차원의 저주

: 샘플의 특성이 많으면 학습이 매우 어려워진다.

- 차원이 커질 수록 두 지점 사이의 거리가 매우 커지는 문제
  - 고차원은 많은 공간을 가지고 있어서, 고차원 데이터셋은 매우 희박할 위험이 있음
  - 특성 수가 아주 많은 경우, 훈련 샘플 사이의 거리가 매우 커서 과대적합 위험도가 커짐

차원 축소

- 특성 수를 줄여서 학습 불가능한 문제를 학습 가능한 문제로 만드는 기법.
- 훈련 속도가 빨라지지만, 일부 정보가 유실되어 성능이 저하될 수 있음.
- 훈련 속도를 높이는 것 외에 데이터 시각화에도 유용함.

차원 축소를 위한 접근 방법

1. 투영:

- n차원 공간에 존재하는 d차원 부분공간을 d차원 공간으로 투영하기(단, d < n)

2. 매니폴드 학습:

- d 차원 매니폴드는 국부적으로 d 차원 초평면으로 보일 수 있는 n 차원 공간의 일부

PCA

- 주성분 분석은 먼저 데이터에 가장 가까운 초평면을 정의한 다음, 데이터를 이 평면에 투영시킴.
- 분산 보존 개념과 주성분 개념이 중요함.

분산보존

- 저차원으로 투영할 때 훈련 세트의 분산이 최대로 보존되는 축을 선택해야 함

주성분

- 주성분 축 찾기
  - 첫 번째 주성분: 훈련 세트에서 분산을 최대한 보존하는 축

- 두 번째 주성분: 첫 번째 주성분과 수직을 이루면서 분산을 최대한 보존하는 축
- 세 번째 주성분: 첫번째, 두번째 주성분과 수직을 이루면서 분산을 최대한 보존하는 축
- 데이터 셋에 있는 차원의 수만큼 네 번째, 다섯 번째 … n 번째 축을 찾음.

적적한 차원 수 선택하기

- 분산 비율의 합이 충분한 분산(95% 정도)이 되도록 하는 차원의 수를 선택

랜덤 PCA

- 랜덤 PCA라 부르는 확률적 알고리즘을 사용해 처음 d개의 주성분에 대해 근삿값을 빠르게 찾는다.

점진적 PCA

- 훈련세트를 미니배치로 나눈 후 IPCA(점진적 PCA)에 하나씩 주입 가능
- 온라인 학습에 적용 가능

커널 PCA

- 커널트릭을 PCA 적용해 차원 축소를 위한 복잡한 비선형 투영을 수행
- 투영된 후에 샘플의 군집을 유지하거나 꼬인 매니폴드에 가까운 데이터 셋을 펼칠 때도 유용함

커널선택과 하이퍼파라미터 튜닝

- kPCA는 비지도 학습
- 커널과 하이퍼파라미터 튜닝을 위한 측정 방식
  - 방식 1 : 전처리 용도로 사용 후 예측기와 연동하는 그리드탐색 등을 활용하여 성능 측정 가능
  - 방식 2 : 가장 낮은 재구성 원상의 오차를 최소화하는 커널과 하이퍼파라미터 선택 가능

LLE(지역선형임베딩)

- 비선형 차원숙소 기법
- 투영이 아닌 매이폴드 학습에 의존한다.
-

1. PCA에서 어떻게 주성분이 선택되어 분산을 최대한 보존하는지?

PCA에서 주성분은 훈련 세트의 분산을 최대한 보존하는 축

첫 번째 주성분은 훈련 세트에서 분산을 최대한 보존하는 축이고,

이후의 주성분들은 이전 주성분과 수직을 이루면서 분산을 최대한 보존하는 방식으로 선택

---

2. PCA와 LLE는 차원 축소를 위한 다른 방법인데 두 알고리즘의 주요 차이점은?

PCA는 주로 선형 차원 축소를 수행, 주성분을 찾아내어 투영

LLE는 비선형 차원 축소를 수행, 매니폴드를 학습

---

3. 차원 축소가 현업에서 어떤 분야에서 주로 사용되고 있으며, 어떤 이점을 가져다 주는지 예시를 들어 설명

차원 축소는 이미지 처리, 음성 인식과 같은 다양한 분야에서 활발하게 사용 중이다.

예를 들어 이미지 쪽에서는 계산 비용을 줄이고, 성능을 향상시킬 수 있지 않을까?