BOAZ Spring Study : Weekly Assignment
(Due date : 2023.04.09)

[필기]

1. 다음 중 비지도학습에 대한 설명이 옳으면 O, 틀리면 X를 하고, 그렇게 생각한 이유를 적으세요. **(40 points)**

(1) PCA(Principle Components Analysis)를 하게 되면, 개별 데이터에 대한 분석을 더 정확하게 할 수 있다.

(2) PCA(Principle Components Analysis)는 데이터의 변수가 너무 많을 때, 차원 축소를 통해 데이터를 시각화해서 볼 수 있다. 이 과정에서 데이터의 정보가 소실될 수 있다.

(3) 거리 기반으로 하는 군집화 과정에서는 데이터의 밀도(분포)에 따라 영향을 받는다.

(4) K-Means Clustering에서는 데이터의 초기 중심점을 어떻게 잡느냐에 따라 다르게 군집이 형성될 수 있다.

(5) 군집화 기법 중 K-Means의 경우가 DBSCAN보다 계산량이 더 적다는 장점이 있다.

(6) A씨가 수집한 데이터 중, 키가 3000cm인 사람으로 기록된 부분이 있다. 이런 경우에는 데이터셋과 무관하다고 볼 수 있으므로 전처리 과정에서 제거하는 것이 더 좋다.

(7) K-Means Clustering에서 k의 값이 3일 때보다 1일 때가 더 이상치에 강건하다고 볼 수 있다.

(8) 어떤 데이터의 구성이 사람들의 연 소득, 키, 나이로 구성되었다고 하자. 이 데이터에 대해 K-Means Clustering을 한다고 할 때, 별도의 scaling을 하지 않아도 scaling을 하는 경우 보다 Clustering이 더 잘 된다.

(9) K-Means Clustering을 이용하여 군집화를 할 때, 최적의 결과(global minimum)을 보장할 수 있다.

(10) K-Means Clustering에서 k의 값이 1일 때 군집화를 시행한다고 가정해보자. 이때, 유클리디안 거리가 아닌 마할라노비스 거리를 사용하면 군집화의 결과가 달라질 수 있다.

2. L씨는 자기가 가져온 데이터에 군집화를 적용시키려 하는데, 데이터의 컬럼 수가 10000000개이다. 이때, 10개의 컬럼으로도 데이터를 99% 설명할 수 있다고 한다. 만약, L씨가 10000000개의 컬럼을 가지고 머신러닝 모델을 돌린다면, 어떠한 문제가 생길 수 있을지 쓰고, 그에 대한 해결책을 써보세요. **(15 points)**

3. 어떤 학교의 학부연구생인 K씨는 교수님과 함께 학생들의 강의를 듣는 패턴과 관련한 연구를 진행하고 있습니다. 수집된 데이터에는 학생들이 강의를 들은 시간, 강의를 들은 횟수, 강의를 언제 들었는 지, 그리고 해당 과목에 대한 학점, 이외에 학생들의 정보(학생의 학점, 학년, 학번 등등)이 있습니다. K씨는 연구 도중, 교수님으로부터 아래와 같은 요청을 받았을 때, K씨는 어떻게 하는 것이 좋을까요? **(15 points)**

■ 학생들을 몇 개의 그룹으로 나누는 게 좋을지 궁금하다. ■ 만약, 해당 그룹으로 나눴을 때, 학생들이 어떤 패턴으로 강의를 듣는 지 궁금하다.

[실습]

4. 노션에 올려놓은 두 가지 중 하나를 선택한 후, 해당 코드를 필사하세요.
필사 시, 코드에 대한 설명을 간략하게 주석 or Markdown의 형식으로 설명하세요.
하나는 Logistic Regression과 PCA에 대한 설명이고, 하나는 PCA와 Clustering에 대한 코드입니다. 평가 방식은 아래와 같습니다. **(30 points)**

■ 주어진 코드를 하나도 빠짐없이 복사/붙여넣기를 하지 않고 잘 필사하였는가?
■ 주어진 코드에 대한 설명이 잘 되어있는가?
■ 결과에 대한 해석을 잘 하였는가?