





(보안데이터분석) 연습문제_08

1. 다음 중 군집 분석의 일반적인 특징으로 가장 **거리가 먼** 것은 무엇인가?
 1. 데이터 내에 숨겨진 의미 있는 패턴이나 구조를 발견하는 것을 목표로 한다.
 2. 비지도 학습 방법으로, 사전에 정의된 클래스 레이블 없이 데이터를 그룹화한다.
 3. 각 군집 내의 개체들은 서로 유사한 속성을 가지며, 다른 군집의 개체들과는 상이한 속성을 갖도록 그룹화한다.
 4. 분석 결과는 항상 유일한 해답을 가지며, 사용자의 주관적인 판단이나 알고리즘 선택에 영향을 받지 않다. 
 5. 데이터 탐색(exploratory data analysis) 단계에서 데이터의 특성을 이해하고 새로운 가설을 설정하는 데 유용하게 활용될 수 있다.
2. 다음 중 군집 분석에서 개체 간의 유사성 또는 비유사성을 측정하는 데 사용되는 방법이 **아닌** 것을 고르시오?
 1. 유클리드 거리 (Euclidean Distance)
 2. 코사인 유사도 (Cosine Similarity)
 3. 마할라노비스 거리 (Mahalanobis Distance)
 4. 주성분 분석 (Principal Component Analysis) 
 5. 맨하탄 거리 (Manhattan Distance)
3. 다음 중 마할라노비스 거리(Mahalanobis Distance)에 대한 설명으로 가장 **부적절한** 것은 무엇인가?
 1. 데이터의 공분산 구조를 고려하여 변수 간의 상관관계를 반영한 거리 측정 방식.
 2. 각 변수의 표준편차로 스케일링(scaling)한 후 유클리드 거리를 계산하는 것과 유사한 효과를 가짐.
 3. 이상치(outlier)에 덜 민감하게 반응하며, 데이터 분포의 형태를 고려하여 거리를 측정.
 4. 두 벡터 간의 거리가 작을수록 통계적으로 더 유사한 분포를 갖는다고 해석할 수 있다.
 5. 변수 간에 상관관계가 존재하지 않는 경우, 마할라노비스 거리는 유클리드 거리와 동일한 값을 갖게 된다. 
일반적인 유클리드거리가 아닌 각 변수의 표준편차로 나눈 표준화된 유클리드 거리가 계산되어 두 값은 다르게 나온다.

4. 다음 중 K-평균(K-means) 군집화 알고리즘에 대한 설명으로 가장 **옳지 않은** 것은 무엇인지 고르시오?

1. K-평균 알고리즘은 비지도 학습 방식으로, 초기 군집 중심(centroid)을 임의로 설정하여 시작한다.
2. 각 데이터 포인트를 가장 가까운 군집 중심에 할당한 후, 각 군집의 평균 벡터를 새로운 군집 중심으로 업데이트하는 과정을 반복.
3. 군집의 개수 K는 알고리즘 실행 전에 사용자가 미리 정의해야 하는 필수적인 파라미터이다.
4. K-평균 알고리즘은 볼록한(convex) 형태의 군집을 찾는 데 효과적이며, 군집의 크기가 균일하지 않거나 복잡한 형태를 갖는 경우 성능이 저하될 수 있다.
5. 알고리즘의 결과는 초기 군집 중심 설정에 영향을 받지 않으므로, 매번 실행하더라도 동일한 군집화 결과를 얻을 수 있다. 

5. 다음 중 계층적 군집화(Hierarchical Clustering)에 대한 설명으로 가장 **적절하지 않은** 것은 무엇가?

1. 계층적 군집화는 데이터 개체 간의 유사성을 기반으로 병합(agglomerative) 또는 분할(divisive) 방식을 사용하여 계층적인 군집 구조를 형성.
2. 병합적 군집화는 각 개체를 하나의 군집으로 시작하여 유사한 군집끼리 순차적으로 병합해 나가는 Bottom-up (상향식) 방식.
3. 분할적 군집화는 전체 데이터를 하나의 군집으로 시작하여 덜 유사한 부분들을 순차적으로 분리해 나가는 Top-down (하향식) 방식이다.
4. 계층적 군집화의 결과는 덴드로그램(dendrogram)이라는 트리 형태의 그림으로 시각화하여 군집 형성 과정을 직관적으로 이해할 수 있습니다.
5. 계층적 군집화는 K-평균과 달리 사전에 군집의 개수를 명확하게 지정해야만 분석을 수행할 수 있다는 특징이 있다. 

6. 다음 중 가우시안 믹스처 모델(Gaussian Mixture Model, GMM)을 사용한 군집 분석에 대한 설명으로 가장 **부적절한** 것을 고르시오?

1. GMM은 각 군집이 다변량 정규 분포(Multivariate Normal Distribution)를 따른다고 가정하고, 데이터가 여러 개의 가우시안 분포의 혼합으로 생성되었다고 모델링한다.
2. 각 데이터 포인트가 특정 군집에 속할 확률을 계산하여, 확률적으로 군집 할당을 수행하는 소프트 클러스터링(soft clustering) 방식이다.
3. GMM은 K-평균과 달리 군집의 형태가 반드시 구형(spherical)일 필요가 없으며, 타원형(elliptical)과 같이 다양한 형태의 군집을 잘 찾아낼 수 있다.
4. GMM의 학습은 일반적으로 기댓값 최대화(Expectation-Maximization, EM) 알고리즘을 통해 수행되며, 이는 초기 파라미터 설정에 민감하게 반응할 수 있다.

EM 알고리즘의 목표는 잠재 변수가 있거나 데이터가 불완전한 상황에서 likelihood 함수를 간접적으로 최대화하는 MLE를 수행하는 방법론이다. E-Step : 현재 파라미터 추정치를 기반으로 변수의 조건부 기댓값을 계산하거나 결측치를

추정하여 전체 데이터 셋을 완성하는 과정. M-step : E-step의 데이터셋을 사용하여 likelihood를 최대화하는 파라미터 추정치를 업데이트 한다. (이단계는 일반적인 MLE와 유사하다)

5. GMM을 사용할 때, 최적의 군집 개수는 데이터의 복잡성과 관계없이 분석가의 주관적인 판단에 의해 결정된다. ☒

하이퍼 파라미터이지만 주관적인 판단에 의존하지 않고 다른 평가 지표등을 사용하여 객관적으로 판단하여야 한다.

7. 다음 중 DBSCAN(Density-Based Spatial Clustering of Applications with Noise) 군집화 알고리즘에 대한 설명으로 가장 **부적절한** 것은 무엇가?

1. DBSCAN은 데이터 포인트의 밀도(density)를 기반으로 군집을 형성하며, 특정 반경(epsilon, ϵ) 내에 최소 개수(minPts) 이상의 이웃을 갖는 핵심(core) 포인트를 중심으로 군집을 확장한다.
2. 밀도 기반 군집화 방법으로, K-평균과 달리 군집의 모양이 원형이 아니어도 잘 작동하며, 임의의 형태를 가진 군집을 효과적으로 찾을 수 있다.
3. ϵ 반경 내에 있는 핵심 포인트의 이웃 포인트는 해당 핵심 포인트와 동일한 군집에 속하는 경계(border) 포인트가 되거나, 다른 핵심 포인트의 이웃이 될 수 있다.
4. ϵ 반경 내에 minPts 미만의 이웃을 가지며 어떤 핵심 포인트의 이웃도 아닌 포인트는 노이즈(noise)로 처리되어 어떤 군집에도 속하지 않는다.
5. DBSCAN은 데이터의 밀도가 균일하지 않은 경우에도 효과적으로 군집을 탐색할 수 있으며, 군집의 개수를 사전에 지정할 필요가 없다는 장점이 있다. ☒ 밀도가 균일하지 않으면 성능이 떨어진다. 밀도가 낮은 구역에서는 군집이 제대로 형성될 않을 수 있다.

8. 다음 중 군집 분석 결과를 평가하는 지표인 실루엣 점수(Silhouette Score)에 대한 설명으로 가장 **올바르지 않은** 것은 무엇인가?

1. 실루엣 점수는 각 데이터 포인트에 대해 계산되며, 해당 포인트가 자신의 군집 내에서 얼마나 밀집되어 있고, 다른 군집과는 얼마나 잘 분리되어 있는지를 나타낸다.
2. 특정 데이터 포인트의 실루엣 계수(silhouette coefficient)는 -1 부터 1 사이의 값을 가지며, 1에 가까울수록 해당 포인트가 자신의 군집에 잘 할당되었고 주변 군집과 잘 분리되어 있음을 의미.
3. 실루엣 계수가 0에 가까운 포인트는 다른 군집과의 경계 부근에 위치하거나, 두 군집에 모두 속할 가능성이 높다는 것을 의미.
4. 실루엣 계수가 -1에 가까운 포인트는 자신의 군집 할당이 잘못되었을 가능성이 높으며, 다른 군집에 더 잘 속할 수 있음을 의미.
5. 전체 실루엣 점수는 모든 데이터 포인트의 실루엣 계수의 최댓값으로 계산되며, 이 값이 클수록 군집화 결과가 좋다고 평가할 수 있다. ☒ 전체 실루엣 점수는 모든 데이터 포인트의 실루엣 계수의 평균값. 따라서 값이 클수록 좋다는 설명도 틀리다.

9. 다음은 GMM을 이용한 군집분석 예제 코드이다.

```

import numpy as np
import matplotlib.pyplot as plt
from sklearn.mixture import GaussianMixture
from sklearn.datasets import make_blobs

# 샘플 데이터 생성
X, y = make_blobs(n_samples=300, centers=3, cluster_std=0.60,
random_state=0)

# GMM 모델 초기화 및 학습
n_components = 3 # 군집 개수 설정
gmm = GaussianMixture(n_components=n_components, random_state=0)
gmm.fit(X)


# 각 데이터 포인트에 대한 군집 할당 예측
labels = gmm.predict(X)

# 군집 중심점 얻기
centers = gmm.means_

# 결과 시각화
plt.scatter(X[:, 0], X[:, 1], c=labels, cmap='viridis', s=50)
plt.scatter(centers[:, 0], centers[:, 1], c='red', s=200, alpha=0.7,
label='Centroids')
plt.title('GMM Clustering')
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.legend()
plt.show()

```

사이킷런(scikit-learn)의 GaussianMixture 클래스를 이용한 군집 분석에 대한 설명으로 가장 **부적절한** 것은 무엇인가?

- 1) GaussianMixture 클래스는 데이터가 여러 개의 가우시안 분포의 혼합으로 생성되었다고 가정하고, 각 군집을 다변량 정규 분포로 모델링한다.
- 2) n_components 파라미터는 군집의 개수를 사용자가 사전에 지정하는 데 사용되며, 이 값은 분석 결과에 중요한 영향을 미친다.
- 3) fit() 메서드를 사용하여 모델을 학습하면 각 가우시안 분포의 평균(means), 공분산(covariance), 그리고 각 군집의 가중치(weights_) 등의 파라미터가 추정된다.
- 4) predict() 메서드는 학습된 모델을 사용하여 각 데이터 포인트가 속할 가능성이 가장 높은 군집을 결정한다.
- 5) GaussianMixture 모델은 K-평균 알고리즘과 동일하게 항상 구 형태의 군집만을 찾아내며, 타원형과 같은 다른 형태의 군집은 탐지할 수 없다. 

10. 다음은 k-means를 이용한 군집분석 예이다.

```

import numpy as np
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.datasets import make_blobs

# 샘플 데이터 생성
X, y = make_blobs(n_samples=300, centers=4, cluster_std=0.60,
random_state=0)

# k-means 모델 초기화 및 학습
n_clusters = 4 # 군집 개수 설정
kmeans = KMeans(n_clusters=n_clusters, random_state=0, n_init='auto')
kmeans.fit(X)

# 각 데이터 포인트에 대한 군집 할당 예측
labels = kmeans.predict(X)

# 군집 중심점 얻기
centers = kmeans.cluster_centers_

# 결과 시각화
plt.scatter(X[:, 0], X[:, 1], c=labels, cmap='viridis', s=50)
plt.scatter(centers[:, 0], centers[:, 1], c='red', s=200, alpha=0.7,
label='Centroids')
plt.title('K-Means Clustering')
plt.xlabel('Feature 1')
plt.ylabel('Feature 2')
plt.legend()
plt.show()

```

사이킷런(scikit-learn)의 KMeans 클래스를 이용한 군집 분석에 대한 설명으로 가장 **부적절한** 것은 무엇인가?

- 1) KMeans 클래스는 각 데이터 포인트를 가장 가까운 군집 중심(centroid)에 할당하고, 할당된 포인트들의 평균으로 군집 중심을 업데이트하는 과정을 반복하여 군집을 형성한다.
- 2) n_clusters 파라미터는 군집의 개수를 사용자가 사전에 지정해야 하며, 적절한 n_clusters 값을 선택하는 것이 중요하다.
- 3) fit() 메서드를 사용하여 모델을 학습하면 각 군집의 중심 좌표 (cluster_centers_)와 각 데이터 포인트에 할당된 군집 레이블을 얻을 수 있다.
- 4) predict() 메서드는 학습된 모델을 사용하여 새로운 데이터 포인트가 어떤 군집에 속할지 예측한다.
- 5) k-means 알고리즘은 초기 군집 중심을 무작위로 설정하기 때문에, 매번 실행하더라도 항상 동일한 군집화 결과를 보장한다. 