Principal Component Analysis

Lưu Trung Tín

Ngày 13 tháng 9 năm 2022

Ma trận hiệp phương sai - Ma trận tương quan

Phân tích thành phần chính

Giới thiệu

- *Phân tích thành phần chính* (**PCA**) là một kỹ thuật trong Thống kê, được thiết kế để tóm tắt các thuộc tính quan trọng nhất trong bộ dữ liệu.
- Mục tiêu chung của **PCA** là giảm chiều dữ liệu. Đôi khi nó còn giúp nhìn ra các mối quan hệ mà dữ liệu gốc không thể hiện được.
- **PCA** có thể làm giảm sự dư thừa trong tập dữ liệu. Cơ bản nhất, sự dư thừa xảy ra khi các biến có tương quan với nhau.

Trung bình cộng - Phương sai

Xét thuộc tính $\mathbf{x} \in \mathbb{R}^{m \times 1}$. Ta có

- **Trung bình cộng** (*mean*)

$$\overline{x} = \frac{1}{m} \sum_{i=1}^{m} x_i.$$

- **Phương sai** (*variance*) là trung bình của bình phương khoảng cách từ các điểm dữ liệu đến kỳ vọng.

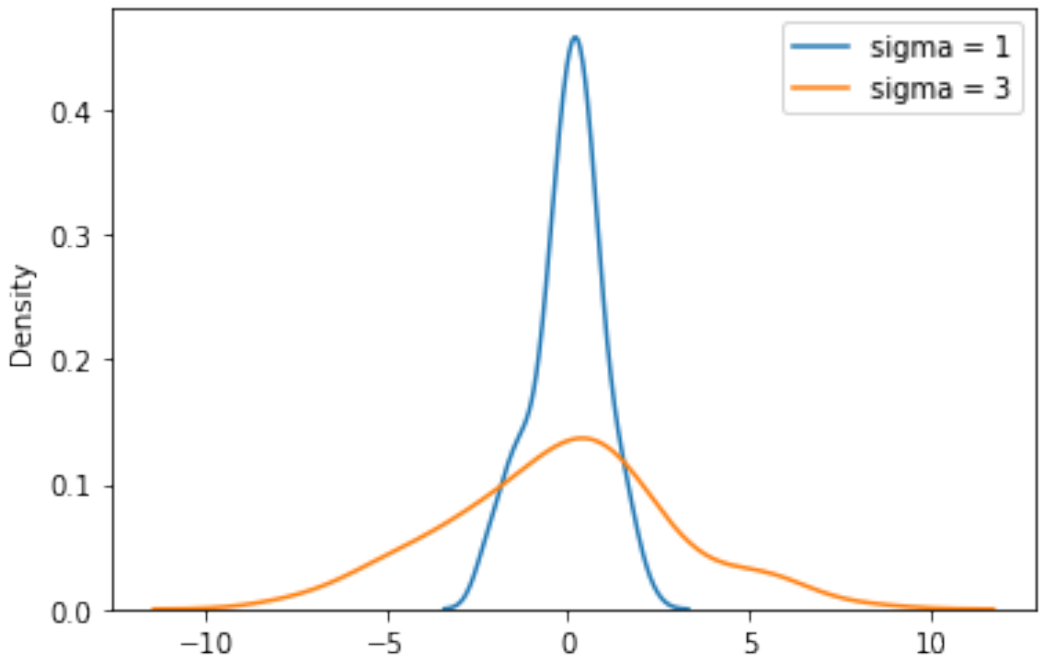
$$\sigma^2 = Var[\mathbf{x}] = \frac{1}{m} \sum_{i=1}^{m} (x_i - \overline{x})^2.$$

Độ lệch chuẩn: σ.

Trung bình cộng - Phương sai

Phương sai là đại lượng để đo lường sự phân tán của dữ liệu xung quanh giá trị trung bình.

- Phương sai nhỏ: các điểm dữ liệu tập trung gần trung bình.
- Phương sai lớn: các điểm dữ liệu phân tán rộng.



Vectơ trung bình - Ma trận hiệp phương sai

Xét bộ dữ liệu $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix} \in \mathbb{R}^{m \times n}$. Ta có

- **Vectơ trung bình** (*mean vector*) là vectơ các giá trị trung bình của mỗi cột

$$\boldsymbol{\mu} = \begin{bmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_n \end{bmatrix}.$$

- **Ma trận hiệp phương sai** (*covariance matrix*)

$$\mathbf{S} = cov(\mathbf{X}) = [\sigma_{ij}]_{n \times n},$$

trong đó

$$\sigma_{ij} = cov(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{n}(\mathbf{x}_i - \overline{x}_i)^T(\mathbf{x}_j - \overline{x}_j).$$

Vectơ trung bình - Ma trận hiệp phương sai

Một số tính chất của ma trận hiệp phương sai:

- $\mathbf{S}$ là ma trận đối xứng, nửa xác định dương.
- $\sigma_{ij} = cov(\mathbf{x}_i, \mathbf{x}_j)$ là hiệp phương sai của $\mathbf{x}_i$ và $\mathbf{x}_j$ $(i \neq j)$, dùng để đo lường sự biến thiên đồng thời của hai biến.
- $\sigma_{ii} = cov(\mathbf{x}_i, \mathbf{x}_i) = Var[\mathbf{x}_i] \geq 0$ là phương sai của $\mathbf{x}_i$, nằm trên đường chéo của $\mathbf{S}$.

Nếu $\mathbf{S}$ là ma trận đường chéo thì các cặp $\mathbf{x}_i$, $\mathbf{x}_j$ hoàn toàn không tương quan với nhau.

Ma trận tương quan

- **Ma trận tương quan** (*correlation matrix*)

$$\mathbf{R} = corr(\mathbf{X}) = [r_{ij}]_{n\times n},$$

trong đó hệ số tương quan

$$r_{ij} = corr(\mathbf{x}_i, \mathbf{x}_j) = \frac{cov(\mathbf{x}_i, \mathbf{x}_j)}{\sigma_{\mathbf{x}_i}\sigma_{\mathbf{x}_j}}$$

đo lường sự tương quan tuyến tính của $\mathbf{x}_i$ và $\mathbf{x}_j$.

Hệ số tương quan r_{ij} là chuẩn hoá của hiệp phương sai σ_{ij}.
Ta có $-1 \leq r_{ij} \leq 1$, $r_{ii} = 1$.

Phân tích thành phần chính (PCA)

- **PCA** tính toán ra các thành phần chính $PC_1, PC_2, \ldots, PC_n$. Chúng chứa đầy đủ thông tin như các thuộc tính ban đầu, nhưng được thể hiện một cách thuận tiện hơn.
- Mục đích của **PCA** là tìm ra k thành phần chính giải thích được hầu hết thông tin trong bộ dữ liệu ($1 \leq k < n$).

Quy trình PCA

Các bước thực hiện **PCA**:

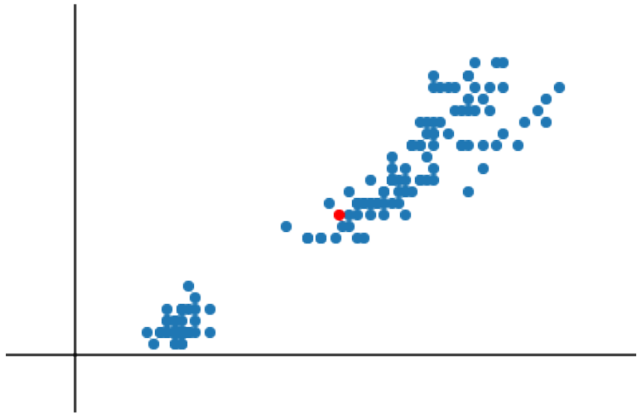
- **B1:** Tính $\hat{\mathbf{X}} = \mathbf{X} - \mu$.
- **B2:** Tìm các cặp trị riêng - vectơ riêng $(\lambda_i, \mathbf{e}_i)$ $(i = 1, \ldots, n)$ của $\mathbf{S} = cov(\mathbf{X}) = cov(\hat{\mathbf{X}})$.
  Sắp xếp chúng sao cho các trị riêng giảm dần: $\lambda_1 \geq \cdots \geq \lambda_n$.
- **B3:** Chọn k cặp có trị riêng lớn nhất. Đặt

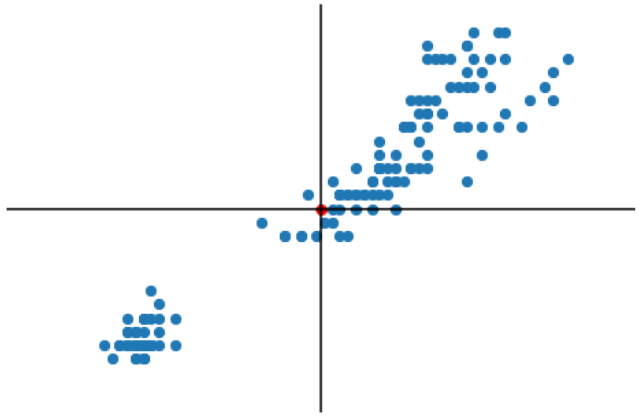
$$\mathbf{B} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \cdots & \mathbf{e}_k \end{bmatrix}.$$

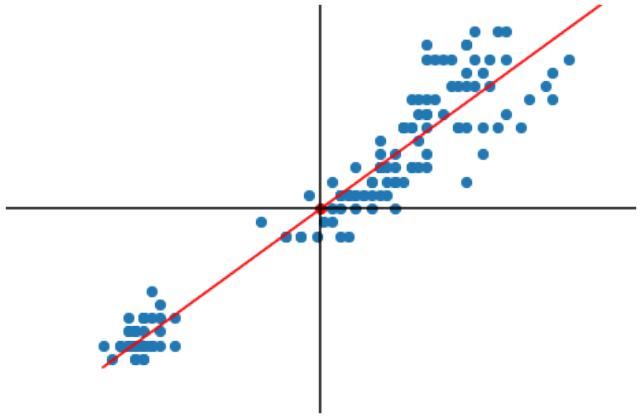
- **B4:** Tính các thành phần chính

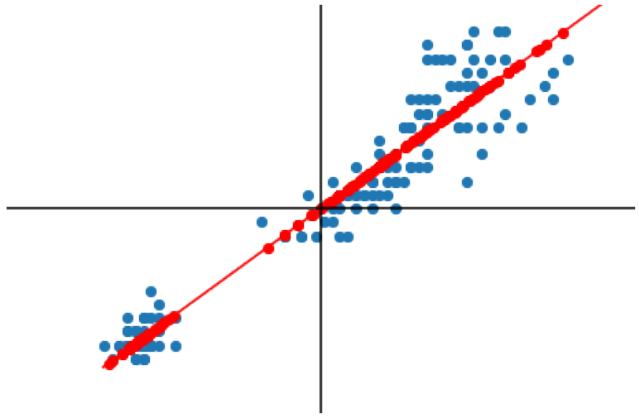
$$\mathbf{PC} = \hat{\mathbf{X}} \cdot \mathbf{B}.$$

Quy trình PCA









Chọn k

- **Chọn k dựa vào tỷ lệ phương sai tích luỹ:**

Tỷ lệ phương sai được giải thích bởi thành phần chính thứ k là

$$\frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}.$$

Tỷ lệ phương sai được giải thích bởi k thành phần chính đầu tiên là

$$\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_n}.$$

Thông thường, ta mong muốn tỷ lệ này trong khoảng $70\% - 80\%$.

Chọn k

- **Chọn k dựa vào phương pháp khuỷu tay (elbow method):**