ĐẠI HỌC QUỐC GIA
THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN 3

Linear Regression

MÔN: TOÁN ỨNG DỤNG VÀ THỐNG KÊ CHO CÔNG NGHỆ THÔNG TIN

SINH VIÊN : **NGUYỄN HOÀI MẪN**
MSSV : **20127561**
LỚP : **20CLC05**
KHOA : **CÔNG NGHỆ THÔNG TIN**

Ngày 02 tháng 08 năm 2022

Mục lục

I. Các chức năng đã hoàn thành:

STT	Yêu cầu	Tỉ lệ hoàn thành(%)
1	1.a	100
2	1.b	100
3	1.c	100

II.Các thư viện đã sử dụng:

- **numpy**: Dùng để thực hiện các thao tác trên ma trận.

- **pandas**: Trích xuất dữ liệu từ csv.

- **KFold**: Dùng để tìm ra mô hình tốt nhất bằng phương pháp 5-fold Cross Validation.

- **mean_squared_error**: Tính RMSE cho mô hình.

III. Mô tả về hàm chức năng:

1 Class OLSLinear Regression.

1.1 fit(self, X, y)

- Hàm tìm ra các giá trị theta.

- Input:

    + X: mảng ma trận các đặc trưng.

    + y: mảng ma trận giá trị.

1.2 get_params(self)

- Hàm nhận các hệ số w của mô hình hồi quy tuyến tính

1.3. Hàm predict(self, X)

- Hàm dự đoán dựa trên mô hình và dữ liệu của các đặc trưng truyền vào.

- Input: Dữ liệu của các đặc trưng truyền vào.

- Output: Mảng gồm các giá trị sau khi dự đoán.

2. Các hàm cài đặt cho cho yêu cầu 1 b:

2.1 ChooseBest_b(X, y)

- Hàm thực hiện tìm ra đặc trưng tốt nhất

- Input

    + Ma trận các đặc trưng.
    + Ma trận giá trị.

- Output:

    + Mảng RMSE tương ứng với các đặc trưng.

    + Vị trí của đặc trưng tốt nhất.

2.2 Cross_Validation(X, y)

- Triển khai phương pháp K-ford Cross Validation

-Input:

    + Ma trận các đặc trưng.
    + Ma trận giá trị.

- Output:

    + Trung bình cộng năm lần của mỗi đặc trưng.

- Ý tưởng:

    + Sử dụng thuật toán Kfold để xáo trộn dữ liệu và chia dữ liệu X_train thành các khung frame với số lượng dòng dữ liệu là X-train.shape[0]/5

    + Lặp qua từng khung frame để huấn luyện theo mô hình theta[10] * Schooling và dự đoán thông qua việc chạy class OLSLinearRegression

    + Sử dụng hàm RMSE để tính độ lệch chuẩn của phần dư

2.3 ChooseBest_b(X, y)

- Hàm thực hiện tìm ra đặc trưng tốt nhất

- Input

+ Ma trận các đặc trưng.
+ Ma trận giá trị.

- Output:

+ Mảng RMSE ứng với các đặc trưng.

+ Vị trí của đặc trưng tốt nhất.

2.4 Table(data, headers)

- Hàm thực hiện in ra bảng RMSE của các đặc trưng.

- Input

+ Mảng RMSE của các đặt trưng.
+Mảng tên của các đặc trưng.

- Output:

+ Bảng kết quả RMSE của các đặc trưng

3. Các hàm cài đặt cho cho yêu cầu 1 b:

3.1 preproces1(x_9, y):

- Hàm tiền xử lý dữ liệu theo mô hình $Y=\theta_0+\theta_{10}\ln(X_{10}\text{^}2)$.

- Input:

+ x_9: Mảng ma trận của đặc trưng Schooling .
+ y: Mảng ma trận giá trị mục tiêu.

- Output:

+ X: Mảng ma trận đặc trưng sau khi biến đổi theo mô hình.

+ y: Mảng ma trận giá trị.

3.2 preproces2(, x_9 , x_{10}, y):

- Hàm tiền xử lý dữ liệu theo mô hình $Y= \theta_0+\theta_9 X_9\text{^}3+\theta_{10}X_{10}\text{^}5$

- Input:

+ x_9: Mảng ma trận của đặc trưng Income composition of resources.
+ x_{10}: Mảng ma trận của đặc trưng Schooling.
+ y: Mảng ma trận giá trị.

- Output:

+ X: Mảng ma trận đặc trưng sau khi biến đổi theo mô hình.
+ y: Mảng ma trận giá trị.

3.3 preproces3(x_8, x_9, y):

- Chức năng: Tiền xử lý dữ liệu theo mô hình $Y = \theta_0 + \theta_9 X_9 + \theta_{10} X_{10}$^5.

- Input:

+ x_9: Mảng ma trận của đặc trưng Income composition of resources.

+ x_{10}: Mảng ma trận của đặc trưng Schooling.

+ y: Mảng ma trận giá trị.

- Output:

+ X: Mảng ma trận đặc trưng sau khi biến đổi theo mô hình.
+ y: Mảng ma trận giá trị.

3.4 ChooseBest_c(X, y)

- Hàm thực hiện tìm ra đặc trưng tốt nhất trong ba đặc trưng tự thiết lập.

- Input

+ x: Mảng ma trận các đặc trưng.
+ y: Mảng ma trận giá trị.

- Output:

+ Mảng RMSE ứng với các đặc trưng.

+ Vị trí của đặc trưng tốt nhất.

3.5 Cross_Validation1(A, b)

- Triển khai phương pháp K-ford Cross Validation

- Input

+ x: Mảng ma trận các đặc trưng.
+ y: Mảng ma trận giá trị.

- Output:

+ Trung bình cộng năm lần của mỗi đặc trưng.

- Ý tưởng:

+ Sử dụng thuật toán Kfold để xáo trộn dữ liệu và chia dữ liệu X_train thành các khung frame với số lượng dòng dữ liệu là X-train.shape[0]/5

+ Lặp qua từng khung frame để huấn luyện theo mô hình đã tiền xử lý ở hàm preproces1 và dự đoán thông qua class OLSLinearRegression

+ Sử dụng hàm RMSE để tính độ lệch chuẩn của phần dư

3.6 Cross_Validation2(A, b)

- Triển khai phương pháp K-ford Cross Validation

- Input

+ x: Mảng ma trận các đặc trưng.
+ y: Mảng ma trận giá trị.

- Output:

+ Trung bình cộng năm lần của mỗi đặc trưng.

- Ý tưởng:

+ Sử dụng thuật toán Kfold để xáo trộn dữ liệu và chia dữ liệu X_train thành các khung frame với số lượng dòng dữ liệu là X-train.shape[0]/5.

+ Lặp qua từng khung frame để huấn luyện theo mô hình đã tiền xử lý ở hàm preproces2 và dự đoán thông qua class OLSLinearRegression.

+ Sử dụng hàm RMSE để tính độ lệch chuẩn của phần dư.

3.7 Cross_Validation3(A, b)

- Triển khai phương pháp K-ford Cross Validation.

- Input

+ x: Mảng ma trận các đặc trưng.
+ y: Mảng ma trận giá trị.

- Output:

+ Trung bình cộng năm lần của mỗi đặc trưng.

- Ý tưởng:

+ Sử dụng thuật toán Kfold để xáo trộn dữ liệu và chia dữ liệu X_train thành các khung frame với số lượng dòng dữ liệu là X-train.shape[0]/5+ Lặp qua từng khung frame để huấn luyện theo mô hình đã tiền xử lý ở hàm preproces3 và dự đoán thông qua class OLSLinearRegression.

+ Sử dụng hàm RMSE để tính độ lệch chuẩn của phần dư.

IV. Kết quả

1. Yêu cầu 1a: Sử dụng toàn bộ 10 đặc trưng đề bài cung cấp

- Công thức hồi quy:

$$\text{Life expectancy} = theta[1]x_1 + theta[2]x_2 + theta[3]x_3 + \ldots + theta[10]x_{10}$$

- Kết quả RMSE trên tập test = 7.064046430584356

2. Yêu cầu 1b: Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất

STT	Mô hình với 1 đặc trưng	RMSE
1	Adult Mortality	46.767300458788725
2	BMI	27.963792700938306
3	Polio	17.912635600733697
4	Diphtheria	16.019287789095134
5	HIV/AIDS	69.0813273795947
6	GDP	60.45039343774032
7	Thinness age 10-19	51.89981477420683

8	Thinness age 5-9	51.7750585703281
9	Income composition of resources	13.29979135150532
10	Schooling	11.820071328899042

- Đặc trưng tốt nhất là: Schooling

- Công thức hồi quy:

$$\text{Life expectancy} = theta[10] * x_{10}$$

- Kết quả RMSE trên tập test = 10.26095039165537

- Giả thuyết cho mô hình đạt kết quả tốt nhất:

+ Theo:https://www.britishcouncil.vn/hoc-tieng-anh/tieng-anh-nguoi-lon/kinh-nghiem/mot-nen-giao-duc-tot-co-the-keo-dai-tuoi-tho: một nền giáo dục tốt có thể giúp kéo dài tuổi thọ của con người.

+ Việc học có thể giúp con người trang bị được các kiến thức về cuộc sống, có thể tránh được rất nhiều rủi ro trong cuộc sống.

+ Việc học giúp nâng cao kiến thức của bản thân, cải thiện trình độ học vấn của bản thân, từ đó giúp cải thiện đợi sống của con người.

3. Yêu cầu 1c: Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất

STT	**Mô hình**	**RMSE**
1	Sử dụng một đặt trưng Schooling	6.197766
2	Sử dụng hai đặt trưng có biến đổi Schooling và Income composition of resources	5.004816
3	Sử dụng hai đặt trưng chỉ biến đổi Schooling(Schooling và Income composition of resources)	5.555097

- Mô hình cho kết quả tốt nhất là Mô hình 2

- Công thức hồi quy:

$$\text{Life expectancy} = theta[0] + theta[9] * x_9^3 + theta[10] * x_{10}^5$$

- Kết quả RMSE trên tập test = 4.926736017593664

- Giả thuyết cho mô hình đạt kết quả tốt nhất:

> + Học vấn với nguồn thu nhập là hai yếu tố quan trọng nhất giúp cải thiện đời sống vật chất, tinh thần của con người. Khi đời sống vật chất, tinh thần, kiến thức đầy đủ thì con người sẽ có nhiều điều kiện chăm sóc sức khỏe, từ đó giúp nâng cao tuổi thọ của con người.

VI. Lý do thiết kế các mô hình đặc trưng:

- Dựa vào bảng RMSE của 10 đặc trưng, ta có thể thấy được trình độ học vấn và nguồn thu nhập có RMSE nhỏ nhất nên việc xây dựng 3 mô hình dựa trên hai đặc trưng này sẽ có độ chính xác rất cao đến cơ sở xác định ảnh hưởng tuổi thọ của con người.

Mô hình 1: $Y=\theta_0+\theta_{10}\ln(X_{10}\text{^}2)$.

Mô hình 2: $Y= \theta_0+\theta_9 X_9\text{^}3+\theta_{10}X_{10}\text{^}5$.

Mô hình 3: $Y= \theta_0+\theta_9 X_9 + \theta_{10}X_{10}\text{^}5$.

V. Tài liệu tham khảo

1. Class OLSLinear Regression: Được tham khảo từ phần Lap4 của cô Phan Thị Phương Uyên.

2. Kfold được tham khảo từ:

https://github.com/phungvhbui/AppMath-Project-LinearRegression/blob/master/18127185.ipynb

https://github.com/kieuconghau/linear-regression/blob/master/Sources/18127259.ipynb