TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN, ĐHQG - HCM

KHOA CÔNG NGHỆ THÔNG TIN

TOÁN ỨNG DỤNG VÀ THỐNG KÊ

BÁO CÁO ĐỒ ÁN 3

< LINEAR REGRESSION >

Lâm Thanh Ngọc - 21127118
Lớp: 21CLC02

Giảng viên:

Vũ Quốc Hoàng

Nguyễn Văn Quang Huy

Lê Thanh Tùng

Phan Thị Phương Uyên

Ngày 24 tháng 8 năm 2023

Mục lục

1 Các thư viện đã sử dụng

pandas

- Thư viện pandas là một công cụ thao tác và phân tích dữ liệu nguồn mở nhanh, mạnh mẽ, linh hoạt và dễ sử dụng [12]
- Lý do sử dụng:
  + Thư viện pandas cung cấp các đối tượng DataFrame và Series.
  + Pandas hỗ trợ nhiều loại định dạng dữ liệu khác nhau, cụ thể là khả năng đọc dữ liệu từ file CSV thông qua hàm `pd.read_csv()` [7] và truy cập các hàng và cột của DataFrame bằng cách sử dụng chỉ số nguyên bằng hàm `pandas.DataFrame.iloc` [11].

numpy

- Thư viện numpy là một thư viện toán học phổ biến và mạnh mẽ của Python. Cho phép làm việc hiệu quả với ma trận và mảng, đặc biệt là dữ liệu ma trận và mảng lớn với tốc độ xử lý nhanh. [14]
- Lý do sử dụng: Numpy hỗ trợ các hàm tính toán với ma trận hiệu quả. Trong đoạn chương trình đã sử dụng các hàm hỗ trợ hồi quy tuyến tính như sau:
  + `numpy.linalg.inv`: tính nghịch đảo của ma trận. [3]
  + `numpy.sum`: tính tổng các phần tử của ma trận. [6]
  + `numpy.mean`: tính giá trị trung bình ma trận dựa theo trục được chỉ định. [4]
  + `numpy.abs`: tính giá trị tuyệt đối theo từng phần tử. [2]
  + `numpy.ravel`: trả về một ma trận phẳng. [5]

sklearn

- Thư viện sklearn là 1 công cụ hỗ trợ phân tích dữ liệu và dự đoán một cách nhanh chóng và hiệu quả được xây dựng trên nền tảng NumPy, SciPy và matplotlib. [8]
- Lý do sử dụng: Thư viện sklearn được sử dụng để hỗ trợ việc lựa chọn các mô hình hình bằng cách sử dụng lớp KFold trong sklearn.model_selection [9] để chia tập dữ liệu thành các tập con để sử dụng trong cross-validation.

2 Các hàm đã sử dụng

class OLSLinearRegression [13]

- `def fit(self, X, y)`: hàm được sử dụng để huấn luyện mô hình hồi quy tuyến tính trên tập dữ liệu X và y bằng cách sử dụng công thức nghịch đảo giả của ma trận X để tính trọng số w cho mô hình (một vector chứa các hệ số của phương trình hồi quy). Hàm này trả về chính đối tượng self được hiểu như mô hình được huấn luyện.
- `def get_params(self)`: hàm được sử dụng để trả về trọng số w đã được tính toán trong hàm fit.
- `def predict(self, X)`: hàm được sử dụng để dự đoán giá trị liên tục từ các đặc trưng bằng phương pháp hồi quy tuyến tính với dữ liệu nhận vào là một ma trận X chứa các đặc trưng của các điểm dữ liệu và trả về một mảng chứa các giá trị dự đoán cho từng điểm dữ liệu bằng cách tính tổng tích vô hướng của vector trọng số w và vector đặc trưng của mỗi điểm dữ liệu.

`def mae(y, y_hat)` [13]

- Hàm được sử dụng để tính trung bình của giá trị tuyệt đối của sự khác biệt giữa y và y_hat bằng cách sử dụng phương thức ravel() để chuyển đổi y và y_hat thành các mảng một chiều, sau đó sử dụng phương thức abs() để tính giá trị tuyệt đối của sự khác biệt. Sử dụng phương thức mean() để tính trung bình.
- Giá trị trả về của hàm mae càng nhỏ thì càng cho thấy mô hình hồi quy có độ chính xác cao.

3 Nhận xét kết quả từ các mô hình

3.1 Yêu cầu 1a: Sử dụng toàn bộ 11 đặc trưng đầu tiên Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain

- Mô tả:
  + Để xây dựng mô hình sử dụng toàn bộ 11 đặc trưng đầu tiên, trước hết chọn các cột từ 0 đến 10 trong cả tập huấn luyện và tập kiểm tra để tạo ra các ma trận X_1a_train và X_1a_test.
  + Sau đó, khởi tạo một đối tượng lr_1a thuộc lớp OLSLinearRegression và gọi phương thức fit để huấn luyện mô hình với X_1a_train và y_train. Phương thức get_params được gọi để xem các tham số của mô hình.
  + Sử dụng phương thức predict để dự đoán giá trị của y cho tập kiểm tra với X_1a_test và lưu kết quả vào Y_1a_test
  + Tính độ lỗi trung bình tuyệt đối của mảng chứa các giá trị thực tế của biến (y_test) và mảng chứa các giá trị dự đoán của biến (Y_1a_test).
- Công thức hồi quy (phần trọng số làm tròn đến 3 chữ số thập phân):

Salary = Gender * (-22756.513) + 10percentage * 804.503 + 12percentage * 1294.655 + CollegeTier * (-91781.898) + Degree * 23182.389 + collegeGPA * 1437.549 + CollegeCityTier * (-8570.662) + English * 147.858 + Logical * 152.888 + Quant * 117.222 + Domain * 34552.286

- MAE mô hình:

$$MAE = 104863.777$$

- Nhận xét mô hình:
  + Mô hình hồi quy tuyến tính đa biến được xây dựng để dự đoán mức lương của sinh viên tốt nghiệp dựa trên 11 đặc trưng đầu tiên: Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain. Có thể thấy, với MAE khá lớn (104863.777) so với giá trị thực tế cho thấy độ chính xác của mô hình trong việc đánh giá salary là không cao. Điều này có thể bị gây ra bởi sự quá khớp (overfitting).
  + Trong thống kê, sự quá khớp là kết quả của một phân tích mà tương ứng với việc đạt độ chính xác quá cao với một tập dữ liệu nào đó, vì vậy điều này có thể thất bại khi so khớp với các dữ liệu bổ sung hoặc dự đoán các quan sát đáng tin cậy trong tương lai. Cụ thể, do sự quá khớp dẫn đến việc mô hình được huấn luyện khá tốt trên tập train nhưng sẽ xảy ra sai lệch đáng kể khi kiểm tra trên tập test. [10]

3.2 Yêu cầu 1b: Xây dựng mô hình sử dụng duy nhất 1 đặc trưng tính cách với các đặc trưng tính cách gồm conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience, tìm mô hình cho kết quả tốt nhất

- Mô tả:
  + Thuật toán k-fold Cross Validation [1]:
    * Thuật toán k-fold Cross Validation là một phương pháp thường được sử dụng để đánh giá hiệu suất của một mô hình trên một tập dữ liệu.
    * Ý tưởng cơ bản của thuật toán này là chia tập dữ liệu thành k phần bằng nhau, gọi là các fold. Sau đó, lặp lại k lần, mỗi lần chọn một fold làm tập kiểm tra, và các fold còn lại làm tập huấn luyện. Kết quả cuối cùng là trung bình của các MAE trên k lần lặp.
    * Cụ thể, quy trình của thuật toán như sau:
      1. Xáo trộn tập dữ liệu một cách ngẫu nhiên.
      2. Chia tập dữ liệu thành k phần.
      3. Đối với mỗi nhóm, chọn một nhóm làm tập kiểm tra và các nhóm còn lại là tập huấn luyện. Huấn luyện mô hình trên tập huấn luyện và đánh giá trên tập kiểm tra.
      4. Tổng kết các kết quả của các mô hình và chọn mô hình cho kết quả tốt nhất.
  + Đầu tiên chọn ra các đặc trưng tính cách của tập train và tập test (5 đặc trưng) để tạo ra các ma trận X_1b_train và X_1b_test.
  + Sử dụng KFold để chia tập train thành 5 phần và lặp qua từng phần, với mỗi phần, lấy ra đặc trưng tính cách thứ i và huấn luyện mô hình hồi quy tuyến tính.
  + Sử dụng mô hình đã huấn luyện để dự đoán điểm số của phần test và tính độ lỗi MAE và lưu lại các giá trị MAE này vào một danh sách.
  + Tính độ lỗi MAE trung bình của mô hình với từng đặc trưng tính cách và tìm ra đặc trưng có độ lỗi thấp nhất.

- MAE từng đặc trưng tính cách:

STT	Mô hình với 1 đặc trưng	MAE
1	conscientiousness	306309.202
2	agreeableness	300912.678
3	extraversion	307030.103
4	neuroticism	299590.050
5	openness_to_experience	302957.692

- MAE mô hình với đặc trưng tốt nhất (nueroticism):

$$MAE = 291019.693$$

- Công thức hồi quy (phần trọng số làm tròn đến 3 chữ số thập phân):

Salary = (-56546.304) * nueroticism

- Nhận xét mô hình:
  + Kết quả mô hình cho thấy đặc trưng tính cách nueroticism là đặc trưng tính cách tốt nhất do MAE của đặc trưng tính cách này là nhỏ nhất. Điều này mang một ý nghĩa rằng lương của nhân viên sẽ giảm khi tính cách nueroticism tăng.
  + Có thể lý giải điều này rằng những người có mức độ nueroticism cao thường có xu hướng lo lắng, bất an và thiếu tự tin. Điều này có thể ảnh hưởng tiêu cực đến hiệu suất làm việc và khả năng đàm phán lương của họ.
  + Ngược lại, những người có mức độ nueroticism thấp thường có tâm trạng ổn định, tự tin và thoải mái. Điều này có thể giúp họ làm việc hiệu quả và đạt được mức lương cao hơn. Do đó, có một mối quan hệ âm giữa nueroticism và mức lương của nhân viên.
  + Tuy nhiên, nhiều yếu tố khác (ví dụ như kinh nghiệm, trình độ học vấn, ngành nghề,...) vẫn cần được xét đến do nueroticism không phải là đặc trưng duy nhất gây ảnh hưởng đến tiền lương. Do đó, mô hình sử dụng duy nhất 1 đặc trưng có thể bỏ qua những yếu tố quan trọng khác và gây ra sai số lớn.

3.3 Yêu cầu 1c: Xây dựng mô hình sử dụng duy nhất 1 đặc trưng English, Logical, Quant, tìm mô hình cho kết quả tốt nhất

- Mô tả:
  + Tương tự với câu 1b, để có thể chọn ra đặc trưng kỹ năng tốt nhất cũng sử dụng thuật toán k-fold Cross Validation.
  + Đầu tiên chọn ra các đặc trưng kỹ năng của tập train và tập test (3 đặc trưng) để tạo ra các ma trận X_1c_train và X_1c_test.
  + Sử dụng KFold để chia tập train thành 5 phần và lặp qua từng phần, với mỗi phần, lấy ra đặc trưng kỹ năng thứ i và huấn luyện mô hình hồi quy tuyến tính.
  + Sử dụng mô hình đã huấn luyện để dự đoán điểm số của phần test và tính độ lỗi MAE và lưu lại các giá trị MAE này vào một danh sách.
  + Tính độ lỗi MAE trung bình của mô hình với từng đặc trưng tính cách và tìm ra đặc trưng có độ lỗi thấp nhất.
- MAE từng đặc trưng tính cách:

STT	Mô hình với 1 đặc trưng	MAE
1	English	121925.884
2	Logical	120274.778
3	Quant	118124.524

- MAE mô hình với đặc trưng tốt nhất (nueroticism):

$$MAE = 300115.555$$

- Công thức hồi quy (phần trọng số làm tròn đến 3 chữ số thập phân):

Salary = 585.895 * Quant

- Nhận xét mô hình:
  + Kết quả mô hình cho thấy đặc trưng kỹ năng Quant là đặc trưng tốt nhất do MAE của đặc trưng này là nhỏ nhất. Điều này mang một ý nghĩa rằng lương của nhân viên sẽ tăng khi Quant tăng. Cụ thể, nếu sinh viên có điểm Quant cao hơn 1 điểm, thì mức lương dự kiến sẽ cao hơn khoảng 585.895 đồng. MAE cho biết sai số trung bình giữa giá trị dự đoán và giá trị thực tế là khoảng 300115.555 đồng.

+ Có thể thấy tồn tại mối quan hệ tuyến tính dương giữa Quant và mức lương của sinh viên. Do Quant phản ánh khả năng định lượng của ứng viên, là một trong những kỹ năng quan trọng liên quan đến nhiều công việc yêu cầu tính toán, phân tích và giải quyết vấn đề trong nhiều ngành nghề, đặc biệt là ngành tài chính, kinh doanh, kỹ thuật... nên ứng với đặc trưng Quant cao cho thấy sinh viên có nhiều tiềm năng và khả năng làm việc tốt hơn với kỹ năng định lượng tốt, và do đó được trả lương cao hơn.
+ Tuy nhiên, cùng với nhiều yếu tố khác (ví dụ như kinh nghiệm, bằng cấp, kỹ năng mềm,...), Quant chỉ là một phần của bài kiểm tra AMCAT, không thể đánh giá toàn diện năng lực của nhân viên. Do đó, mô hình sử dụng duy nhất 1 đặc trưng có thể bỏ qua những yếu tố quan trọng khác và gây ra sai số lớn.

3.4 Yêu cầu 1d: Tự xây dựng mô hình, tìm mô hình cho kết quả tốt nhất

Mô hình 1

Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain, ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg, conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience

Quy trình tìm ra mô hình:

Mô hình 1 được tìm ra bằng cách sử dụng toàn bộ 23 đặc trưng để đưa ra dự đoán mức lương của kỹ sư dựa trên giả thuyết rằng mức lương có sự phụ thuộc vào toàn bộ các đặc trưng. Ví dụ cụ thể như:

- Gender có thể phản ánh sự chênh lệch lương giữa nam và nữ trong một số ngành nghề.
- 10percentage và 12percentage có thể cho biết khả năng học tập và năng lực của các kỹ sư từ cấp hai và cấp ba.
- CollegeTier và Degree có thể cho biết chất lượng và uy tín của trường đại học và bằng cấp của các kỹ sư.
- collegeGPA và CollegeCityTier có thể cho biết thành tích học tập và môi trường sống của các kỹ sư trong quá trình đào tạo.
- English, Logical, Quant và Domain có thể cho biết kỹ năng ngôn ngữ, logic, toán học và chuyên môn của các kỹ sư.
- ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg và CivilEngg có thể cho biết

khả năng lập trình và kiến thức về các ngành kỹ thuật của các kỹ sư.

- conscientiousness, agreeableness, extraversion, nueroticism và openess_to_experience có thể cho biết tính cách và thái độ làm việc của các kỹ sư.

Mô hình 2

CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain, ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg, conscientiousness, agreeableness, extraversion, nueroticism, openess_to_experience

Quy trình tìm ra mô hình:

Mô hình 2 được tìm ra bằng cách sử dụng 20 đặc trưng (bỏ 3 đặc trưng Gender, 10percentage, 12percentage) dựa trên giả thuyết loại bỏ các đặc trưng ít có sự liên quan nhất đến tiền lương và việc bỏ 3 đặc trưng này sẽ giúp mô hình tập trung vào những yếu tố quan trọng hơn trong việc dự đoán tiền lương. Cụ thể, theo phân tích:

- Đặc trưng Gender có thể phản ánh sự phân biệt giới tính dẫn đến những vấn đề gây tranh cãi trong thị trường lao động và không liên quan nhiều đến năng lực của ứng viên.
- Đặc trưng 10percentage và 12percentage chỉ ra điểm số của ứng viên ở cấp 3 và cấp 2, nhưng chúng không phản ánh được kiến thức chuyên môn và kỹ năng mềm của ứng viên.

Mô hình 3

Gender, 10percentage, 12percentage, CollegeTier, Degree, collegeGPA, CollegeCityTier, English, Logical, Quant, Domain, conscientiousness, agreeableness, extraversion,nueroticism,openess_to_experience

Quy trình tìm ra mô hình: Qua quá trình thực hiện, các câu 1a, 1b và 1c đều sử dụng các đặc trưng có nhiều sự liên quan đến mức lương của kỹ sư. Tuy nhiên các đặc trưng câu 1a chứa cả các đặc trưng 1c nên mô hình 3 được chọn ra từ việc kết hợp 1a và 1b.

Mô tả việc xây dựng mô hình

- Đầu tiên, khởi tạo X_1d_cv_train và X_1d_cv_test_1 chứa lần lượt toàn bộ 23 đặc trưng trong tập huấn luyện và tập kiểm tra.
- Sử dụng thuật toán k-fold Cross Validation để tìm ra danh sách MAE. Sau đó, tính trung bình các MAE trong danh sách và gán giá trị vào mean_mae_1

Kết quả các mô hình

STT	Mô hình	MAE
1	Sử dụng toàn bộ đặc trưng	110420.414
2	Bỏ 3 đặc tính Gender, 10percentage, 12percentage	111724.327
3	Gộp các đặc trưng câu 1a và 1b	113088.918

Qua bảng kết quả các mô hình có thể thấy các mô hình được xây dựng có kết quả cải thiện rõ rệt so với câu 1b và 1c. Trong đó, mô hình nhỏ nhất có kết quả MAE nhỏ hơn câu 1a là **mô hình 1: sử dụng toàn bộ đặc trưng.**

Công thức hồi quy (phần trọng số làm tròn đến 3 chữ số thập phân)

Salary = -23874.542 * Gender + 898.576 * 10percentage + 1203.496 * 12percentage - 83592.388 * CollegeTier + 11515.431 * Degree + 1626.519 * CollegeGPA - 5717.734 * CollegeCityTier + 153.435 * English + 120.511 * Logical + 102.581 * Quant + 27939.640 * Domain + 76.730 * ComputerProgramming - 47.747 * ElectronicsAndSemicon - 177.387 * ComputerScience + 33.932 * MechanicalEngg - 151.471 * ElectricalEngg - 64.198 * TelecomEngg + 145.894 * CivilEngg - 19814.830 * conscientiousness + 15503.266 * agreeableness + 4908.582 * extraversion - 10661.029 * nueroticism - 5815.021 * openess_to_experience

MAE mô hình tốt nhất

$$MAE = 101872.210$$

Nhận xét mô hình tốt nhất

- Kết quả cho thấy mô hình tốt nhất là mô hình 1 khi sử dụng toàn bộ 23 đặc trưng để đưa đến dự đoán kết quả tiền lương.
- Công thức hồi quy cho biết mối quan hệ giữa các đặc trưng và tiền lương được dự đoán. Các hệ số của công thức hồi quy cho thấy mức độ ảnh hưởng của từng đặc trưng lên tiền lương. Ví dụ, Domain có hệ số dương cao nhất, có nghĩa là nếu ứng viên có điểm số cao về lĩnh vực chuyên môn thì tiền lương sẽ tăng nhiều nhất.
- Một số đặc trưng khác có hệ số gần bằng không, có nghĩa là không có ảnh hưởng rõ rệt lên tiền lương, ví dụ như MechanicalEngg hay CivilEngg.
- Mô hình này có MAE bằng 101872.210, tức là sai số trung bình giữa tiền lương thực tế và tiền lương dự đoán là khoảng 100 nghìn. Đây là một sai số khá cao, cho thấy mô hình này chưa thể dự đoán chính xác tiền lương của ứng viên do lượng dữ liệu được lấy khá nhiều sẽ dẫn đến trường hợp overfitting được đề cập ở mô hình 1a.

Tài liệu tham khảo

[1] Jason Brownlee. *A Gentle Introduction to k-fold Cross-Validation*. 2020. URL: https://machinelearningmastery.com/k-fold-cross-validation/.

[2] NumPy Developers. *numpy.abs — NumPy v1.21 Manual*. https://numpy.org/doc/stable/reference/generated/numpy.abs.html. 2021.

[3] NumPy Developers. *numpy.linalg.inv — NumPy v1.21 Manual*. https://numpy.org/doc/stable/reference/generated/numpy.linalg.inv.html. [Online documentation]. 2021.

[4] NumPy Developers. *numpy.mean — NumPy v1.21 Manual*. https://numpy.org/doc/stable/reference/generated/numpy.mean.html. 2021.

[5] NumPy Developers. *numpy.ravel — NumPy v1.21 Manual*. https://numpy.org/doc/stable/reference/generated/numpy.ravel.html. 2021.

[6] NumPy Developers. *numpy.sum — NumPy v1.21 Manual*. https://numpy.org/doc/stable/reference/generated/numpy.sum.html. 2021.

[7] Pandas Developers. $pandas.read_csv^-pandas1.4.0.dev0 + 1155.gd9c9f3a2adocumentation$. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.read_csv.html. 2021.

[8] Scikit-learn Developers. *scikit-learn: machine learning in Python*. 2007. URL: https://scikit-learn.org/stable/.

[9] Scikit-learn Developers. $sklearn.model_selection.KFold^-scikit-learn0.24.2documentation$. https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html. 2021.

[10] *Overfitting*. Accessed on 2023-08-23. URL: https://en.wikipedia.org/wiki/Overfitting.

[11] pandas development team. *pandas.DataFrame.iloc — pandas 2.0.3 documentation*. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.iloc.html. Accessed: 2023-08-23. 2021.

[12] The pandas development team. *pandas-dev/pandas: Pandas*. **version** latest. **february** 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[13] GV. Phan Thị Phương Uyên. "Hướng dẫn đồ án 03". **in***Học kỳ 3 - Năm 2*: (2023).

[14] Vimentor. *Chi tiết bài học 22. Giới thiệu Numpy*. https://vimentor.com/vi/lesson/22-gioi-thieu-numpy. [Online forum post]. n.d.