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.

PCA는 개별 데이터에 대한 분석을 정확하게 하기 힘듭니다. PCA를 하는 이유는 차원이 너무 많아져서 차원의 저주에 빠질 수 있기 때문에, 일부만 축소해서 보자는 의미를 가집니다. 즉, 전체적인 분포를 차원 축소를 통해 어떻게 분포가 되어 있는 지를 보기 위해 차원 축소를 하는 것이지, 개별 데이터에 대한 분석을 더 정확하게 하기는 힘듭니다. (X)

(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.

기존 컬럼에 있던 정보를 압축하는 과정이기 때문에, 정보는 당연히 소실될 수 있습니다. (O)

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

거리 기반으로 하는 군집화 과정에는 K-Means Clustering 방법이 있습니다. K-Means Clustering은 데이터의 밀도에 상관없이 오로지 거리를 기준으로 삼습니다. 즉, 밀도가 변하면 거리가 달라질 수 있기 때문에, 데이터의 밀도(분포)에 따라 영향을 받는다고 할 수 있습니다. (O)

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

K-Means Clustering을 할 때는 초기 중심점을 잡습니다. 초기 중심점을 기준으로 거리를 계산한 후 군집화를 진행하기 때문에, 만약 초기 중심점이 이상하게 잡힌다면, 군집화가 잘 이뤄지지 않습니다. (O)

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

K-Means 군집화 과정에서는 모든 데이터에 대해서 거리를 계산해야 합니다. 또한, 계산 결과 후 수정된 군집 중심점을 기준으로 또 한번 계산합니다. 반면에, DBSCAN의 경우에는 K-Means처럼 군집 개수를 지정하지 않아도 되기 때문에, 반복적인 계산을 하지는 않습니다. (X)

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

노이즈(Noise)와 이상치(Outlier)에 대한 문항입니다. 키가 3000cm인 사람은 데이터가 수집되는 과정에서 잘못된 정보가 수집되었기 때문에, 이상치가 아닌 노이즈에 해당이 됩니다. 이상치는 수집된 정보에는 오류가 없지만, 전체적인 분포에서 상위 0.x%, 하위 0.x%에 해당이 될 때 이상치라고 합니다.(0.x%는 제가 낸 기준일 뿐, 통상적으로 q3 + 1.5 * iqr이상이면 이상치라고 판단합니다.) (O)

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

K-means clustering에서 군집 개수를 1로 한다면, 이상치에 대한 부분을 잘못 계산할 수 있기 때문에, 강건하다고 볼 수 없습니다. (X)

(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.
scaling을 해야 합니다. 왜냐하면, 각각의 컬럼에 대한 범위는 계속 달라질 수 있고, 특히나 거리를 기반으로 한 군집화에서는 컬럼의 범위가 다르면, 거리 계산 시에 특정 컬럼에 대한 고평가를 할 수 있기 때문에, scaling을 하는 것이 좋습니다. (X)
(9) K-Means Clustering을 이용하여 군집화를 할 때, 최적의 결과(global minimum)을 보장할 수 있다.
앞선 내용과 비슷하게, K-Means clustering은 초기점을 어떻게 잡냐에 따라 결과가 달라지기 때문에 최적의 결과(Global minimum)을 보장할 수 없습니다. (X)
(10) K-Means Clustering에서 k의 값이 1일 때 군집화를 시행한다고 가정해보자.
이때, 유클리디안 거리가 아닌 마할라노비스 거리를 사용하면 군집화의 결과가 달라질 수 있다.
마할라노비스와 유클리디안 거리에서 거리의 차를 제곱한다는 점에서 동일하지만, 두 컬럼 간의 공분산(Covariance)를 곱한다는 점에서 차이가 있습니다.
하지만, K = 1일 때는 개별 데이터를 군집으로 묶습니다. 자기 것을 하나의 군집으로 묶는다는 점입니다. 따라서, 거리 척도를 다르게 가져가도, 결과가 달라지지 않습니다.

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)**
수많은 컬럼을 가지고 모델을 돌린다면, 돌아가지 않을 것이고, 차원의 저주에 빠질 수 있습니다. 너무 많은 컬럼을 포함하는 것이 좋지 않습니다. 더군다나, 이 상황은 10개의 컬럼으로 전체 데이터의 99%를 설명할 수 있기 때문에, 1%를 설명하기 위해서 나머지 컬럼들을 모두 사용하는 것은 좋은 방법이 아닙니다. 따라서, 성능이 조금은 낮아지더라도 10개의 컬럼을 이용해서 빠른 속도의 학습을 하는 것이 더 좋습니다. 그러기 위해서는, 10000000개를 10개의 컬럼으로 차원 축소(Dimension reduction)를 진행을 해서 머신러닝 모델을 돌려주는 것이 좋습니다.

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.

학생들을 몇 개의 그룹으로 나누는 것이 좋을 지에 대한 부분은, 먼저 K-Means clustering을 진행을 해서, 몇 개의 군집이 최적인 지를 결정하는 plot을 그려봅니다. 해당 plot에서 elbow point를 정한 다음에 군집화를 진행할 k를 정하는 것이 좋습니다.

또한, k가 정해지고 나서, 해당 군집의 학생들의 강의 듣는 패턴을 보기 위해서는 차원 축소를 해서 시각화를 한 다음에 어떤 요인이 차원 축소를 하는 데 있어서 가장 많은 영향을 주었는 지를 보면서 분석하는 것이 좋습니다.