

Phân tích hiệu suất nhân viên và phân loại ảnh ANN

Thời gian: 2.5 giờ - Ngày: 04/04/2025

Lưu ý: Sinh viên được phép sử dụng tài liệu, nhưng phải giải thích rõ ràng logic, lý do và ngữ cảnh thực tế cho từng bước. Đáp án chỉ có code hoặc kết quả mà không có giải thích sẽ không được chấm điểm.

Phần A: Bộ dữ liệu hiệu suất nhân viên

Dưới đây là bộ dữ liệu giả định về hiệu suất làm việc của 20 nhân viên, bao gồm các cột: **Employee_ID**, **Department**, **Performance** (0-100), **Hours_Worked**, **Training_Hours**, **Gender**, **Location**.

Bảng 1: Dữ liệu hiệu suất

Employee_ID	Department	Performance	Hours_Worked	Training_Hours	Gender	Location
E001	Sales	88	40	5	Male	Site_A
E002	IT		45	3	Female	Site_B
E003	HR	75	38		Male	Site_A
E004	Sales	92	42	6	Female	Site_C
E005	IT	68		4		Site_B
E006	HR	85	39	5	Male	Site_A
E007	Sales		41	3	Female	Site_C
E008	IT	79	44	4	Male	Site_B
E009	HR	90	37		Female	Site_A
E010	Sales	73	40	5	Male	Site_C
E011	IT	86	46	2	Female	Site_B
E012	HR		38	6	Male	Site_A
E013	Sales	95	43	7	Female	Site_C
E014	IT	70		3	Male	Site_B
E015	HR	82	39	5		Site_A
E016	Sales	78	41	4	Female	Site_C
E017	IT	89	45		Male	Site_B
E018	HR	71	37	3	Female	Site_A
E019	Sales	87		5	Male	Site_C
E020	IT	93	44	6	Female	Site_B

Phần B: Câu hỏi kiểm tra

Câu hỏi EDA (60 điểm)

- (4 điểm) Tính tỷ lệ phần trăm giá trị thiếu trong từng cột bằng Pandas. Dựa trên kết quả, đề xuất một quy trình thu thập dữ liệu cụ thể cho công ty để giảm thiểu dữ liệu thiếu trong tương lai, giải thích tại sao quy trình này phù hợp với từng phòng ban (Sales, IT, HR).
- (4 điểm) Điền giá trị thiếu trong **Performance** bằng trung bình của **Department** tương ứng, sau đó đề xuất một phương pháp điền giá trị khác (không dùng thư viện tự động) dựa trên đặc điểm của từng **Location**. Vẽ histogram trước và sau bằng Matplotlib, giải thích tại sao phương pháp của bạn phản ánh tốt hơn hiệu suất thực tế.

3. (4 điểm) Tính độ lệch chuẩn của `Hours_Worked` bằng NumPy. Đề xuất một chính sách quản lý thời gian làm việc cho công ty dựa trên phân tích độ lệch chuẩn và trung vị, giải thích tại sao chính sách này có thể cải thiện hiệu suất tổng thể.
4. (4 điểm) Tính correlation giữa `Hours_Worked` và `Performance` cho từng `Location`. Dựa trên kết quả, đề xuất một chiến lược phân bổ công việc khác nhau cho từng `Location` để tối ưu hóa hiệu suất, giải thích tại sao chiến lược này phù hợp với đặc điểm nhân viên tại mỗi địa điểm.
5. (4 điểm) Vẽ boxplot của `Performance` theo `Department` và `Gender` (kết hợp) bằng Seaborn. Xác định outlier bằng IQR, sau đó đề xuất một kế hoạch phỏng vấn cá nhân với các nhân viên outlier để tìm hiểu nguyên nhân, giải thích cách kế hoạch này cải thiện quản trị nhân sự.
6. (4 điểm) Tạo cột mới `Efficiency = Performance / Hours_Worked`. Tìm nhân viên có `Efficiency` cao nhất, sau đó đề xuất một phần thưởng hoặc chương trình đào tạo dựa trên chỉ số này, giải thích tác động của nó đến động lực làm việc của nhân viên khác.
7. (4 điểm) Tính tỷ lệ nhân viên nữ (`Gender = Female`) trong từng `Department` sau khi điền giá trị thiếu bằng mode. Dựa trên kết quả, đề xuất một chính sách đa dạng giới tính cho công ty, giải thích cách chính sách này ảnh hưởng đến văn hóa tổ chức.
8. (4 điểm) Vẽ scatter plot giữa `Training_Hours` và `Performance`, tô màu theo `Location`. Đề xuất một kế hoạch đào tạo cụ thể cho từng `Location` dựa trên phân bố dữ liệu, giải thích tại sao kế hoạch này tối ưu hóa hiệu suất.
9. (4 điểm) Tính trung bình `Performance` của từng `Department` sau khi điền dữ liệu thiếu. Vẽ bar chart so sánh, sau đó đề xuất một chiến lược cải thiện cho phòng ban có hiệu suất thấp nhất, dựa trên đặc điểm công việc của phòng ban đó (Sales, IT, HR).
10. (4 điểm) Tìm các nhân viên có `Hours_Worked` dưới 40 nhưng `Performance` trên 85 bằng Pandas. Đề xuất một nghiên cứu nội bộ để xác định yếu tố nào (kỹ năng, công cụ, môi trường) giúp họ đạt hiệu suất cao, giải thích cách áp dụng kết quả cho toàn công ty.
11. (4 điểm) Điền giá trị thiếu trong `Performance` bằng hồi quy tuyến tính thủ công dựa trên `Hours_Worked` và `Training_Hours`. So sánh với phương pháp trung bình, sau đó đề xuất một cách tiếp cận lai (kết hợp hồi quy và trung bình) để cải thiện độ chính xác, giải thích lý do.
12. (4 điểm) Tính skewness của `Performance` bằng SciPy. Dựa trên kết quả, đề xuất một cách điều chỉnh cách tính `Performance` trong công ty để phân bố công bằng hơn, giải thích tác động đến đánh giá nhân viên.
13. (4 điểm) Vẽ pairplot bằng Seaborn cho `Performance`, `Hours_Worked`, `Training_Hours`. Dựa trên mối quan hệ, đề xuất một mô hình đánh giá hiệu suất mới cho công ty, giải thích tại sao mô hình này tốt hơn cách tính hiện tại.
14. (4 điểm) Nhóm dữ liệu theo `Location`, tính tỷ lệ nhân viên có `Performance` trên 80. Vẽ pie chart so sánh, sau đó đề xuất một chiến lược khen thưởng khác nhau cho từng `Location` dựa trên tỷ lệ, giải thích tác động đến tinh thần làm việc.
15. (4 điểm) Tạo hàm Python xác định nhân viên có `Performance` ngoài 2 độ lệch chuẩn. Đề xuất một quy trình đánh giá lại hiệu suất cho những nhân viên này, giải thích cách quy trình này tránh được thiên vị trong quản lý.

Câu hỏi ANN (40 điểm)

Sử dụng bộ dữ liệu ảnh MNIST (có sẵn trong PyTorch) để xây dựng ANN phân loại chữ số viết tay (0-9). MNIST gồm 60,000 ảnh train và 10,000 ảnh test, mỗi ảnh kích thước 28x28 pixel.

16. (10 điểm) Tải dữ liệu MNIST từ PyTorch (`torchvision.datasets.MNIST`). Chuẩn hóa pixel về `[0, 1]`, flatten thành vector 784 chiều. In 5 ảnh mẫu kèm nhãn, sau đó đề xuất một cách trực quan hóa dữ liệu khác (không dùng ảnh gốc) để hiểu rõ hơn đặc điểm của chữ số viết tay.
17. (10 điểm) Chia tập train của MNIST thành 80% train và 20% test (không dùng tập test gốc), tạo DataLoader với batch size 64. Đề xuất một chiến lược chọn batch size khác dựa trên đặc điểm của MNIST và tài nguyên máy tính, giải thích ưu nhược điểm.

18. (15 điểm) Xây dựng ANN bằng PyTorch với kiến trúc như hình sau. Huấn luyện 20 epochs với CrossEntropyLoss và Adam (lr=0.001), dùng early stopping (patience=5) dựa trên loss của tập test. Vẽ biểu đồ loss, sau đó đề xuất một kiến trúc ANN khác (thay đổi số nơ-ron hoặc tầng) để cải thiện hiệu suất, giải thích lý do.
19. (5 điểm) Đánh giá mô hình bằng accuracy và confusion matrix trên tập test tự chia. Nếu accuracy dưới 95%, phân tích các chữ số bị nhầm lẫn nhiều nhất trong confusion matrix, đề xuất một cách cải thiện mô hình dựa trên đặc điểm của những chữ số này (không chỉ dùng Dropout hay tăng epochs).

Hình: Kiến trúc ANN

