6장 차원 축소

1. 차원 축소란?

- 고차원 데이터의 특징을 잃지 않으면서 **더 낮은 차원으로 변환**하는 기법
- 목적:
    - 데이터 시각화 용이
    - 계산 효율 향상
    - 노이즈 제거 및 과적합 방지
    - 상관관계 높은 변수 압축

>> 차원이 증가할수록 데이터가 **희소**해지고 거리 기반 알고리즘 성능 저하

2. 접근 방식

- **피처 선택**: 불필요한 변수 제거
- **피처 추출**: 기존 피처를 조합하여 새로운 축 생성(예: PCA, SVD)

3. 매니폴드 학습 관점

- 고차원 데이터가 실제로는 더 낮은 차원 공간(매니폴드)에 놓여 있다는 가정
- 예: 스위스롤 데이터 → 펼치면 평면 구조

4. 주요 차원 축소 알고리즘

1) PCA (주성분 분석)

- 목표: **데이터 분산을 가장 많이 설명하는 축(주성분) 찾기**
- 방법:
    1. 데이터 스케일링
    2. 공분산 행렬 계산
    3. 고유값/고유벡터 분해
    4. 주성분 방향으로 투영

- 특징:
    - 비지도학습
    - 선형 변환
    - 가장 큰 분산 방향 우선
- 장점: 빠름, 해석 용이
- 단점: 비선형 구조에는 취약
- 예시:
    - Iris 데이터 4D → 2D
    - 첫 PC가 **72.9%**, 두 번째 PC가 **22.8%** 설명 (총 95% 유지)

2) LDA (선형 판별 분석)

- 목표: **클래스 간 분리 극대화**
- 비교:

PCA	LDA
비지도	지도학습
분산 최대	클래스 분리 최대
전체 데이터	라벨 필요

3) SVD (특이값 분해)

- 행렬을 $U\Sigma V^T$로 분해하는 방법
- PCA와 유사 (PCA는 SVD 기반 구현)
- **Truncated SVD**:
    - 큰 특이값만 사용 → 차원 축소
    - 희소행렬 처리 가능 (PCA는 불가)
- 활용: 이미지 압축, 텍스트 토픽 모델링(LSA)

4) NMF (비음수 행렬 분해)

- 모든 값이 0 이상(양수)일 때 사용
- 행렬을 $W \times H$로 분해하여 **잠재요소 학습**
- 특징: 해석 용이 (음수가 없어 의미 해석 가능)
- 활용: 텍스트 토픽 모델링, 추천 시스템, 이미지 패턴 분석

5. 추가 차원 축소 기법 비교

기법	특징	장점	단점
Random Projection	무작위 투영	매우 빠름	정확도 낮을 수 있음
MDS	거리 보존	시각화에 강함	느림
Isomap	지오데식 거리 보존	비선형 구조 학습	계산비용 큼
t-SNE	이웃 구조 보존	군집 시각화 최고	거리/축 해석 어려움, 느림

6. 실전/캐글 예시 정리

: 유방암 데이터(30D)를 2D로 축소

- PCA: 전체 구조 파악에 유리
- MDS: 거리 관계 보존
- t-SNE: 군집(악성/양성) 매우 명확하게 분리

>> 데이터 목적에 맞는 알고리즘 선택이 중요

- 전역 구조 → PCA, MDS
- 국소 군집 시각화 → t-SNE
- 지도학습 + 분류 성능 향상 → LDA
- 희소 텍스트, 추천 시스템 → SVD, NMF