TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

ĐẠI HỌC QUỐC GIA TP.HCM

BÁO CÁO ĐỒ ÁN

TOÁN ỨNG DỤNG VÀ THỐNG KÊ CHO CÔNG NGHỆ THÔNG TIN

Họ và tên sinh viên: Nguyễn Thuận Phát

MSSV: 21127665

Giảng viên lý thuyết:Vũ Quốc Hoàng

Giảng viên thực hành:

- Phan Thị Phương Uyên
- Nguyễn Văn Quang Huy

Mục lục

1. Mức độ hoàn thành

STT	Tên chức năng	Tiến độ
1	Yêu cầu 1a	Hoàn thành
2	Yêu cầu 1b	Hoàn thành
3	Yêu cầu 1c	Hoàn thành
4	Yêu cầu 1d	Hoàn thành

2. Lý Thuyết và thuật toán

2.1 OLS Linear Regression

-Ta cần tìm nghiệm của phương trình $Ax \approx b$

-Xét ma trận A có kích thước m x n (m > n) và vector (cột) b có kích thước m. Ta có chuẩn Euclidean của bình phương phần dư r của Ax – b như sau:

$$r = ||Ax - b||^2 \quad (1)$$

-Để giải được nghiệm x cho hệ phương trình, ta thực hiện tối thiểu hóa công thức (1) được nghiệm x của hệ phương trình được tính như sau:

$$x = (A^T A)^{-1} A^T b$$

-Note: $(A^T A)^{-1} A^T$ là ma trận giả nghịch đảo của A*

-Bài thực hành này sử dụng tên gọi khác cho đầu vào, đầu ra và tham số trong hồi quy tuyến tính như sau:

$A \rightarrow X$

$b \rightarrow y$

$x \rightarrow w$ (w: weight)

$Ax \approx b \rightarrow Xw \approx y$ hay $Xw = y$ (y được gọi là đường hồi quy (regression line))

2.2 k-fold Cross Validation

-Cross validation là một kỹ thuật lấy mẫu để đánh giá mô hình học máy trong trường hợp dữ liệu không được dồi dào cho lắm.

-Tham số quan trọng trong kỹ thuật này là k, đại diện cho số nhóm mà dữ liệu sẽ được chia ra. Vì lý do đó, nó được mang tên k-fold cross-validation. Khi giá trị của k được lựa chọn, người ta sử dụng trực tiếp giá trị đó trong tên của phương pháp đánh giá. Ví dụ với k = 10, phương pháp sẽ mang tên 10-fold Cross Validation.

-Kỹ thuật này thường bao gồm các bước như sau:

1. Xáo trộn dataset một cách ngẫu nhiên
2. Chia dataset thành k nhóm
3. Với mỗi nhóm:
    3.1. Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình
    3.2. Các nhóm còn lại được sử dụng để huấn luyện mô hình
    3.3. Huấn luyện mô hình
    3.4. Đánh giá và sau đó hủy mô hình
4. Tổng hợp hiệu quả của mô hình dựa từ các số liệu đánh giá

-Lưu ý: mỗi mẫu chỉ được gán cho duy nhất một nhóm và phải ở nguyên trong nhóm đó cho đến hết quá trình.

-Kết quả tổng hợp thường là trung bình của các lần đánh giá.

2.3 Phương pháp đo lường độ lỗi: MAE - Mean Absolute Error

-MAE được dùng để ước **lượng trung bình của sai số (độ lỗi) bình phương**, được tính bằng công thức:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Trong đó:

- n: số lượng mẫu quan sát

- y_i: giá trị mục tiêu của mẫu thứ i

- x_i: giá trị mục tiêu của mẫu thứ i được dự đoán từ mô hình hồi quy tuyến tính

3. Các thư viện sử dụng và hàm cần thiết

3.1 Numpy

-Hàm numpy.mean(): sử dụng để tính độ lỗi mae trong hàm mae(), sử dụng để tính trung bình độ lỗi để đánh giá trong hàm k_foldCrossValidation.

-Hàm numpy.abs(): sử dụng để tính độ lỗi mae trong hàm mae().

-Hàm numpy.linalg.inv(): sử dụng để trong phương thức fit của lớp OLSLinearRegression để tính các tham số.

-Hàm numpy.sum(): sử dụng trong phương thức predict của lớp OLSLinearRegression.

3.2 Pandas

-Hàm pandas.read_csv(): sử dụng để đọc dữ liệu từ file csv.

-Hàm pandas.concat(): sử dụng để kết hợp một số cột để tạo thành dataset cho các yêu cầu 1c, 1d; sử dụng để kết hợp các fold khác fold đang duyệt thành dataset dùng để train trong hàm k_foldCrossValidation().
-Hàm pandas.DataFrame(): dùng để chuyển list kết quả thành data frame để hiển thị, dùng trong câu 1d.

3.3 Math

-math.ceil(): sử dụng trong lúc tính kích thước của một fold data trong hàm k_foldCrossValidation().

4. Mô tả các hàm

4.1 lớp OLSLinearRegression

- Lấy từ file lab4.ipynb.

-Cách phương thức của lớp thực hiện các chức năng theo thuật toán hồi quy tuyến tính OLS Linear Regression.

-Các phương thức của lớp:

+fit(): có 2 tham số đầu vào là X và y (tương đương A và b trong phần **2.1**). Phương thức sẽ thực hiện tính toán tham số cho mô hình hồi quy tuyến tính rồi lưu mảng tham số vào thuộc tính w. Là phương thức để huấn luyện mô hình.

+get_params(): là getter cho thuộc tính w (mảng các tham số) của mô hình đã được train bằng phương thức fit() trước đó.

+predict(): Mô hình sau khi đã train bằng phương thức fit() có thể gọi phương thức này và truyền vào test dataset để cho ra kết quả dự đoán y theo test dataset. Kết quả này dùng để đánh giá độ lỗi của mô hình (sử dụng dàm mae()).

4.2 Hàm mae()

-Lấy từ file lab4.ipynb.

-Là hàm tính toán độ lỗi mae của mô hình, cách thức thực hiên như công thức ở mục **2.3.**

-Tham số truyền vào là y và y_hat, lần lượt là giá trị thực của test dataset và giá trị dự đoán nhận được từ phương thức predict() của mô hình hồi quy tuyến tính (lớp OLSLinearRegression).

4.3 Hàm k_foldCrossValidation()

-Là hàm thực hiện kỹ thuật k-fold Cross Validation cho một mô hình hồi quy tuyến tính OLS Linear Regression.

-Tham số: dataset là dữ liệu truyền vào để kiểm tra mô hình. k là số fold sẽ được chia, mặc định là 10.

-Hàm mặc đinh cột cuối cùng trong dataset truyền vào là cột giá trị kiểm tra y, các cột còn lại là X.

-Quy trình thực hiên của hàm như sau:

+Đầu tiên hàm sẽ xáo trộn các dòng của dataset bằng phương thức sample() của padas data_frame.

+Tính số dòng của mỗi fold: foldSize = số lượng dòng của dataset / k. Kết quả foldSize sẽ được làm tròn lên (sử dụng ceil() trong thư viện math), như vậy fold cuối cùng có thể sẽ có kích thước nhỏ hơn các fold khác, nhưng không đáng kể (nếu làm tròn xuống thì fold cuối cùng sẽ có kích thước nhỏ hơn rất nhiều so với các fold còn lại, VD: số dòng trong dataset là 504 và k = 5, thì nếu kích thước fold làm tròn xuống thì foldSize = 100 => fold cuối chỉ có 4 row, chưa kể chia như vậy được 6 fold tất cả).

+Chia dataset thành các fold và lưu vào trong list folds. Cách thực hiện:tạo list folds trống sau đó chạy vòng lặp i từ 0 đến k -1, mỗi lần lặp sẽ append các dòng có index có giá trị từ i * foldSize đến i * foldSize + foldSize vào list folds. Như vậy sẽ chia dataset thành k fold có kích thước chính xác như đã định.

+Chạy vòng lặp từng fold trong list folds. Mỗi lần lặp sẽ dùng fold hiện tại để test mô hình, các fold còn lại ghép thành train dataset dùng để train mô hình. Cụ thể, vì quy ước cột cuối của dataset truyền vào là cột giá trị y nên với mỗi lần lặp fold hiện tại sẽ tách cột cuối ra thành fold_y_test, các cột còn lại của fold hiện tại sẽ thành fold_X_test. Các fold còn lại sẽ ghép thành train dataset(bằng pd.concat()) và tách cột cuối của train dataset thành fold_y_train, các cột còn lại thành fold_X_train. Mô hình sẽ được train bằng fold_X_train và fold_y_train, sau đó gọi phương thức predict() truyền vào fold_X_test để nhận về giá trị dự đoán. Dùng giá trị dự đoán này và fold_y_test đề đánh giá độ lỗi của lần lặp này (gọi hàm mae()). Độ lỗi của mỗi lần lặp sẽ được lưu vào một list, cuối cùng hàm sẽ tính trung bình các giá trị độ lỗi trong list này rồi trả về.

5. Đánh giá và nhận xét

5.1 yêu cầu 1a

- Độ lỗi MAE trung bình trên tập test khi sử dụng 11 đặc trưng đầu tiên: 104863.778

-Công thức hồi quy:

Salary = - 22756.513 x Gender + 804.503 x 10percentage +1294.655 x 12percentage - 91781.898 x CollegeTier + 23182.389 x Degree + 1437.549 x

collegeGPA - 8570.662 x CollegeCityTier + 147.858 x English + 152.888 x Logical + 117.222 x Quant + 34552.286 x Domain

-Nhận xét:

+Mô hình sử dụng 11 đặc trưng đầu dự đoán dữ liệu khá tốt khi cho độ lệch chấp nhận được so với giá trị của cột lương.

5.2 yêu cầu 1b

-Kết quả(do xáo trộn dataset nên mỗi lần chạy kết quả khác nhau đôi chút):

STT	Mô hình với 1 đặc trưng	MAE
1	conscientiousness	306253.738
2	agreeableness	300761.536
3	extraversion	307037.108
4	nueroticism	299321.269
5	openess_to_experience	303046.571

-Nhận xét:

+Đặc trưng ảnh hưởng đến lương nhất trong 5 đặc trưng tính cách là nueroticism.

+Các đặc trưng tính các không ảnh hưởng quá nhiều đến lương khi độ lệch khá lớn

-Công thức hồi quy:

Salary = -56546.304 x nueroticism

-Độ lỗi MAE trung bình trên dữ liệu của test.csv khi chỉ sử dụng 1 đặc trưng nueroticism: 291019.693

5.3 yêu cầu 1c

-Kết quả(do xáo trộn dataset nên mỗi lần chạy kết quả khác nhau đôi chút):

STT	Mô hình với 1 đặc trưng	MAE
1	English	121909.565
2	Logical	120221.348
3	Quant	118054.307

-Nhận xét:

+Đặc trưng ảnh hưởng đến lương nhất trong 3 đặc trưng trên là Quant

+Ba đặc trưng english, logical và quant có ảnh hưởng đến lương nhiều hơn các đặc trưng tính cách

-Công thức hồi quy:

$$Salary = 585.895 \times Quant$$

-Độ lỗi MAE trên dữ liệu của test.csv khi chỉ sử dụng 1 đặc trưng Quant: 106819.578

5.4 yêu cầu 1d

-Tìm mô hình:

+Sử dụng k-fold Cross Validation với mô hình chỉ sử dụng 1 đặc trưng tương tự như yêu cầu 1b và 1c nhưng lần này là cho tất cả các đặc trưng.

+Sau khi chạy thử thấy các đặc trưng Quant, 10percentage, 12percentage, Logical, collegeGPA, English, CollegeTier, Degree, Gender, ComputerProgramming có độ lỗi thấp nhất và giá trị độ lỗi gần nhau nên em cho cho mô hình 1 sử dụng 10 đặc trưng này.

+Với mô hình 2, em chỉ sử dụng 2 đặc trưng có độ lỗi thấp nhất là Quant và 10percentage, nhưng khi chạy kiểm tra thử thấy độ lỗi của mô hình sử dụng 2 đặc trưng này vẫn còn lớn hơn độ lỗi trong các mô hình của yêu cầu trước nên em bình phương các giá trị của đặc trưng, sau đó kiểm tra lại thấy độ lệch đã nhỏ hơn các mô hình của yêu cầu trước.

+Tương tự mô hình 2, mô hình 3 sử dụng 3 đặc trưng có độ lỗi nhỏ nhất là Quant, 10percentage và 12percentage. Chạy kiểm tra thử thấy độ lỗi của mô hình sử dụng 3 đặc trưng này vẫn lớn hơn các mô hình này vẫn lớn hơn mô hình trong các đặc trưng trước nên em tiếp tục bình phương các giá trị của 3 đặc trưng đó, sau đó kiểm tra lại thấy độ lệch đã nhỏ hơn các mô hình của yêu cầu trước.

-Mô hình 1:

$$Salary = a_1 x Quant + a_2 \times 10percentage + a_3 \times 12percentage + a_4 \times Logical + a_5 \times collegeGPA + a_6 \times English + a_7 \times CollegeTier + a_8 \times Degree + a_9 \times Gender + a_{10} \times ComputerProgramming$$

với a_i là hệ số, $i \in \{1, 2, ..., 10\}$

-Mô hình 2

$$Salary = a_1 \times Quant^2 + a_2 \times 10percentage^2$$

với a_i là hệ số, $i \in \{1, 2\}$

-Mô hình 3:

$$Salary = a_1 \times Quant^2 + a_2 \times 10percentage^2 + a_3 \times 12percentage^2$$

với a_i là hệ số, $i \in \{1, 2, 3\}$

-Kết quả:

STT	Mô hình	MAE
1	Mô hình 1	113422.138
2	Mô hình 2	115840.607
3	Mô hình 3	115233.531

-Nhận xét:

+Mô hình tốt nhất là mô hình 1

+Do sử dung 10 đặc trưng có độ lỗi thấp nhất, nên mô hình dự đoán dữ liệu khá tốt.

+Mô hình 2 và mô hình 3 chỉ khác nhau rất ít, sự khác biệt này đến từ việc mô hình 3 có thêm đặc trưng $12percentage^2$

-Độ lỗi MAE trung bình trên dữ liệu của test.csv của mô hình 1: 104621.272

-Công thức hồi quy:

Salary = 133.45 x Quant + 899.534 x 10percentage + 1146.371 x 12percentage + 149.242 x Logical + 1307.67 x collegeGPA + 136.434 x English - 89598.216 x CollegeTier + 16709.546 x Degree - 23151.196 x Gender + 85.302 x ComputerProgramming

6. Tham khảo

1. pandas.pydata.org
2. 1trituenhantao.io, Giới thiệu về k-fold cross-validation <https://trituenhantao.io/kien-thuc/gioi-thieu-ve-k-fold-cross-validation/>, [ngày truy cập: 20 tháng 08 năm 2023].
3. Piyush Raj, Randomly Shuffle Pandas DataFrame Rows<https://datascienceparichay.com/article/randomly-shuffle-pandas-dataframe-rows/>[ngày truy cập: 20 tháng 08 năm 2023].
4. www.datacamp.com, Python Select Columns Tutorial <https://www.datacamp.com/tutorial/python-select-columns?utm_source=google&utm_medium=paid_search&utm_campaignid=19589720824&utm_adgroupid=143216588537&utm_device=c&utm_keyword=&utm_matchtype=&utm_network=g&utm_adpostion=&utm_creative=661628555495&utm_targetid=aud-438999696879:dsa-1947282172981&utm_loc_interest_ms=&utm_loc_physical_ms=1028581&utm_content=dsa~page~community-tuto&utm_campaign=230119_1-sea~dsa~tutorials_2-b2c_3-row-p2_4-prc_5-na_6-na_7-le_8-pdsh-go_9-na_10-na_11-na&gclid=Cj0KCQjw84anBhCtARIsAISI-xdIfhJMRpb-gg2nIU0kwpyF9jiqGHP67CEkV-BjunSrL3Y5aPV4TYAaAtaBEALw_wcB>[ngày truy cập: 21 tháng 08 năm 2023].

5. Edward Lin, 2018, 'DataFrame' object has no attribute 'ravel' when transforming target variable , *stackoverflow* <https://stackoverflow.com/questions/48841624/dataframe-object-has-no-attribute-ravel-when-transforming-target-variable>[ngày truy cập: 21 tháng 08 năm 2023].
6. ankthon, Sort rows or columns in Pandas Dataframe based on values<https://www.geeksforgeeks.org/sort-rows-or-columns-in-pandas-dataframe-based-on-values/>[ngày truy cập: 22 tháng 08 năm 2023].
7. Iserlohn, 2014, *stackoverflow*< https://stackoverflow.com/questions/24644656/how-to-print-pandas-dataframe-without-index>[ngày truy cập: 22 tháng 08 năm 2023].