UNIVERSITY OF SCIENCE
FALCUTY OF INFORMATION TECHNOLOGY

SUBJECT: Applied Mathematics and Statistics

PROJECT 3: Linear Regression

Name: Nguyễn Quốc Huy
Class: 20CLC02
Student ID: 20127188

Lecturers: **VŨ QUỐC HOÀNG**
NGUYỄN VĂN QUANG HUY
LÊ THANH TÙNG
PHAN THỊ PHƯƠNG UYÊN

Mục lục

I. Yêu cầu đồ án

- Xây dựng mô hình dự đoán tuổi thọ trung bình sử dụng hồi quy tuyến tính
- Yêu cầu 1a: Sử dụng toàn bộ 10 đặc trưng đề bài cung cấp
  - Huấn luyện 1 lần duy nhất cho 10 đặc trưng trên toàn bộ tập huấn luyện (`train.csv`)
  - Thể hiện công thức cho mô hình hồi quy (tính theo 10 đặc trưng trong)
  - Báo cáo 1 kết quả trên tập kiểm tra (`test.csv`) cho mô hình vừa huấn luyện được

- Yêu cầu 1b: Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất
  - Thử nghiệm trên toàn bộ (10) đặc trưng đề bài cung cấp
  - Yêu cầu sử dụng phương pháp 5-fold Cross Validation để tìm ra đặc trưng tốt nhất
  - Báo cáo 10 kết quả tương ứng cho 10 mô hình từ 5-fold Cross Validation (lấy trung bình)
  - Thể hiện công thức cho mô hình hồi quy theo đặc trưng tốt nhất (tính theo đặc trưng tốt nhất tìm được)
  - Báo cáo 1 kết quả trên tập kiểm tra (`test.csv`) cho mô hình tốt nhất tìm được

- Yêu cầu 1c: Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất
  - Xây dựng `m` mô hình khác nhau (tối thiểu 3), đồng thời khác mô hình ở 1a và 1b
  - Mô hình có thể là sự kết hợp của 2 hoặc nhiều đặc trưng
  - Mô hình có thể sử dụng đặc trưng đã được chuẩn hóa hoặc biến đổi (bình phương, lập phương...)
  - Mô hình có thể sử dụng đặc trưng được tạo ra từ 2 hoặc nhiều đặc trưng khác nhau (cộng 2 đặc trưng, nhân 2 đặc trưng

II. Ý tưởng thực hiện

- Với yêu cầu 1a
  - Do đã được cô giới thiệu về OLS Linear Regression và công thức tính MSE (Sai số bình phương trung bình) trong buổi học trên lớp nên ở yêu cầu này em sử dụng lại kiến thức đó đồng thời đề bài yêu cầu tính RMSE nên em sẽ sử dụng $\sqrt{MSE}$ để ra kết quả mong muốn.
- Với yêu cầu 1b
  - Ở yêu cầu này để bài muốn dùng 1 đặc trưng và trả ra đặc trưng tốt nhất, sử dụng phương pháp 5 fold cross validation[1] để chia tập train 5 lần, mỗi lần chia thành 2 phần, một phần làm tập test còn lại làm tập để train. Với mỗi đặc trưng em tính ra làm 5 RMSE, sau cùng em tính trung bình cộng của 5 RMSE này, so sánh với các đặc trưng còn lại để ra được RMSE nhỏ nhất từ đó suy ra đặc trưng tốt nhất. Sau

khi có được đặc trưng tốt nhất ta lấy cột đó train lại với tập test để ra giá trị RMSE tốt nhất.

- Với yêu cầu 1c
  - Với yêu cầu này đề bài yêu cầu xây dựng ra mô hình tự chọn, sau khi có được mô hình tự chọn thực hiện giống với câu b để ra được RMSE, thực hiện trên các mô hình khác từ đó đưa ra được mô hình nào có RMSE nhỏ nhất nghĩa là mô hình này tốt nhất, từ đó train lại với tập test để ra được giá trị RMSE tốt nhất.

III. Thư viện

Giới thiệu thư viện

Ở đây em dùng 3 thư viện chính cho đồ án:

- Pandas (Dùng để đọc các file .csv và đưa dữ liệu về dạng Dataframe (Khung dữ liệu) để dễ dàng xử lý dữ liệu)
- Numpy (Dùng để tính toán trung bình, tính giá trị nhỏ nhất, xử lý mảng)
- Sklearn import model_selection (Dùng để xáo trộn dữ liệu và chia dữ liệu thành các phần)

Các hàm dùng trong thư viện

- np.mean(): Tính toán trung bình
- np.min(): Tìm giá trị nhỏ nhất
- np.zeros(): Tạo ra ma trận có giá trị bằng 0 với kích thước tùy chỉnh
- np.array(): Đưa giá trị về dạng mảng
- np.sqrt(): Tính giá trị căn

- model_selection.Kfold()[2]: Dùng để xáo trộn dữ liệu và chia dữ liệu thành các phần mong muốn

- iloc[][3]: Dùng để duyệt cũng như lấy giá trị/cột mong muốn

IV. Mô tả các hàm

Xây dựng hàm cần thiết

- Class OLSLinearRegression và hàm mse()
  - Hàm mse() này cô đã giới thiệu trên lớp, công dụng chính là tính toán sai số trung bình bởi công thức sau

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

- Trong class OLSLinearRegression có các hàm: fit(), get_params(), predict()
  - Hàm fit có chức năng tính toán giá trị x (với đề bài sẽ trọng số w) thông qua công thức w = $(A^T * A)^{-1} * A^T *$ B
  - Hàm get_params sẽ trả về w
  - Hàm predict tính toán kết quả dự đoán thông qua việc lấy trọng số w nhân cho mô hình đặc trưng

- Các hàm chức năng
  - Hàm task_a():

- Đầu tiên chúng ta tính toán trọng số w thông qua hàm fit trong class OLSLinearRegression đã khai báo bên trên, hai thông số truyền vào sẽ là X_train (10 đặc trưng của tập train) và y_train (Cột mục tiêu Life expectancy).
- Sau khi có được trọng số w chúng ta tính toán giá trị y_predict thông qua việc gọi hàm predict và 2 thông số truyền vào sẽ là trọng số w vừa tính và X_test (10 đặc trưng của tập test)
- Sau khi có được giá trị y_predict dùng hàm mse(), truyền vào 2 thông số là tập y_test (giá trị mục tiêu của tập test) và y_predict vừa tính bên trên đồng thời lấy căn để ra được giá trị RMSE mà đề bài yêu cầu, dùng hàm get_params để ra được giá trị của w dùng để tính được công thức Life expectancy.

  - Hàm task_b():

- Đầu tiên em sẽ tạo ra 2 mảng arr_w và arr_r để chứa 2 giá trị Weight và RMSE, kích thước của mảng nào bao gồm tổng số các cột của tập train (train.shape[1] – 1)
- Sau đó sẽ chạy vòng lặp để chia tập train thành tập X_train thành 5 phần và xáo trộn dữ liệu, trong mỗi phần em sẽ tính toán ra các giá trị train_feature, train_label, test_feature, test_label tương đương với các giá trị X_train, y_train, X_test, y_test từ đó tính toán RMSE (sử dụng hàm task_a() đã cài đặt bên trên), cộng dồn các giá trị RMSE của 1 cột đặc trưng lại sau đó lấy chia cho 5 sẽ ra giá trị RMSE trung bình của đặc trưng đó, làm tương tự cho các đặc trưng còn lại sau đó tìm min và đưa ra đặc trưng tốt nhất.

- Sau khi đã biết được cột đặc trưng nào tốt nhất chúng ta sẽ lấy đặc trưng đó đi tính toán lại giá trị RSME của cột đặc trưng đó bên tập train và cột đặc trưng đó bên tập test từ đó đưa ra được giá trị RMSE tốt nhất

- Hàm task_c():

- Về yêu cầu của c chúng ta thực hiện cũng khá giống với câu b, câu c yêu cầu xây dựng mô hình riêng đó có thể là một cột đặc trưng bình phương, 3 cột đặc trưng tự chọn hoặc tạo thêm 1 đặc trưng mới dựa trên các đặc trưng đã có trước đó, sau khi có được mô hình thì chúng ta thực hiện giống câu b từ đó cũng tính toán được giá trị RMSE của nó, huấn luyện trên nhiều mô hình khác nhau sẽ ra được nhiều giá trị RMSE khác nhau từ đó dùng hàm min() để tính toán ra giá trị RMSE nhỏ nhất tương ứng với mô hình tốt nhất
- Sau khi có mô hình tốt nhất chúng ta lại dùng các cột của mô hình đó đi tính toán giá trị RMSE với tập test(các cột X_test tương ứng) để ra được giá trị RMSE tốt nhất.

  - Lý do chọn mô hình
- Em dùng vòng lặp để thử các kết quả với từ 2 đặc trưng đến 9 đặc trưng(10 đặc trưng trùng với câu a) và thử cũng từ 2 đặc trưng đến 9 nhưng mũ 3 lên xem có chênh lệch nhau nhiều hay không
- Em lựa chọn như vậy bởi vì em thấy ở câu a dùng 10 đặc trưng số cũng đã khá nhỏ nên em muốn thử xem với 9 đặc trưng thì có khác nhau không

V. Báo cáo kết quả và nhận xét

- Báo cáo kết quả
  - Câu a

Loại mô hình	Kết quả RMSE
Mô hình 10 đặc trưng	7.0640464305840505

Công thức hồi quy

Life expectancy = 0.0151013627 * Adult Mortality + 0.0902199807 * BMI + 0.0429218175 * Polio + 0.139289117 * Diphtheria + (-0.567332827) * HIV/AIDS + (-0.000100765115) * GDP + 0.740713438 * Thinness age 10-19 + 0.190935798 * Thinness age 5-9 + 0.245059736 * Income composition of resources + 2.39351661 * Schooling

- Câu b

STT	Đặc trưng	Kết quả RMSE
0	Adult Mortality	46.219096
1	BMI	27.963911
2	Polio	18.027210
3	Diphtheria	16.027598
4	HIV/AIDS	67.178847
5	GDP	60.231533
6	Thinness age 10-19	51.793758
7	Thinness age 5-9	51.700257
8	Income composition of resources	13.194252
9	Schooling	11.789461

Sau khi train với tập test:
RMSE = 10.260950391655376
Weight = [5.5573994]

Công thức hồi quy
Life expectancy = 5.5573994 * Schooling

- Câu c

STT	Mô hình	Kết quả RMSE
0	2 đặc trưng(Adult Mortality,BMI)	22.623468
1	3 đặc trưng(Adult Mortality,BMI,Polio)	14.964050
2	4 đặc trưng(Adult Mortality,BMI,Polio,Diphtheria)	13.147579
3	5 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS)	12.990777
4	6 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS, GDP)	12.891247
5	7 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS, GDP, Thinness age 10-19)	11.675221
6	8 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS, GDP, Thinness age 10-19, Thinness age 5-9)	11.539536

7	9 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS, GDP, Thinness age 10-19, Thinness age 5-9, Income composition of resources)	8.735078
8	2 đặc trưng(Adult Mortality,BMI) mũ 3	39.418300
9	3 đặc trưng(Adult Mortality,BMI,Polio) mũ 3	23.246128
10	4 đặc trưng(Adult Mortality,BMI,Polio,Diphtheria) mũ 3	21.572479
11	5 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS) mũ 3	21.456716
12	6 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS, GDP) mũ 3	21.466164
13	7 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS, GDP, Thinness age 10-19) mũ 3	20.378760
14	8 đặc trưng(Adult Mortality, BMI, Polio, Diphtheria, HIV/AIDS, GDP, Thinness age 10-19, Thinness age 5-9) mũ 3	20.301298

Sau khi train với tập test
RMSE = 7.3379852626993385
Weight = [1.73671084e-02 1.54066222e-01 7.80981602e-02 1.82294109e-01
-4.32676946e-01 -5.40471095e-06 7.09662791e-01 2.60263436e-01
5.36120768e+01]

Công thức hồi quy

Life expectancy = 0.0173671084 * Adult Mortality + 0.0154066222 * BMI + 0.0780981602 * Polio + 0.182294109 * Diphtheria + (-0.432676946) * HIV/AIDS + (-0.00000540471095) * GDP + 0.709662791 * Thinness age 10-19 + 0. 260263436 * Thinness age 5-9 + 0.536120768 * Income composition of resources

VI. BẢNG CÔNG VIỆC

Công việc	Mức độ hoàn thành	Ghi chú
Câu 1a	100%	
Câu 1b	100%	
Câu 1c	90%	Em không nghĩ mô hình câu c của em chọn sẽ là tốt nhất

VII. NGUỒN THAM KHẢO

[1]: https://scikit-learn.org/stable/modules/cross_validation.html
[2]: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html
[3]: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html