Pair Programing

11주차

1조
201984014
성도윤
201904126
허준혁

Q1. 군집이란?

A1. 비슷한 샘플을 클러스터로 모으는 것
 -> 데이터 분석, 고객 분류, 추천 시스템, 검색 엔진 등에 사용하는 훌륭한 도구이다

Q2. 이상치 탐지란?

A2. 정상 데이터가 어떻게 보이는지 학습하고 비정상 샘플을 감지하는 것이다

Q3. 밀도 추정이란?

A3. 데이터의 밀도를 예측하는 걸까?
 -> 데이터셋 생성 확률 과정의 확률 밀도 함수를 추정
 -> 밀도가 매우 낮은 영역에 놓인 샘플이 이상치일 가능성이 높다

Q4. k-평균 알고리즘은 무엇일까?

A4. 반복 몇 번으로 레이블이 없는 데이터셋을 빠르고 효율적으로 클러스터로 묶는 알고리즘
 -> 단점은 없을까?
 -> 군집의 크기가 서로 많이 다르면 잘 작동이 안된다
 -> 하드 군집과 소프트 군집 이란??

-> 하드 군집은 각 샘플에 가장 가까운 클러스터를 선택
-> 소프트 군집은 클러스터마다 샘플에 점수를 부여
-> 샘플별로 각 군집 센트로이드와의 거리를 측정

Q5. 좋은 모델을 선택하는 방법은?

A5. 다양한 초기화 과정을 실험한 후에 가장 좋은 것을 선택
-> n_init = 10 이 기본값으로 사용된다
-> 10번 학습 후 가장 낮은 관성을 갖는 모델을 선택

Q6. elkan 알고리즘은?

A6. algorithm=elkan: 불필요한 거리 계산을 많이 피함으로 학습 속도가 향상된다

Q7. 미니배치 k-평균은?

A7. 미니배치를 이용해서 센트로이드를 조금씩 이동한다
-> 특징은 뭐가 있을까?
-> k-평균 알고리즘보다 속도가 훨씬 빠르다

Q8. 실루엣 점수는?

A8. 모든 샘플에 대한 실루엣 계수의 평균

Q9. k-평균의 한계는?

A9. 속도가 빠르고 확장이 용이하지만 완벽하지 않다
 -> 최적이 아닌 방법을 피하려면 알고리즘을 여러번 실행해야 한다
 -> 클러스터 개수를 미리 지정해야한다
 -> 클러스터의 크기나 밀집도가 다르거나 특이한 모영일 경우 잘 작동하지 않는다

Q10. 군집을 이용한 이미지 분할에는 뭐가 있을까?

A10. 이미지 분할, 시멘틱 분할, 색상 분할

Q11. 군집을 사용한 준지도 학습은 언제하는가?

A11. 레이블이 없는 샘플이 많고 레이블이 있는 샘플이 적을 때 사용한다