Họ và tên: Đặng Anh Tiến
MSSV: 20520800
Môn: CS116.N11
Bài tập LT01-Chuẩn hoá giá trị liên tục
Các phương pháp chuẩn hoá giá trị liên tục trong scikit-learn? Công thức? Khi nào sử dụng?

Bài làm

StandardScaler

Công thức:

$$X\_scale = \frac{X - \mu}{\sigma}$$

Trong đó:

μ: mean.
σ: standard deviation.

Sử dụng khi các feature tuân theo Normal Distribution, các feature được scale về mean bằng 0 và standard deviation bằng 1.

MinMaxScaler

Công thức:

$$X_{std} = \frac{X - X\_min}{X_{max} - X\_min}$$
$$X_{scaled} = X_{std} * (\max - \min) + min$$

Trong đó:

X_min: Min của feature.
X_max: Max của feature.
max, min: feature range.

Sử dụng khi đã biết trước được khoảng trên dưới của feature. Feature sẽ được scale về khoảng đã biết trước.

MaxAbsScaler

Công thức:

$$X_{scale} = \frac{X}{abs(X\_max)}$$

Trong đó:

X_max là giá trị lớn nhất.

Tương tự MinMaxScaler nhưng scale các feature trong [-1, 1].

RobustScaler

Công thức:

$$X_{scale} = \frac{X - X\_med}{Q3(x) - Q1(x)}$$

Trong đó:

X_med là median của X.

Q1(x) là 1st quantile của X.

Q2(x) là 3rd quantile của X.

Trong StandardScaler, giá trị mean và standard deviation có thể bị ảnh hưởng xấu bởi outlier. RobustScaler tính dựa trên median và IQR sẽ hạn chế ảnh hưởng của outlier.

Normalizer

Công thức:

- **L2 norm:**

$$x_{scale} = \frac{x}{||x||_2}$$

Trong đó:

$$\left|\left|x\right|\right|_2 = \sqrt{\sum x^2}$$

- **L1 norm:**

$$x_{scale} = \frac{x}{||x||_1}$$

Trong đó:

$$||x||_1 = \sum |x|$$

- **Max**

$$x_{scale} = \frac{x}{\max(x)}$$

Dùng để scale mỗi sample sao cho các sample có norm bằng 1.