Practical Session 1: Dimensionality Reduction Course

Principal Components Analysis

Give your answer handwritten or typed in LaTeX; save it in pdf or jpeg files. Your code must be written in an ipynb file. Compress all your files in a zip file named *YourName-PS1.zip* and send to email: *le.tntran1107@gmail.com* before the deadline **23h59 Tuesday, April 4th, 2023**.

1 Method

Step 1 Mean normalization.

Step 2 Compute Covariance matrix.

Step 3 Compute eigenvectors/values.

Step 4 Select top $k-$ eigenvalues and their eigenvectors.

Step 5 Create an orthogonal base with the eigenvectors.

Step 6 Transform data by multiplying with said base.

2 Example

Applying the PCA method to find the transformed data for the following dataset:
$D = \{(126, 78), (128, 80), (128, 82), (130, 82), (130, 84), (132, 86)\}$.

Solution

We call $A = \begin{pmatrix} 126 & 78 \\ 128 & 80 \\ 128 & 82 \\ 130 & 82 \\ 130 & 84 \\ 132 & 86 \end{pmatrix}$ be a matrix that each row is a data point in D.

Step 1 Mean normalization.
$\overline{x} = 129, \overline{y} = 82$
Subtract each data point for mean, we obtain matrix $\overline{A}$

$$\overline{A} = \begin{pmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{pmatrix}$$

Step 2 Compute Covariance matrix.

$$S = Cov(\overline{A}) = \begin{pmatrix} 4.4 & 5.6 \\ 5.6 & 8 \end{pmatrix}$$

Step 3 Compute eigenvectors/values.
$det(S - \lambda \mathbb{I}) = 0$

$$\Leftrightarrow \begin{vmatrix} 4.4 - \lambda & 5.6 \\ 5.6 & 8 - \lambda \end{vmatrix} = 0$$

$\Leftrightarrow (4.4-\lambda)(8-\lambda)-5.6^2=0$
$\Leftrightarrow \lambda_1=12.08, \lambda_2=0.32$

Step 4 Select top $k-$ eigenvalues and their eigenvectors.
$eig_1=(x_1,y_1), eig_2=(x_2,y_2)$
eigenvectors must be unit vectors.
• With $\lambda_1=12.08$
$Sv_1=\lambda_1 v_1$
$\Leftrightarrow \begin{pmatrix}4.4 & 5.6\\ 5.6 & 8\end{pmatrix}\begin{pmatrix}v_{11}\\ v_{12}\end{pmatrix}=12.08\begin{pmatrix}v_{11}\\ v_{12}\end{pmatrix}$

$\Leftrightarrow \begin{cases}4.4v_{11}+5.6v_{12}=12.08v_{11}\\ 5.6v_{11}+8v_{12}=12.08v_{12}\end{cases}$
$\Longrightarrow v_{12}=1.37v_{11}$
By chosing $v_{11}=1$, we have $v_1=\begin{pmatrix}1\\ 1.37\end{pmatrix}$
v_1 must be a unit vector $\Longrightarrow v_1=\begin{pmatrix}0.59\\ 0.81\end{pmatrix}$
• With $\lambda_2=0.32$ We found $v_2=\begin{pmatrix}-0.81\\ 0.59\end{pmatrix}$

Step 5 Create an orthogonal base with the eigenvectors.
$V=(eig_1, eig_2)$
$V=\begin{pmatrix}0.59 & -0.81\\ 0.81 & 0.59\end{pmatrix}$

Step 6 Transform data by multiplying with said base.
Transformed data $=\overline{A}*V$
Transformed data $=\begin{pmatrix}-3 & -4\\ -1 & -2\\ -1 & 0\\ 1 & 0\\ 1 & 2\\ 3 & 4\end{pmatrix}\begin{pmatrix}0.59 & -0.81\\ 0.81 & 0.59\end{pmatrix}=\begin{pmatrix}-5.01 & 0.07\\ -2.21 & -0.37\\ -0.59 & 0.81\\ 0.59 & -0.81\\ 2.21 & 0.37\\ 5.01 & -0.07\end{pmatrix}$

3 Exercises

3.1 Maximum variance formulation

Consider a data set of observations $\{x_n\}$ where $n=1,...,N$ and x_n is a Euclidean variable with dimensionality D. Our goal is to project the data onto a space having dimensionality $M \leqslant D$ while maximizing the variance of the projected data.
Give detailed proof for PCA using maximum variance formulation with $M=2$. Address the following questions:

a) State the optimization problem to maximum variance, indicating the form of the Covariance matrix.

b) Using the Lagrange function to solve the problem - finding the formula of v_1, v_2 (two principal components).

Note: Give the reason why we need to maximize $v_1^T S v_1$ (what is the exact variance, do some transformation to get $v_1^T S v_1$). With $M=1$, we already have the proof in Bishop, so what additional constraints do we need for the second optimization problem?

3.2 Write your calculations

Applying the PCA method to find the transformed data for the following dataset. Check and compare the result to Python code and scikit-learn libraries, visualize the transformed data and give your comments.
$D=\{(1,0),(2,0),(3,0),(5,6),(6,6),(7,6)\}$.

3.3 Cực đại hóa phương sai

Xét một tập dữ liệu gồm các quan sát $\{x_n\}$ trong đó $n = 1, ..., N$ và x_n là một biến Euclide có số chiều D. Mục tiêu là chiếu dữ liệu lên một không gian có số chiều $M \leqslant D$ bằng cực đại hóa phương sai của dữ liệu được chiếu.

Đưa ra một chứng minh chi tiết của PCA bằng phương pháp cực đại hóa phương sai với số chiều $M = 2$. Trả lời các câu hỏi sau để làm rõ chứng minh:

a) Chỉ ra bài toán tối ưu cụ thể, ghi rõ công thức của ma trận Hiệp phương sai.

b) Sử dụng hàm Lagrange để giải bài toán tối ưu đã nêu - ghi rõ biểu diễn của v_1, v_2 (hai thành phần chính).

Note: Chỉ ra lý do bằng cách thực hiện một số biến đổi để dẫn đến cần cực đại hóa $v_1^T S v_1$. Với số chiều bằng 1, chứng minh đã có trong sách Bishop, khi giải bài toán tối ưu cho v_2 cần thêm điều kiện gì?

3.4 Tính toán

Sử dụng phương pháp PCA để tìm transformed data cho tập dữ liệu sau $D = \{(1, 0), (2, 0), (3, 0), (5, 6), (6, 6), (7, 6)\}$. Kiểm tra và so sánh kết quả đã tính bằng tay với kết quả chạy bằng code Python (hàm do các bạn viết), kết quả từ PCA của Scikit-learn trả về. Visualize raw data, centered data và transformed data và viết một số nhận xét (ghi nhận xét trong file ipynb).