8주차 과제

박민서

1. 차원 축소
   -차원의 저주: 샘플의 특성이 너무 많으면 학습이 매우 어려움.
   -차원 축소: 특성 수를 줄여 학습이 불가능한 문제를 학습이 가능한 문제로 만드는 기법
   -차원 축소로 인한 정보 손실을 어느 정도 감안하면서 훈련 속도와 성능을 최대로 유지하는 것이 목표
2. 차원의 저주: 벡터의 차원에 해당하는 특성 수가 커질수록 두 샘플 사이의 거리가 매우 커져서 과대적합 위험이 커짐
   -이유: 새로운 데이터 샘플이 주어졌을 때, 훈련셋에 포함된 샘플과의 거리가 일반적으로 매우 멀어 기존 값들을 이용한 추정이 어렵기 때문.
   -해결책: 샘플 수 늘리기, 하지만 고차원의 경우 충분히 많은 샘플 수를 준비하는 일이 매우 어렵거나 사실상 불가능
3. 차원 축소 기법
   -기본 아이디어: 모든 훈련 샘플이 고차원 공간의 일부인 저차원 부분공간에 가깝게 놓여 있는 경우가 일반적으로 발생
   -사영 기법: n차원 공간에 존재하는 데이터셋을 낮은 d차원 공간으로 사영하기
   -다양체 학습: 롤케이크의 경우 사영보다는 말린 것을 펼치면 보다 적절한 2차원 이미지를 얻을 수 있음.
4. PCA(주성분 분석)
   -아이디어: 훈련 데이터에 가장 가까운 초평면에 데이터셋을 사영하는 기법
   -분산 보존: 저차원으로 사영할 때 데이터셋의 분산이 최대한 유지되도록 축을 지정해야 함.
   -주성분
   (1) 첫째 주성분: 분산을 최대한 보존하는 축
   (2) 둘째 주성분: 첫째 주성분과 수직을 이루면서 첫째 주성분이 담당하지 않는 분산을 최대한 보존하는 축
   (3) 셋째 주성분: 첫째, 둘째 주성분과 수직을 이루면서, 첫째, 둘째 주성분이 담당하지 않는 분산을 최대한 보존하는 축

-특잇값 분해: 데이터셋의 주성분은 특잇값 분해 기법을 이용하면 쉽게 계산 가능, 찾아진 초평면으로의 사영 또한 쉽게 계산됨, 데이터셋이 크거나 특성이 많으면 계산이 오래 걸릴 수 있음.

-적절한 차원: 밝혀진 분산 비율의 합이 95% 정도 되도록 하는 주성분들로 구성, 데이터 시각화 목적의 경우 보통 2개 또는 3개로 한다.

-랜덤 PCA: 주성분 선택을 위해 사용되는 SVD 알고리즘을 확률적으로 작동하도록 만드는 기법, 보다 빠르게 지정된 개수의 주성분에 대한 근삿값을 찾음

-점진적 PCA: 훈련세트를 미니배치로 나눈 후 IPCA에 하나씩 주입 가능, 온라인 학습에 적용 가능, partial_fit() 활용에 주의할 것.

5. 임의 사영

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

6. LLE(국소적 선형 임베딩)

-기본 아이디어: 대표적 다양체 학습 기법, 롤케이크 데이터셋의 경우처럼 전체적으로 비선형인 다양체이지만, 국소적으로는 데이터가 선형적으로 연관되어 있음, 국소적 관계가 가장 잘 보존되는 훈련 세트의 저차원 표현을 찾을 수 있음, 사영이 아닌 다양체 학습에 의존

-사이킷런에서 지원하는 모델: 다차원 스케일링 등