Báo cáo đồ án Linear Regession

I. Đánh giá kết quả:

CHỨC NĂNG	KẾT QUẢ
ĐỌC VÀ TÁCH DỮ LIỆU	Hoàn thành
LINEAR REGRESSION	Hoàn thành
CROSS VALIDATION	Hoàn thành
XÂY DỰNG MÔ HÌNH TRÊN 11 TÍNH CHẤT	Hoàn thành
CHỌN TÍNH CHẤT TỐT NHẤT ĐỂ XÂY DỰNG MÔ HÌNH	Hoàn thành
XÂY DỰNG MÔ HÌNH RIÊNG	Hoàn thành
BÁO CÁO ĐỒ ÁN	Hoàn thành

II. Mô tả bài làm:

1. Các module sử dụng:

Bộ thư viện pandas, numpy và sklearn. Trong đó sklearn có sử dụng để xây dựng mô hình hồi quy tuyến tính.

2. Đọc và tách dữ liệu:

Sử dụng các hàm cài đặt sẵn.

3. Linear Regression:

Xây dựng mô hình theo công thức Ax = b:

- A là ma trận dữ liệu.
- B là header của mỗi dòng dữ liệu trong ma trận A.

Với mô hình trên, ta có thể tìm mô hình theo công thức: $\hat{x} = A^{\dagger} \cdot b$.

Khi đó, mô hình sẽ đi qua gốc tọa độ của đồ thị nên bị hạn chế.

Sau đó, chọn mô hình theo công thức $Ax + b_0 = b$

4. Cross Validation:

Dùng K-Fold Cross Validation trong module sklearn.

- Chia data thành k nhóm.
- Với mỗi nhóm:
  - Sử dụng nhóm hiện tại để đánh giá kết quả của mô hình.
  - Các nhóm còn lại dùng để huấn luyện (train) mô hình.
  - Huấn luyện (train) mô hình.
  - Đánh giá.
- Tổng hợp kết quả qua các đánh giá.

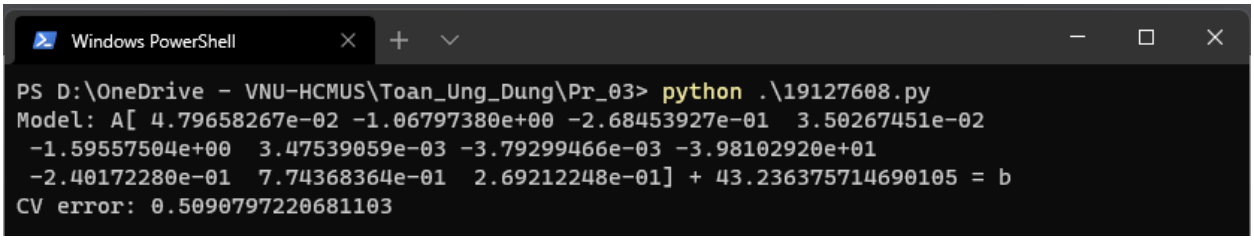
Dùng hàm sklearn.model_selection.Kfold để chia tập dữ liệu ra làm k nhóm và phân ra làm k bộ dữ liệu với bộ train/test khác nhau.

- Xây dựng mô hình trên tập train, ta được $\hat{x}$, b_0
- Áp dụng mô hình này lên tập test, ta được $A_{test} . \hat{x} = b'$
- Tính sai số so với header của tập test $|b' - b_{test}|$. Kết quả tìm được sẽ là một ma trận có kiểu giống b_{test}, khi đó ta tính trung bình của ma trận này để được sai số của mô hình trên tập train/test đó.
- Chạy hết tất cả các tập train/test được chia ra ở trên, tính trung bình các sai số này để lấy sai số trung bình.

Số nhóm thường được dùng là k = 10

5. **Xây dựng mô hình trên 11 tính chất:**

Dùng LinearRegression() đã cài đặt để tìm modle trên tất cả tính chất



6. **Chọn tính chất tốt nhất để xây dựng mô hình:**

Chạy CrossValidation() để tìm sai số từng cột của mô hình dựa trên các tính chất.
Tìm tính chất có giá trị sai số bé nhất, tính chất đó sẽ là tính chất tốt nhất.
Chạy LinearRegression() để tìm model dựa theo tính chất vừa tìm được.

```
Best property: alcohol
Model: A[0.37403439] + 1.7807151719965795 = b
CV error: 0.5689866007906972
```

7. **Xây dựng mô hình riêng:**

Chọn các tính chất tốt nhất (top 10) rồi chạy CrossValidation()
Chọn ra bộ có sai số thấp nhất để xây dựng mô hình.
Chạy LinearRegression() để tìm model dựa trên các tính chất này.

```
Best properties: ['alcohol', 'volatile acidity', 'total sulfur dioxide', 'citric acid',
'sulphates', 'density', 'fixed acidity', 'chlorides', 'free sulfur dioxide']
Model: A[ 2.79189114e-01 -1.08542522e+00 -3.27611003e-03 -2.50107046e-01
  7.55280875e-01 -3.10636950e+01  5.89809754e-02 -1.44110179e+00
  2.85695898e-03] + 33.60217206120657 = b
CV error: 0.5100839496676576
```

So sánh sai số giữa mô hình dựa trên 11 tính chất và mô hình tự xây dựng:
0.5689866007906972 và **0.5100839496676576**
Có thể thấy sự thay đổi rõ rệt (từ 0.56 xuống 0.51) cho thấy sự hiệu quả của mô hình này.

III. Tham khảo:

Linear Regression in Python (Real Python)

Choice of K in K-Fold cross-validation (Stack exchange)

In the LinearRegression method in sklearn, what exactly is the fit_intercept parameter doing? (Stack overflow)