머신러닝 세션 8주차

2024170928 노연경

# 차원축소

샘플의 특성이 너무 많으면 학습이 어렵기 때문에 특성 수를 줄여 학습이 가능한 문제로 만든다

차원 축소로 인한 정보손실을 어느 정도 감안하면서 훈련 속도와 성능을 최대로 유지하는 것이 목표

# 차원 숙소 기법

모든 훈련 샘플이 고차원 공간의 일부인 저차원 부분공간에 가깝게 놓여 있는 경우가 일반적인데 이런 경우 고차원의 데이터셋을 저차원의 데이터셋으로 변환시켜도 정보의 손실이 크지 않음

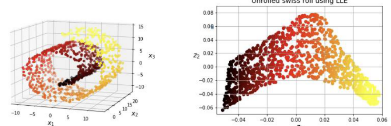
대표적 차원 축소 기법 : 사영, 다양체학습

1. 사영 기법

n차원 공간에 존재하는 데이터셋을 낮은 d차원 공간으로 사영하는 방법

사영은 경우에 따라 복잡한 결과를 낼 수도 있음

Ex. 롤케이크를 x1과 x2축으로 사영하면 샘플 구분이 보다 어려워짐 🡪 말린 것을 펼쳐야됨



2. 다양체 가설

대부분의 고차원 데이터셋이 더 낮은 차원의 다양체에 가깝다는 사실

# PCA (주성분 분석)

훈련 데이터에 가장 가까운 초평면에 데이터셋을 사영하는 기법

분산 보존 개념, 주성분 개념을 활용함

- 분산 보존 : 저차원으로 사영할 때 데이터셋의 분산이 최대한 유지되도록 축을 지정해야 함

- 주성분 : 첫째 주성분 : 분산을 최대한 보존하는 축, 둘째 주성분 : 첫째 주성분과 수직을 이루면서 첫째 주성분이 담당하지 않는 분산을 최대한 보존하는 축, …

데이터셋의 주성분은 SVD (특잇값 분해) 기법을 이용하면 쉽게 계산 가능

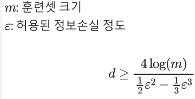
파일 압축 용도로 PCA를 활용할 수 있음

랜덤 PCA : 주성분 선택을 위해 사용되는 SVD 알고리즘을 확률적으로 작동하도록 만드는 기법

점진적 PCA : 훈련세트를 미니배치로 나눈 후 IPCA에 하나씩 주입 가능

# 임의 사영

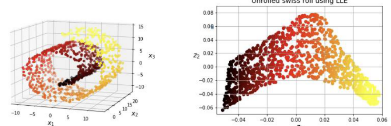
존슨-린덴슈트라우스 정리 : 고차원의 데이터를 적절한 크기의 저차원으로 임의로 사영하더라도 데이터셋의 정보를 많이 잃어버리지 않음을 보장

 를 만족하는 d를 사영공간의 차원으로 지정

# LLE(국소적 선형 임베딩)

대표적인 다양체학습 기법

전체적으로 비선형인 다양체이지만 국소적으로는 데이터가 선형적으로 연관되어 있음을 기본 아이디어로 한다. Ex. 롤케이크 데이터셋



사영이 아닌 다양체 학습에 의존