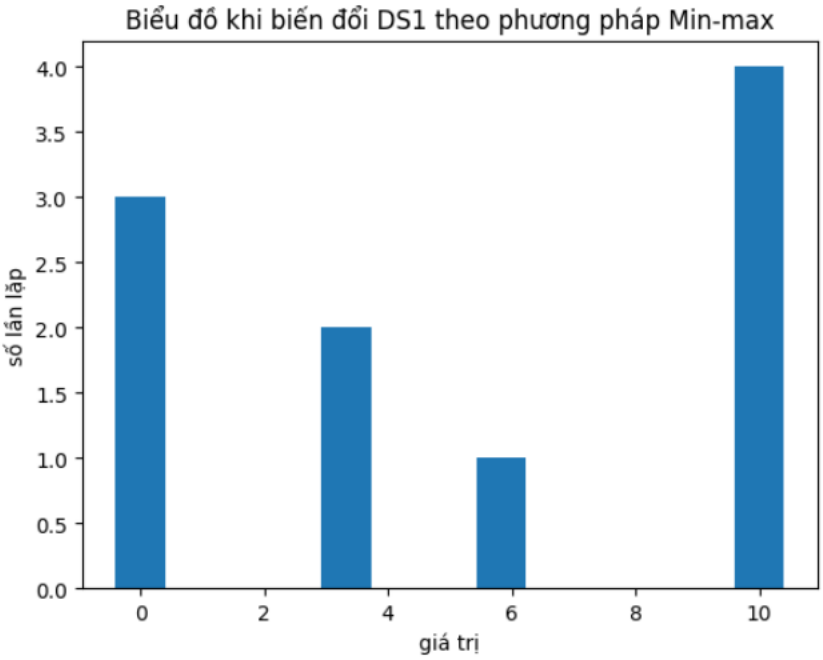
DS11: SD(DS11) = 4.25

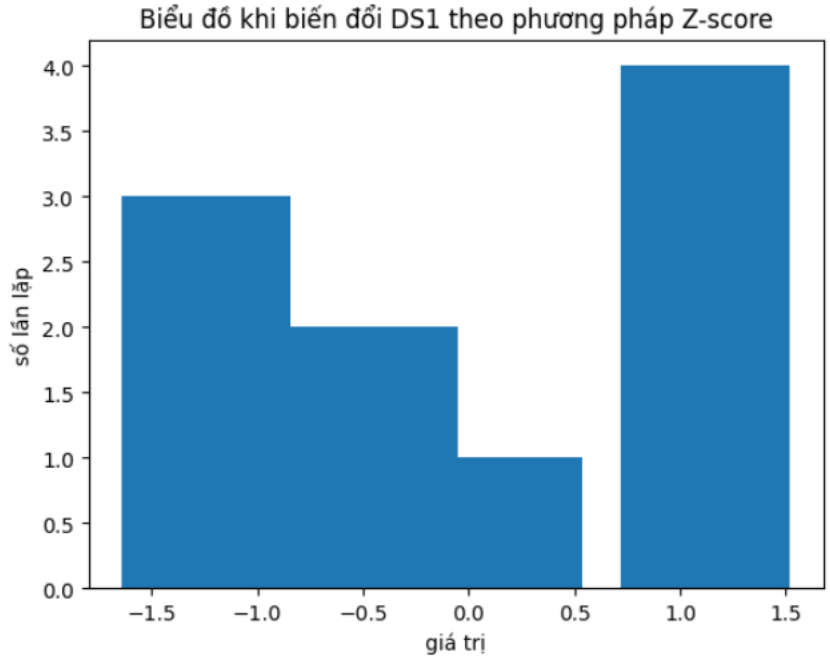
DS12: SD(DS12) = 1.00

DS21: SD(DS21) = 3.26

DS22: SD(DS22) = 1.00

Vẽ 3 biểu đồ cột trước và sau khi biến đổi DS1. Biểu đồ gồm 2 trục Ox: Giá trị, Oy: số lần lặp [vd: (70, 4); (10, 3)]





Nhận xét:

- Phân phối của các giá trị không thay đổi, gồm 4 giá trị khác nhau và mỗi giá trị có số lần lặp lại lần lượt là 3, 2, 1, 4.

- Các giá trị chỉ thay đổi khi t chuẩn hóa theo các phương pháp khác nhau.

Câu 3: (Vấn đề 3)

Khi nào dùng min-max, z-score, hoặc không dùng cả 2, hoặc dùng cả 2 như là tham số

để chọn mô hình dự báo?

Min-max và z-score là hai kỹ thuật chuẩn hóa dữ liệu thường được sử dụng để chuẩn hóa các biến đầu vào trong các mô hình dự báo. Tuy nhiên, việc sử dụng loại chuẩn hóa nào tốt hơn tùy thuộc vào bộ dữ liệu cụ thể và mục tiêu của mô hình dự báo.

Thường thì min-max và z-score được sử dụng để chuẩn hóa các biến đầu vào để đưa chúng về cùng một phạm vi hoặc đơn vị đo lường. Tuy nhiên, một số mô hình dự báo có thể không cần chuẩn hóa dữ liệu và có thể hoạt động tốt với các giá trị không chuẩn hóa.

Khi lựa chọn giữa hai phương pháp chuẩn hóa này, chúng ta nên cân nhắc các yếu tố sau đây:

- Đối với các biến đầu vào có phân phối chuẩn, z-score là một phương pháp tốt hơn vì nó trung tâm dữ liệu quanh giá trị trung bình và giảm độ lệch chuẩn.

- Đối với các biến đầu vào không phân phối chuẩn, min-max có thể là phương pháp tốt hơn, vì nó giúp đưa các giá trị đầu vào về cùng một phạm vi.

- Nếu mục tiêu của mô hình dự báo là dự đoán giá trị tuyệt đối của biến đầu vào, thì không cần chuẩn hóa dữ liệu.

Trong một số trường hợp, việc kết hợp cả hai phương pháp chuẩn hóa min-max và z-score có thể cải thiện hiệu suất của mô hình.

Tóm lại, lựa chọn phương pháp chuẩn hóa nào tốt hơn phụ thuộc vào loại dữ liệu và mục tiêu của mô hình dự báo. Chúng ta có thể thử nghiệm nhiều phương pháp khác nhau và so sánh hiệu suất của mô hình để tìm ra phương pháp phù hợp nhất cho bộ dữ liệu của mình.