TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN ĐHQG-HCM

KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN 3 – Linear Regression

Môn Học: Toán ứng dụng & thống kê

Giảng viên:

Trần Hà Sơn

Nguyễn Văn Quang Huy

Nguyễn Đình Thúc

Nguyễn Ngọc Toàn

Sinh viên thực hiện:

Nguyễn Lê Hùng 22127135

I. Các thư viện sử dụng

Thư viện	Lý do sử dụng
numpy	Cung cấp các công cụ mạnh mẽ cho tính toán số học với mảng và ma trận
panda	Thao tác với dữ liệu đầu vào dạng DataFrame
Sklearn	Sử dụng các hàm có sẵn như LinearRegression, KFold, tính MAE
Matplotlib	Tạo ra các biểu đồ và hình minh hoạ, trực quan hoá dữ liệu và kết quả
Seaborn	Tạo ra các biểu đồ như heatmap, box plot, và line plot

II. Các hàm sử dụng và mô tả hàm

1. Yêu cầu 1: Phân tích khám phá dữ liệu

-*Đầu vào*: 5 đặc trưng trong tập dữ liệu train đề bài cung cấp.

- *Mô tả các hàm khai phá dữ liệu:*

`describe()` pandas	Tạo thống kê mô tả cho một DataFrame, chẳng hạn như giá trị trung bình (mean), độ lệch chuẩn (standard deviation), giá trị nhỏ nhất (min) và giá trị lớn nhất (max)[1]
`histplot()` seaborn	Vẽ biểu đồ histogram với tùy chọn đường ước lượng mật độ hạt nhân (KDE). Hàm này giúp hiển thị phân phối của một đặc trưng trong tập dữ liệu.[2]
`countplot()` seaborn	Tạo ra một biểu đồ thanh (bar plot), hiển thị số lượng các quan sát trong mỗi nhóm danh mục bằng các thanh.[2]
`boxplot()` seaborn	Tạo ra một biểu đồ hộp, biểu diễn các nhóm dữ liệu số thông qua các tứ phân vị (quartile). Nó cũng có thể làm nổi bật các giá trị ngoại lai.
`heatmap()` seaborn	Tạo ra một bản đồ nhiệt (heatmap), một biểu diễn đồ họa của dữ liệu trong đó các giá trị riêng lẻ được thể hiện bằng các màu sắc.

	Thông thường, nó được sử dụng để trực quan hóa ma trận tương quan.
`scatterplot()` seaborn	Tạo ra một biểu đồ phân tán (scatter plot), là một loại biểu đồ hiển thị mối quan hệ giữa hai biến bằng cách biểu diễn các điểm tại giao điểm của trục x và y.
`figure()` `matplotlib`	Tạo một figure mới trong Matplotlib. Nó cho phép tùy chỉnh figure, bao gồm kích thước và các thuộc tính khác.[3]

Những hàm này kết hợp với nhau tạo thành một quá trình phân tích dữ liệu khám phá toàn diện, rất quan trọng để hiểu rõ dữ liệu trước khi xây dựng các mô hình dự đoán. Mỗi hàm đều được chọn lựa để làm nổi bật các khía cạnh cụ thể của tập dữ liệu, giúp đưa ra những thông tin quan trọng dẫn đến các bước tiếp theo trong việc xây dựng mô hình.

2. **Yêu cầu 2a: Xây dựng mô hình sử dụng toàn bộ 5 đặc trưng đề bài cung cấp**
   -Đầu vào: Các đặc trưng: Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, Sample Question Papers Practiced
   -Mô tả hàm tìm ra hệ số: Để xây dựng mô hình hồi quy tuyến tính với 5 đặc trưng, sử dụng lớp `LinearRegression()[4]` từ thư viện sklearn.linear_model. Sau đó sử dụng phương thức `fit()` dựa trên phương pháp bình phương nhỏ nhất (Ordinary Least Squares – OLS) để tìm ra các hệ số hồi quy tối ưu. phương thức này sẽ tìm các hệ số β0,β1 ,…,βn sao cho hàm hồi quy tuyến tính dưới đây.[5] Đây cũng chính là phương thức huấn luyện chính trong đồ án này.

$$Y_i = \underbrace{\beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}}_{\text{predictable}} + \underbrace{\varepsilon_i}_{\text{unpredictable}}$$

where $E\varepsilon_i = 0$, or equivalently

$$\mathbf{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip}$$

Least squares estimation

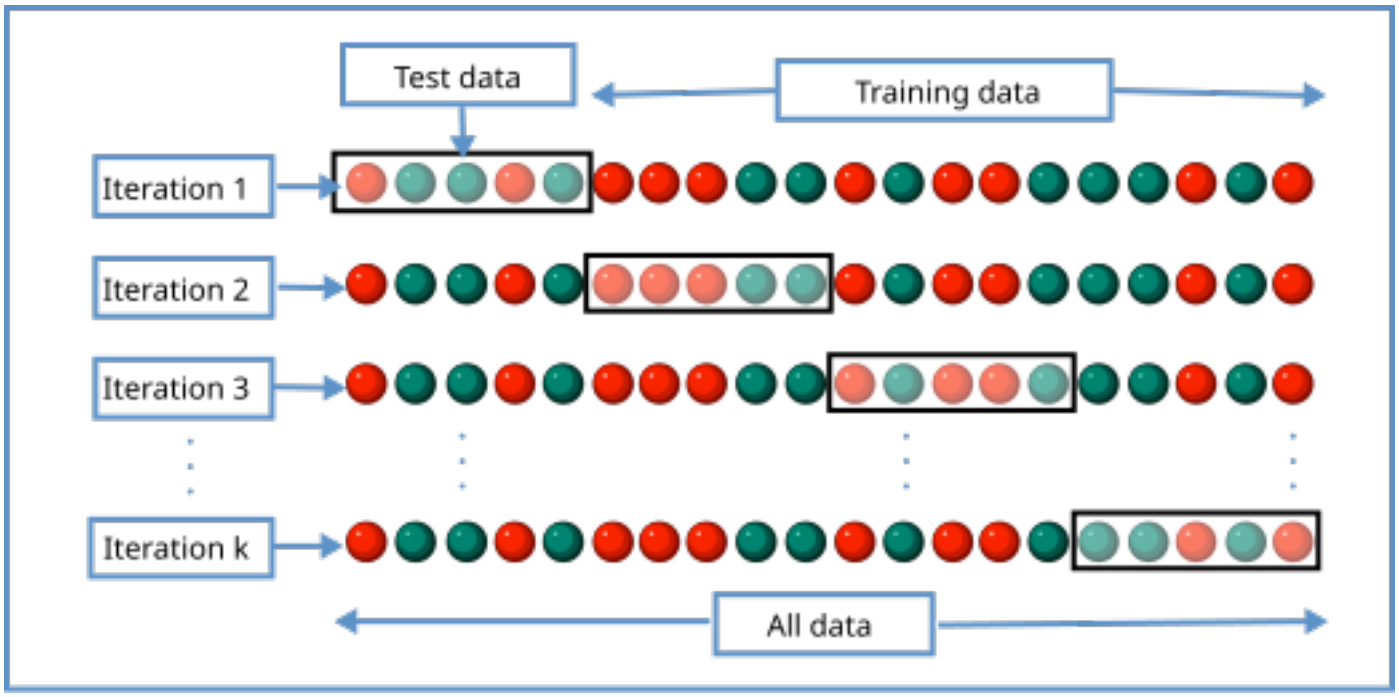
- Minimize $Q(b_0, ..., b_p) = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - ... - b_p X_{i,p})^2$

Trong đó:

- Y_i là giá trị thực tế của mẫu i
- Q là giá trị dự đoán của mẫu i.
- b_0 là hệ số của đặc trưng .

-Dự đoán: Sau khi đã tìm được hệ số cần thiết, tiến hành sử dụng các giá trị trong 5 đặc trưng trong *test* để tiến hành dự đoán. Sử dụng phương thức `predict()` từ thư viện sklearn.linear.model[4] để thực hiện tiến trình, gán giá trị dự đoán là *y_pred*
.-Tạo giá trị MAE để kiểm định: sử dung hàm `mean_absolute_error()` của thư viện sklearn[6], đầu vào là giá trị y_test và y_pred là giá trị dự đoán trẻ về sau khi thực hiện các tiến trình trên, hàm sẽ trả về giá trị MAE
-Lấy các hệ số của mô hình để tạo công thức hồi quy: Mô hình sẽ tính toán các hệ số hồi quy (coefficients) và hệ số tự do (intercept) dựa trên phương pháp bình phương nhỏ nhất (OLS)

3. **Yêu cầu 2b: Sử dụng duy nhất 1 đặc trưng, tìm mô hình cho kết quả tốt nhất**
-Đầu vào: Các đặc trưng: Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, Sample Question Papers Practiced
Khác với yêu cầu 2a, mỗi đặc trưng này sẽ được thử nghiệm riêng lẻ với mô hình hồi quy tuyến tính để tìm ra đặc trưng mang lại kết quả dự đoán tốt nhất sử dụng kỹ thuật K-Fold
-Mô tả hàm và phân tích kỹ thuật K-Fold Cross-Validation[7]
Cross validation là một phương pháp thống kê được sử dụng để ước lượng hiệu quả của các mô hình học máy. Nó thường được sử dụng để so sánh và chọn ra mô hình tốt nhất cho một bài toán. Kỹ thuật này dễ hiểu, dễ thực hiện và cho ra các ước lượng tin cậy hơn so với các phương pháp khác.
Cách thức hoạt động của K-Fold Cross-Validation:
- **Bước 1**: Chia tập dữ liệu thành K phần bằng nhau.
- **Bước 2**: Trong mỗi lần lặp, sử dụng K-1 phần để huấn luyện mô hình và phần còn lại để kiểm tra mô hình.
- **Bước 3**: Lặp lại quá trình trên K lần, mỗi lần chọn một phần khác nhau làm tập kiểm tra.
- **Bước 4**: Tính toán giá trị trung bình của một số liệu đánh giá (ví dụ: MAE) trên tất cả K lần lặp để có được một ước lượng tổng quát về hiệu suất của mô hình.



Hình minh hoạ

- -Sử dụng `shuffle` và `random_state`, `n_splits` trong hàm khởi tạo KFold nhằm đảm bảo dữ liệu được xáo trộn trước khi chia thành các phần, giúp đảm bảo mỗi phần là đại diện tốt cho toàn bộ tập dữ liệu.
- -Thực hiện tính toán trên từng đặc trưng sử dụng hàm `cross_val_score()`
- -Tính giá trị trung bình MAE sử dụng `mae_scores.mean()`

Ý nghĩa của kết quả

Kết quả cuối cùng của đoạn code trên là đặc trưng tốt nhất cho mô hình hồi quy tuyến tính, dựa trên giá trị MAE trung bình thấp nhất khi sử dụng K-Fold Cross-Validation. Đặc trưng này sau đó sẽ được sử dụng để huấn luyện lại mô hình trên toàn bộ tập dữ liệu và dự đoán trên tập kiểm tra.

-Dự đoán: sử dụng phương thức huấn luyện và dự đoán giống với yêu cầu 2a
-Tạo giá trị MAE để kiểm định: giống yêu cầu 2a
-Lấy các hệ số của mô hình để tạo công thức hồi quy: giống yêu cầu 2a

4. **Yêu cầu 2c: Sinh viên tự xây dựng/thiết kế mô hình, tìm mô hình cho kết quả tốt nhất**

-*Đầu vào*: Các đặc trưng và thiết kế mô hình

Khác với yêu cầu 2a 2b thì 2c thiết kế và thử nghiệm năm mô hình khác nhau dựa trên các đặc trưng trong tập dữ liệu. Mỗi mô hình được thiết kế với một mục tiêu cụ thể, nhằm tìm ra mối quan hệ tốt nhất giữa các đặc trưng và chỉ số thành tích học tập (Performance Index). Các mô hình này được thiết kế để kiểm tra cả mối quan hệ tuyến tính và phi tuyến tính,[8] cũng như sự tương tác giữa các đặc trưng. Tiếp tục sử dụng kỹ thuật K-Fold Cross-Validation

Danh sách các mô hình thiết kế và giải thích

1. **Model 1: Sử dụng tất cả các đặc trưng gốc**:
   - *Mô tả*: Mô hình này sử dụng toàn bộ năm đặc trưng ban đầu mà không có bất kỳ biến đổi nào.
   - *Giả thuyết*: Sử dụng tất cả các đặc trưng sẽ cung cấp đầy đủ thông tin cho mô hình, từ đó giúp mô hình có khả năng dự đoán tốt hơn.
2. **Model 2: Sử dụng bình phương của 'Hours Studied':**
   - *Mô tả*: Mô hình này thêm một đặc trưng mới là bình phương của 'Hours Studied'. Đây là một biến đổi phi tuyến tính của đặc trưng này.
   - *Giả thuyết*: Mối quan hệ giữa số giờ học và thành tích có thể không phải là tuyến tính. Bình phương của 'Hours Studied' có thể giúp mô hình phát hiện những ảnh hưởng phi tuyến tính đến kết quả học tập.
3. **Model 3: Sử dụng tổng của ('Hours Studied' và 'Previous Scores'):**
   - *Mô tả*: Mô hình này thêm một đặc trưng mới là tổng của 'Hours Studied' và 'Previous Scores'.
   - *Giả thuyế:t* Tổng của số giờ học và điểm số trước đó có thể đại diện cho tổng nỗ lực học tập của học sinh. Việc tổng hợp này có thể cải thiện khả năng dự đoán.
4. **Model 4: Sử dụng 'Hours Studied', 'Previous Scores', và tương tác giữa hai đặc trưng này:**
   - *Mô tả*: Mô hình này thêm một đặc trưng mới là tổng của 'Hours Studied' và 'Previous Scores'.
   - *Giả thuyết:* Tổng của số giờ học và điểm số trước đó có thể đại diện cho tổng nỗ lực học tập của học sinh. Việc tổng hợp này có thể cải thiện khả năng dự đoán.

5. **Model 5: Sử dụng 'Sleep Hours' và 'Sample Question Papers Practiced':**
   - *Mô tả*: Mô hình này thêm một đặc trưng mới là tổng của 'Hours Studied' và 'Previous Scores'.
   - *Giả thuyết:* Tổng của số giờ học và điểm số trước đó có thể đại diện cho tổng nỗ lực học tập của học sinh. Việc tổng hợp này có thể cải thiện khả năng dự đoán.

III. Báo cáo kết quả

1. **Yêu cầu 1:**
   Trong file mã nguồn.
2. **Yêu cầu 2a:**
   -Công thức hồi quy tuyến tính:
   Student Performance = 2.852 * Hours Studied + 1.018 * Previous Scores + 0.604 * Extracurricular Activities + 0.474 * Sleep Hours + 0.192 * Sample Question Papers Practiced + -33.969
   -MAE trên tập test: 1.595648688476298

Actual Performance Index		Predicted Performance Index
65.0	0	65.298108
79.0	1	79.665047
60.0	2	58.979376
52.0	3	54.430307
39.0	4	36.863759

3. **Yêu cầu 2b:**
   -Đặc trưng tốt nhất: Previous Scores
   -Công thức hồi quy tuyến tính cho đặc trưng tốt nhất:
   Student Performance = 1.011 * Previous Scores + -14.989
   -MAE các đặc trưng sử dụng mô hình K-Fold Cross Validation:

```
Feature: Hours Studied, MAE: 15.450840250192456
Feature: Previous Scores, MAE: 6.618829055214323
Feature: Extracurricular Activities, MAE: 16.19698077678371
Feature: Sleep Hours, MAE: 16.190870831540032
Feature: Sample Question Papers Practiced, MAE: 16.188351141715593
```

   -MAE của Previous Scores trên tập test: 6.544277293452478
4. **Yêu cầu 2c:**
   -Đặc trưng tốt nhất: + Mô hình tất cả các features tốt nhất

+Mô hình 2 features 'Hours Studied', 'Previous Scores, và sự tương tác của 2 đặc trưng này.

-Công thức hồi quy tuyến tính cho đặc trưng tốt nhất:

Student Performance = 1.011 * Previous Scores + -14.989

-MAE các đặc trưng sử dụng mô hình K-Fold Cross Validation:

```
Model 1 (All features) average MAE from cross-validation: 1.6210767924783938
Model 2 (Hours Studied^2) average MAE from cross-validation: 15.508817336101135
Model 3 (Hours + Scores) average MAE from cross-validation: 4.403828645415788
Model 4 (Hours Studied, Previous Scores, Interaction) average MAE from cross-validation: 1.8163124697842
Model 5 (Sleep Hours + Sample Question Papers) average MAE from cross-validation: 16.181756069565036

Best model based on cross-validation: Model 1 (All features), Best MAE: 1.6210767924783938
```

-MAE của Previous Scores trên tập test: 1.595648688476298

IV. Nhận xét kết quả

1. **Yêu cầu 2a:**

   -Nhân xét: Việc sử dụng tất cả các đặc trưng gốc cho phép mô hình hồi quy tuyến tính khai thác tối đa thông tin từ dữ liệu. Điều này hợp lý vì mỗi đặc trưng có thể đóng góp một phần vào dự đoán chỉ số thành tích học tập. Tuy nhiên, mô hình này cũng có thể gặp phải vấn đề quá khớp (overfitting) nếu một hoặc nhiều đặc trưng không thực sự cần thiết. Trong trường hợp này, việc loại bỏ các đặc trưng ít quan trọng có thể cải thiện khả năng tổng quát hóa của mô hình.

2. **Yêu cầu 2b:**

   -Nhân xét: **Previous Scores** là một chỉ số hợp lý dự đoán kết quả học tập, vì điểm số trước đó thường phản ánh chính xác năng lực học tập của học sinh trong các kỳ thi tiếp theo.Mặc dù mô hình với một đặc trưng tốt nhất có thể đạt được kết quả tốt, nhưng nó có thể bỏ qua các yếu tố khác có thể ảnh hưởng đến thành tích học tập (như số giờ học, hoạt động ngoại khóa, v.v.). Điều này làm cho mô hình đơn giản nhưng có thể thiếu chính xác trong một số trường hợp

3. **Yêu cầu 2c:**

   -Nhận xét:

   - **Model 1 (All features)**: Đây là mô hình sử dụng tất cả các đặc trưng gốc, và như đã thấy từ kết quả, nó có khả năng dự đoán tốt nhất. Điều này cho thấy rằng khi tất cả các thông tin có sẵn được sử dụng, mô hình có thể khai thác toàn bộ dữ liệu để đưa ra dự đoán chính xác.
   - **Model 2 (Hours Studied^2)**: Mô hình này có MAE rất cao, cho thấy rằng mối quan hệ giữa Hours Studied và kết quả học tập không phải là phi tuyến tính đơn giản. Việc chỉ sử dụng biến đổi phi tuyến tính này không mang lại hiệu quả tốt.

- **Model 3 (Hours + Scores)** và **Model 4 (Interaction)**: Cả hai mô hình này đều có MAE thấp hơn so với Model 2 nhưng vẫn không tốt bằng Model 1. Điều này cho thấy rằng mặc dù việc kết hợp và tương tác giữa các đặc trưng có thể mang lại lợi ích, nhưng nó không đủ mạnh để vượt qua việc sử dụng tất cả các đặc trưng gốc.
-**Model 5 (Sleep Hours + Sample Question Papers Practiced)**: MAE của mô hình này cao nhất, cho thấy rằng hai đặc trưng này không phải là những yếu tố quyết định chính đến kết quả học tập.

V. Tài liệu tham khảo

[1] Thư viện pandas tại:

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.describe.html

[2] Thư viện seaborn tại:

https://seaborn.pydata.org/generated/seaborn.histplot.html

[3] Thư viện trực quan hoá dữ liệu matplot.lib

https://matplotlib.org/

[4] Thực hiện Linear Regression với Scikit-learn tại

https://vndataproduct.org/huong-dan-chay-simple-linear-regression-trong-sklearn/

[5]Công thức hàm hồi quy tuyến tính, Chapter 2 Multiple Regression I (Part I) NewJersy Institute of Technology.

https://web.njit.edu/~wguo/Math644_2012/Math644_Chapter%202_part1.pdf

[6] ‘ Đánh giá mô hình học máy’ by Trung Đức on Viblo

https://viblo.asia/p/danh-gia-cac-mo-hinh-hoc-may-RnB5pp4D5PG

[7] Giới thiệu về k-fold cross-validation

https://trituenhantao.io/kien-thuc/gioi-thieu-ve-k-fold-cross-validation/

[8] Hồi quy phi tuyến tính

Wikipedia

**********HẾT**********