Lâm Minh Tuấn – 20520843 – CS116.N11 – LT01

Sklearn có các phương pháp chuẩn hóa sau đây:

StandardScaler: chuẩn hóa bằng cách đưa dữ liệu về dạng có trung bình bằng 0 và độ lệch chuẩn bằng 1.

Công thức:

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

Trong đó μ là trung bình của dữ liệu và σ là độ lệch chuẩn của dữ liệu

Dùng StandarScaler khi sử dụng các mô hình tuyến tính như Logistic Regression hay SVM, v.v bởi vì các mô hình này thường khởi tạo trọng số bằng 0 hoặc gần bằng 0. Dùng StandarScaler, ta có thể chuẩn hoá dữ liệu về dạng phân phối chuẩn làm cho mô hình có thể cập nhật trọng số nhanh hơn. Tuy nhiên StandarScaler mặc định cho rằng dữ liệu đã có dạng phân phối chuẩn, do đó không nên dùng phương pháp này khi dữ liệu ban đầu không có dạng phân bố chuẩn. Ngoài ra nên dùng StandarScaler khi dữ liệu không có quá nhiều điểm ngoại lai vì phương pháp chuẩn hoá này rất nhạy cảm với các điểm ngoại lai.

MinMaxScaler: chuẩn hoá bằng cách đưa dữ liệu về 1 khoảng cố định (thông thường là [0, 1]).

Công thức:

$$x_{scaled} = \frac{x - x_{\min}}{x_{max} - x_{min}}$$

Dùng MinMaxScaler khi ta cần đưa dữ liệu về 1 khoảng cố định (ví dụ như chuẩn hoá các giá trị pixel trong hình ảnh về 0 đến 255). Tương tự StandarScaler, chỉ nên dùng MinMaxScaler khi dữ liệu không có quá nhiều điểm ngoại lai vì phương pháp chuẩn hoá này rất nhạy cảm với các điểm ngoại lai.

MaxAbsScaler: phương pháp chuẩn hoá này tương tự như MinMaxScaler khi đưa dữ liệu về 1 khoảng cố định. Nếu dữ liệu đều mang giá trị dương, khoảng là [0, 1]. Nếu dữ liệu đều mang giá trị âm, khoảng là [-1, 0]. Nếu có cả giá trị dương và giá trị âm, khoảng là [-1, 1].

Công thức:

$$x_{scaled} = \frac{x}{\max(|x|)}$$

Tương tự như MinMaxScaler, chỉ nên dùng MaxAbsScaler khi dữ liệu không có quá nhiều điểm ngoại lai.

RobustScaler: chuẩn hoá dữ liệu bằng độ trải giữa (IQR: một đại lượng đo lường mức độ phân tán của dữ liệu).

Công thức

$$x_{scaled} = \frac{x - x_{med}}{x_{75} - x_{25}}$$

Trong đó x_{med} là trung vị, x_{75} là tứ phân vị thứ ba và x_{25} là tứ phân vị thứ nhất.

Không như các phương pháp chuẩn hoá ở trên rất nhạy cảm với các điểm ngoại lai, RobustScaler không bị ảnh hưởng bởi các điểm ngoại lai có giá trị lớn. Vì vậy nên sử dụng RobustScaler khi dữ liệu có nhiều điểm ngoại lai khó xử lý.

PowerTransformer: chuẩn hoá dữ liệu phi tuyến tính bằng cách đưa dữ liệu về dạng phân phối chuẩn hoặc gần về dạng phân phối chuẩn (phân phối Gaussian).

Công thức:

Biến đổi Box-Cox yêu cầu dữ liệu toàn giá trị dương trong khi biến đổi Yeo-Johnson hỗ trợ dữ liệu giá trị dương và giá trị âm.

Công thức biến đổi Box-Cox:

$$y_i^{(\lambda)} = \begin{cases} \frac{y_i^{\lambda} - 1}{\lambda} \; nếu \; \lambda \neq 0 \\ \ln y_i \; nếu \; \lambda = 0 \end{cases}$$

Công thức biến đổi Yeo-Johnson:

$$y_i^{(\lambda)} = \begin{cases} \frac{(y_i + 1)^{\lambda} - 1}{\lambda} nếu \; \lambda \neq 0, y \geq 0 \\ \log(y_i + 1) \; nếu \; \lambda = 0, y \geq 0 \\ -\frac{(-y_i + 1)^{(2-\lambda)} - 1}{2 - \lambda} \; nếu \; \lambda \neq 2, y < 0 \\ -\log(-y_1 + 1) \; nếu \; \lambda = 2, y < 0 \end{cases}$$

Dùng PowerTransformer khi mô hình có hiệp phương sai không đồng nhất (phương sai không phải là hằng số) và khi mô hình yêu cầu dữ liệu có dạng phân phối chuẩn (đối xứng), ví dụ các mô hình dựa trên khoảng cách như KNN hay K-means, DBSCAN.

QuantileTransformer (dữ liệu đầu ra có dạng phân phối đồng nhất/chuẩn): chuẩn hoá dữ liệu phi tuyến tính bằng cách đưa dữ liệu về dạng phân phối đồng nhất/chuẩn dựa trên thông tin lượng tử.

Các bước chuẩn hoá của phương pháp này:

1. Tính toán hàm phân phối tích luỹ của biến.
2. Dùng hàm này để ánh xạ giá trị thành phân phối đồng nhất/chuẩn.
3. Ánh xạ các giá trị thu được tới phân phối đầu ra mong muốn bằng cách sử dụng hàm lượng tử liên quan.

Vì phương pháp chuẩn hoá này thay đổi phân phối của các biến nên các mối quan hệ tuyến tính giữa các biến có thể bị sai lệch. Vì vậy tốt nhất dùng phương pháp chuẩn hoá này cho dữ liệu phi tuyến tính. Ngoài ra, phương pháp này còn có thể xử lý tốt các điểm ngoại lai.

Normalizer: chuẩn hoá từng mẫu (theo dòng) dựa trên norm của mẫu đó. Trong Sklearn hỗ trợ l1, l2 và max norm.

Công thức:

$$x_{scaled} = \frac{x}{\|x\|_p}$$

Trong đó:

$$\|x\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

L1 norm: p=1, L2 norm: p=2, max norm: p = ∞.

Bởi vì chỉ chuẩn hoá theo từng mẫu một cách độc lập nên không sử dụng Normalizer như một bước tiền xử lý dữ liệu. Normalizer chủ yếu được sử dụng để kiểm soát kích thước vector trong quá trình lặp lại, ví dụ như vector tham số trong quá trình train để tránh sự không ổn định do giá trị quá lớn hay quá nhỏ.