유런 ML 분과 권혜수 10주차 요약본

1. 차원 축소(Dimensionality Reduction)
- 매우 많은 피처로 구성된 다차원 데이터 세트의 차원을 축소하여 새로운 차원의 데이터 세트 생성
- 데이터의 차원이 많아질수록 데이터 포인트 간 거리가 멀어지고 분석이 어려워짐
- 피처가 많을 때 개별 피처 간 상관관계가 높을 가능성이 큼
(1) 피처 선택: 불필요하거나 상관관계가 높은 피처를 제거하고 핵심 피처만 남김
(2) 피처 추출: 피처를 함축적으로 더 잘 설명할 수 있는 다른 공간으로 매핑하여 저차원의 중요 피처로 압출해서 추출. 완전히 다른 피처 값이 됨.
(3) 차원 축소 알고리즘: PCA, SVD, NMF등

2. PCA
- 여러 변수 간의 상관관계 -> 데이터의 분산이 가장 큰 방향 -> 데이터를 축소하는 기법
- 공분산 행렬의 고유값분해를 통해 주성분 벡터를 도출
- 고유벡터로 입력 데이터를 선형 변환
- 비지도
- 붓꽃(Iris) 데이터의 4차원 속성을 2차원으로 축소하더라도 95% 이상의 정보를 유지할 수 있으며, 상관성이 높은 변수를 두세 개의 주성분으로 압축하여 유사한 성능을 얻을 수 있음

3. LDA
- PCA와 달리 지도학습 분류에서 사용하도록 클래스 간 분산을 최대화하고 클래스 내 분산을 최소화하는 방향
-즉, 단순히 데이터의 분산이 아니라 클래스 구분을 명확히 할 수 있는 축을 찾는 데 초점**
-PCA가 데이터의 구조적 분산 & LDA는 클래스 간 경계를 뚜렷하게
- 클래스 간 분산과 클래스 내부 분산의 비율을 최대화

4. SVD
- 임의의 행렬을 세 개의 행렬로 분해하는 방식으로, PCA가 공분산 행렬과 유사
- 특이 벡터: 서로 직교하는 성질
- 그러나 SVD는 정방행렬뿐 아니라 여러 크기의 행렬에도 적용 가능
**Truncated SVD는 상위 몇 개의 특이값만 추출
- 정확하게 다시 복구 X 하지만 상당한 수준으로 근사 가능
- 원래 차원 차수에 가깝게 자를수록 원본 행렬에 더 가깝게 복원
- 사이파이에서만 지원(넘파이 안됨)
- 희소행렬로만 지원. **scipy.sparse.linalg.svds 이용**

5. NMF
- 모든 데이터가 양수일 때 사용할 수 있는 행렬 분해 기법
- 원본 행렬을 두 개의 양수 행렬 W와 H로 분해하여 **잠재 요소** 추출

- W는 원본 행에 대해 잠재 요소의 값이 얼마나 되는지, H는 잠재 요소가 어떻게 구성됐는지
- 중배 갱신 규칙: 행렬 W, H 값 무작위로 설정 -> R와 W*H 간의 거리 최소화 -> 함수가 달라짐
- PCA는 다변량 통계 분석 방법, NMF는 대량의 정보 의미 특징 & 의미 변수
-> 이미지 압축 패턴 인식, 텍스트 토픽 모델링, 추천 영역

**랜덤 투영: 랜덤 선형 투영을 통해 거리 정보를 보존하면서 빠르게 차원을 줄이는 방법
**MDS: 샘플 간 거리를 최대한 유지하면서 차원을 축소
**Isomap: 비선형 데이터의 매니폴드 구조를 유지, 가까운 이웃과 연결해서 그래프 -> 지오데식 거리를 유지하며 축소
**t-SNE: 고차원 데이터에서 비슷한 샘플만 가깝게 군집 시각화에 매우 뛰어남.
**선형 판별 분석: 클래스 간 분산 최대화, 클래스 내 분산 최소화하는 축 학습. 분류인데 차원 축소에도 쓰임. 전처리용으로 적합