TOÁN ỨNG DỤNG VÀ THỐNG KÊ CHO CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN MÔN HỌC #3

REPORT

Tài liệu này mô tả nội dung đồ án môn học cho môn học Toán ứng dụng và thống kê cho Công nghệ thông tin

Họ tên sinh viên: Nguyễn Văn Đạt

MSSV: 20127132

Khoa Công nghệ Thông tin
Đại học Khoa học Tự nhiên TP HCM
Tháng Thg7-22

MỤC LỤC

1

Tổng quan

Thông tin sinh viên

MSSV	Họ tên	Email
20127132	Nguyễn Văn Đạt	20127132@student.hcmus.edu.vn

Thông tin đồ án

Tên đồ án: Linear Regression	
Công cụ hướng dẫn	Visual Studio Code Jupyter Notebook

2

Nội dung đồ án

2.1

Giới thiệu

-Các thư viện đã sử dụng:

+pandas: Đọc dữ liệu file csv cho sẵn;

+numpy: Xử lý các dữ liệu được cho trước từ file csv;

+shuffle: Xáo trộn dữ liệu có đánh dấu sao cho không bị khác dòng giữa hai đặc trưng.

-Các hàm sử dụng lại từ Lab 4:

+Lớp OLSLinearRegression bao gồm các hàm:

• def fit(self, X, y);

• def get_params(self);

• def predict(self, X).

+def plot_regression(lr, x, y): Đã được sửa lại các giá trị đầu vào plt.scatter, x_para và X_para để phù hợp với đầu vào của X_train và y_train.

-Các hàm đã sử dụng:

+ def RMSE(y_test, y_test_pred): Tính RMSE (1)

2.2
Nội dung đồ án

Xây dựng mô hình dự đoán tuổi thọ trung bình sử dụng hồi quy tuyến tính

2.2.1. Câu 1a:

2.2.1.1. Yêu cầu:

- 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 y theo 10 đặc trưng trong X)

+ Báo cáo 1 kết quả trên tập kiểm tra (test.csv) cho mô hình vừa huấn luyện được

2.2.1.2. Hàm sử dụng

-def main1a(X_train, y_train, X_test) là hàm chạy chính cho câu 1a:

+Đầu vào:

• X_train: 10 cột đầu của tập train

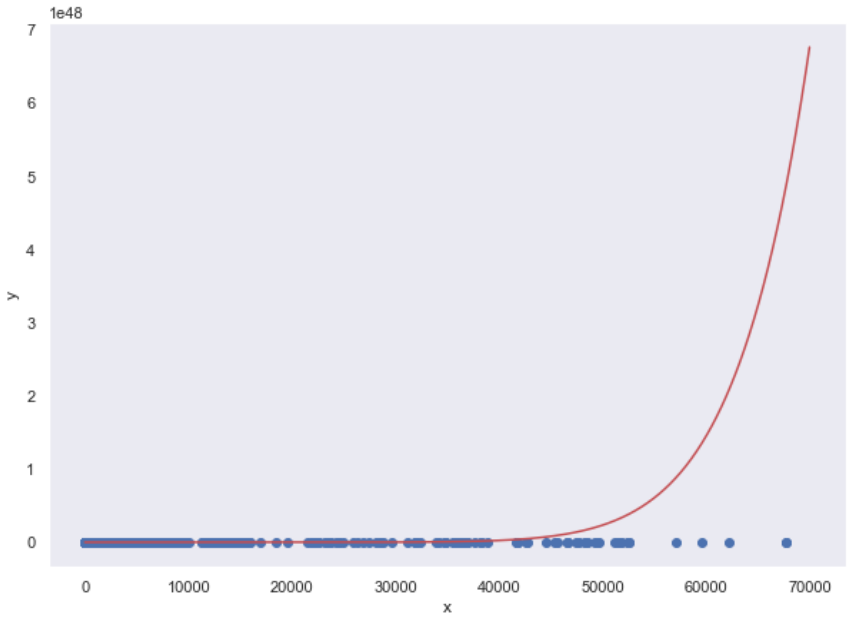
• y_train: Cột Life expectancy của tập train

• X_test: 10 cột đầu của tập test

+Đầu ra: y_test_pred: Kết quả kiểm tra trên tập test

2.2.1.3. Báo cáo và nhận xét

-Kết quả sau huấn luyện:



RMSE = 7.06404

Nhận xét:

Kết quả có giá trị tốt dựa trên biểu đồ đã được trực quan. Nó phản ánh rõ để dự đoán được tuổi thọ trung bình.

2.2.2. Câu 1b:

2.2.2.1. Yêu cầu:

- 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)

STT	Mô hình với 1 đặc trưng	RMSE
1	Adult Mortality	
2	BMI	
...	...	
10	Schooling	

+Thể hiện công thức cho mô hình hồi quy theo đặc trưng tốt nhất (tính y 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.

2.2.2.2. Hàm sử dụng

-def main1b(X_train, y_train) là hàm chạy chính cho câu 1a:

+Đầu vào:

• X_train: 10 cột đầu của tập train

• y_train: Cột Life expectancy của tập train

+Đầu ra: rmse_mean: Mảng RMSE của các cột trong X_train sau khi lấy trung bình

2.2.2.3. Giải thích logic

Bài toán được thực hiện dựa trên phương pháp 5-fold Cross Validation (2) chia mô hình thành 5 phần bằng nhau và tính RMSE trên từng phần đó, sau cùng là lấy trung bình.

Ta lần lượt chạy phần cột X_train cùng với cột y_train đã được chia thành 5 phần và phải chạy hết 5 phần đó. Ta thực hiện chọn 1 phần để test và 4 phần còn lại sẽ train.

X_train	y_train
1	1
2	2
3	3
4	4
5	5

X_train	y_train

1	1
2	2
3	3
4	4
5	5

X_train	y_train
1	1
2	2
3	3
4	4
5	5

X_train	y_train
1	1
2	2
3	3
4	4
5	5

X_train	y_train
1	1
2	2
3	3
4	4
5	5

Test	Train

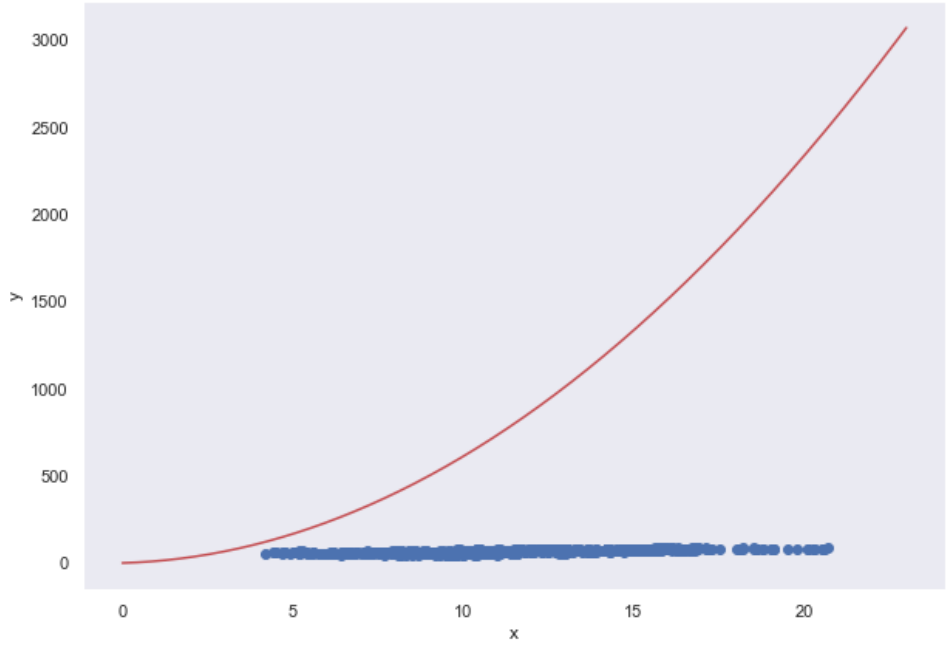
Từ đó suy ra được 5 RMSE cho mỗi cặp X_train, y_train.

2.2.2.4. Báo cáo và nhận xét

-Kết quả sau huấn luyện:

STT	Mô hình với 1 đặc trưng	RMSE	Nhận xét
1	Adult Mortality	41.9200	Độ lệch trung bình quá lớn nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
2	BMI	33.2783	Độ lệch trung bình quá lớn nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
3	Polio	21.1875	Độ lệch trung bình tốt nhưng chưa phải thấp nhất nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
4	Diphtheria	19.5753	Độ lệch trung bình tốt nhưng chưa phải thấp nhất nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
5	HIV/AIDS	65.6792	Độ lệch trung bình quá lớn nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
6	GDP	62.4901	Độ lệch trung bình quá lớn nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
7	Thinness age 10-19	48.9019	Độ lệch trung bình quá lớn nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
8	Thinness age 5-9	48.8179	Độ lệch trung bình quá lớn nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
9	Income composition of resources	20.9910	Độ lệch trung bình tốt nhưng chưa phải thấp nhất nên không thể chọn làm đặc trưng tốt nhất dự đoán tuổi thọ trung bình
10	Schooling	19.1516	Đặc trưng tốt nhất để chọn dự đoán tuổi thọ trung bình

-Kết quả trên tập kiểm tra:



RMSE = 10.26095

Nhận xét:

Kết quả có giá trị không tốt dựa trên biểu đồ đã được trực quan. Nó vẫn chưa phản ánh rõ để dự đoán được tuổi thọ trung bình.

2.2.3. Câu 1c:

2.2.3.1. Yêu cầu:

- 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...)

• ...

+ Yêu cầu sử dụng phương pháp 5-fold Cross Validation để tìm ra mô hình tốt nhất

+ Báo cáo m kết quả tương ứng cho m mô hình từ 5-fold Cross Validation (lấy trung bình)

STT	Mô hình	RMSE
1	Sử dụng 2 đặc trưng (a, b)	
…	…	
m	…	

+ Thể hiện công thức cho mô hình hồi quy tốt nhất mà sinh viên tìm được

+ Báo cáo 1 kết quả trên tập kiểm tra (test.csv) cho mô hình tốt nhất tìm được.

2.2.3.2. Mô hình chọn

-Mô hình 1: a * "Tên cột" + b * "Tên cột"

-Mô hình 2: a * "Tên cột"+ b * "Tên cột"+ c * "Tên cột"

-Mô hình 3: ("Tên cột" * "Tên cột") * a + b * "Tên cột"

2.2.3.3. Hàm sử dụng

-def model1(X_train, i, j): Lưu mô hình 1

+Đầu vào:

- X_train: 10 cột đầu của tập train
- i: Cột thứ i của tập train
- j: Cột thứ j của tập train

+Đầu ra: Numpyarrray có 2 cột i, j và số dòng trùng với X_train

-def model2(X_train, i, j, k): Lưu mô hình 2

+Đầu vào:

- X_train: 10 cột đầu của tập train
- i: Cột thứ i của tập train

• j: Cột thứ j của tập train

• k: Cột thứ k của tập train

+Đầu ra: Numpyarrray có 3 cột i, j, k và số dòng trùng với X_train

-def model3(X_train, i, j, k): Lưu mô hình 3

+Đầu vào:

• X_train: 10 cột đầu của tập train

• i: Cột thứ i của tập train

• j: Cột thứ j của tập train

• k: Cột thứ k của tập train

+Đầu ra: Numpyarrray có 3 cột i, j, k và số dòng trùng với X_train

-def main1c_1(X_train, y_train): Thực hiện mô hình 1

+Đầu vào:

• X_train: 10 cột đầu của tập train

• y_train: Cột Life expectancy của tập train

+Đầu ra:

• rmse_mean: Mảng RMSE của các cột trong X_train sau khi lấy trung bình

• rmse_mean_pos: Mảng vị trí đánh dấu từng cột sao cho tương ứng với mảng

-def main1c_2(X_train, y_train) : Thực hiện mô hình 2

+Đầu vào:

• X_train: 10 cột đầu của tập train

• y_train: Cột Life expectancy của tập train

+Đầu ra:

• rmse_mean: Mảng RMSE của các cột trong X_train sau khi lấy trung bình

• rmse_mean_pos: Mảng vị trí đánh dấu từng cột sao cho tương ứng với mảng rmse_mean

-def main1c_3(X_train, y_train) : Thực hiện mô hình 3

+Đầu vào:

• X_train: 10 cột đầu của tập train

• y_train: Cột Life expectancy của tập train

+Đầu ra:

• rmse_mean: Mảng RMSE của các cột trong X_train sau khi lấy trung bình

• rmse_mean_pos: Mảng vị trí đánh dấu từng cột sao cho tương ứng với mảng rmse_mean

2.2.3.4. Giả thuyết mô hình

Mô hình 1: Mô hình được thực hiện trên 2 đặc trưng được chọn từ X_train, thực hiện huấn luyện mô hình trên toàn tập X_train và chọn ra mô hình có RMSE nhỏ nhất. Mô hình được tạo nên với mong muốn tạo ra mô hình kết hợp 2 đặc trưng đưa ra kết quả tốt nhất để xác nhận tuổi thọ trung bình.

Mô hình 2: Mô hình được thực hiện trên 3 đặc trưng được chọn từ X_train, thực hiện huấn luyện mô hình trên toàn tập X_train và chọn ra mô hình có RMSE nhỏ nhất. Mô hình được tạo nên với mong muốn tạo ra mô hình kết hợp 3 đặc trưng đưa ra kết quả tốt nhất để xác nhận tuổi thọ trung bình.

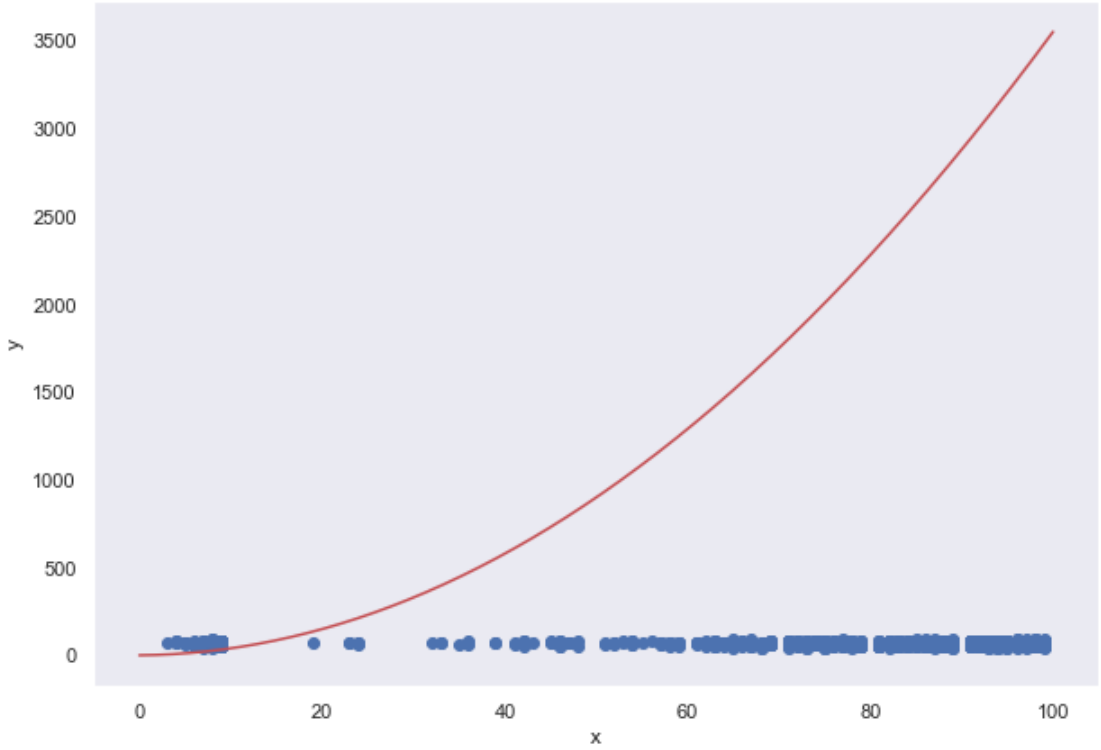
Mô hình 3: Mô hình được thực hiện trên 3 đặc trưng được chọn từ X_train, thực hiện huấn luyện mô hình trên toàn tập X_train bằng cách nhân 2 đặc trưng thành 1 cột rồi kết hợp với đặc trưng còn lại và chọn ra mô hình có RMSE nhỏ nhất. Mô hình được tạo nên với mong muốn tạo ra mô hình kết hợp 2 đặc trưng đưa ra kết quả tốt nhất để xác nhận tuổi thọ trung bình.

2.2.2.5. Báo cáo và nhận xét

-Kết quả mô hình sau khi huấn luyện:

STT	Mô hình	RMSE	Nhận xét
1	Sử dụng 2 đặc trưng (Polio, Diphtheria)	16.0253	Mô hình đưa ra có RMSE khá tốt và tốt nhất trong 3 mô hình nên là mô hình tốt nhất cho việc dự đoán tuổi thọ trung bình
2	Sử dụng 3 đặc trưng (BMI, Polio, Diphtheria)	24.8921	Mô hình đưa ra có RMSE khá tốt nhưng không đủ để trở thành mô hình tốt nhất cho việc dự đoán tuổi thọ trung bình
3	…Sử dụng 3 đặc trưng (Diphtheria, Income composition of resources , Polio)	23.9041	Mô hình đưa ra có RMSE khá tốt nhưng không đủ để trở thành mô hình tốt nhất cho việc dự đoán tuổi thọ trung bình

-Kết quả tập kiểm tra trên mô hình tốt nhất:



RMSE = 16.94119

Life expectancy = 0.4398 * Polio + 0.3505 * Diphtheria

Nhận xét:

Tuy là mô hình tốt nhất để dự đoán tuổi thọ trung bình nhưng qua trực quan các giá trị trên tập test thì các giá trị không tập trung quá gần Regression Line nên không đạt kết quả như mong muốn.

3

Mức độ hoàn thành

Câu	Yêu cầu	Đã/chưa hoàn thành	Mức độ hoàn thành
1a	Huấn luyện 1 lần duy nhất cho 10 đặc trưng trên toàn bộ tập huấn luyện (train.csv)	Đã hoàn thành	100%
	Thể hiện công thức cho mô hình hồi quy (tính y theo 10 đặc trưng trong X)	Đã hoàn thành	100%
	Báo cáo 1 kết quả trên tập kiểm tra (test.csv) cho mô hình vừa huấn luyện được	Đã hoàn thành	100%
1b	Thử nghiệm trên toàn bộ (10) đặc trưng đề bài cung cấp	Đã hoàn thành	100%
	Yêu cầu sử dụng phương pháp 5-fold Cross Validation để tìm ra đặc trưng tốt nhất	Đã hoàn thành	100%
	Báo cáo 10 kết quả tương ứng cho 10 mô hình từ 5-fold Cross Validation (lấy trung bình)	Đã hoàn thành	100%
	Thể hiện công thức cho mô hình hồi quy theo đặc trưng tốt nhất (tính y theo đặc trưng tốt nhất tìm được)	Đã hoàn thành	100%
	Báo cáo 1 kết quả trên tập kiểm tra (test.csv) cho mô hình tốt nhất tìm được	Đã hoàn thành	100%

1c	Xây dựng m mô hình khác nhau (tối thiểu 3), đồng thời khác mô hình ở 1a và 1b	Đã hoàn thành	100%
	Yêu cầu sử dụng phương pháp 5-fold Cross Validation để tìm ra mô hình tốt nhất	Đã hoàn thành	100%
	Báo cáo m kết quả tương ứng cho m mô hình từ 5-fold Cross Validation (lấy trung bình)	Đã hoàn thành	100%
	Thể hiện công thức cho mô hình hồi quy tốt nhất mà sinh viên tìm được	Đã hoàn thành	100%
	Báo cáo 1 kết quả trên tập kiểm tra (test.csv) cho mô hình tốt nhất tìm được	Đã hoàn thành	100%

4

Tài liệu tham khảo

- Tài liệu từ web:

https://numpy.org/

(1): https://www.delftstack.com/howto/python/rmse-python/

(2): https://miai.vn/2021/01/18/k-fold-cross-validation-tuyet-chieu-train-khi-it-du-lieu/