BÀI 1 - PHÂN TÍCH THÀNH PHẦN CHÍNH (Classical PCA)

Hạn nộp bài: hết ngày 25/03/2024.

1 Bài tập lý thuyết

Bài 1. Cho dữ liệu có ma trận hiệp phương sai $S = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$.

a. Xác định các thành phần chính.

b. Tính tỷ trọng phương sai mà mỗi thành phần chính đại diện. Thành phần chính nào là quan trọng hơn?

Bài 2. Cho dữ liệu có ma trận hiệp phương sai $S = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$, trong đó $-1 \leq \rho \leq 1$.

a. Xác định các thành phần chính.

b. Tính tỷ trọng phương sai mà mỗi thành phần chính đại diện. Thành phần chính nào là quan trọng hơn?

2 Bài tập thực hành

Bài 3. Tập dữ liệu Iris (trong `sklearn.datasets.load_iris`) chứa 3 lớp, trong đó mỗi lớp có 50 quan trắc về các bông hoa diên vĩ.

a. Dùng PCA để giảm chiều dữ liệu.

b. Trực quan hoá dữ liệu đã giảm chiều với màu là loài hoa. Nhận xét.

Bài 4. Tập tin turtle chứa dữ liệu về kích thước mai và giới tính của 48 cá thể rùa Chrysemys picta.

a. Trực quan hoá các biến kích thước mai (riêng lẻ và theo cặp) với màu là `Gender`. Nhận xét xem có sự tách biệt giữa hai giới tính không.

b. Dùng PCA để giảm chiều dữ liệu kích thước mai rùa.

c. Trực quan hóa dữ liệu đã giảm chiều với màu là `Gender`. Nhận xét và so sánh với kết quả câu a.

Bài 5. Tập dữ liệu stock chứa thông tin về tỷ suất lợi nhuận hàng tuần của năm cổ phiếu JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell và ExxonMobil niêm yết trên Sở giao dịch chứng khoán New York trong giai đoạn từ tháng 1 năm 2004 đến tháng 12 năm 2005. Hãy dùng PCA để giảm chiều dữ liệu và nhận xét.

Bài 6. Dùng PCA để nén ảnh `flower` trong `sklearn.datasets.load_sample_image`.

a. Tách ảnh gốc thành các ảnh đơn sắc xanh dương, xanh lá và đỏ. Trực quan hoá.

b. Dùng PCA để giảm chiều các ảnh đơn sắc còn 50 chiều. Nhận xét về tỷ lệ phương sai giải thích.

c. Trả dữ liệu về số chiều gốc. Trực quan hoá và nhận xét.

Bài 7. Dùng PCA để nén một ảnh tự chọn theo các bước tương tự bài trên.