1. 차원 축소 개요

- 차원 축소는 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소해 새로운 차원의 데이터 세트를 생성하는 것
- 차원 축소
    - 피처 선택(feature selection): 특정 피처에 종속성이 강한 불필요한 피처는 아예 제거하고, 데이터의 특징을 잘 나타내는 주요 피처만 선택하는 것
    - 피처 추출(feature extraction): 기존 피처를 저차원의 중요 피처로 압축해서 추출하는데, 단순 압축이 아닌 피처를 함축적으로 더 잘 설명할 수 있는 또 다른 공간으로 매핑해 추출한다.
- 차원 축소의 의미는 이를 통해 좀 더 데이터를 잘 설명할 수 있는 잠재적인 요소를 추출하는 데에 있다. – PCA, SVD, NMF는 이처럼 잠재적인 요소를 찾는 대표적인 차원 축소 알고리즘이다.
- 차원 축소의 예시:
    - 이미지 분류: 매우 많은 픽셀로 이루어진 이미지 데이터에서 잠재된 특성을 피처로 도출해 함축적 형태의 이미지 변환과 압축을 수행
    - 텍스트 문서의 숨겨진 의미를 추출

2. PCA (Principal Component Analysis)

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

- 기존 데이터의 정보 유실이 최소화, 그러므로 가장 높은 분산을 가지는 데이터의 축을 찾아 이 축으로 차원을 축소한다

- 축 생성:

    - 가장 큰 데이터 변동성을 기반으로 첫 번째 벡터 축 생성
    - 첫 번째 벡터 축에 직각이 되는 벡터를 두 번째 벡터로 생성
    - 두 번째 벡터와 직각이 되는 벡터를 세번째 벡터로 생성

- -> 이처럼 원본 데이터의 피처 개수에 비해 매우 작은 주성분으로 원본 데이터의 총 변동성을 대부분 설명할 수 있는 분석법이다.

- 선형대수식을 이용해 설명하자면 입력 데이터의 공분산 행렬이 고유벡터와 고유값으로 분해될 수 있으며, 이렇게 분해된 고유벡터를 이용해 입력 데이터를 선형 변환하는 방식이 PCA이다.

    - 1. 입력 데이터 세트의 공분산 행렬을 생성
    - 2. 공분산 행렬의 고유벡터와 고유값을 계산
    - 3. 고유값이 가장 큰 순으로 K개만큼 고유벡터 추출
    - 4. 고유값이 가장 큰 순으로 추출된 고유벡터를 이용해 새롭게 입력 데이터를 변환

3. LDA (Linear Discriminant Analysis)

- LDA는 선형 판별 분석법으로 불리며, 지도학습의 분류에서 사용하기 쉽도록 개별 클래스를 분별할 수 있는 기준을 최대한 유지하면서 차원을 축소한다. -> 입력 데이터의 결정 값

클래스를 최대한으로 분리할 수 있는 축을 찾는다

- 클래스 간 분산과 클래스 내부 분산의 비율을 최대화하는 방식으로 차원을 축소한다. 클래스 간 분산은 최대한 크게 가져가고 클래스 내부의 분산은 최대한 작게 가져감.

4. SVD (Singular Value Decomposition)

- SVD는 PCA와 유사한 행렬 분해 기법을 이용하는데 정방행렬뿐만 아니라 행과 열의 크기가 다른 행렬에도 적용할 수 있다. 그래서 특이값 분해라고 불린다.

5. NMF (Non-Negative Matrix Factorization)

- NMF는 Truncated SVD와 같이 낮은 랭크를 통한 행렬 근사 방식의 변형. 원본 행렬 내의 모든 원소 값이 모두 양수라는 게 보장되면 간단하게 두개의 기반 양수 행렬로 분해될 수 있다.