ĐẠI HỌC QUỐC GIA TP.HCM
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
□¶□

LAB 05 – CANONICAL CORRELATION ANALYSIS
MÔN: PHÂN TÍCH THỐNG KÊ DỮ LIỆU NHIỀU BIẾN

Giáo viên Lý thuyết: Lý Quốc Ngọc
Giáo viên Thực hành: Nguyễn Mạnh Hùng
Sinh viên thực hiện: Lê Quang Vĩnh Quyền

TP. Hồ Chí Minh – Năm 2025

Mục lục

1. Bảng tự đánh giá

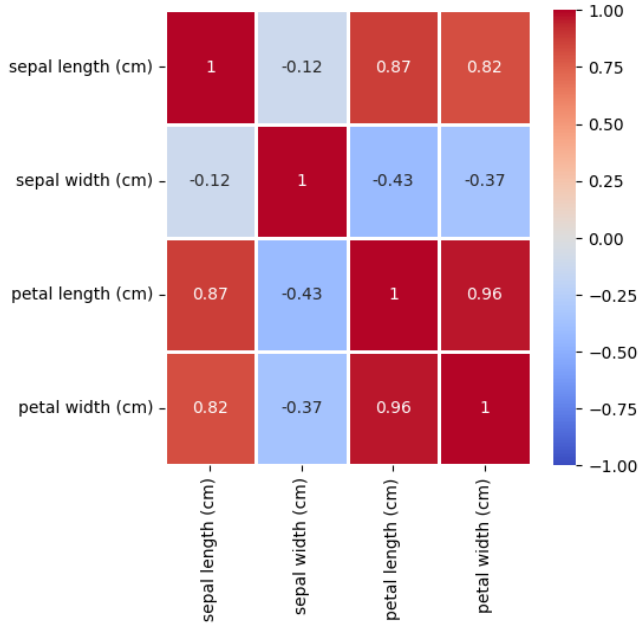
Nội dung	Mức độ hoàn thành
1. Cài đặt	
1.1. Yêu cầu 01	100%
1.2. Yêu cầu 02	100%
1.3. Yêu cầu 03	100%
1.4. Yêu cầu 04	100%
1.5. Yêu cầu 05	100%
2. Nghiên cứu	
2.1. Phân tích Tương quan Chính tắc (Canonical Correlation Analysis) là gì, và nó khác với phân tích tương quan truyền thống như thế nào?	100%
2.2. Giải thích khái niệm biến chính tắc (canonical variables) và ý nghĩa của chúng trong CCA.	100%
2.3. Các tập dữ liệu có cần cùng số chiều (dimensionality) để thực hiện CCA không?	100%

2. Cài đặt

2.1. Yêu cầu 01

Yêu cầu: Giải thích sự khác biệt về tương quan giữa chiều rộng đài hoa (sepal width) / chiều dài đài hoa (sepal length) với các đặc trưng liên quan đến cánh hoa (petal related value - tức là chiều dài và chiều rộng cánh hoa).

Thực hiện phân tích mối tương quan giữa các đặc trưng ta thu được kết quả như sau:



- **Tương quan giữa chiều dài đài hoa (sepal length) với các đặc trưng cánh hoa:**
    - Chiều dài đài hoa (sepal length) và chiều dài cánh hoa (petal length) có hệ số tương quan là 0.87. Đây là một mối tương quan dương mạnh, khi chiều dài đài hoa tăng, chiều dài cánh hoa có xu hướng tăng đáng kể theo, và ngược lại.

    - Chiều dài đài hoa (sepal length) và chiều rộng cánh hoa (petal width) có hệ số tương quan là 0.82. Đây cũng là một mối tương quan dương mạnh, khi chiều dài đài hoa tăng, chiều rộng cánh hoa cũng có xu hướng tăng mạnh.

  Chiều dài đài hoa có mối quan hệ tương quan dương rất mạnh với cả chiều dài và chiều rộng của cánh hoa. Cho thấy các loài hoa có đài hoa dài hơn thường cũng có cánh hoa lớn hơn cả về chiều dài và chiều rộng.
- **Tương quan giữa chiều rộng đài hoa (sepal width) với các đặc trưng cánh hoa:**
    - Chiều rộng đài hoa (sepal width) và chiều dài cánh hoa (petal length) có hệ số tương quan là -0.43. Đây là một mối tương quan âm trung bình, khi chiều rộng đài hoa tăng, chiều dài cánh hoa có xu hướng giảm nhẹ, và ngược lại.
    - Chiều rộng đài hoa (sepal width) và chiều rộng cánh hoa (petal width) có hệ số tương quan là -0.37. Đây là một mối tương quan âm nhẹ đến vừa, khi chiều rộng đài hoa tăng, chiều rộng cánh hoa cũng có xu hướng giảm nhẹ.

  Chiều rộng đài hoa có mối quan hệ tương quan âm với cả chiều dài và chiều rộng của cánh hoa. Cho thấy rằng các loài hoa có đài hoa rộng hơn một chút có thể có cánh hoa hơi nhỏ hơn.

2.2. Yêu cầu 02

Yêu cầu: Giải thích tầm quan trọng của việc chia tỷ lệ dữ liệu (scaling data).

Quá trình chia tỷ lệ được xem như là một bước tiền xử lý dữ liệu trước khi phân tích. Quá trình đó có vai trò rất quan trọng bởi:

- Giúp đảm bảo tính công bằng giữa các đặc trưng bởi các đặc trưng trong dữ liệu có thể có đơn vị và phạm vi giá trị khác nhau. Do đó việc chia tỷ lệ dữ liệu đảm bảo tất cả các đặc trưng có ảnh hưởng công bằng, không bị chi phối bởi các đặc trưng có giá trị lớn hơn.
- Giúp tăng hiệu suất, tăng cường độ chính xác. Khi các đặc trưng không được chia tỷ lệ, đặc trưng có phạm vi giá trị lớn hơn sẽ chiếm ưu thế trong việc xác định khoảng cách, dẫn đến kết quả kém chính xác hoặc không phù hợp. Ví dụ, một sự thay đổi nhỏ ở một đặc trưng có giá trị lớn có thể có tác động lớn hơn nhiều so với một thay đổi lớn ở một đặc trưng có giá trị nhỏ.
- Tăng tốc độ hội tụ của các thuật toán dựa trên gradient descent. Khi các đặc trưng không đồng nhất về tỷ lệ, hàm mất mát sẽ có hình dạng kéo dài và không đối xứng, làm chậm quá trình hội tụ của thuật toán và gây khó khăn trong việc tìm ra điểm cực tiểu. Chia tỷ lệ dữ liệu giúp bề mặt hàm mất mát trở nên đối xứng hơn, từ đó đẩy nhanh tốc độ hội tụ và nâng cao hiệu quả của quá trình tối ưu hóa.
- Tránh ảnh hưởng của các biến có độ lớn lớn hơn. Nếu dữ liệu không được chia tỷ lệ, các đặc trưng có phương sai lớn hơn sẽ chiếm ưu thế trong việc xác định các thành phần chính hoặc biến chính tắc, dù có thể chúng không phải là đặc trưng quan trọng nhất về mặt thông tin.

Trong chương trình được thực thi quá trình chia tỷ lệ nằm được thực hiện bằng cách sử dụng StandardScaler để chuẩn hóa tất cả các đặc trưng về cùng một thang đo (giá trị trung bình = 0 và độ lệch chuẩn = 1), đảm bảo mỗi đặc trưng có ảnh hưởng công bằng trong quá trình phân tích.

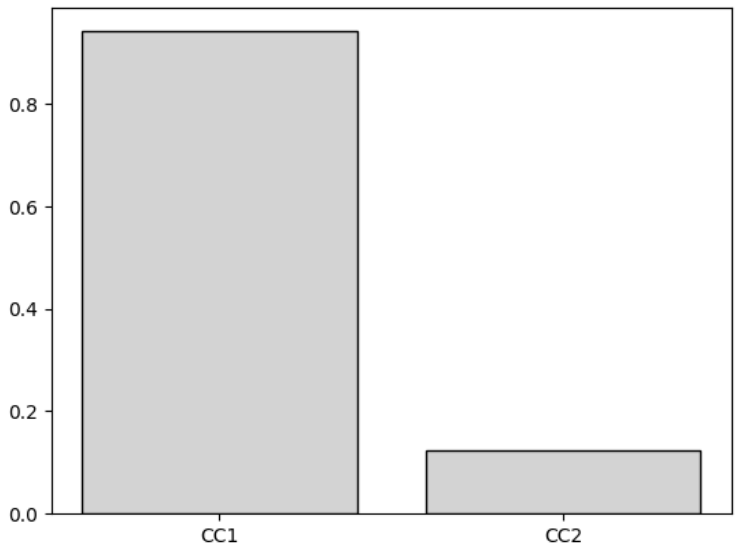
```
scaler = StandardScaler()
X1_sc = scaler.fit_transform(X1)
X2_sc = scaler.fit_transform(X2)
```

- X1_sc = scaler.fit_transform(X1): Thực hiện quá trình chia tỷ lệ cho tập dữ liệu X1 (đặc trưng đài hoa).
- X2_sc = scaler.fit_transform(X2): Thực hiện quá trình chia tỷ lệ cho tập dữ liệu X2 (đặc trưng cánh hoa).

2.3. Yêu cầu 03

Yêu cầu: So sánh cặp biến chính tắc đầu tiên (U1, V1) và cặp biến chính tắc thứ hai (U2, V2). Nên phân tích cặp nào?

Cặp biến chính tắc đầu tiên (U1, V1): Tương ứng với X1_c[:, 0] và X2_c[:, 0]. Trong biểu đồ, đây là CC1.
Cặp biến chính tắc thứ hai (U2, V2): Tương ứng với X1_c[:, 1] và X2_c[:, 1]. Trong biểu đồ, đây là CC2.



Dựa vào biểu đồ và giá trị comp_corr tính toán được ta thấy:

- Cặp CC1 (U1, V1) có giá trị tương quan của CC1 rất cao, khoảng 0.87. Điều này biểu thị một mối quan hệ tuyến tính rất mạnh mẽ và tích cực giữa tổ hợp tuyến tính đầu tiên của các đặc trưng đài hoa (U1) và tổ hợp tuyến tính đầu tiên của các đặc trưng cánh hoa (V1).
- Cặp CC2 (U2, V2) thì ngược lại, giá trị tương quan của CC2 rất thấp, khoảng 0.15. Điều này cho thấy mối quan hệ tuyến tính giữa tổ hợp tuyến tính thứ hai của các đặc trưng đài hoa (U2) và cánh hoa (V2) là rất yếu, gần như không có ý nghĩa.

Dựa trên kết quả đó thì ta nên tập trung phân tích cặp biến chính tắc đầu tiên (U1, V1) bởi nó có hệ số tương quan rất cao là 0.87, biểu thị mối liên hệ tuyến tính mạnh mẽ và có ý nghĩa nhất giữa các đặc trưng đài hoa và cánh hoa. Còn cặp biến chính tắc thứ hai (U2, V2) có hệ

số tương quan rất thấp là 0.15, cho thấy mối quan hệ rất yếu và có thể không có ý nghĩa thực tiễn hay thống kê. Vậy nên cặp này thường không được phân tích chuyên sâu.

2.4. Yêu cầu 04

Yêu cầu: Rút ra kết luận dựa trên bảng hệ số tải (loadings table).

Bảng Hệ số tải cho đặc trưng đài hoa (X1)

	CC1/U1	CC2/U2
sepal length (cm)	0.892246	0.388008
sepal width (cm)	-0.457866	0.921656

Dựa vào bảng Hệ số tải cho đặc trưng đài hoa ta thấy:

- Đối với Biến chính tắc 1 (CC1/U1): Chiều dài đài hoa (sepal length) có hệ số tải dương rất cao là 0.892, cho thấy nó đóng góp mạnh mẽ và cùng chiều vào U1. Ngược lại, chiều rộng đài hoa (sepal width) có hệ số tải âm trung bình là -0.458, cho thấy đóng góp ngược chiều. Như vậy, U1 chủ yếu đại diện cho sự kết hợp của chiều dài đài hoa lớn và chiều rộng đài hoa tương đối hẹp.
- Đối với Biến chính tắc 2 (CC2/U2): Chiều rộng đài hoa (sepal width) có hệ số tải dương rất cao là 0.922, là yếu tố đóng góp chính. Điều này có nghĩa là U2 chủ yếu phản ánh chiều rộng đài hoa lớn.

Bảng Hệ số tải cho đặc trưng cánh hoa (X2)

	CC1/V1	CC2/V2
petal length (cm)	1.573225	0.332706
petal width (cm)	1.453533	0.943031

Dựa vào bảng Hệ số tải cho đặc trưng cánh hoa ta thấy:

- Đối với Biến chính tắc 1 (CC1/V1): Cả chiều dài cánh hoa (petal length) và chiều rộng cánh hoa (petal width) đều có hệ số tải dương cực kỳ cao lần lượt là 1.573 và 1.454. Điều này cho thấy V1 được định nghĩa rất mạnh bởi cả hai đặc trưng kích thước của cánh hoa.
- Đối với Biến chính tắc 2 (CC2/V2): Chiều rộng cánh hoa (petal width) có hệ số tải dương rất cao là 0.943 và đóng góp chính. Điều này có nghĩa là V2 chủ yếu phản ánh chiều rộng cánh hoa lớn.

2.5. Yêu cầu 05

Yêu cầu: So sánh biểu đồ nhiệt (heatmap) ở bước 2 với các hệ số CCA này

Bảng Hệ số CCA

	petal length (cm)	petal width (cm)
sepal length (cm)	1.60	-0.31
sepal width (cm)	1.77	0.28

Dựa vào bảng Hệ số CCA ta thấy rằng:

- petal length có mối liên hệ dương rất mạnh với cả sepal length là 1.60 và đặc biệt là sepal width là 1.77. Cho thấy chiều dài cánh hoa là đặc trưng cánh hoa chính được định hình bởi các đặc trưng của đài hoa trong mối quan hệ chính tắc.
- petal width thể hiện mối liên hệ yếu với các đặc trưng đài hoa, âm với sepal length là -0.31 và dương với sepal width là 0.28.

Bảng Hệ số CCA so với biểu đồ biểu diễn mối tương quan giữa các đặc trưng thì:

- Mối quan hệ giữa sepal length và petal length thể hiện sự nhất quán cao. Cả tương quan ban đầu là 0.87 và hệ số CCA là 1.60 đều cho thấy một mối liên hệ mạnh mẽ. Điều này chứng tỏ chiều dài của đài hoa và cánh hoa là những đặc trưng gắn kết với nhau.
- Tuy nhiên ở mối quan hệ giữa sepal width và petal length. Biểu đồ biểu diễn tương quan ban đầu chỉ ra một tương quan âm trung bình là -0.43, còn ở bảng Hệ số CCA lại cho thấy một hệ số dương cực kỳ mạnh là 1.77. Bởi ở biểu đồ tương quan đơn giản chỉ xem xét mối quan hệ giữa hai biến riêng lẻ. Trong khi hệ số CCA lại phản ánh trọng số biến đổi của sepal width khi nó kết hợp với các đặc trưng khác của đài hoa để tạo thành biến chính tắc, nhằm tối đa hóa tương quan với các đặc trưng của cánh hoa. Do đó trong một cấu trúc đa biến phức tạp, sepal width đóng một vai trò tích cực mạnh mẽ trong việc xác định petal length khi CCA tìm kiếm mối liên hệ mạnh nhất giữa hai tập hợp biến.

3. Nghiên cứu

3.1. Phân tích Tương quan Chính tắc (Canonical Correlation Analysis) là gì, và nó khác với phân tích tương quan truyền thống như thế nào?

Phân tích Tương quan Chính tắc (CCA) là được sử dụng để định lượng mối quan hệ tuyến tính giữa hai tập hợp các biến liên tục. Mục tiêu của CCA là tạo ra các tổ hợp tuyến tính mới của các biến trong mỗi tập hợp gọi là biến chính tắc sao cho mối tương quan giữa các cặp biến chính tắc này là lớn nhất. Bằng cách này, CCA giúp xác định mức độ và cách thức mà các biến trong một tập hợp liên quan đến các biến trong tập hợp kia, đồng thời làm rõ cấu trúc tiềm ẩn của mối quan hệ đa biến.

Sự khác nhau giữa CCA và phân tích tương quan truyền thống:

Tiêu chí	Phân tích Tương quan Truyền thống	Phân tích Tương quan Chính tắc (CCA)
Mục tiêu	Đo lường mối quan hệ tuyến tính giữa hai biến đơn lẻ.	Đo lường mối quan hệ tuyến tính giữa hai tập hợp các biến (mỗi tập hợp có nhiều biến).
Đầu vào	Hai biến định lượng (ví dụ: Chiều dài đài hoa và Chiều dài cánh hoa).	Hai tập hợp biến định lượng (ví dụ: Tập các biến đài hoa {chiều dài, chiều rộng} và Tập các biến cánh hoa {chiều dài, chiều rộng}).
Đầu ra	Một hệ số tương quan	Nhiều cặp biến chính tắc và các hệ số tương quan chính tắc tương ứng. Ngoài ra còn có các hệ số tải và hệ số biến đổi.
Khả năng phân tích	Chỉ có thể phân tích mối quan hệ giữa từng cặp biến một cách độc lập. Không tính đến mối quan hệ nội tại giữa các biến trong cùng một tập hợp.	Có khả năng tìm ra tổ hợp tuyến tính tối ưu của các biến trong mỗi tập hợp để tối đa hóa tương quan giữa chúng. Có thể phát hiện các mối quan hệ đa chiều và phức tạp mà tương quan đơn lẻ không thấy được.
Ứng dụng	Đánh giá mối liên hệ trực tiếp, đơn giản giữa hai yếu tố.	Thường được sử dụng để hiểu cấu trúc phức tạp, chẳng hạn như mối quan hệ giữa các bộ đo lường tâm lý và sinh lý.

3.2. Giải thích khái niệm biến chính tắc (canonical variables) và ý nghĩa của chúng trong CCA.

Trong CCA, biến chính tắc (canonical variates) là các biến mới, không thể quan sát trực tiếp, được tạo ra dưới dạng các tổ hợp tuyến tính của các biến gốc trong mỗi tập hợp.

Ta có một tập hợp các biến $X = \{X_1, X_2, \dots, X_p\}$ và một tập các biến $Y = \{Y_1, Y_2, \dots, Y_q\}$, CCA sẽ tìm các biến chính tắc như sau:

- Biến chính tắc của tập X, kí hiệu là U: $U = a_{k1}.X_1 + a_{k2}.X_2 + \cdots + a_{kp}.X_p$
- Biến chính tắc của tập Y, kí hiệu là V: $V = b_{k1}.Y_1 + b_{k2}.Y_2 + \cdots + b_{kq}.Y_q$

Trong đó, U_k và V_k là cặp biến chính tắc thứ k. a_{ki} và b_{kj} là các hệ số biến đổi.

Ý nghĩa của biến chính tắc trong CCA:

- Xác định mối tương quan giữa các tập hợp biến thông qua các cặp biến chính tắc (U, V) đại diện cho những mối liên hệ tuyến tính mạnh mẽ nhất giữa hai tập hợp biến.
- Giảm chiều dữ liệu: Các biến chính tắc sẽ tóm gọn thông tin từ nhiều biến gốc thành một số lượng nhỏ hơn các biến mới có ý nghĩa. Nếu chỉ một vài cặp biến chính tắc đầu tiên có tương quan đáng kể, điều đó có nghĩa là phần lớn mối quan hệ giữa hai tập hợp biến có thể được giải thích bởi những chiều kích này.
- Diễn giải cấu trúc tiềm ẩn thông qua các hệ số tải bằng cách xem xét các hệ số tải lớn, chúng ta có thể hiểu rõ hơn về những đặc trưng nào của các biến gốc đang định hình các biến chính tắc và do đó, định hình mối quan hệ giữa hai tập hợp.
- Phát hiện mối quan hệ đa biến phức tạp giữa các nhóm biến mà phân tích đơn lẻ không thể thấy được.

3.3. Các tập dữ liệu có cần cùng số chiều (dimensionality) để thực hiện CCA không?

Trong CCA, các tập dữ liệu không nhất thiết phải có cùng số chiều (dimensionality), miễn là phải có cùng số mẫu. Bởi vì số lượng cặp biến chính tắc tối đa mà CCA có thể trích xuất được sẽ bằng số lượng đặc trưng tối thiểu của một trong hai tập hợp biến. Ví dụ, nếu một tập hợp biến có 5 đặc trưng và tập hợp còn lại có 3 đặc trưng, CCA sẽ tìm được tối đa 3 cặp biến chính tắc. Điều này đảm bảo rằng CCA có thể được áp dụng hiệu quả ngay cả khi các nhóm biến được phân tích có kích thước không đồng đều.

4. Tài liệu tham khảo

- https://www.sciencedirect.com/topics/mathematics/canonical-correlation-analysis#:~:text=Canonical%20correlation%20analysis%20(CCA)%20is,the%20correlation%20technique%20%5B43%5D.
- https://www.geeksforgeeks.org/what-is-canonical-correlation-analysis/
- https://en.wikipedia.org/wiki/Canonical_correlation
- https://stats.oarc.ucla.edu/stata/dae/canonical-correlation-analysis/