LAB 04 – PROJECT 03: LINEAR REGRESSION

- Họ tên: Lê Yến Nhi

- MSSV: 19127498

- Đã hoàn thành: 100%

Mô hình đánh giá chất lượng rượu sử dụng phương pháp hồi quy tuyến tính

a. Sử dụng toàn bộ 11 đặc trưng đề bài cung cấp

1. Ý TƯỞNG THỰC HIỆN

- Dùng thư viện Pandas để đọc file 'wine.csv' (dữ liệu của một dòng ngăn cách nhau bởi dấu ';'), lưu dự liệu vào dataframe `df`.

- Tạo dataframe `x` chỉ chứa data làm biến phụ thuộc.

- Tạo dataframe `y` chứa data của đặc trưng 'quality' làm biến độc lập (biến mục tiêu).

- Tạo mô hình với `x` và `y`: `reg.fit(x, y)`

- In hệ số hồi quy (Coefficients) và sai số (Normal error)

2. KẾT QUẢ

Phương trình hồi quy:

[quality] = 0.047966 × [fixed acidity] + (-1.067974) × [volatile acidity]
+ (-0.268454) × [citric acid] + 0.035027 × [residual sugar]
+ (-1.595575) × [chlorides] + 0.003475 × [free sulfur dioxide]
+ (-0.003793) × [total sulfur dioxide] + (-39.810292) × [density]
+ (-0.240172) × [pH] + 0.774368 × [sulphates]
+ 0.269212 × [alcohol] + 43.23637571469012

3. NHẬN XÉT

- Độ sai lệch nhỏ, mô hình này tốt nhất trong tất cả các loại mô hình sử dụng phương pháp hồi quy tuyến tính (11 đặc trưng).

b. Sử dụng duy nhất 1 đặc trưng cho kết quả tốt nhất (Gợi ý: Phương pháp Cross Validation)

1. Ý TƯỞNG THỰC HIỆN

- Chia dữ liệu thành 5 set, mỗi set gồm 5 phần: 4 phần để huấn luyện (train) và 1 phần để kiểm tra (test).

- Dùng `i` duyệt qua từng đặc trưng của `x` (gồm 11 đặc trưng), với mỗi đặc trưng:

    + Duyệt qua từng set, với mỗi set:

_ Tạo `X_train, y_train` là phần train của một set và `X_test, y_test` là phần test của một set.

_ Tạo `X_train_child, X_test_child` là phần train thứ `i` và phần test thứ `i` của một set.

_ Tạo mô hình với `X_train_child` và `y_train.`

_ Tiên đoán giá trị `y_pred` của `X_test_child` trong mô hình trên.

_ Tính độ sai lệch bình phương `mse` giữa `y_test` và `y_pred` của từng đặc trưng.

+ So sánh độ sai lệch giữa các đặc trưng.

- Đặc trưng nào có sự sai lệch nhỏ nhất sẽ cho kết quả tốt nhất (đặc trưng này nằm ở cột `min_i`)

2. KẾT QUẢ

- Đặc trưng cho kết quả tốt nhất nằm ở cột thứ **10** là "**alcohol**"
- Phương trình hồi quy:

$$[quality] = 0.37403439 \times [alcohol] + 1.780715171996576$$

3. NHẬN XÉT

- Đặc trưng có sự sai lệch nhỏ nhất sẽ tạo mô hình cho kết quả gần với giá trị thực nhất.

+ "alcohol" cho kết quả tốt nhất, "density" cho kết quả xấu nhất.

- Ý tưởng thực hiện này sử dụng nhiều vòng lặp, nhưng nhìn chung chạy khá nhanh.

c. Mô hình riêng cho kết quả tốt nhất.

1. Ý TƯỞNG THỰC HIỆN

- Chia dữ liệu thành 5 **set**, mỗi set gồm 5 phần: 4 phần để huấn luyện (train) và 1 phần để kiểm tra (test).
- Duyệt qua từng **set**, với mỗi set:

_ Tạo `X1_train, y1_train` là phần train của một set và `X1_test, y1_test` là phần test của một set.

_ Tạo mô hình với `X1_train` và `y1_train.`

_ Tiên đoán giá trị `y_pred` của `X1_test` trong mô hình trên.

_ Tính độ sai lệch bình phương `mse` giữa `y1_test` và `y_pred` của từng **set**.

_ So sánh độ sai lệch giữa các **set**.

_ **set** nào có sự sai lệch nhỏ nhất sẽ được chọn để tạo mô hình. Gán set này vào `X1_trainMin, y1_trainMin.`

- Tạo mô hình với `X1_trainMin` và `y1_trainMin.`
- In hệ số hồi quy (Coefficients) và sai số (Normal error).

2. KẾT QUẢ

Phương trình hồi quy:

[quality] = 0.043252 × [fixed acidity] + (-1.016627) × [volatile acidity]
+ (-0.203611) × [citric acid] + 0.027407 × [residual sugar]
+ (-1.447932) × [chlorides] + 0.004248 × [free sulfur dioxide]
+ (-0.004393) × [total sulfur dioxide] + (-33.550069) × [density]
+ (-0.136899) × [pH] + 0.726958 × [sulphates]
+ 0.270375 × [alcohol] + 36.68193527097975

3. NHẬN XÉT

- Mô hình này chọn **set** có sự sai lệch nhỏ nhất nên sẽ cho kết quả gần với giá trị thực nhất.
- Độ sai lệch nhỏ, mô hình này cho kết quả tốt.

TÀI LIỆU THAM KHẢO:

https://codetudau.com/posts/hoi-quy-tuyen-tinh/

https://towardsdatascience.com/what-are-the-best-metrics-to-evaluate-your-regression-model-418ca481755b

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html#sklearn.model_selection.KFold