Bài 2: Mô hình Hồi quy tuyến tính

Hướng dẫn 1: Sử dụng Scaler

1. Phân tích Toán học

Cho tập dữ liệu ban đầu:

$$D = \{(X, y) \mid X \in R^{m \times n} \text{ và } y \in R^m\}$$

Giá trị của các phần tử $x_{ij} \in X$ và $y_i \in y$ có thể rất phân tán và ảnh hưởng đến việc tìm bộ thông số tối ưu w^* của mô hình - $h_w(X)$. Do vậy các giá trị x_{ij} và y_i cần được điều chỉnh để giúp quá trình tìm w^* được thuận lợi hơn.

Lưu ý:

- Có một số ý kiến trái ngược về việc điều chỉnh (hoặc không cần điều chỉnh) vector *y*;
- *Gợi ý: bạn có thể thực hiện cả 2 thao tác này và phân tích các chỉ số (ví dụ: MSE, r^2, ...) để có lựa chọn phù hợp cho từng trường hợp dữ liệu cụ thể.*

2. Một số phương pháp điều chỉnh giá trị dữ liệu (data scaling)

Một cách tổng quát, phương pháp điều chỉnh dữ liệu (data scaling) có thể xem là:

$$f(X): R^{m \times n} \rightarrow R^{m \times n}$$

Với hàm $f(x)$ có nhiều định nghĩa khác nhau như:

- Chuẩn hóa dữ liệu (***Data normalization***) là thao tác đưa các giá trị của tập dữ liệu ban đầu về khoảng $[0,1]$: $x_i = \frac{x_i - min}{\max - min}$ *(Trong scikit-learn, hàm tương đương là MinMaxScaler)*;
- Chuẩn hóa dữ liệu theo phương pháp điều chỉnh hướng tâm (***center scaling***) được định nghĩa là: $x_i = \frac{x_i - \bar{x}}{\delta}$ với $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$ và $\delta = \sqrt{\frac{\sum(x_i - \bar{x})^2}{m}}$ *(Trong scikit-learn, hàm tương đương là StandardScaler)*

3. Điều chỉnh dữ liệu trong sklearn

Thư viện sklearn.preprocessing cung cấp các phương pháp tiền xử lý dữ llieeujtrong đó bao gồm các phương pháp điều chỉnh dữ liệu (data scaling): MinMaxScaler, StandardScaler và các phương pháp khác.

Quy tắc điều chỉnh dữ liệu khi sử dụng sklearn là:

- Bước 1: Khởi tạo bộ điều chỉnh dữ liệu
- Bước 2: Đưa dữ liệu vào bộ điều chỉnh thông qua hàm fit()
- Bước 3: Thực hiện điều chỉnh dữ liệu bằng hàm transform()

Ví dụ về điều chỉnh dữ liệu bằng MinMaxScaler

```
import os
import numpy as np
from sklearn.preprocessing import MinMaxScaler

D = np.loadtxt(os.path.join("D:/data/hocmay", "ex1data2.txt"),
               delimiter=",")
print('Kích thước của tập dữ liệu: ', D.shape)
print('Giá trị của tập dữ liệu: ')
print(D)
print('Thực hiện MinMaxScaler')
#Khởi tạo bộ điều chỉnh dữ liệu
scaler = MinMaxScaler()
#Phải thực hiện thao tác fit(data) trước khi điều chỉnh dữ liệu
scaler.fit(D)
#Thực hiện điều chỉnh dữ liệu
D = scaler.transform(D)
print('Kích thước của tập dữ liệu: ', D.shape)
print('Giá trị của tập dữ liệu: ')
print(D)
print('Lấy ra tập dữ liệu X, y')
X, y = D[:,:-1], D[:, -1]
print('Kích thước tập X: ', X.shape)
print('Kích thước vector y: ', y.shape)
```