Euron 10주차 예습과제_개념정리_한송희

Chapter6 차원축소

01. 차원축소(Dimension Reduction)

차원축소: 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성

차원이 증가→ 데이터 포인트 간 거리 기하급수적으로 멀어짐, sparse한 구조→ 신뢰도 떨어짐/개별 피처간 상관 관계 높아짐→ 다중 공선성 문제로 모델 예측 성능 저하

⇒ 다차원의 피처를 차원축소해 피처 수를 줄이면 직관적으로 데이터 해석 가능

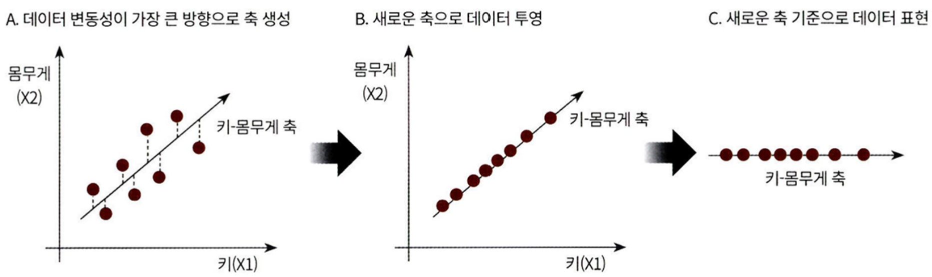
차원축소

-피처 선택(feature seletion): 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거하고 데이터의 특징을 잘 나타내는 주요 피처만 선택→ 새롭게 추출된 중요 특성은 기존의 피처가 압축된 것으로 기존 피처와 완전히 다른값

-피처 추출(feature extraction): 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출→ 기존 피처가 인지하기 어려웠던 잠재적 요소(Latent Factor)를 추출 e.g. PCA,SVD,NMF

02. PCA(Principal Component Analysis)

PCA: 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 주성분(Principal Component)을 추출해 차원을 축소



2개의 피처→ 1개의 주성분을 가진 데이터 세트로 차원축소

선형대수 관점=입력 데이터의 공분산 행렬(Covariance Matrix)을 고유값 분해하고, 이렇게 구한 고유벡터에 입력 데이터를 선형 변환. 고유값(eigenvalue)가 고유 벡터의 크기

<선형변환,공분산 행렬, 고유벡터>

선형변환: 특정 벡터에 행렬 A를 곱해 새로운 벡터로 변환=특정 벡터를 하나의 공간에서 다른 공간으로 투영(행렬=공간)

공분산 행렬: 정방행렬(열과 행이 같은 행렬)+대칭행렬(정방행렬 중 대각 원소를 중심으로 원소 값이 대칭)

$$C = P \sum P^T$$

$$C = [e_1 \cdots e_n] \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} e_1^t \\ \cdots \\ e_n^t \end{bmatrix}$$

*입력 데이터의 공분산 행렬이 고유벡터와 고유값으로 분해 가능하며 이렇게 분해된 고유벡터를 이용해 입력 데이터를 선형 변환하는 방식이 PCA이다

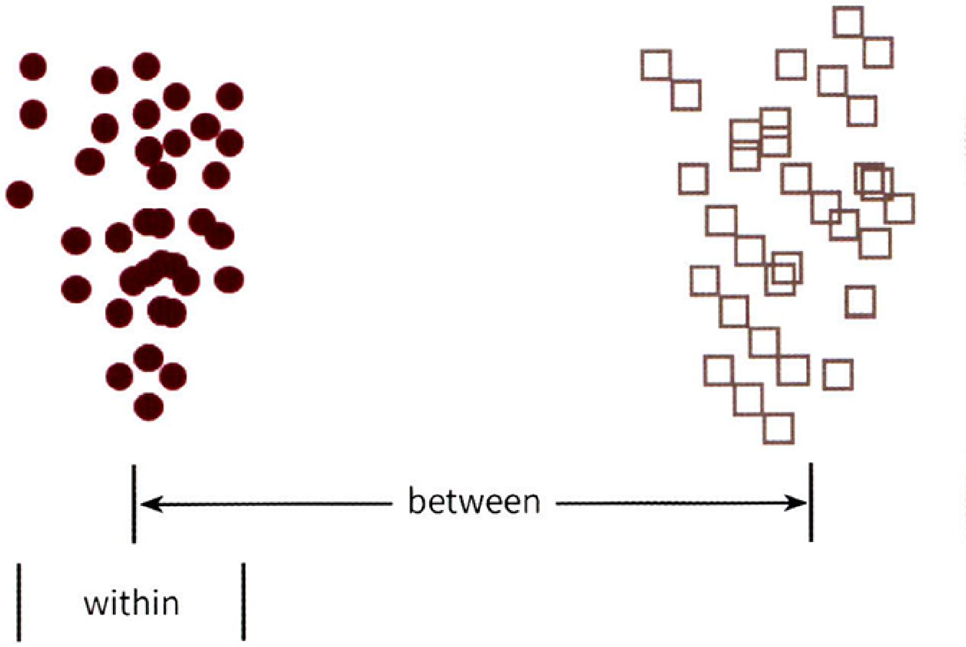
PCA순서

1. 입력 데이터 세트의 공분산 행렬을 생성
2. 공분산 행렬이 고유벡터와 고유값을 계산
3. 고유값이 가장 큰 순으로 K개(PCA 변환 차수만큼) 고유벡터 추출
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터 변환

03. LDA(Linear Discriminant Analysis)

LDA: 선형 판별 분석법. PCA와 유사하지만 지도학습의 classification에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원을 축소함.

클래스간 분산(between-calss scatter)과 클래스 내부 분산(within-class scatter)의 비율을 최대화 하는 방식으로 차원을 축소=클래스 간 분산은 최대한 크게! 클래스 내부의 분산은 최대한 작게!



$$S_W^T S_B = \begin{bmatrix} e_1 & \cdots & e_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} e_1^T \\ \cdots \\ e_n^T \end{bmatrix}$$

1. 클래스 내부와 클래스간 분산 행렬을 구함(mean vector 기반)
2. 위의 식처럼 두 행렬을 고유벡터로 분해
3. 고유값이 가장 큰 순으로 K개 추출
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환

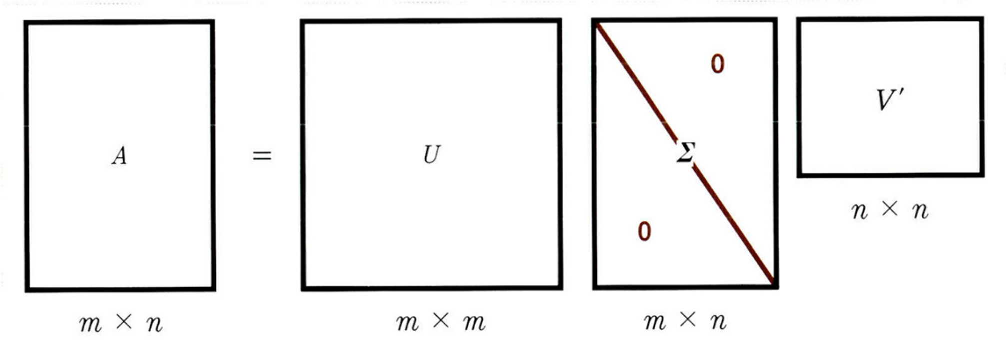
04. SVD(Singular Value Decomposition)

SVD: PCA와 유사하지만 SVD는 정방행렬뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용할 수 있다.

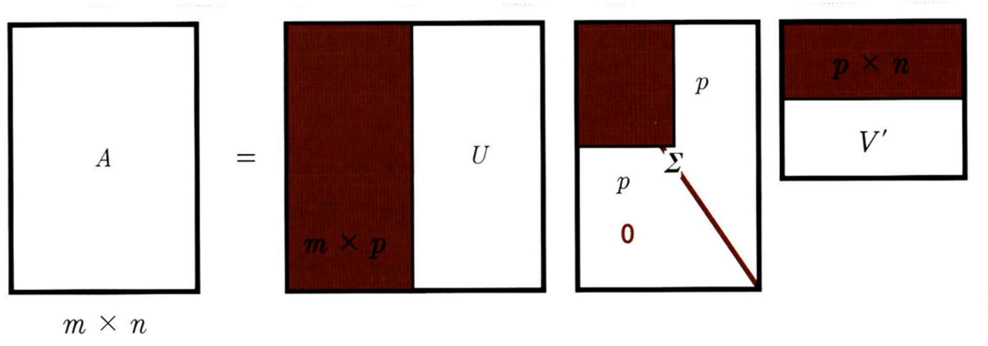
$$A = U \sum V^T$$

SVD:특이값 분해

행렬 U와 V에 속한 벡터는 특이벡터(*모든 특이 벡터는 서로 직교)



A의 차원이 mxn일때 U의 차원이 mxm, 시그마의 차원이 mxn, VT의 차원이 nxn 으로 분해

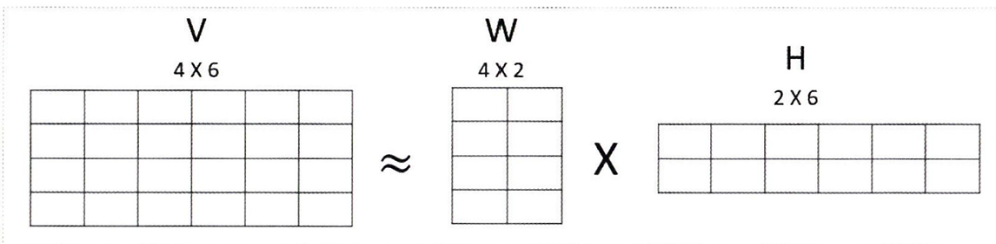


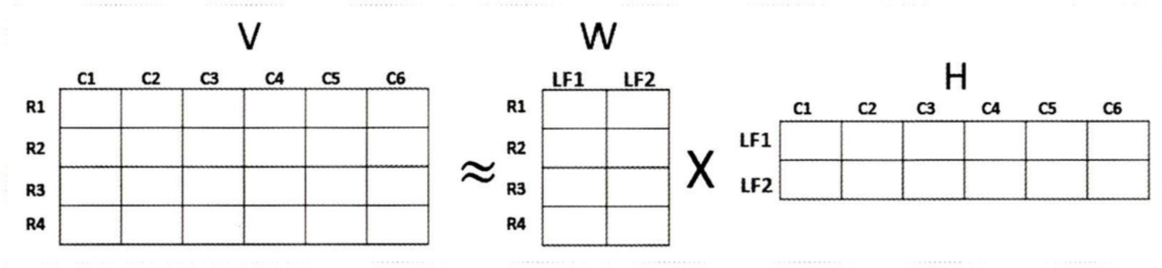
특이값이 0인 부분이 모두 제거된 컴팩트한 SVD

Truncated SVD: 시그마의 대각원소 중 상위 몇개만 추출하여 여기에 대응하는 U와 V의 원소도 함께 제거하여 차원을 축소

05. NMF(Non-Negative Matrix Factorization)

NMF: Truncated SVD와 같이 낮은 랭크를 통한 행렬 근사 방식의 변형으로 원본 행렬 내 모든 원소 값이 양수라는게 보장되면 그림처럼 간단하게 두 개의 기반 양수 행렬로 분해될 수 있는 기법





SVD처럼 행렬분해를 하게 되면 그림과 같이 분해되고 이 행렬은 잠재 요소로 특성으로 가지게 됨. 분해행렬 W는 원본 행에 대해서 잠재 요소 값이 얼마나 되는지에 대응하고 분해행렬 H는 잠재 요소가 원본 열로 어떻게 구성되었는지 나타냄.