chap 6 차원 축소

1. 차원 축소의 개요

차원 축소: 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것

- 차원이 증가할수록 / 데이터 포인트 간의 거리가 기하급수적으로 멀어지고, 희소한 구조를 가지게 된다.
- 피처가 많을 경우 개별 피처 간의 상관관계가 높을 가능성이 크다
- 차원 축소를 할 경우 학습 데이터의 크기가 줄어들어서 학습에 필요한 처리 능력을 줄일 수 있다

차원 축소 - 피처 선택, 피처 추출

- 피처 선택: 특성 선택은 말 그대로 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거하고, 데이터의 특징을 잘 나타내는 주요 피처만 선택하는 것
- 피처 추출: 기존 피처를 저차원의 중요 피처로 압축해서 추출하는 것,. 단순 압축이 아닌, 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출하는 것

차원 축소 알고리즘 : PCA, SVD, NMF

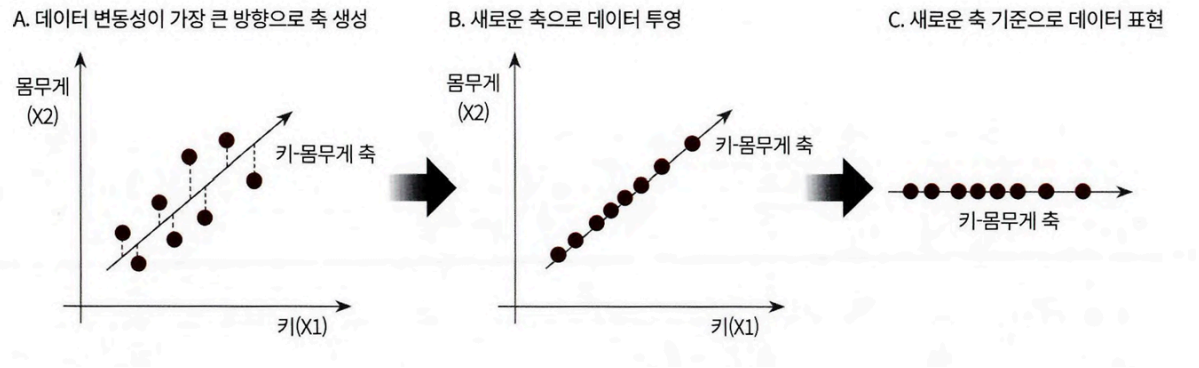
- 데이터에 잠재된 특성을 피처로 도출해 함축적 형태의 이미지 변환과 압축 수행
- 문서 내 단어들의 구성에서 숨겨져 있는 시맨틱(Semantic) 의미나 토픽을 잠재 요소로 간주하고 이를 찾아냄

.....

2. PCA (Principal Component Anaylsis)

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

기존 데이터의 정보 유실을 최소화하기 위해 가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원을 축소함.



PCA 선형대수 관점에서 해석: 입력 데이터의 공분산 행렬을 고유값 분해하고, 이렇게 구한 고유벡터에 입력 데이터를 선형 변환하는 것

PCA 스텝

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

3. LDA(Linear Discriminant Anaylsis)

: 입력 데이터를 저차원 공간에 투영해 차원을 축소하는 기법, 지도학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원 축소

vs PCA : PCA는 입력 데이터의 변동성의 가장 큰 축을 찾지만, LDA는 입력 데이터의 결정값 클래스를 최대한으로 분리할 수 있는 축을 찾음

클래스 간 분산(between-class scatter)과 클래스 내부 분산(within-class scatter)의 비율을 최대화하는 방식으로 차원 축소

STEP

1. 클래스 내부와 클래스 간 분산 행렬을 구한다. 이 두 개의 행렬은 입력 데이터의 결정 값 클래스별로 개별 피처의 평균 벡터를 기반으로 구한다.
2. 클래스 내부 분산 행렬을 Sw, 클래스 간 분산 행렬을 SB라고 하면 다음 식으로 두 행렬을 고유벡터로 분해할 수 있다

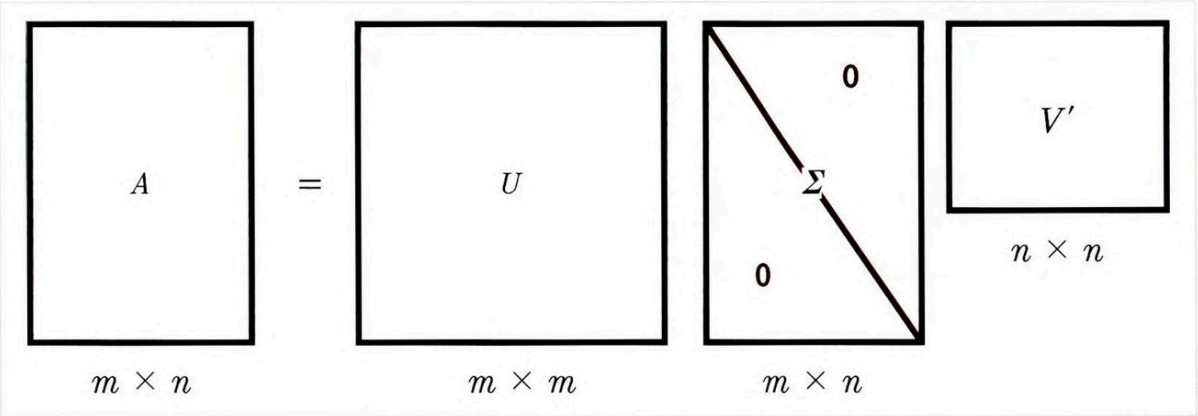
$$S_W^T S_B = \begin{bmatrix} e_1 & \cdots & e_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \cdots & \cdots & \cdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} e_1^T \\ \cdots \\ e_n^T \end{bmatrix}$$

3. 고유값이 가장 큰 순으로 K개(LDA변환 차수만큼) 추출한다.
4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환한다.

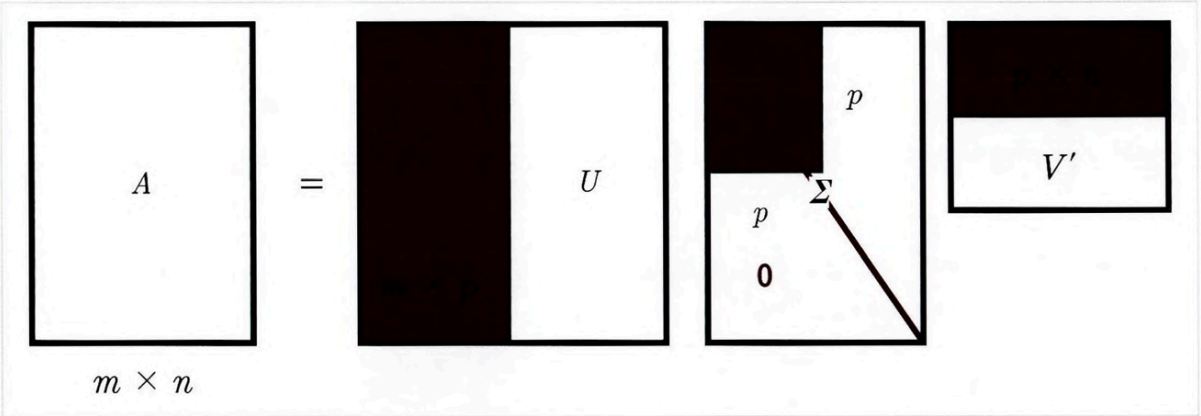
4. SVD(Singular Value Decomposition)

m x n 크기의 행렬 A

A의 차원이 m x n 일 때 U의 차원이 m x m, ∑의 차원이 m x n, V^T의 차원이 n x n으로 분해



하지만 일반적으로는 다음과 같이 Σ의 비대각인 부분과 대각원소 중에 특이값이 0인 부분도 모두 제거하고 제거된 Σ에 대응되는 U와 V 원소도 함께 제거해 차원을 줄인 형태로 SVD를 적용합니다. 이렇게 컴팩트한 형태로 SVD를 적용하면 A의 차원이 m × n일 때, U의 차원을 m × p, Σ의 차원을 p × p, V^T의 차원을 p × n으로 분해합니다.



SVD(특이값 분해)는 PCA와 유사한 행렬 분해 기법으로, 정방행렬뿐 아니라 행과 열의 크기가 다른 임의의 행렬에도 적용 가능.

행렬 A(m×n)을 SVD로 분해하면 $A = U\Sigma V^T$ 형태가 되며,

- **U**: m×m 직교 행렬 (왼쪽 특이벡터)
- **Σ**: m×n 대각행렬 (특이값이 대각에 위치)
- **V^T**: n×n 직교 행렬 (오른쪽 특이벡터)

Σ의 대각 원소 중 0이 아닌 값이 **특이값**이며, 모든 특이벡터는 서로 직교함.

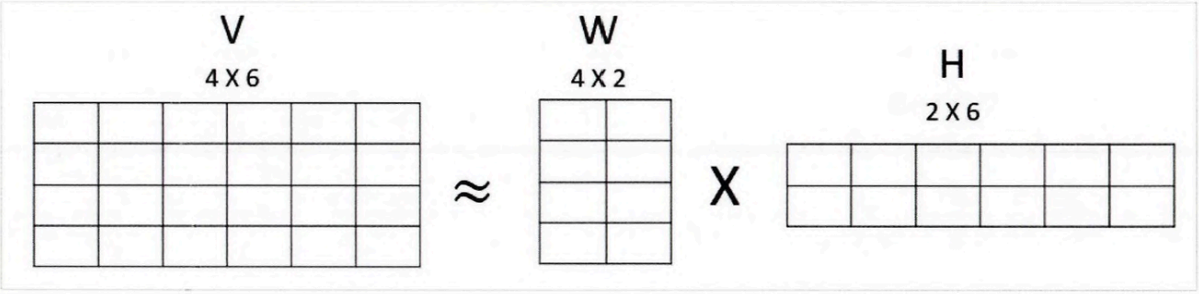
일반적으로는 0인 특이값과 그에 대응하는 U, V 원소를 제거한 **축소형(Compact SVD)** 을 사용하며,

이 경우 차원은 U(m×p), Σ(p×p), V^T(p×n)으로 줄어듦.

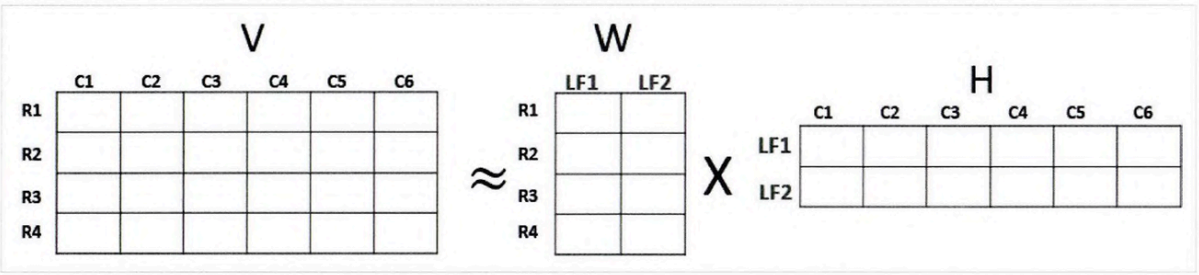
SVD: 컴퓨터 비전 영역에서 이미지 압축을 통한 패턴 인식과 신호 처리 분야에 사용, 텍스트의 토픽 모델링 기법인 LSA의 기반 알고리즘.

5. NMF(Non-Negative Matrix Factorization)

: 낮은 랭크를 통한 행렬 근사 방식의 변형



원본 행렬 내의 모든 원소 값이 모두 양수(0 이상)라는 것이 보장되면 이렇게 간단하게 두 개 기반 양수 행렬로 분해될 수 있는 기법을 의미.



일반적으로 W는 길고 가는 행렬(원본 행렬의 행 크기는 같고 열 크기보다는 작은),

H는 작고 넓은 행렬(원본 행렬의 행 크기보다 작고 열 크기와는 같은)이다.

NMF: 이미지 압축을 통한 패턴 인식, 텍스트의 토픽 모델링 기법, 문서 유사도 및 클러스터링, 추천 영역에 사용.