10주차 예습과제

01. 차원 축소 개요

- 차원축소: 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것
  - 차원이 증가할수록 데이터 간의 거리가 먼 희소한 구조를 가짐 → 예측 신뢰도 떨어짐
  - 입력 변수 간의 상관관계가 높을 가능성도 큼
  - 차원 축소를 통해 더 직관적으로 데이터를 해석할 수 있음
- 피처 선택: 종속성이 강한 피처는 아예 제거하고 주요 피처만 선택
- 피처 추출: 기존 피처를 저차원의 중요 피처로 압축해서 추출
  - 피처를 함축적으로 더 잘 설명하는 다른 공간으로 매핑해 추출하는 것
  - 잠재적인 요소를 추출하는 것을 의미
- 활용:
  - 이미지 데이터에서 잠재된 특성을 피처로 도출해 이미지를 변환하여 이미지 분류 수행 시 과적합 영향력을 낮춤
  - 문서 내 단어들의 구성에서 숨겨진 시맨틱이나 토픽을 찾아냄 (SVD, NMF 사용)

02. PCA(Principal Component Analysis)

- 여러 변수 간의 상관관계를 이용해 이를 대표하는 주성분을 추출하는 기법
- 기존 데이터의 정보 유실을 최소화하기 위해 가장 높은 분산을 가지는 데이터의 축으로 차원을 축소함
  - 첫 번째 축: 데이터 변동성이 가장 큰 방향으로 생성
  - 두 번째 축: 첫 번째 축에 직각이 되는 벡터
  - 세 번째 축: 두 번째 축과 직각이 되는 벡터
  - 이렇게 생성된 벡터 축에 원본 데이터를 투영 → 축의 개수만큼의 차원으로 데이터가 축소됨
- 선형대수 관점에서 PCA는 입력 데이터의 공분산 행렬을 고유벡터와 고윳값으로 분해하고, 구해진 고유벡터에 입력 데이터를 선형 변환하는 것임.

- 고유 벡터가 PCA의 주성분 벡터
- 고윳값은 고유벡터의 크기이면서 입력 데이터의 분산

[PCA 과정]

1. 입력 데이터 세트의 공분산 행렬 생성
2. 공분산 행렬의 고유벡터와 고유값 계산
3. 고유값이 가장 큰 순으로 변환 차수 K개 만큼 고유 벡터를 추출
4. 추출된 고유벡터를 이용해 입력 데이터를 변환

- PCA는 속성의 스케일에 영향을 받기 때문에 압축 전 스케일 변환 필요
- PCA 클래스 제공: `from sklearn.decomposition import PCA`
  - n_components=__ : 변환할 차원의 수 지정
  - explained_variancec_ratio_ : 전체 변동성에서 개별 pca 속성 별로 차지하는 변동성 비율
- 원본 데이터 대비 예측 성능은 떨어질 수 밖에 없음
- 컴퓨터 비전 분야에서 활발하게 적용됨

03. LDA(Linear Discriminant Analysis)

- LDA(=선형 판별 분석법): PCA와 유사하지만 지도학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원을 축소
  - 입력 데이터의 결정 값 클래스를 최대한으로 분리할 수 있는 축을 찾음
  - 즉, 클래스 간 분산은 크게하고 클래스 내부 분산은 최소화하는 방식으로 차원을 축소

[LDA 과정]

1. 클래스 내부와 클래스 간 분산 행렬 구함. (개별 피처의 평균 벡터를 기반으로)
2. 두 행렬을 고유 벡터로 분해

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

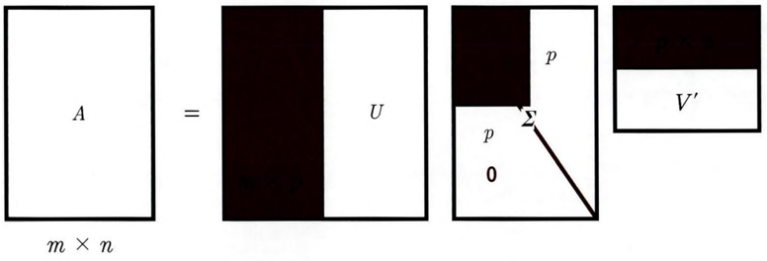
3. 고유값이 가장 큰 순으로 변환 차수 K개 만큼 추출
4. 추출된 고유벡터를 이용해 입력 데이터 변환

- LDA는 지도학습으로, 변환 시 결정값 입력해야 함.

04. SVD(Singular Value Decomposition)

- SVD(=특이값 분해): PCA와 달리 행과 열의 크기가 다른 행렬에도 적용할 수 있음

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

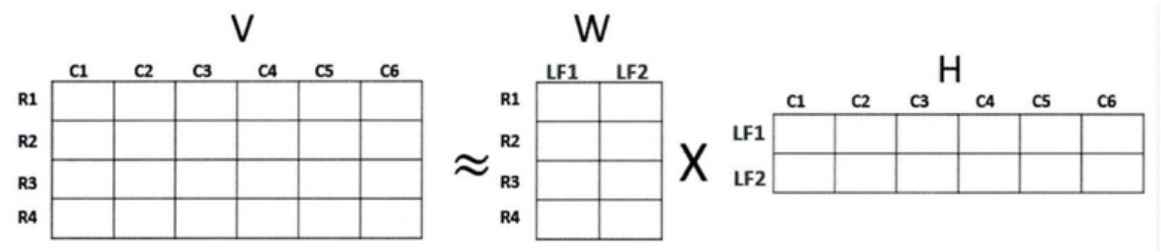


  - U와 V에 속한 벡터는 특이 벡터
  - 모든 특이 벡터는 서로 직교함
  - 시그마는 대각 행렬
  - Truncated SVD: 특이값 중 상위 일부 데이터만 추출해 분해하는 방식
    → 원본 행렬 다시 원복할수는 없지만 상당한 수준으로 근사할 수 있음

- `from sklearn.decomposition import TruncatedSVD`
- 스케일링을 통해 데이터 중심이 동일해지면 SVD와 PCA는 동일한 변환을 수행함.
- 하지만 PCA는 밀집행렬에 대한 변환만 가능한데 SVD는 희소 행렬에 대한 변환도 가능

05. NMF(Non-Negative Matrix Factorization)

- 모든 값이 0 이상으로 보장되면 두 개의 기반 양수 행렬의 곱으로 분해할 수 있음



- W = 각 샘플이 어떤 잠재 요소를 얼마나 갖고 있는지
- H = 잠재 요소가 어떤 속성의 조합으로 이루어졌는지
- `from sklearn.decomposition import NMF`