ML 9기 1팀 박혜린

## 차원 축소 개요

매우 많은 피처로 구성된 다차원 데이터셋의 차원을 축소 -> 새로운 차원의 데이터셋 생성

차원 축소를 위한 접근 방법: 투영, 매니폴드 학습

## PCA

PCA: 여러 변수 간에 존재하는 상관관계를 이용해 이를 대표하는 Principal Component를 추출해 차원을 축소하는 기법 (가장 높은 분산을 가지는 데이터 축을 찾아 이 축으로 차원 축소함)

방법

1. 가장 큰 데이터 Variance 기반으로 첫번째 벡터 축 생성
2. 두번째 축은 이 벡터 축에 직각이 되는 벡터(직교 벡터)를 축으로 함
3. 세번째 축은 다시 두번째 축과 직각이 되는 벡터를 설정하는 방식으로 축 생성
4. 이렇게 생성된 벡터 축에 원본 데이터를 투영하면 벡터 축 개수만큼의 차원으로 원본 데이터가 차원 축소

## LDA

특정 공간상에서 클래스 분리를 최대화하는 축을 찾기 위해 클래스 간 분산(between-class scatter)과 클래스 내부 분산(within-class-scatter)의 비율을 최대화하는 방식으로 차원 축소

## SVD

SVD: PCA와 유사한 행렬 분해 기법 이용. 정방행렬뿐만 아니라 m*n 크기의 행렬 특이값도 분해함.

Truncated SVD: 시그마의 대각원소 중 상위 몇 개만 추출해 대응하는 U와 V의 원소도 함께 제거해서 차원을 줄인 형태로 분해하는 것

## NMF

Truncated SVD처럼 낮은 랭크 통한 행렬근사 방식의 변형. 원본 행렬 내의 모든 원소 값

이 모두 양수라는게 보장되면, 더 간단하게 두 개의 기반 양수 행렬로 분해될 수 있음.

다른 차원 축소 기법들

1. 랜덤 투영: 랜덤한 선형 투영을 사용해 고차원 데이터를 저차원으로 변환
2. 다차원 스케일링: 샘플 간의 거리를 최대한 보존하면서 차원 축소
3. IsoMap: 각 샘플을 가까운 이웃과 연결해 그래프를 만들고, 그 그래프상의 geodesic distance를 유지하면서 차원 축소
4. t-SNE: 비슷한 샘플은 가깝게, 다른 샘플은 멀리 배치되도록 확률적 방식으로 임베딩
5. LDA: 클래스 간 분산을 최대화하고 클래스 내 분산을 최소화하는 축 학습