B3. PCA

Bổ sung thêm cho giáo trình

2021

Nội dung bổ sung

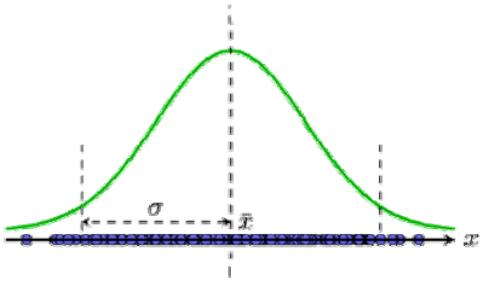
1. **Ma trận hiệp phương sai**
2. **Principal Component Analysis**

1. Ma trận hiệp phương sai

❑ **Kỳ vọng** (*expectation*)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

trung bình cộng (*mean*)



❑ **Phương sai** (*variance*) **và độ lệch chuẩn** (*standard deviation*)

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

trung bình khoảng cách đến kỳ vọng

- σ: độ lệch chuẩn
- phương sai càng NHỎ thì các điểm dữ liệu càng gần kỳ vọng
- phương sai càng LỚN thì các điểm dữ liệu càng phân tán

1. Ma trận hiệp phương sai (tt.)

❑ **Vectơ cột $x_1, x_2, \ldots, x_n \in R^m$, ma trận $X = (x_1\ x_2 \ldots x_n) \in R_{m,n}$**

- ma trận trung tâm (*center matrix* ≠ *centering matrix*) $\hat{X} \in R_{m,n}$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

có thể tính trung bình trên mỗi cột

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \Rightarrow \hat{x}_{ij} = (x_{ij} - \bar{x}_j)$$

$$\hat{x}_{ij} = (x_{ij} - \bar{x})$$

$$\hat{X} = (X - \bar{x}) = \begin{pmatrix}(x_1 - \bar{x}) & (x_2 - \bar{x}) & \cdots & (x_n - \bar{x})\end{pmatrix}$$

- ma trận hiệp phương sai (*covariance matrix*) **của X**

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^T.(x_i - \bar{x}) = \frac{1}{n}\hat{X}^T.\hat{X} \equiv V$$

1. Ma trận hiệp phương sai (tt.)

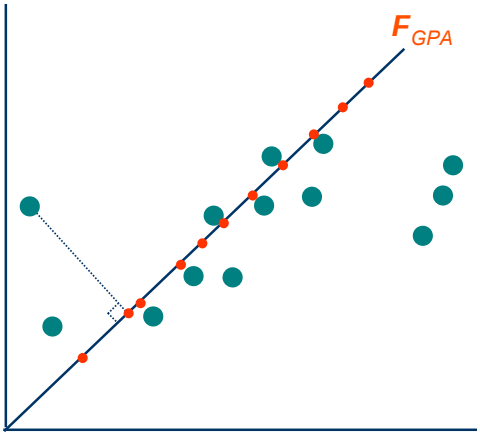
- **Một số tính chất của ma trận hiệp phương sai V**
  - ma trận đối xứng
  - ma trận nửa xác định dương
  - hệ số không âm trên đường chéo: phương sai trên từng chiều
  - hiệp phương sai v_{ij} ($i \neq j$): mối tương quan giữa x_i và x_j
  - nếu V là ma trận đường chéo $\Rightarrow$ hoàn toàn không tương quan

Nội dung bổ sung

1. Ma trận hiệp phương sai
2. **Principal Component Analysis**

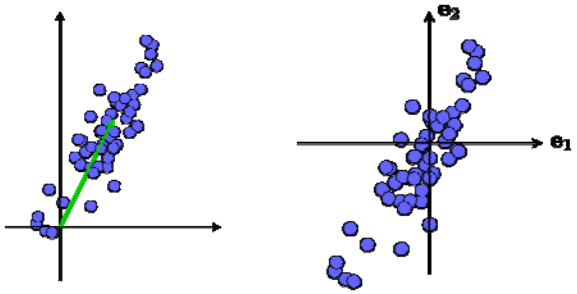
2. Principal Component Analysis (tt.)

- Tìm <u>không gian đặc trưng</u> mới ***F'*** tạo phân hoạch trên items tốt hơn không gian đặc trưng ban đầu ***F***



2. Principal Component Analysis

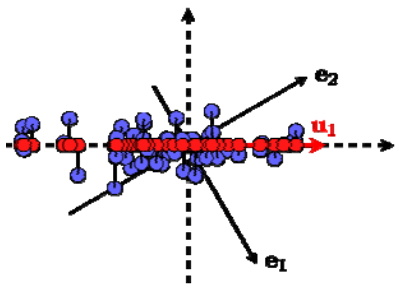
B1. Tính $\bar{x}$ → **B2.** Tạo ma trận $\hat{X}$ → **B3.** Tạo ma trận hiệp phương sai V

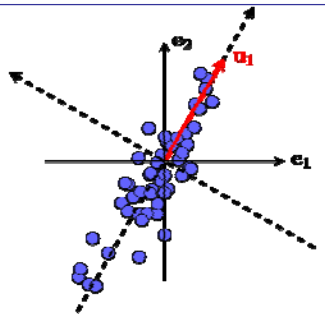


B4. Tính eigenvalues $\lambda_1, \ldots, \lambda_n$ và eigenvectors $u_1, \ldots, u_n$ của V

B5. Chọn k eigenvectors $u_1, \ldots, u_k$ có eigenvalues lớn nhất

B6. Thực hiện phép chiếu $x_1, \ldots, x_n$ và xuống cơ sở $u_1, \ldots, u_n$





2. Principal Component Analysis (tt.)

- **Hệ cơ sở – Tọa độ trong không gian vectơ V**

Cơ sở "có thứ tự" B gồm các vectơ độc lập tuyến tính:

$B = \{ u_1, u_2, \ldots, u_n \}$

$\forall v \in V$: $v = \alpha_1 u_1 + \alpha_2 u_2 + \ldots + \alpha_n u_n \qquad \alpha_i \in \}$

Ma trận cơ sở của không gian V: $B = \begin{pmatrix} u_1^T & u_2^T & \cdots & u_n^T \end{pmatrix}$

Tọa độ của v theo B:

$$[v]_B = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix}$$

2. Principal Component Analysis (tt.)

- **Hệ cơ sở – Tọa độ trong không gian vectơ V**

Cơ sở "có thứ tự" B' gồm các vectơ độc lập tuyến tính:

$B = \{ u'_1, u'_2, \ldots, u'_n \}$

Ma trận chuyển đổi cơ sở từ B sang B':

$$(B \to B') = \begin{pmatrix} [u'_1]_B & [u'_2]_B & \cdots & [u'_n]_B \end{pmatrix}$$

$(B \to B')$ khả nghịch

Công thức chuyển đổi tọa độ:

$$[v]_{B'} = (B \to B')^{-1} [v]_B$$

$$[v]_B = (B \to B')[v]_{B'}$$

Tài liệu tham khảo

Vũ Hữu Tiệp, *Machine Learning cơ bản*, 2018