ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

MÔN HỌC : Toán ứng dụng và thống kê cho Công nghệ thông tin

ĐỒ ÁN THỰC HÀNH 3

BÁO CÁO

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.

Giảng viên hướng dẫn

Phan Thị Phương Uyên

Nguyễn Văn Quang Huy

Sinh viên thực hiện :

Phan Trí Nguyên - 20127578

Thành phố Hồ Chí Minh, tháng 7 năm 2022

MỤC LỤC

1

Tổng quan

1.1 Thông tin sinh viên

MSSV	Họ và tên	Email	Lớp
20127578	Phan Trí Nguyên	20127578@student.hcmus.edu.vn phantringuyen2002@gmail.com	20CLC05

1.2 Thông tin đồ án

Tên đồ án	Môi trường lập trình
Linear Regression	Phần mềm: Jupyther Notebook, Anaconda , Python 3.10, PyCharm Community Edition 2022.1.2.
	Ngôn ngữ: Python

Mô tả bài toán:

- Xây dựng mô hình dự đoán tuổi thọ trung bình sử dụng hồi quy tuyến tính (7 điểm)
  - Yêu cầu 1a: Sử dụng toàn bộ 10 đặc trưng đề bài cung cấp (2 điểm)
    - ✓ 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
  - Yêu cầu 1b: Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất (2 điểm)
    - ✓ 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 yy 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 (3 điểm)
  - ✓ Xây dựng m mô hình khác nhau (tối thiểu 3), đồng thời khác mô hình ở 1a và 1b
  - ✓ 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)
  - ✓ 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

Ý tưởng thực hiện và mô tả chi tiết các hàm

2.1 Ý tưởng thực hiện

2.1.1 Huấn luyện 1 lần duy nhất cho 10 đặc trưng trên toàn bộ tập huấn luyện (train.csv)

- Thực hiện tính toán theo công thức thống kê bình phương nhỏ nhất thông thường (*OLSLinearRegression*). [1]
- Tính toán một giá trị y theo biến cột theo cách dự đoán.
- Dựa theo công thức sai số toàn phương trung bình (MSE)

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

2.1.3 Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất

- Xây dựng hàm *CrossValidation* để chia nhỏ các dữ liệu nhằm tạo sự công bằng. [3]
- Tạo một ma trận 3 chiều dựa vào ma trận X_train với các dữ liệu từ file excel train.
- Ma trận 10 dòng 10 cột toàn 0 sẽ là kết quả tính toán của RMSE ban đầu.
- Ban đầu thực hiện việc xáo trộn ma trận.
- Nối từ ma trận rỗng vừa tạo với ma trận 2 chiều x_train_2d với thứ tự phần tử danh sách từ ma trận toàn 0.
- Thực hiện tính toán RMSE.
- Tìm ra 1 đặc trưng 10 đặc trưng dựa vào công thức sai số toàn phương trung bình, đây sẽ là đặc trưng tốt nhất.
- Dựa vào đặc trưng vừa tìm được, tính RMSE và in kết quả ra màn hình

2.1.4 Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất

- Tính toán để chọn ra Hệ số tương quan, những số liệu đặc trưng nào có Hệ số tương quan gần nhau thì có chỉ số thống kê đo lường mức độ mạnh yếu của mối quan hệ giữa các biến số.
- Xây dựng hàm tính *Hệ số tương quan - Correlation Coefficient*. [4]

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$ [5]

- 2 đặc trưng cho mô hình thứ 1: Thinness age 5-9, Income composition of resources.
- Thực hiện tính toán hàm nối 2 cột với nhau và tính rmse1 dựa trên công thức kiểm chứng chéo.
- 3 đặc trưng cho mô hình thứ 2: Polio, Diphtheria, GDP.
- Thực hiện tính toán hàm nối 3 cột với nhau và tính rmse2 dựa trên công thức kiểm chứng chéo.
- 5 đặc trưng cho mô hình thứ 3: BMI, Polio, Diphtheria, GDP, Schooling
- Thực hiện tính toán hàm nối 5 cột với nhau và tính rmse3 dựa trên công thức kiểm chứng chéo.
- So sánh 3 kết quả rmse vừa tính được trên từ 3 mô hình, mô hình nào có chỉ số rmse thấp nhất thì đó chính là mô hình tốt nhất.
- Sử dụng mô hình vừa chọn lọc để huấn luyện và tính ra RMSE, song in kết quả ra màn hình.

2.2 Mô tả các hàm

Tên hàm	Mô tả

`import pandas as pd`
`import numpy as np`
`from sklearn.metrics import mean_squared_error`
`import random`
	Import các thư viện cần thiết trong đó:
	• pandas dùng để việc dễ dàng và trực quan với dữ liệu có cấu trúc (dạng bảng, đa chiều, không đồng nhất) và dữ liệu chuỗi thời gian
	• numpy: dùng cho tính toán trên ma trận.
	• random: dùng để xáo trộn dữ liệu trong ma trận.
	• sklearn.metrics: dùng để sử dụng hàm viết sẵn để tính RMSE.

`def CrossValidation(x_test, y_test, fold):`	• Dùng để đảm bảo tránh được kiểm tra tổng hợp từ dữ liệu, chia nhỏ các dữ liệu nhằm tạo sự công bằng. • **Input:** x_test<np.array, y_test<np.array>, fold<int> fold = 5 để tính theo phương pháp 5-fold Cross Validation nhằm tìm ra đặc trưng tốt nhất. • **Output**: rmse<float>.
`def shuffle_matrix():`	• Xáo trộn các dữ liệu trong ma trận • **Input:** ma trận cần thực hiện<np.array>. • **Output**: ma trận sau khi được xáo trộn dữ liệu <np.array>.
`def append_data(myList):`	• Nối các ma trận gốc myList với ma trận chứa 10 đặc trưng huấn luyện • **Input:** myList <np.array> ma trận gốc ban đầu. • **Output:** x, y<np.array> ma trận sau khi được nối thành công.
`def fit(self, X, y):`	• $x = (A^T A)^{-1} A^T b$ • **Input:** X, Y<np.array> • **Output:** self<np.array>
`def predict(self, X):`	• Dự đoán mô hình tốt nhất.

	• **Input:** X< np.array > mảng 3 chiều ma trận ảnh gốc (X_test). • **Output:** y_test_preds < np.array > mảng 3 chiều ma trận với mô hình như yêu cầu.
`def RMSE_3(y, y_preds):`	• Hàm tính toán RMSE theo công thức [6] $RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$ • **Input:** y_test, y_test_preds< np.array > ma trận test theo cột và ma trận test predict từ hàm trên • **Output** <float> kết quả của RMSE. •
`def calculate_RMSE():`	• Tính giá trị Cross validation từ 10 đặc trưng file train • **Input:** x_train_feature, y_train_feature< np.array > địa chỉ ảnh thứ 1, địa chỉ ảnh thứ 2. • **Output:** float<> ảnh sau khi chồng 2 ảnh lên nhau.
`def calculate_RMSE_2feature():`	• Tính giá trị Cross validation từ 2 đặc trưng: Thinness age 5-9 và Income composition of resources • **Input:** x_train_feature_age_5_9, x_train_feature_Income < np.array > ma trận của từng đặc trưng 7 và 8 • **Output:** X_train_feature1 < np.array > và rmse1<float> lần lượt là ma trận được nối và giá trị crossvalidation.
`def calculate_RMSE_3feature():`	• Tính giá trị Cross validation từ 3 đặc trưng: Polio, Diphtheria, GDP. • **Input:** x_train_feature_Polio, x_train_feature_Diphtheria và x_train_feature_GDP< np.array > ma trận của từng đặc trưng 7 và 8

	• **Output:** X_train_feature2 < np.array > và rmse2<float> lần lượt là ma trận được nối và giá trị crossvalidation.
`def calculate_RMSE_5feature():`	• Tính giá trị Cross validation từ 5 đặc trưng: BMI, Polio, Diphtheria, GDP, Schooling. • **Input:** X_train_feature_BMI, X_train_feature_Polio, X_train_feature_Diphtheria, X_train_feature_GDP, X_train_feature_Schooling < np.array > ma trận của từng đặc trưng 7 và 8 • **Output:** X_train_feature3 < np.array > và rmse2<float> lần lượt là ma trận được nối và giá trị crossvalidation.
`def find_min_rmse():`	• Tìm ra rmse nhỏ nhất từ 3 mô hình trên • **Input:** rmse1, rmse2, rmse3 <float> giá trị crossvalidation từ 3 mô hình vừa xử lí trên. • **Output:** min< float > min của rmse

3

Demo kết quả chạy chương trình

3.1 Sử dụng toàn bộ 10 đặc trưng đề bài cung cấp

3.1.1 Huấn luyện 1 lần duy nhất cho 10 đặc trưng trên toàn bộ tập huấn luyện (train.csv)

```
[0.015101362735318279, 0.09021998065775627, 0.04292181752549435, 0.13928911689488216, -0.5673328270884068, -0.00010076511487489
53, 0.7407134377587112, 0.19093579767396474, 24.505973591149445, 2.393516607832779]
```

3.1.2 Thể hiện công thức cho mô hình hồi quy (tính y theo 10 đặc trưng trong X)

$$\text{Life expectancy} = (0.0151 * AdultMortality) + (0.0902 * BMI) + (0.0429 * Polio) + (0.1392 * Diphtheria) + (-0.5673 * HIV/AIDS) + (-0.0001 * GDP) + (0.7407 * Thinnessage10-19) + (0.1909 * Thinnessage5-9) + (24.5059 * Incomecompositionofresources) + (2.3935 * Schooling)$$

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

```
Dùng thư viện có sẵn với squared là False:
 RMSE =  7.064046430584705

Dùng thư viện có sẵn với squared là True:
 RMSE 2 =  7.064046430584705

Hàm xử lí tính toán:
 RMSE 3 =  7.064046430584705
```

3.2 Thay đổi độ tương phản cho ảnh

3.2.1 Báo cáo 10 kết quả tương ứng cho 10 mô hình từ 5-fold Cross Validation (lấy trung bình)

	Life expectancy	Mô hình với 1 đặc trưng	RMSE
0	65.0	Adult Mortality	46.159141
1	59.9	BMI	27.856910
2	77.8	Polio	18.025099
3	77.5	Diphtheria	15.751841
4	75.4	HIV/AIDS	66.899127
5	51.7	GDP	60.062601
6	76.2	Thinness age 10-19	51.681615
7	74.6	Thinness age 5-9	51.599730
8	82.7	Income composition of resources	13.304504
9	81.4	Schooling	11.775565

```
Index of the best feature:  9
```

3.2.2 Thể hiện công thức cho mô hình hồi quy theo đặc trưng tốt nhất (tính y theo đặc trưng tốt nhất tìm được)

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

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

```
RMSE =  10.26095039165538
```

3.3 Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất

3.3.1 Xây dựng m mô hình khác nhau (tối thiểu 3), đồng thời khác mô hình ở 1a và 1b

- 2 đặc trưng cho mô hình 1: Thinness age 5-9, Income composition of resources (đặc trưng thứ: 7 và 8)

- 3 đặc trưng cho mô hình: Polio, Diphtheria, GDP (đặc trưng thứ: 2, 3 và 5)

- 5 đặc trưng cho mô hình: BMI, Polio, Diphtheria, GDP, Schooling (đặc trưng thứ: 1, 2, 3, 5 và 9)

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

```
rmse1 = CrossValidation(x_train_feature1, y_train_feature, 5)
rmse2 = CrossValidation(x_train_feature2, y_train_feature, 5)
rmse3 = CrossValidation(X_train_feature3, y_train_feature, 5)
```

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

```
RMSE 1 =  11.408144308539155
RMSE 2 =  14.679466310634249
RMSE 3 =  9.623684761151651
```

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

$$\text{Life expectancy} = (BMI * 0.0351) + (Polio * 0.0786) + (Diphtheria * 0.1866) + (GDP * (-0.0002)) + (Schooling * 3.7819)$$

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

```
RMSE =  8.559353765137361
```

4

Nhận xét kết quả demo chạy chương trình

STT		Chức năng	Mức độ hoàn thành
	Sử dụng toàn bộ 10 đặc trưng đề bài cung cấp (2 điểm)		
1		Huấn luyện 1 lần duy nhất cho 10 đặc trưng trên toàn bộ tập huấn luyện (train.csv)	100%
2		Thể hiện công thức cho mô hình hồi quy (tính y theo 10 đặc trưng trong X)	100%
3		Báo cáo 1 kết quả trên tập kiểm tra (test.csv) cho mô hình vừa huấn luyện được	100%
	Xây dựng mô hình sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất (2 điểm)		
4		Thử nghiệm trên toàn bộ (10) đặc trưng đề bài cung cấp	100%
5		Yêu cầu sử dụng phương pháp 5-fold Cross Validation để tìm ra đặc trưng tốt nhất	100%
6		Báo cáo 10 kết quả tương ứng cho 10 mô hình từ 5-fold Cross Validation (lấy trung bình)	100%
7		Thể hiện công thức cho mô hình hồi quy theo đặc trưng tốt nhất (tính y theo đặc trưng tốt nhất tìm được)	100%
8		Báo cáo 1 kết quả trên tập kiểm tra (test.csv) cho mô hình tốt nhất tìm được	100%
	Sinh viên tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất (3điểm)		

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

4.1 Đánh giá kết quả

- Em thực hiện chạy tính độ tương quan cho những đặc trưng trên là thấy hợp lí nhất nên em chọn chạy hàm độ tương quan trên.
- Huấn luyện với các đặc trưng có hệ số tương quan gần nhau sẽ đưa ra số liệu RMSE tương đối nhỏ

5

Tài liệu tham khảo

[1] tài liệu này tham khảo tại https://en.wikipedia.org/wiki/Ordinary_least_squares dùng để để ước tính các tham số chưa biết trong mô hình hồi quy tuyến tính

[2] tài liệu này tham khảo tại https://vi.wikipedia.org/wiki/Sai_s%E1%BB%91_to%C3%A0n_ph%C6%B0%C6%A1ng_trung_b%C3%ACnh dùng để tham khảo công thức MSE trong thống kê học – sai số toàn phương trung bình.

[3] tài liệu này tham khảo tại https://vi.wikipedia.org/wiki/Ki%E1%BB%83m_ch%E1%BB%A9ng_ch%C3%A9o dùng để tham khảo hàm Cross validation trong thống kê - kiểm chứng chéo.

[4] tài liệu này tham khảo tại https://www.careerlink.vn/en/careertools/economic-knowledge/he-so-tuong-quan-correlation-coefficient-la-gi-va-ung-dung#:~:text=H%E1%BB%87%20s%E1%BB%91%20t%C6%B0%C6%A1ng%20quan%20l%C3%A0,trong%20ph%C3%A9p%20%C4%91o%20t%C6%B0%C6%A1ng%20quan dùng để đo thống kê về độ mạnh yếu của mối quan hệ giữa các chuyển động tương đối của hai biến.

[5] tài liệu này tham khảo tại https://phantichspss.com/he-so-tuong-quan-pearson-cach-thao-tac-phan-tich-tuong-quan-trong-spss.html dùng để tham khảo công thức tính độ tương quan.

[6] tài liệu này tham khảo tại https://towardsdatascience.com/what-does-rmse-really-mean-806b65f2e48e dùng để tham khảo công thức tính độ tương quan.

https://www.enjoyalgorithms.com/blog/life-expectancy-prediction-using-linear-regression để tham khảo về thuật toán Linear Regression