

Week 9 - PCA - 특징선택 - 손익

① PCA (Principal Component Analysis)

주요한 것들 (특성)을 찾아내는 방법.

- 여러 Feature를 \Rightarrow 'Feature Selection'.
- 여러 Feature를 여러 Feature로 감소 \Rightarrow Feature Dimension Reduction / 차원 축소.

- 1) 데이터의 분포 (특성)를 파악하는 것
- 2) 데이터에서 중요한 것들, 고차원, 고차원 데이터
- 3) 고차원 데이터 (특성)를 데이터의 분포, 가장 큰 분포를 찾아냄.

② SVD (Singular Value Decomposition) \rightarrow 분해 (m x n 행렬에 대해 3개의 행렬로 나누는 데 사용됨)

행렬이 주어지면 그 행렬의 분해를 (m x n)

\star 데이터의 방향성, 중요도를 나타내어 준다!

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

$$\begin{bmatrix} A \\ m \times n \end{bmatrix} = \begin{bmatrix} U \\ m \times n \end{bmatrix} \begin{bmatrix} \Sigma \\ m \times n \end{bmatrix} \begin{bmatrix} V^T \\ n \times n \end{bmatrix} \quad (m > n)$$

Σ 는 주 성분들의 $\sqrt{\lambda_i}$ 로 이루어진 $m \times n$ 차원 대각행렬.

Q. 왜 이런 식일까?

SVD 특징!

행렬의 차원 축소 및 노이즈 제거 가능. \rightarrow 행렬의 노이즈도 제거 가능.

\rightarrow 행렬의 노이즈도 제거 가능.

\rightarrow 행렬의 노이즈도 제거 가능.

1) Full SVD

$$A = U \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_r \\ & & & 0 \end{bmatrix} V^T$$

2) Thin SVD

$$A = U_s \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_r \end{bmatrix} V_s^T$$

이런 식으로 U 와 V 를 잘라내면 U_s 와 V_s 가 된다. U_s 는 U 의 첫 r 개 열, V_s 는 V 의 첫 r 개 열이다.

3) Compact SVD

$$A = U \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_r \end{bmatrix} V^T$$

이런 식으로 U 와 V 를 잘라내면 U_s 와 V_s 가 된다.

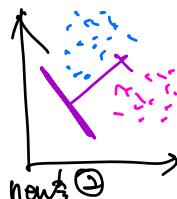
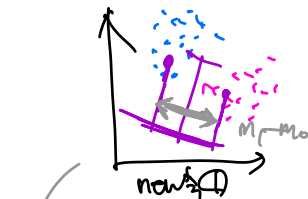
4) Truncated SVD

$$A = U \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_r \end{bmatrix} V^T$$

이런 식으로 U 와 V 를 잘라내면 U_s 와 V_s 가 된다. U_s 는 U 의 첫 r 개 열, V_s 는 V 의 첫 r 개 열이다. \Rightarrow 이미지 영상 관련시 용량 ↓.

3. LDA (Linear Discriminant Analysis)

2 클래스를 가장 잘 구분할 수 있는 사선(선)을 찾는 것이다.



이런 식으로 $m_1 - m_0$ 를 구한다.

두 클래스의 중심점의 차이와 각 클래스의 레이어들의 분산의

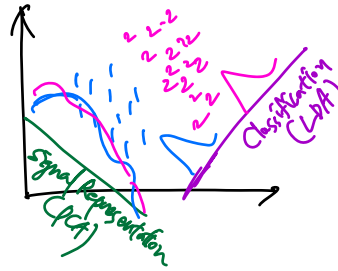
$m_1 - m_0$ 를 구한다!

$$J(w) = \frac{(m_1 - m_0)^2}{S_1 + S_2}$$

- 용도: 분류 (이진/다중 클래스). 사선(선)을 구하는 방법, Feature Selection 가능.

- 한계: 두 클래스의 중심에 레이어가 위치하는 경우 \rightarrow $\frac{m_1 - m_0}{S_1 + S_2}$ 가 0이 된다. \Rightarrow $\frac{m_1 - m_0}{S_1 + S_2}$ 가 0이 된다.

Q. PCA와 LDA 차이?



- PCA: 데이터 **원래 (불변)** 되어있는 축 찾는 것
- LDA: **클래스**를 가장 잘 구별할 수 있는 새로운 축 찾는 것

④ ICA (Independent Component Analysis)

관측된 신호는 복수 개 독립적인 신호의 합으로 표현될 수 있다.

가장) 각각은 어떻게든 섞여있는 신호를 서로 통계적으로 독립된 것으로 가짐. → **원래 신호**

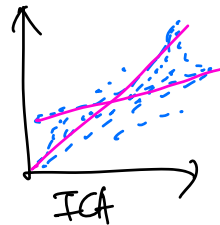
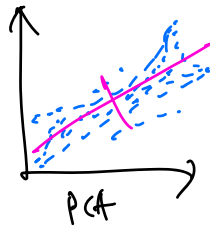
④ 지베르스 X 독립 가정.

독립성이 데이터 모든 특성으로 찾

이것들은 각각은 모두 X 데이터.

PCA와 달리 가장 독립적인 축을 찾는다. = **독립성**이 데이터 모든 특성에 대해

ICA의 관측에 의해



⑤ t-SNE

관측된 데이터는 **비선형**하게 섞여

관측된 데이터를 가장 잘 표현할 수 있는

기술한다.

(관측된 데이터가 N^2 정도 상관관계)

4) Factor Analysis

인공지능 → FA (요인분석) : 레이어의 여러 변수의 1/2 정도는 공통적으로 설명되는 것.
→ PCA (주성분분석) : 최대한 많은 변수를 가장 높은 차원으로 변환하는 것.
신경망에서 새로운 변수 생성