**8주차 요약: 차원 축소 (Dimensionality Reduction)**

**1. 차원 축소 개념**

* 고차원의 데이터는 모델 학습을 어렵게 만들 수 있으며 이를 \*\*차원의 저주(curse of dimensionality)\*\*라고 함.
* 차원 축소는 데이터의 **특성 수를 줄여** 계산 효율성과 모델 성능을 높이는 기법으로, **정보 손실을 최소화하며 데이터 구조를 보존**하는 것이 목적임.

**2. 차원의 저주**

* 차원이 높아질수록 두 점 사이의 거리가 멀어져 예측이 부정확해질 수 있음.
* 해결책: 훈련 샘플 수를 늘리는 것이 가장 직접적인 방법이나 현실적으로 어려움 → 따라서 차원 축소가 필요함.

**3. 차원 축소 방법**

**사영 (Projection)**

* 고차원 데이터를 저차원 평면에 \*\*투영(projection)\*\*하는 방식.
* 예: 3차원 데이터를 가장 잘 설명할 수 있는 2차원 평면으로 사영하여 시각화나 분석에 활용.

**다양체 학습 (Manifold Learning)**

* 데이터가 국소적으로는 선형 구조를 가진다는 가정 하에 **복잡한 고차원 구조를 펼치는 방식**.
* 예: 말려 있는 롤케이크 구조를 펼치듯이, 데이터의 본질적인 구조를 드러냄.

**4. PCA (주성분 분석)**

* 고차원 데이터에서 가장 큰 분산을 가지는 축을 찾아 해당 축으로 데이터를 사영함.
* **분산을 최대한 보존**하는 방향으로 차원을 줄이므로 정보 손실이 적음.
* explained\_variance\_ratio\_: 각 주성분이 설명하는 분산의 비율을 의미.
* 전체 분산의 95% 이상을 설명하는 주성분 수를 기준으로 적절한 차원을 선택할 수 있음.
* 활용 예시: MNIST 데이터에서 784차원을 154차원으로 줄여도 학습 성능 유지 가능.

**5. PCA의 변형 기법**

* **랜덤 PCA**: SVD 과정을 확률적으로 처리해 계산 속도를 높인 방법.
* **점진적 PCA(IPCA)**: 데이터가 한꺼번에 메모리에 들어가지 않을 경우 **배치 단위로 학습**할 수 있도록 함. partial\_fit() 메서드 사용 시 유의 필요.

**6. 임의 사영(Random Projection)**

* **존슨-린덴슈트라우스(JL) 정리** 기반: 고차원 공간의 데이터를 무작위로 사영해도 데이터 간 거리의 왜곡이 크지 않다는 수학적 보장 제공.
* 사이킷런에서 제공하는 모델:
  + GaussianRandomProjection
  + SparseRandomProjection (희소 행렬 기반으로 대규모 데이터에 적합)

**7. LLE (국소적 선형 임베딩)**

* 전체적으로는 비선형 구조를 가지지만 **국소적으로는 선형 구조를 갖는 데이터셋**에 적합.
* 주변 이웃 간의 거리를 유지하면서 저차원 공간에 표현하는 방식.
* 사영보다는 **다양체 학습 방식에 가까움**.

**8. 기타 차원 축소 기법**

* 사이킷런에서 제공하는 주요 모델:
  + **MDS (다차원 스케일링)**
  + **Isomap**
  + **t-SNE**: 군집 시각화 등에 자주 활용됨
  + **LDA (선형 판별 분석)**
  + **커널 PCA**