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 học: Toán ứng dụng và thống kê

GIÁO VIÊN HƯỚNG DẪN

Vũ Quốc Hoàng
Nguyễn Văn Quang Huy
Lê Thanh Tùng
Phan Thị Phương Uyên

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 học: Toán ứng dụng và thống kê

SINH VIÊN THỰC HIỆN

20127662 - Nguyễn Đình Văn

MỤC LỤC

Mục lục

THÔNG TIN THÀNH VIÊN

Mã số sinh viên	Họ và tên	Chú thích
20127662	Nguyễn Đình Văn	20127662@student.hcmus.edu.vn

CÁC CHỨC NĂNG ĐÃ HOÀN THÀNH

STT	Yêu cầu		Mức độ hoàn thành	Ghi chú
1	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`)	100%	
		Thể hiện công thức cho mô hình hồi quy (tính y theo 10 đặc trưng trong X)	100%	
		Báo cáo 1 kết quả trên tập kiểm tra (`test.csv`) cho mô hình vừa huấn luyện được	100%	
2	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	100%	
		Yêu cầu sử dụng phương pháp 5-fold Cross Validation để tìm ra đặc trưng tốt nhất	100%	
		Báo cáo 10 kết quả tương ứng cho 10 mô hình từ 5-fold Cross Validation (lấy trung bình)	100%	
		Thể hiện công thức cho mô hình hồi quy theo đặc trưng tốt nhất (tính y theo đặc trưng tốt nhất tìm được)	100%	
		Báo cáo 1 kết quả trên tập kiểm tra (`test.csv`) cho mô hình tốt nhất tìm được	100%	
	Yêu cầu 1c: Sinh viên tự xây dựng mô hình, tìm mô hình cho	Xây dựng m mô hình khác nhau (tối thiểu 3), đồng thời khác mô hình ở 1a và 1b	100%	

3	kết quả tốt nhất	Yêu cầu sử dụng phương pháp 5-fold Cross Validation để tìm ra mô hình tốt nhất	100%	
		Báo cáo `m` kết quả tương ứng cho `m` mô hình từ 5-fold Cross Validation (lấy trung bình)	100%	
		Thể hiện công thức cho mô hình hồi quy tốt nhất mà sinh viên tìm được	100%	
		Báo cáo1 kết quả trên tập kiểm tra (`test.csv`) cho mô hình tốt nhất tìm được	100%	

CÁC HÀM TRONG CHƯƠNG TRÌNH

1. **Các thư viện sử dụng:**
   - Thư viện *Pandas*: Đọc dữ liệu lấy các đặc trưng để huấn luyện.
   - Thư viện *Sklearn*: Sử dụng hàm Shuffle của thư viện để xáo trộn dữ liệu.
   - Thư viện *Numpy:* Sử dụng để thực hiện các tính toán với ma trận.
2. **Thư viện OLSLinearRegression:**
   - Khai báo: OLSLinearRegression()*[1]*
   - Ý tưởng: Sử dụng thư viện OLSLinearRegression cô đã cho ở Lab4.
   - Các hàm con: fit(), get_pamas(), predict()

3. **Hàm tính RMSE:**
   - Khai báo: RMSE()
   - Ý tưởng: Dựa vào công thức sai số bình phương trung bình
   $MSE = \frac{1}{n}\sum_{n}^{i=1}(yi - yi\hat{})^2$ và $RMSE = \sqrt{MSE}$ *[1]*
   - Input: Giá trị mục tiêu dự đoán hồi quy tuyến tính và giá trị mục tiêu
   - Output: RMSE

4. **Hàm tìm kiếm đặc trưng tốt nhất:**
   - Khai báo: find_best_feature()
   - Ý tưởng: Xáo trộn mô hình, chạy qua lần lượt từng đặc trưng, ở mỗi đặc trưng áp dụng phương pháp 5-fold Cross Validation chia mô hình thành 5 phần nhỏ huấn luyện và tính giá trị RMSE ở mỗi phần, sau cùng lấy giá trị RMSE nhỏ nhất từ 5 phần trên. Sau khi áp dụng lần lượt với 10 đặc trưng ta thu được một list chưa các giá trị RMSE tương ứng với từng mô hình. Và mô hình tốt nhất là mô hình có giá trị RMSE nhỏ nhất.
   - Input: Dataframe.
   - Output: List RMSE với mỗi đặc trưng.

5. **Hàm tìm kiếm mô hình tốt nhất:**
   - Khai báo: find_best_model()
   - Ý tưởng: Em lần lượt kiểm ra với 3 mô hình sau để tìm ra mô hình tốt nhất.
     - $y = Adult\ Mortality * x1 + Schooling * x2$
     - $y = Adult\ Mortality * x1 + BMI * x2 + Schooling * x3$
     - $y = Adult\ Mortality * x1 + Schooling^2 * x2$
   - Xáo trộn mô hình, với mỗi mô hình tìm ra các X_train và Y_train tương ứng. Sau đó áp dụng phương pháp 5-fold Cross Validation chia mô hình thành 5 phần nhỏ huấn luyện và tính giá trị RMSE ở mỗi phần, sau cùng lấy giá trị RMSE nhỏ nhất từ 5 phần trên. Sau khi áp dụng lần lượt với 3 mô hình ta thu được một list chưa các giá trị RMSE tương ứng với từng mô hình. Và mô hình tốt nhất là mô hình có giá trị RMSE nhỏ nhất.
   - Input: Dataframe
   - Output: List RMSE ứng với mỗi mô hình

THỰC THI VÀ KẾT QUẢ

❖ Với yêu cầu 1a

- RMSE: 7.06404643058411
- Ta có công thức hồi quy:

Công thức hồi quy

$$\text{Life expectancy} = AdultMortality * 0.015101 + BMI * 0.090220 + Polio * 0.042922 + Diphtheria * 0.139289 - HIV/AIDS * 0.567333$$
$$-GDP * 0.000101 + Thinnessage10 - 19 * 0.740713 + Thinnessage5 - 9 * 0.190936 + Incomecompositionofresources * 24.505974$$
$$+Schooling * 2.393517$$

❖ Với yêu cầu 1b

- Ta có bảng sau khi chạy xong hàm find_best_feature()
  Mô hình tốt nhất là mô hình có RMSE bé nhất

STT	Mô hình với 1 đặc trưng	RMSE
1	Adult Mortality	46.261459
2	BMI	27.965586
3	Polio	18.061084
4	Diphtheria	16.023751
5	HIV/AIDS	67.170939
6	GDP	60.212443
7	Thinness age 10-19	51.846121
8	Thinness age 5-9	51.729883
9	Income composition of resources	13.348367
10	Schooling	11.818677

- Đặc trưng tốt nhất là: Schooling
- RMSE: 10.260950391655376
- Công thức hồi quy:

Công thức hồi quy

$$\text{Life expectancy} = Schooling * 5.5573994$$

❖ Với yêu cầu 1c

- Em chọn các mô hình:
  - $y = Adult\ Mortality * x1 + Schooling * x2$
  - $y = Adult\ Mortality * x1 + BMI * x2 + Schooling * x3$
  - $y = Adult\ Mortality * x1 + Schooling^2 * x2$

- Ta có bảng sau khi chạy hàm find_best_model():
  Mô hình tốt nhất là mô hình cho kết quả RMSE bé nhất.

STT	Mô hình	RMSE
1	Sử dụng 2 đặc trưng (Adult Mortality, Schooling)	11.151860198801495
2	Sử dụng 3 đặc trưng (Adult Mortality, BMI, Schooling)	11.132938430508839
3	Sử dụng 2 đặc trưng (Adult Mortality, Schooling^2)	19.318036726199303

⇨ Chọn mô hình thứ 2

- Mô hình tốt nhất: $y = Adult\ Mortality * x1 + BMI * x2 + Schooling * x3$
- RMSE: 9.312027694213468
- Công thức hồi quy:

Công thức hồi quy

$$\text{Life expectancy} = AdultMortality * 0.028012 + BMI * 0.066049 + Schooling * 5.004187$$

Nhận xét chung: Với các mô hình được huấn luyện từ 3 yêu cầu a, b, c thì mô hình chứa 10 đặc trưng là mô hình tối ưu nhất, cho ra kết quả tốt nhất.

TÀI LIỆU THAM KHẢO

[1] Thư viện OLSLinearRegression ở Lab4
[2] https://pandas.pydata.org/docs/user_guide/10min.html