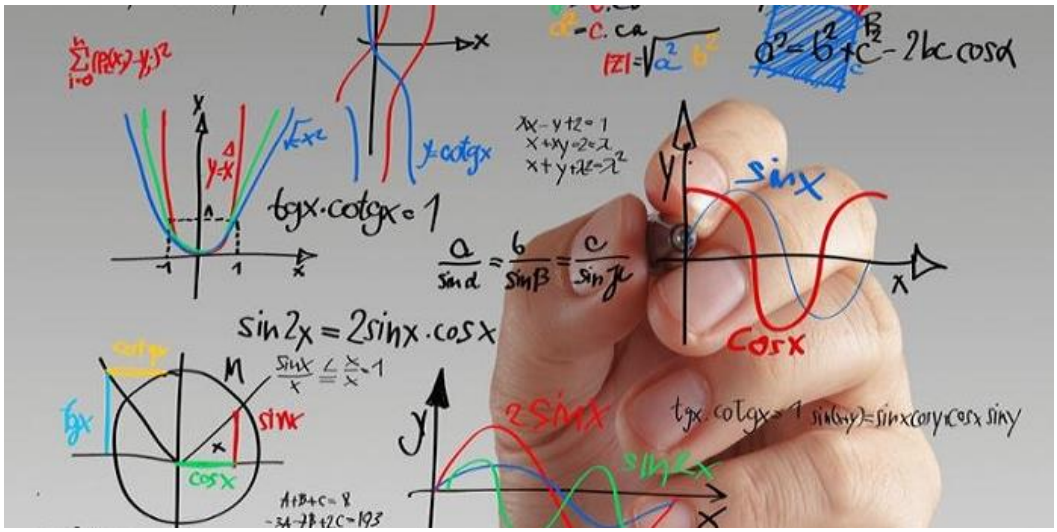
KHOA CÔNG NGHỆ THÔNG TIN

Project 3: Linear Regression

GV hướng dẫn: Thầy Vũ Quốc Hoàng
Thầy Nguyễn Văn Quang Huy
Thầy Lê Thanh Tùng
Cô Phan Thị Phương Uyên



Thông tin của sinh viên:

Huỳnh Tấn Vinh – 2012766

Lớp: 20CLC08

Trường Đại học Khoa Học Tự Nhiên 227 Nguyễn Văn Cừ, Quận 5, TP Hồ Chí Minh

Table of Contents

I. Thông tin tổng quát

Đồ án 3 - Linear Regression

Giới thiệu: dữ liệu Tuổi thọ trung bình (Life expectancy) được thu thập từ tổ chức WHO và trang web United Nations từ năm 2000 đến 2015 trên tất cả quốc gia.

Sau quá trình xử lý dữ liệu có:

- 1180 dòng dữ liệu
- 11 cột dữ liệu

Trong đồ án này, ta sẽ thực hiện Xây dựng mô hình dự đoán tuổi thọ trung bình sử dụng hồi quy tuyến tính, báo cáo về kết quả, đánh giá và nhận xét các mô hình đã xây dựng.

Môi trường thực hiện: Sử dụng Jupyter Notebook. Python 3.9.12

Đánh giá mức độ hoàn thành: 100%

Tổng số trang báo cáo: 8

II. Kết quả, đánh giá và nhận xét các mô hình đã xây dựng

1. Mô tả các hàm và ý tưởng đã sử dụng.

a) Các thư viện đã sử dụng.

- pandas: Trích xuất dữ liệu từ csv vào DataFrame
- numpy: xử lý dữ liệu trên ma trận, các toán hạng, …
- matplotlib.pyplot: in dữ liệu trực quan lên tọa độ
- random: random dữ liệu.

b) Các hàm phục vụ cho đồ án.

- Class OLSLinearRegression: Triển khai mô hình Linear Regression. Chức các làm: fit, get_params, predict.
- def **fit**(self, X, y): để thực hiện train.
- def **get_params**(self): để lấy hệ số.
- def **predict**(self, X): thực hiện một dự đoán cho mỗi trường hợp thử nghiệm, đối với bộ phân loại và bộ hồi quy, giá trị dự đoán sẽ nằm trong cùng không gian với giá trị được thấy trong tập huấn luyện.
- def **RMSE**(y, y_hat): thực hiện tính toán dựa trên công thức của phương pháp "Lỗi trung bình bình phương gốc" (RMSE). Nó là thước đo mức độ hiệu quả của mô hình. Nó thực hiện điều này bằng cách đo sự khác biệt giữa các giá trị dự đoán và giá trị thực tế. Nó thực hiện điều này bằng cách đo sự khác biệt giữa các giá trị dự đoán và giá trị thực tế. R-MSE càng nhỏ tức là sai số càng bé thì mức độ ước lượng cho thấy độ tin cậy của mô hình có thể đạt cao nhất.

Công thức RMSE:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

- def **CrossValidationHandle**(X_train_feature, y_train_feature, split_fold): thống kê ước lượng, đánh giá hiệu quả của các mô hình. Tham số quan trọng trong hàm (kỹ thuật) này là **split_fold** đại diện cho số nhóm mà dữ liệu sẽ được chia ra.

  Kỹ thuật này bao gồm các bước:

  B1. Xáo trộn dataset một cách ngẫu nhiên.

  B2. Chia dataset thành k nhóm.

  B3. Với mỗi nhóm

  - Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình
  - Các nhóm còn lại được sử dụng để huấn luyện mô hình
  - Huấn luyện mô hình
  - Đánh giá mô hình

  B4. Tổng hợp hiệu quả của mô hình dựa từ các số liệu đã đánh giá

- def plot_data(x, y): Biểu diễn dữ liệu trực quan trên đồ thị

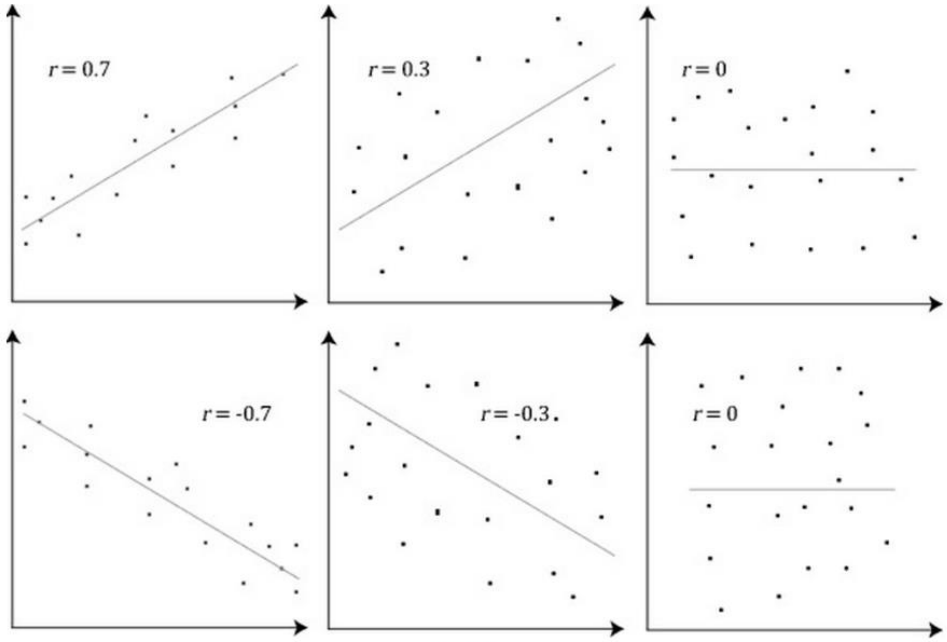
c) Ý tưởng thực hiện yêu cầu 1c:

Xây dựng mô hình dựa vào giá trị hệ số tương quan của từng mô hình trong 10 mô hình so với mô hình mục tiêu kiểm tra (**Life expectancy**).

Hệ số tương quan hay Correlation Coefficient là chỉ số đo thống kê giữa 2 biến số với nhau. Qua đó có thể thấy được độ mạnh yếu của mối quan hệ. Giá trị của hệ số nằm trong khoảng -1,0 đến 1,0. Hệ số tương quan Pearson ký hiệu R là loại phổ biến nhất. Mối tương quan tuyến tính càng mạnh nếu hệ số R gần về 1 hoặc -1. Tương quan tuyến tính yếu dần nếu R càng tiến về 0.

Nếu giá trị bằng 1,0 thì giữa 2 biến số có mối quan hệ dương tuyệt đối, ngược lại nếu bằng -1,0 là âm tuyệt đối. Nhìn vào đây có thể thấy được sự chuyển động ngược chiều nhau giữa các biến. Nếu biến này tăng dương thì biến kia sẽ giảm xuống. Trường hợp giá trị hệ số R bằng 0 ta có thể kết luận rằng không có quan hệ tuyến tính giữa 2 biến.

Hình ảnh ví dụ về độ tương quan:



Hình minh họa

Cách tính độ tương quan:

Sử dụng hàm **corrcoef** trong numpy. Cú pháp trong code: **np.corrcoef(list1, list2)[0, 1]**. Hàm trả về hệ số tương quan Pearson product-moment.

Hệ số tương quan Pearson được ước tính bằng công thức sau đây:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Kết quả tính độ tương quan:

```
# print(X_train_feature)
corr = 0
for i in range (10):
    corr = np.corrcoef(xtrain[:,i], ytrain)[0, 1]
    print(corr)
```

```
-0.685290066876047
0.6106869049325891
0.36570463313391016
0.40679787281341867
-0.5865777200247837
0.4867608561696108
-0.5036523346464049
-0.5002309457544405
0.779403572108283
0.7548637211053559
```

Từ kết quả tính độ tương quan, xây dựng được 3 mô hình:

- Mô hình 1: Sử dụng 2 đặc trưng "Income composition of resources", "Schooling".
  Giải thích: Giá trị tương quan của "Income composition of resources" và "Schooling". lần lượt là 0.779403572108283 và 0.7548637211053559. Hệ số tương quan cao nhất và có mối quan hệ tương quan dương.

Ta có mô hình hồi quy tuyến tính tổng quát:

$$\mathbf{y = w1 * } Incomecompositionofresources \mathbf{+ w2 * } Schooling$$

- Mô hình 2: Sử dụng 3 đặc trưng "BMI", "Diphtheria", "Income composition of resources" và "Schooling".
  Giải thích: Giá trị tương quan của "BMI", "Diphtheria", "Income composition of resources" và "Schooling" lần lượt là 0.6774872626372263, 0.40679787281341867, 0.779403572108283 và 0.7548637211053559. Hệ số tương quan cao nhất và có mối quan hệ tương quan dương.

Ta có mô hình hồi quy tuyến tính tổng quát:

$$\mathbf{y = w1 * BMI + w2 * Diphtheria + w3 * } Incomecompositionofresources \mathbf{ + w4 * } Schooling$$

- Mô hình 3: Sử dụng 4 đặc trưng **"**Adult Mortality", "HIV/AIDS", "Thinness age 10-19", "Thinness age 5-9".
  Giải thích: Hệ số tương quan của "Adult Mortality", "HIV/AIDS", "Thinness age 10-19", "Thinness age 5-9" lần lượt là -0.685290066876047, -0.5865777200247837, -0.5036523346464049 và -0.5002309457544405. Hệ số tương quan của mỗi đặc trưng này đều âm nên 4 đặc trưng có mối quan hệ tương quan âm.

Ta có mô hình hồi quy tuyến tính tổng quát:

y = w1 * Adult Mortality + w2 * HIV/AIDS + w3 * Thinness age 10-19 + w4 * Thinness age 5-9

2. Báo cáo và nhận xét kết quả từ toàn bộ các mô hình xây dựng được.

a) Báo cáo và nhận xét kết quả từ yêu cầu 1a:

Kết quả:

- Một kết quả trên tập kiểm tra (test.csv) cho mô hình: 7.064046430584466

```
Out[8]:  7.064046430584477
```

Nhận xét:

- Dựa vào kết quả trên cho ta biết mức độ tập trung của dữ liệu xung quanh dòng phù hợp nhất.
- Công thức quy hồi:

$$\begin{aligned}\text{Life expectancy} = (1.51013627e-02 * AdultMortality) + (9.02199807e-02 * BMI) + (4.29218175e-02 * Polio) \\ + (1.39289117e-01 * Diphtheria) + (-5.67332827e-01 * HIV/AIDS) + (-1.00765115e-04 * GDP) \\ + (7.40713438e-01 * Thinnessage10-19) + (1.90935798e-01 * Thinnessage5-9) \\ + (2.45059736e+01 * Incomecompositionofresources) + (2.39351661e+00 * Schooling)\end{aligned}$$

b) Báo cáo và nhận xét kết quả từ yêu cầu 1b:

Kết quả:

❖ Kết quả tương ứng cho **10 mô hình từ 5-fold Cross Validation**:

```
RMSE:  [46.14962884 27.94486472 17.98047167 16.02368229 67.08114439 60.19097752
 51.79660157 51.69414562 13.32327506 11.79050192]
```

```
Vị trí của đặc trưng tốt nhất:  9
```

Sử dụng phương pháp 5-fold Cross Validation để tìm ra vị trí của đặc trưng tốt nhất:

Hay đặc trưng tốt nhất là: **Schooling.**

```
Out[13]: 10.260950391655376
```

❖ Một kết quả cho tập kiểm tra:

Nhận xét:

- R-MSE càng nhỏ tức là sai số càng bé thì mức độ ước lượng cho thấy độ tin cậy của mô hình. Với 10 kết quả tương ứng cho 10 mô hình ta thấy được lỗi trung bình bình phương gốc RMSE thấp nhất là 11.79050192 (mô hình với đặc trưng Schooling), vì vậy mô hình với đặc trưng Schooling là đáng tin cậy nhất. Ta cũng tìm được đặc trưng Schooling là tốt nhất.
- Dựa trên kết quả trên tập kiểm tra (test.csv) với mô hình best_feature_model cho ta biết mức độ tập trung dữ liệu xung quanh dòng phù hợp nhất với RMSE bằng 10.260950391655376.

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

- Công thức hồi quy:

Giải thích hoặc nêu giả thuyết cho mô hình đạt mô hình tốt nhất (mô hình với đặc trưng Schooling):

Schooling là đặc trưng tốt nhất bởi vì trong thực tế cuộc sống, một nền giáo dục tốt có thể kéo dài tuổi thọ. Các nghiên cứu của UNESCO đã chỉ ra mối liên kết chặt chẽ giữa giáo dục và tuổi thọ. Theo báo cáo giám sát Toàn cầu (Global Monituring Report), mỗi năm có thêm một bà mẹ được đến trường sẽ giảm khả năng tử vong ở trẻ sơ sinh từ 5 đến 10%. Dịch vụ giáo dục và y tế đã góp phần tích cực vào việc kéo dài tuổi thọ ở nhiều quốc gia. [1]

c) Báo cáo và nhận xét kết quả từ yêu cầu 1c.

Có tổng cộng 3 mô hình được xây dựng dựa trên hệ số tương quan.

Kết quả:

- Mô hình 1:

```
RMSE 1:  11.226909591744127
```

- Mô hình 2:

```
RMSE 2:  9.664871124985922
```

- Mô hình 3:

```
RMSE 3:  41.54545157986963
```

Kết quả cho tập kiểm tra:

```
Out[20]: 8.280332718815753
```

Nhận xét:

- Ta thấy được mô hình 2 cho kết quả tốt nhất.
- Công thức hồi quy:

$$\text{Life expectancy} = (BMI * -6.40259278e-03) + (Diphtheria * 2.50004203e-01) + (Incomecompositionofresources * 2.78218706e+01) + (Schooling * 2.44185275e+00)$$

Giải thích hoặc nêu giả thuyết cho mô hình đạt mô hình tốt nhất:

Trên thực tế bệnh bạch hầu khá phổ biến ở trẻ em từ độ tuổi từ 1 đến 15. [2] Tỉ lệ tiêm chủng giải độc uốn ván và ho gà (DTP3) ở trẻ 1 tuổi (%) liên quan tới BMI Chỉ số khối lượng cơ thể (BMI) trung bình của toàn bộ dân số. Nếu không được tiêm chủng khả năng nhiễm bệnh cao dẫn đến tình trạng suy nhược, thiếu dinh dưỡng do uốn ván. Đối với nhiều người, BMI thừa cân lại là một yếu tố làm giảm nguy cơ tử vong khoảng 13 năm theo công bố gần đây trên tạp chí Obesity[3]. Schooling (học tập, giáo dục) ảnh hưởng tới chỉ số phát triển của con người. Đồng thời chỉ số phát triển con người (HDI) tính theo thành phần thu nhập tài nguyên cũng ảnh hưởng quan trọng tới tỉ lệ tử vong của con người. Như ở câu b, giáo dục và sự phát triển của con người có mối liên kết chặt chẽ với tuổi thọ. [1]

IV. Tài liệu tham khảo

1. Tài liệu tham khảo cho yêu cầu 1a

- https://solieu.vip/mse-va-rmse-la-gi-va-cach-tinh-tren-stata/
- https://viblo.asia/p/tim-hieu-cong-thuc-toan-cua-phuong-phap-hoi-quy-tuyen-tinh-qua-bai-toan-du-bao-xa-lu-understanding-the-linear-regression-E375z7mdKGW
- https://viblo.asia/p/linear-regression-hoi-quy-tuyen-tinh-trong-machine-learning-4P856akRlY3

2. Tài liệu tham khảo cho yêu cầu 1b

- https://trituenhantao.io/kien-thuc/gioi-thieu-ve-k-fold-cross-validation/
- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_numpy.html
- [1] http://xn--https-ix3b//www.britishcouncil.vn/hoc-tieng-anh/tieng-anh-nguoi-lon/kinh-nghiem/mot-nen-giao-duc-tot-co-the-keo-dai-tuoi-tho?fbclid=IwAR0g2Rd_qgCX_dFF1KT7KzGX2dsJV1GNChANlGMWepJF-h5VmRgQ7EGB-Bo

3. Tài liệu tham khảo cho yêu cầu 1c

- https://helpex.vn/question/tinh-toan-tuong-quan-va-y-nghia-cua-pearson-trong-python-5cbbb2d9ae03f60a1cce05d5
- https://numpy.org/doc/stable/reference/generated/numpy.corrcoef.html
- https://toptradingforex.com/he-so-tuong-quan-la-gi-y-nghia-cach-tinh-va-ung-dung/
- [2] https://www.vinmec.com/vi/vac-xin/kien-thuc-tiem-chung/ai-co-the-mac-benh-bach-hau/?link_type=related_posts
- [3] https://hellobacsi.com/an-uong-lanh-manh/beo-phi/chi-so-bmi-co-the-bat-mi-tuoi-tho-cua-ban/