Pair Programing

10주차

1조
201984014
성도윤
201904126
허준혁

Q1. 차원의 저주는 무엇인가?

A1. 학습이 어려워지는 것?
-> 샘플의 특성이 많아진다
-> 그러면 어떻게 해결해 나가는건지?
-> 특성 수를 줄인다
-> 이게 차원 축소!!
-> 하지만 훈련 속도는 빨라지고, 데이터 시각화에도 유용하지만 일부 정보가 유실되어 성능이 저하 될 수 있다

Q2. 특성 수가 많아지면 학습이 어려워지는 이유는?

A2. 테스트해야할 것이 많나?
-> 샘플들의 사이가 멀어진다
-> 과대적합!
-> 거리가 멀면 예측이 어렵다
-> 해결 방법은?
-> 샘플 수를 늘린다
-> 샘플 수를 많이 준비하기 어려운 고차원이면 불가능..

Q3. PCA는 뭘까?

A3. 다차원 데이터의 차원을 줄이고 데이터를 새로운 좌표계로 변환하는 통계적 기법

-> 분산 보존, 주성분 개념이 중요하다
-> 분산이 최대로 보존되는 축 선택해야한다
-> 왜지..?
-> 손실이 적다

Q4. 적절한 차원수는?

A4. 분산 비율의 합이 충분한 분산이 되도록 하는 값
-> 차원을 축소하면 훈련세트의 크기가 줄어든다

Q5. pca = PCA(n_components=154)의 의미는?

A5. MNIST 데이터셋을 154차원으로 압축한다
-> 차원 축소 코드
-> 그러면 차원 복원 코드는?
-> pca.inverse_transform(X_reduced)
-> inverse_transform 함수!

Q6. 랜덤 PCA는?

A6. 확률적 알고리즘 사용할 것 같다
-> 주성분의 근삿값을 빠르게 찾는다

-> d가 n보다 많이 작으면 완전 SVD보다 훨씬 빠르다
-> 코드로 이걸 적용하는 방법은?
-> svd_solver='randomized'로 지정
-> 코드 어렵다...