2019-01-31

주간보고

Man Machine Interface Lab.

This Week

- 책의 목차

3. 비지도 학습과 데이터 전처리

3.4 차원 축소

- PCA
- NMF
- t-SNE

3.5 군집

4. 데이터 표현과 특성 공학

4.5 특성 자동 선택

- ANOVA

This Week

- 발표 목차

1. Covariance
2. Dimension reduction
    1) Feature Extraction
        a. PCA
    2) Feature Selection
        a. SVM
        b. T-test, ANOVA

김양곤

Dimension reduction
= Feature

Feature Extraction — PCA

E_1 E_2 … E_m

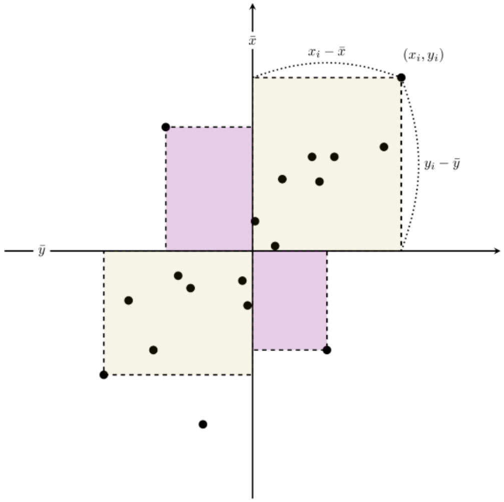
X_1 X_2 X_2 … X_{d-1} X_d

Feature Selection — SVD; t-Test, ANOVA

X_1 X_2 X_2 … X_{d-1} X_d

Covariance

$$\mathrm{cov}(X,Y) = \sigma_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{x})(y_i - \hat{y})$$
$$= \frac{1}{N-1}X^T Y$$



Correlation Coefficient

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}, \quad -1 \le \rho \le 1$$

Correlation Coefficient

$$\rho(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \hat{x}}{s_x}\right)\left(\frac{y_i - \hat{y}}{s_y}\right) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{x_i - \hat{x}}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{x})^2}}\right)\left(\frac{y_i - \hat{y}}{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \hat{y})^2}}\right)$$

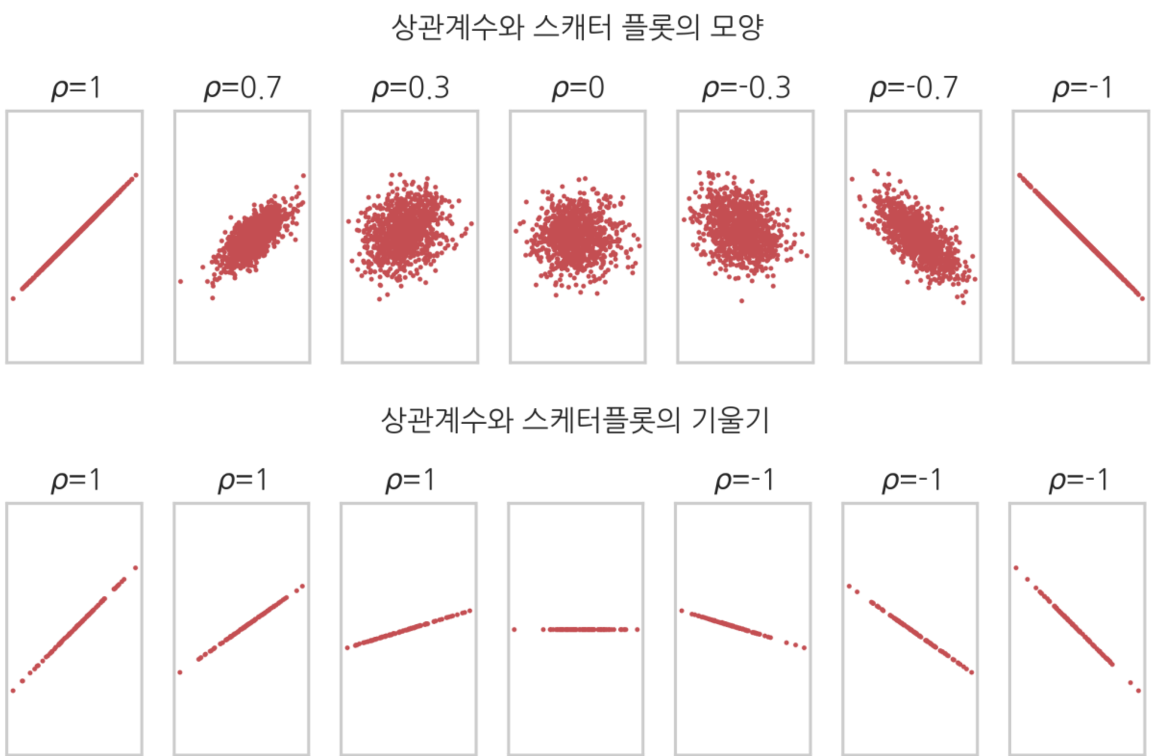
$$= \sum_{i=1}^{N}\frac{(x_i - \hat{x})}{\{\sum_{i=1}^{N}(x_i - \hat{x})^2}\frac{(y_i - \hat{y})}{\sum_{i=1}^{N}(y_i - \hat{y})^2\}^{1/2}} = \frac{\vec{a}\cdot\vec{b}}{\sqrt{|\vec{a}|^2 \cdot |\vec{b}|^2}} = \frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\cdot|\vec{b}|} \qquad -1 \le \rho \le 1$$

벡터의 내적

$$\vec{a}\cdot\vec{b} = |\vec{a}|\,\mathrm{Cos}\,\theta\,|\vec{b}| \qquad \frac{\vec{a}\cdot\vec{b}}{|\vec{a}||\vec{b}|} = \cos\theta, \qquad -1 \le \cos\theta \le 1$$

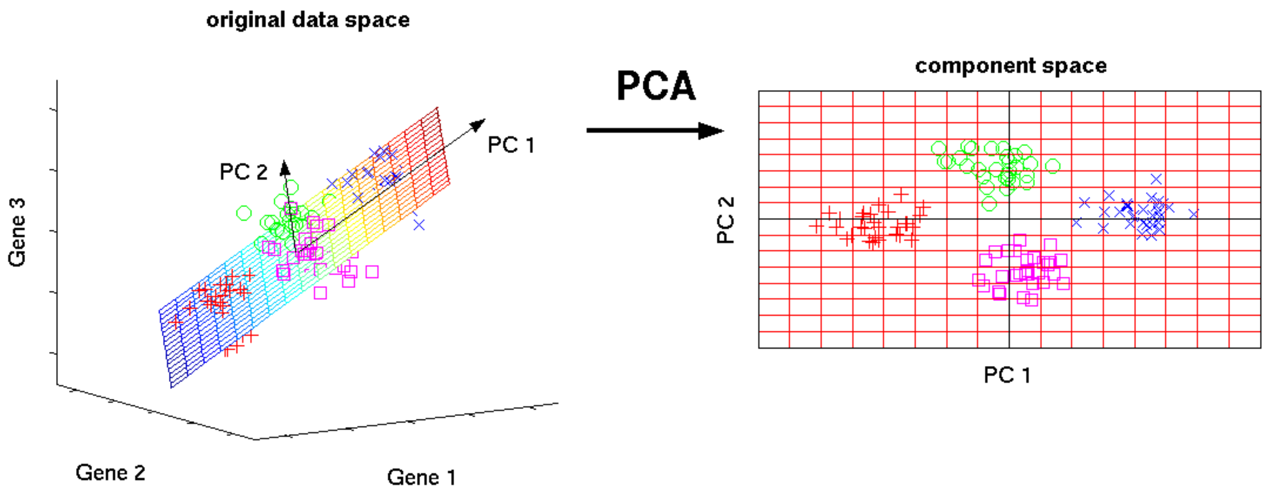
Correlation Coefficient = 벡터의 내적

→ $x_i - \widehat{x},\ y_i - \widehat{y}$ 의 관계를 $-1 \le \rho \le 1$ 범위의 수로 표현한 것



출처 : https://datascienceschool.net/view-notebook/4cab41c0d9cd4eafaff8a45f590592c5/

PCA (Principal Component Analysis) *for extraction*



기존에 관찰한 features(basis : x, y z)에 놓인 데이터를
분산이 최대가 되도록 하는 새로운 feature(basis : PC1, 2)로 Projection

$$Z = XA$$

(5*2) (5*3) (3*2)

X : 입력, Z = 출력
(Data * Features)

PCA (Principal Component Analysis)

$$\max_a\{Var(Z)\} \rightarrow \frac{1}{n}\sum_{i=0}\sum_{j=0}(x_{ij}a_{ij}-u)^2 \xrightarrow{\text{미분}} \frac{\partial(\mathrm{var}(Z))}{\partial a_e}$$

$$\frac{2}{n}\sum_{i=1}(\sum_{j=1}x_{ij}a_j)x_{ie}-2\lambda a_e=0 \xrightarrow{\text{정리}} 2\sum_{j=1}a_j\underline{(\frac{1}{n}\sum_{i=1}x_{ie})x_{ij}}=2\lambda a_e$$

$$\therefore \Sigma A=\lambda A$$

$$\Sigma\begin{Bmatrix} a_1 & a_2 & \cdots & a_m \end{Bmatrix}=\begin{Bmatrix} a_1 & a_2 & \cdots & a_m \end{Bmatrix}^T\begin{Bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_m \end{Bmatrix}$$

$$\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_m$$

a 가 unit length

Lagrange Multiplier

$$\mathrm{cov}(X)=\Sigma=\frac{1}{N-1}X^TX$$

eigen value

eigen vector

λ_i *: Variance of projected data*

$$\frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_m}\geq 0.9$$

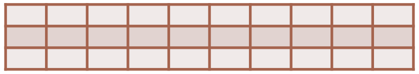
λ

PCA (Eigenface)

87

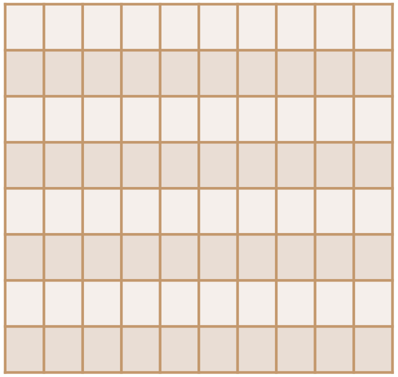
65

X



$N \times 5655$

$\mathrm{cov}(X) = \Sigma$



5666×5655

$$\Sigma A = \lambda A \qquad \begin{bmatrix} a_1 & a_2 & \cdots & a_K \end{bmatrix} \qquad \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \cdots & \\ & & & \lambda_f \end{bmatrix}$$

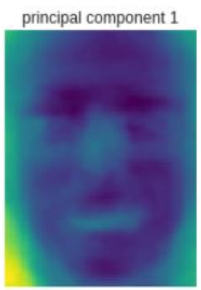
5666×5655 $\quad$ 5666×5655

PCA (Eigenface)

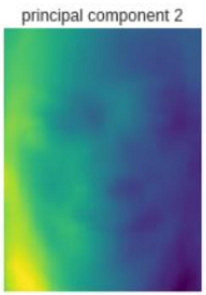
$\begin{Bmatrix} a_1 & a_2 & \cdots & a_K \end{Bmatrix}$ 5655×K (5655 행 중 K개만 선택)

87

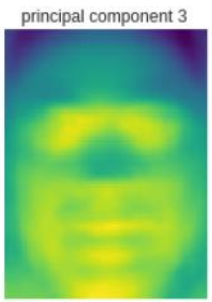
65



$*Z_{11} +$



$*Z_{12} + \cdots +$



$*Z_{1K} =$

즉, 5666×K 개의 행렬과 Z만 있으면, N개의 얼굴을 복원 가능

$$\begin{Bmatrix} Z_{11} & Z_{12} & \dots & Z_{1K} \\ Z_{21} & & & \\ \dots & & & \\ Z_{K1} & Z_{K2} & \dots & Z_{NK} \end{Bmatrix}$$

$$Z = XA$$

Z 의 한 row은 K차원에서의 한 점을 의미

ex) (2, 3, 1)은 3차원에서의 한 점

This Week

- 발표 목차

1. Covariance
2. Dimension reduction
   1) Feature Extraction
      a. PCA
   2) Feature Selection
      a. SVM
      b. T-test, ANOVA

SVD (Singular Value Decomposition) *for selection*

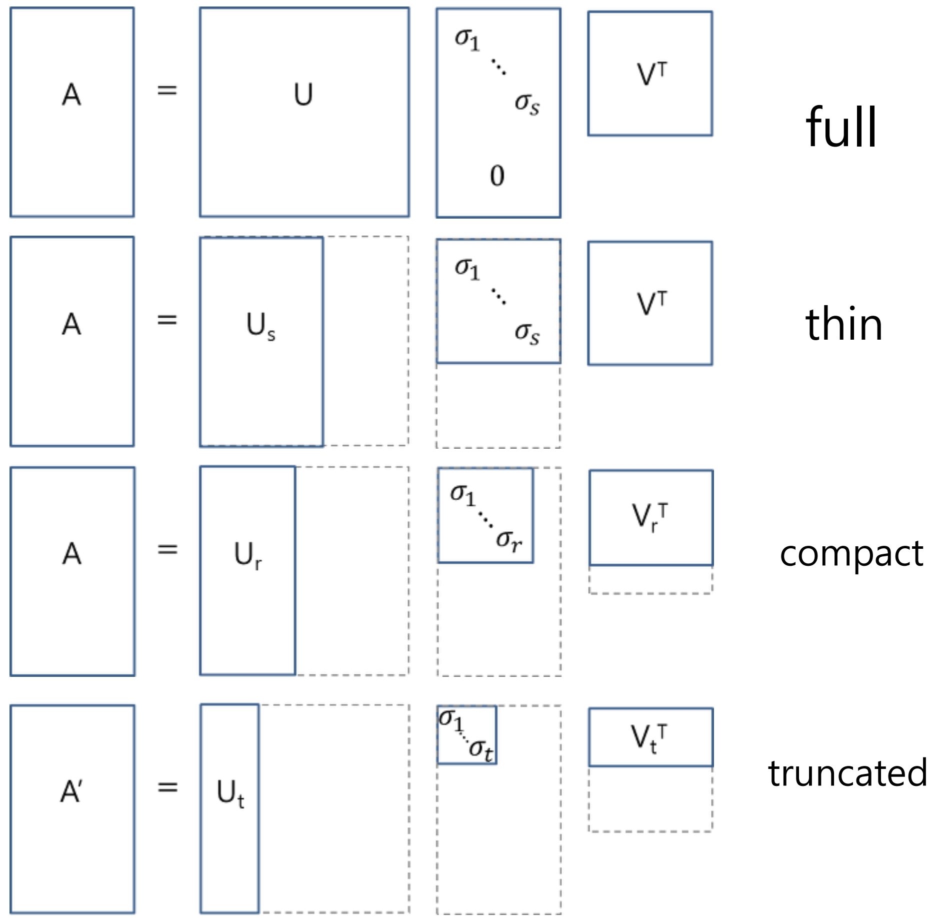
$$A = U \Sigma V^T$$

$$A^T A = V(\Sigma^T \Sigma)V^T$$

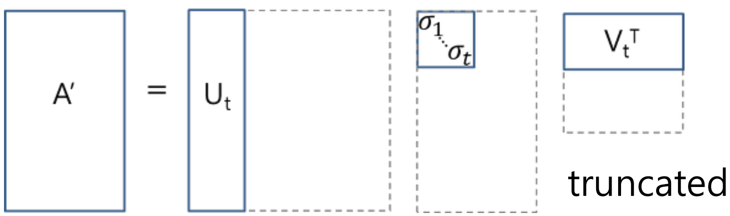
$$AA^T = U(\Sigma \Sigma^T)U^T$$

$$\Sigma^T \Sigma = \Sigma \Sigma^T = \sqrt{\lambda}$$

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_s$$



SVD (Singular Value Decomposition) *for selection*



출처 https://darkpgmr.tistory.com/106

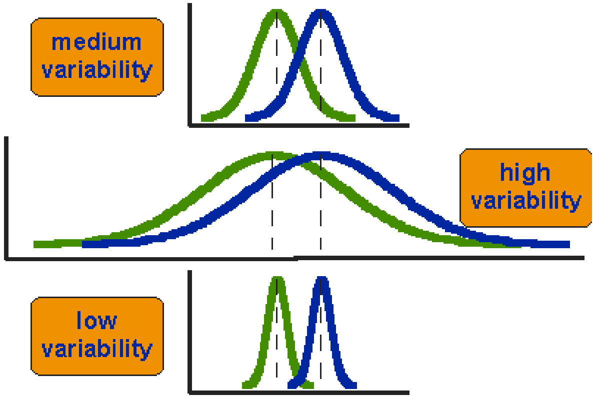
t=50

t=20

ANOVA (Analysis of Variance)

두 분포

$$t-test=\frac{|\widehat{x}_1-\widehat{x}_2|}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$$



출처 http://www.socialresearchmethods.net/kb/stat_t.htm

두 분포 이상

$$ANOVA=\frac{MeanSumofSquareBetween}{MeanSumofSquareWithin}$$

ANOVA (Analysis of Variance)

$$ANOVA = \frac{MeanSumofSquareBetween}{MeanSumofSquareWithin} = \frac{\text{집단 간 변동의 평균}}{\text{각 집단 내 변동의 평균}}$$

$$\text{분자} = \frac{\sum n_i \cdot (x_i - \widehat{\mathrm{X}})^2}{i-1}$$

$$\text{분모} = \frac{\sum\sum (x_{ij} - \widehat{\mathrm{X}}_i)^2}{n-i}$$

➡ t-test와 ANOVA가 작을 수록 두 집단의 분포가 유사하다

Next Week

- 프로젝트 과제 구체화, 사전 조사, 시작?
- 기본 공부가 중요하다.
  : 선대(행렬 곱도 똑바로 모른다, projection의 의미도 모른다)
  : 확률(Matrix로 어떻게 계산하나, 어떤 의미를 갖고 있나)
  : 프로젝트는 간단한 거 하면서 그 속에 담긴 기본 공부 시간?