ĐẠI HỌC XÂY DỰNG HÀ NỘI
KHOA CÔNG NGHỆ THÔNG TIN

BÀI TẬP LỚN HỌC MÁY

Sinh viên thực hiện:

Đặng Hoàng Hải - 0047167
Đỗ Văn Dũng - 0300467
Nguyễn Văn Đông - 0225167

Giảng viên hướng dẫn: Nguyễn Đình Quý

Hà Nội, 11-2024

Mục lục

1 Giới thiệu

1.1 Mục tiêu

Xây dựng mô hình áp dụng thuật toán Naive Bayes để nhận dạng chữ viết tay và K-means để phân nhóm khách hàng. Đây là những bài toán thực tế có ứng dụng quan trọng trong nhận dạng hình ảnh và phân tích dữ liệu kinh doanh.

1.2 Tóm tắt nội dung

1.2.1 Xử lý dữ liệu

q

- Naive Bayes: Chuyển đổi hình ảnh chữ số (28x28 pixel) thành ma trận nhị phân dựa trên màu sắc pixel.
- K-means: Chuẩn hóa dữ liệu khách hàng (tuổi, thu nhập, điểm chi tiêu) để đảm bảo hiệu quả phân cụm.

1.2.2 Triển khai thuật toán

- Viết mã Python để huấn luyện mô hình Naive Bayes (đọc dữ liệu, tính xác suất điều kiện, dự đoán).
- Triển khai K-means bằng cách tính toán cụm trung tâm (centroid) và phân nhóm các khách hàng.

1.2.3 Đánh giá kết quả

- Sử dụng ma trận nhầm lẫn để kiểm tra độ chính xác của Naive Bayes.
- Sử dụng các chỉ số đánh giá F1-score, Precision, Recall để đánh giá mô hình huấn luyện.
- Với bài B-KMeans, sử dụng phương pháp elbow để tìm được số lượng cụm tối ưu và vẽ biểu đồ phân tán (scatter plot) để đánh giá hiệu quả phân cụm.

1.2.4 Thư viện sử dụng

Các thư viện chính trong Python được sử dụng:

- `numpy`: Xử lý ma trận và tính toán số học.
- `pandas`: Quản lý dữ liệu đầu vào.
- `matplotlib`, `seaborn`: Trực quan hóa dữ liệu và kết quả.

2 A-Naive Bayes

2.1 Xử lý dữ liệu

Dữ liệu hình ảnh chữ viết tay được biểu diễn dưới dạng tệp văn bản ASCII, trong đó:

- Mỗi hình ảnh là một ma trận 28×28 pixel.
- Màu sắc các pixel được mã hóa:
  - `#`: Pixel đen (foreground).
  - `+`: Pixel xám (foreground).
  - ` `: Pixel trắng (background).

Chuyển đổi dữ liệu: Hình ảnh được chuyển đổi thành ma trận nhị phân với giá trị 0 và 1:

- Pixel là foreground (`#` hoặc `+`): Gán giá trị 1.
- Pixel là background ( ): Gán giá trị 0.

Sau khi xử lý, mỗi hình ảnh được biểu diễn dưới dạng một vector đặc trưng có 784 phần tử (tương ứng với 28×28 pixel).

Kết quả xử lý dữ liệu: Mỗi hình ảnh đầu vào được biểu diễn bằng vector nhị phân 784 chiều.

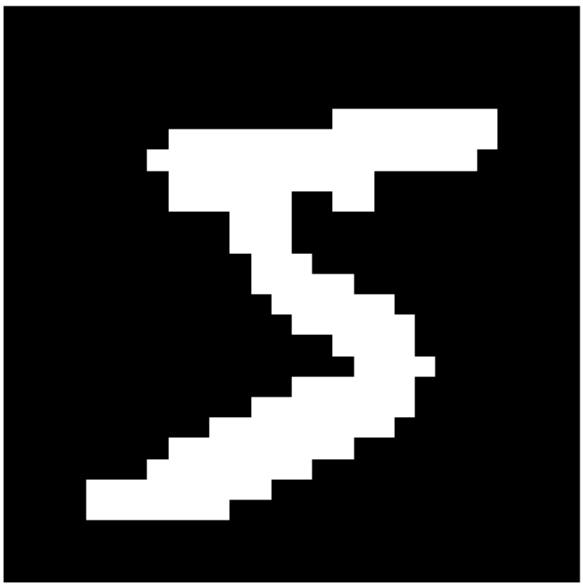

Figure 1: Dữ liệu sau khi được chuyển đổi.

2.2 Kết quả Naive Bayes

2.2.1 Hiệu suất của mô hình

Để đánh giá hiệu suất của thuật toán Naive Bayes, sử dụng **ma trận nhầm lẫn** (*Confusion Matrix*). Ma trận này biểu diễn tần suất các mẫu từ mỗi lớp được phân loại đúng hoặc nhầm sang lớp khác.

Biểu đồ trực quan hóa: Hình dưới đây hiển thị ma trận nhầm lẫn dưới dạng biểu đồ nhiệt, giúp minh họa rõ hơn tỷ lệ phân loại đúng và nhầm lẫn giữa các lớp.

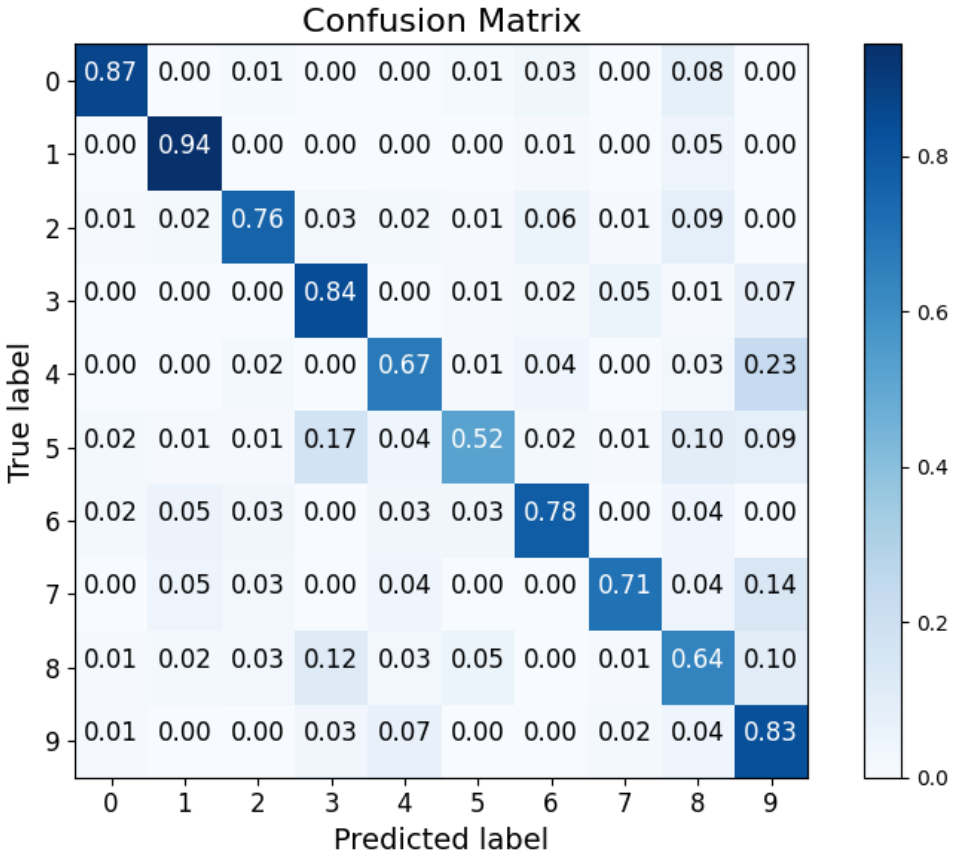

Figure 2: Biểu đồ nhiệt của ma trận nhầm lẫn.

2.2.2 Phân tích kết quả

- **Độ chính xác tổng thể:** đạt được **75,7%** với giá trị k tìm được là 0,1.
- Chỉ số **Precision**, **Recall** và **F1-score** lần lượt đạt được **0,7747**, **0,7562** và **0,7575** cho thấy khả năng dự đoán của mô hình khá tốt
- Chỉ số F1-score cho thấy mô hình có sự cân bằng khá tốt giữa Precision và Recall.
- **Độ chính xác từng lớp:** Các giá trị trên đường chéo của ma trận nhầm lẫn thể hiện tỷ lệ dự đoán đúng của từng lớp. Hầu hết các lớp có độ chính xác cao (trên 0.7), đặc biệt là các lớp **0**, **1**, và **3** với giá trị lần lượt là 0.87, 0.94, và 0.84.

2.2.3 Phân tích cụ thể từng lớp

- **Lớp 0 và 1:**
  - Chính xác cao: Tỷ lệ dự đoán đúng lần lượt là 87% và 94%.
  - Nhầm lẫn nhỏ: Cả 2 lớp bị nhầm lẫn một chút thành lớp 8 (8%, 5%).
- **Lớp 4**:
  - Có sự nhầm lẫn lớn với lớp 9 (24%)
- **Lớp 5:**
  - Hiệu suất thấp nhất: Tỷ lệ dự đoán đúng chỉ đạt 52%.
  - Nhầm lẫn phổ biến: Chủ yếu nhầm với lớp 3 (17%) và lớp 8 (10%), lớp 9 (9%).
- **Lớp 7:**
  - Khá chính xác: Tỷ lệ dự đoán đúng là 71%.
  - Nhầm lẫn phổ biến: Với lớp 9 (14%).

3 B-K-means

3.1 Xử lý dữ liệu

Trước khi đưa dữ liệu vào thuật toán K-means, dữ liệu cần được xử lý để đảm bảo phù hợp với yêu cầu của từng thuật toán. Dưới đây là các bước xử lý dữ liệu:

Dữ liệu khách hàng bao gồm các thông tin:

- **ID khách hàng:** Số định danh duy nhất.
- **Tuổi:** Tuổi của khách hàng (năm).
- **Giới tính:** Nam hoặc nữ.
- **Thu nhập hàng năm:** Tính theo nghìn USD.
- **Điểm chi tiêu:** Thang điểm từ 1 đến 100.

Chuẩn hóa Z-score: Để đảm bảo các đặc trưng có đơn vị đo tương đồng, dữ liệu được chuẩn hóa theo công thức:

$$x' = \frac{x - \mu}{\sigma} \tag{1}$$

Trong đó:

- x: Giá trị ban đầu.
- μ: Trung bình của đặc trưng.
- σ: Độ lệch chuẩn của đặc trưng.

Label encoding: Các đặc trưng rời rạc(Gender) của dữ liệu được chuẩn hóa bằng Label encoding. Giá trị 'Female' được chuẩn hóa thành 1, giá trị 'Male' được chuẩn hóa thành 2.

Trực quan hóa dữ liệu: Dữ liệu sau khi chuẩn hóa có thể được hiển thị bằng biểu đồ phân tán (scatter plot) để minh họa mối quan hệ giữa các đặc trưng.

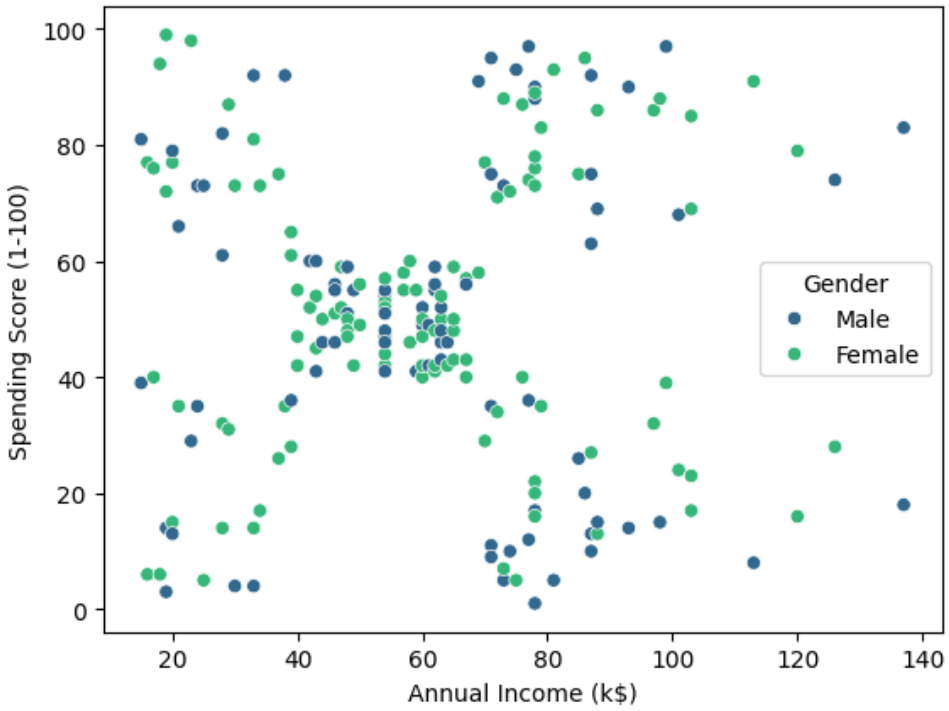

Figure 3: Biểu đồ phân tán dữ liệu khách hàng.

Kết quả xử lý dữ liệu

- **K-means:** Dữ liệu khách hàng được chuẩn hóa và chuyển đổi thành các điểm trong không gian nhiều chiều, sẵn sàng cho quá trình phân cụm.

3.2 Kết quả KMeans

3.2.1 Hiệu suất của mô hình

Đối với thuật toán K-means, kết quả được đánh giá dựa trên việc phân cụm dữ liệu khách hàng và hiển thị trực quan cụm trong không gian hai chiều.

Biểu đồ Elbow: Biểu đồ dưới đây minh họa giá trị *WCSS* (*Within-Cluster Sum of Squares*) theo số cụm K. Điểm gãy (elbow) cho thấy $K = 5$ là giá trị tối ưu.

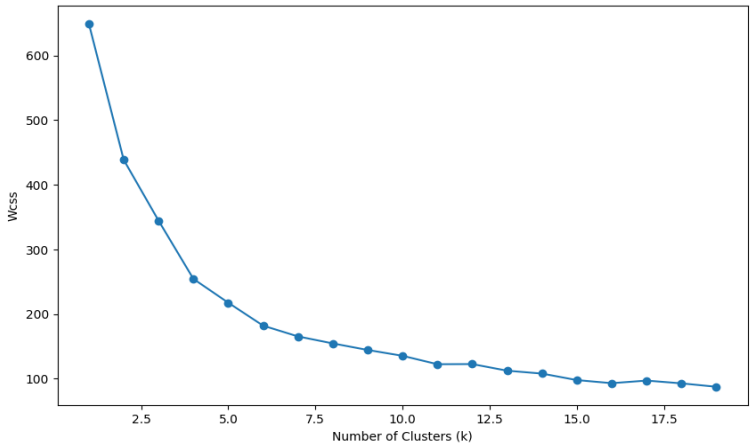

Figure 4: Biểu đồ Elbow để xác định số cụm tối ưu.

Biểu đồ phân cụm: Sau khi xác định $K = 5$, dữ liệu được phân cụm và hiển thị bằng biểu đồ phân tán. Các cụm được đánh dấu bằng màu sắc khác nhau, với trung tâm cụm (centroid) được biểu diễn bằng các dấu sao ($\star$).

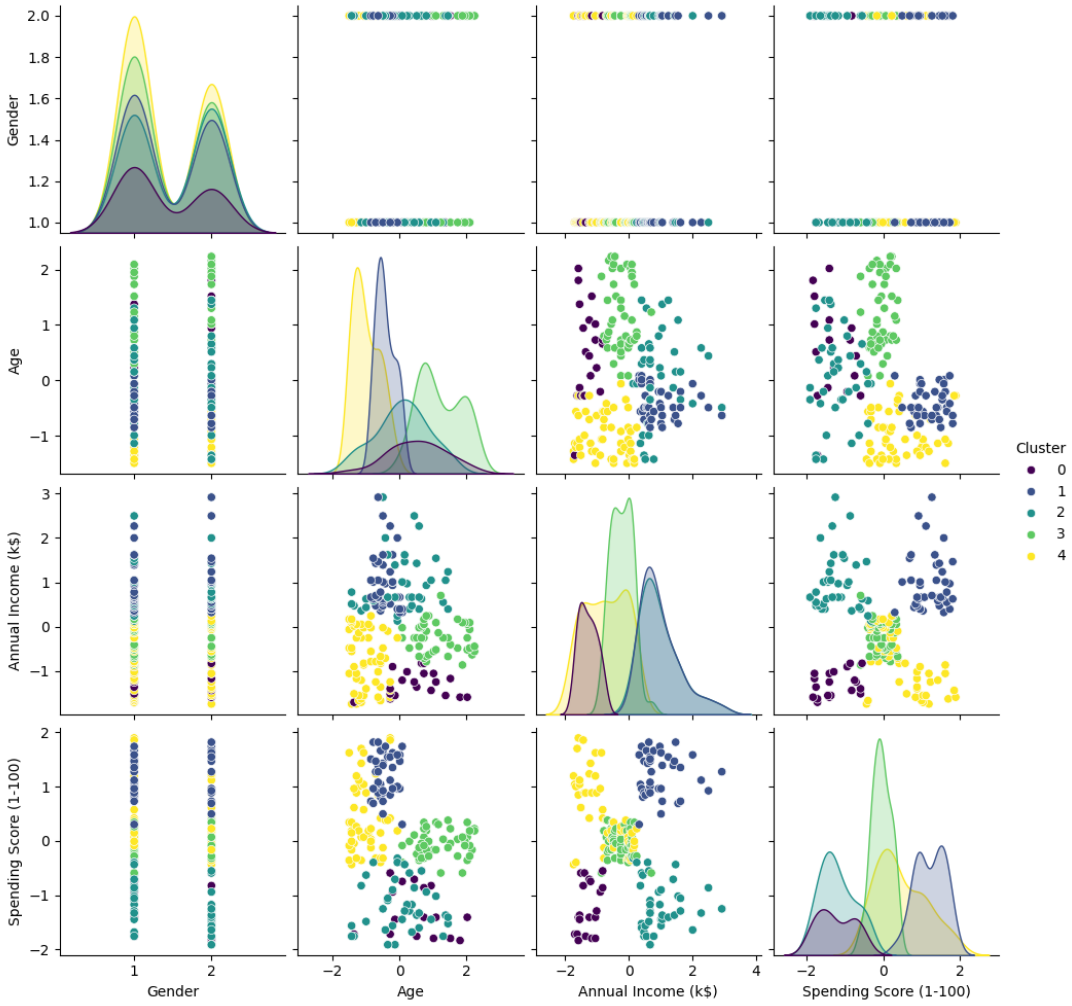

Figure 5: Biểu đồ phân cụm khách hàng với $K = 5$.

3.2.2 Phân tích kết quả

Kết quả phân cụm sau khi áp dụng thuật toán KMeans với 5 cụm cho thấy mối quan hệ giữa các yếu tố `Gender`, `Age`, `Annual Income (k$)`, và `Spending Score (1-100)`. Dưới đây là phân tích chi tiết về từng yếu tố ảnh hưởng đến các cụm:

- **Gender:** Biến `Gender` chỉ có hai giá trị (1 và 2), nên không có sự phân biệt rõ rệt giữa các cụm dựa trên giới tính. Các cụm phân bố khá đồng đều giữa hai giới tính.
- **Age:** Biến `Age` có sự phân tách rõ rệt giữa các cụm.
  - Cụm 0 (màu tím) chủ yếu tập trung vào độ tuổi lớn hơn.
  - Cụm 4 (màu vàng) đại diện cho nhóm khách hàng trẻ tuổi.
  - Cụm 2 (màu xanh lá) trải đều trên các độ tuổi, với sự tập trung vào nhóm tuổi trung bình.
- **Annual Income (Thu nhập hàng năm):** Cụm phân loại theo mức thu nhập có sự phân hóa rõ rệt.
  - Cụm 1 (màu xanh dương) chứa nhóm khách hàng có thu nhập cao, nhưng không có sự phân chia rõ ràng với các nhóm khác.
  - Cụm 3 (màu xanh lá đậm) có thu nhập trung bình, với sự phân bố đồng đều.
  - Cụm 4 (màu vàng) có thu nhập thấp hơn, nhưng lại chi tiêu cao.
- **Spending Score (Điểm chi tiêu):** Điểm chi tiêu là yếu tố quan trọng để phân biệt các cụm.
  - Cụm 4 (màu vàng) có điểm chi tiêu cao, nhưng thu nhập lại thấp.
  - Cụm 2 (màu xanh lá) có điểm chi tiêu trung bình và thu nhập cũng ở mức trung bình.

3.2.3 Kết luận

- Các kết quả phân cụm cho thấy mối quan hệ chặt chẽ giữa các yếu tố `Age`, `Annual Income`, và `Spending Score`.
- Cụm 4 (màu vàng) là nhóm có thu nhập thấp nhưng lại có Spending Score cao, cho thấy họ là những khách hàng chi tiêu nhiều.
- Cụm 0 (màu tím) có thu nhập cao và chi tiêu ít, cho thấy những khách hàng này có xu hướng tiết kiệm.
- Các nhóm khác phân bổ theo các mức độ khác nhau của thu nhập và chi tiêu.
- Thông tin này có thể giúp tối ưu hóa các chiến lược marketing, đặc biệt đối với những nhóm có Spending Score cao như Cụm 4.