Họ và tên: Nguyễn Nhật Trường

MSSV: 20522087

Bài tập lí thuyết 01: Chuẩn hóa giá trị liên tục

Có những phương pháp chuẩn hóa giá trị liên tục nào trong Sklearn? Công thức từng phương pháp chuẩn hóa là gì? Khi nào ta nên sử dụng hàm chuẩn hóa nào?

Phương pháp chuẩn hóa	Công thức	Khi nào nên sử dụng
MaxAbsScaler	$x_{scaled} = \frac{x}{\max(\lvert x \rvert)}$ $x_{scaled}[-1,1]$	Sử dụng hàm chuẩn hóa MaxAbsScaler khi độ lệnh chuẩn của các đặc trưng quá nhỏ và muốn giữ lại giá trị 0 trong ma trận thưa thớt (sparse data).
MinMaxScaler	$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$ $x_{scaled}[0,1]$	Khi muốn giữ nguyên hình dạng phân phối ban đầu. Không làm giảm tầm quan trọng của các yếu tố ngoại lai. Ít làm gián đoạn thông tin của dữ liệu gốc.
Normalizer	Mỗi hàng được chuẩn hóa bằng cách áp dụng chuẩn hóa l2 (Euclidian). Nếu mỗi phần tử được bình phương và tính tổng thì tổng sẽ bằng 1. Cũng có thể dùng chuẩn hóa l1 (Mahattan). $y = x/z$ Với z có thể là: • $l1$: $z = \lVert x \rVert_1 = \sum_{i=1}^{n} \lvert x_i \rvert$ • $l2$: $z = \lVert x \rVert_2 = \sum_{i=1}^{n} \lvert {x_i}^2 \rvert$	+Dùng khi muốn chuẩn hóa mỗi dòng, không dùng để chuẩn hóa các cột. +Hữu ích khi dùng để kiểm soát kích thước của một vector trong quá trình lặp đi lặp lại để tránh sự không ổn định do các giá trị số quá lớn. Hữu ích trong

	• $max: z = \max(x_i)$ • x: giá trị chưa chuẩn hóa	hồi quy hơn phân loại +Hiếm khi được sử dụng.
RobustScaler	$x_{scaled} = \frac{x - median}{IQR}$ Với: • x: giá trị chưa chuẩn hóa • $IQR = Q75 - Q25$ • Median: giá trị trung vị	Sử dụng RobustScaler khi dữ liệu có các ngoại lai (outliers) và mong muốn làm giảm tầm ảnh hưởng của các outliers với dữ liệu của mình.
StandardScaler	$z = \frac{x - \mu}{\sigma}$ Với: • μ: giá trị trung bình • σ: độ lệch chuẩn • x: giá trị chưa chuẩn hóa	+ Sử dụng StandardScaler khi mong muốn mỗi đặc trưng có giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1. Nếu muốn dữ liệu được phân phối ổn định với việc chuyển đổi dữ liệu. + StandardScaler không có phạm vi giới hạn nghĩa là kể cả có những giá trị ngoại lai trong dữ liệu chúng cũng không bị ảnh hưởng bởi quá trình standardlization.
FunctionTransformer	$x_{scaled} = \log(x)$	FunctionTransformer phù hợp với bộ dữ liệu chứa giá trị ngoại lai lớn (giá trị khác biệt so với phần còn lại).

Tài liệu tham khảo:

[Min Max Normalization in data mining | T4Tutorials.com]

API Reference — scikit-learn 1.1.2 documentation