







Multivariate Data Analysis Assignment #8

[Q1] 데이터셋 선정하기

Dataset: Kaggle Clustering 데이터셋 중 1개 선택

Kaggle 사이트의 Datasets 항목에서 “clustering”을 키워드로 검색하면 총 412개의 데이터셋이 아래와 같이 검색됩니다.

<https://www.kaggle.com/datasets?search=clustering>

412 Datasets		Hotness ▾	📄 🗪
	Credit Card Dataset for Clustering Arjun Bhasin · Updated 3 years ago Usability 5.9 · 1 File (CSV) · 340 KB · 1 Task	253	Silver ...
	Household Electric Power Consumption UCI Machine Learning · Updated 5 years ago Usability 7.5 · 1 File (other) · 19 MB	196	Silver ...
	Online Retail K-means & Hierarchical Clustering Manish Kumar · Updated 2 years ago Usability 7.1 · 1 File (CSV) · 7 MB · 1 Task	55	Bronze ...
	Unsupervised Learning on Country Data Rohan kokkula · Updated a year ago Usability 8.2 · 2 Files (CSV) · 5 KB · 1 Task	88	Silver ...
	Wine Dataset for Clustering Harry Wang · Updated a year ago Usability 10.0 · 1 File (CSV) · 4 KB	19	...
	Sample Sales Data Gus Segura · Updated 5 years ago Usability 7.1 · 1 File (CSV) · 78 KB	328	Bronze ...

이 중에서 군집화 후 각 군집에 대한 속성 분석이 유의미할(또는 재미있을) 것으로 판단되는 데이터셋 하나를 선정하고 본인이 해당 데이터셋을 선정한 이유를 설명하시오.

[K-Means Clustering]

[Q2] `clValid()` 함수를 사용하여 K-Means Clustering의 군집 수를 2개부터 K(데이터의 규모에 따라서 본인이 적절히 선정할 것)개까지 증가시켜 가면서 internal 및 stability 관련 타당성 지표 값들을 산출하시오. 총 소요 시간은 얼마인가? Dunn index와 Silhouette index 기준으로 가장 최적의 군집 수는 몇 개로 판별이 되었는가?

[Q3] [Q2]에서 선택된 군집의 수를 사용하여 K-Means Clustering을 10회 반복하고 회차마다 각 군집의 Centroid와 Size를 확인해보시오. 10회 반복 시 몇 가지 경우의 군집화 결과물이 도출되었으며 각 경우의 군집화는 몇번 반복되어 발생하는지 확인해보시오.

[Q4] [Q3]에서 가장 빈번하게 발생한 군집화 결과물에 대해서 각 군집별 변수들의 Radar Chart를 도시해보시오. Radar Chart상으로 판단할 때, 군집의 속성이 가장 상이할 것으로 예상되는 두 군집(군집 A와 군집 B로 명명)과, 가장 유사할 것으로 예상되는 두 군집(군집 X와 군집 Y로 명명)을 각각 선택하고 선택 이유를 설명하시오.

[Q5] [Q4]에서 선택된 군집 A와 군집 B에 대해 각 변수별 평균값 차이에 대한 통계적 검정을 수행하시오. 전체 변수 중에서 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중은 얼마인가? 또한 [Q4]에서 선택된 군집 X와 군집 Y에 대해서도 각 변수별 평균값 차이에 대한 통계적 검정을 수행하시오. 이 경우, 전체 변수 중에서 유의수준 0.05에서 값의 차이가 나타나는 변수의 비중은 얼마인가?

[Hierarchical Clustering]

[Q6] Distance matrix는 실습 자료에서 제공하는 spearman correlation을 이용한 방식으로 산출한 뒤, hclust() 함수의 method 옵션을 “single”과 “complete” 두 가지로 설정한 후 각각에 대한 dendrogram을 그려보고 비교해보시오.

[Q7] 각 Linkage로부터 생성된 Dendrogram으로부터 [Q2]에서 선정한 최적 개수의 군집을 각각 찾은 후 각 군집들의 변수값을 활용하여 Radar Chart를 도시해보시오. Radar Chart를 바탕으로 판단할 때, K-Means Clustering과 보다 유사한 결과물이 나오는 방식은 어떤 Linkage인지 본인의 생각을 바탕으로 서술해보시오.

[Q8] Single Linkage와 Complete Linkage을 사용한 군집 결과물 중 어떤 결과물이 [Q4]에서 찾은 K-Means Clustering과 더 유사한지를 정량적으로 측정할 수 있는 아이디어를 제시하고 해당 아이디어를 구현해서 비교해보시오.

[DBSCAN]

[Q9] dbscan() 함수의 eps 옵션과 minPts 옵션을 조정해가면서 [Q2]에서 선정한 최적 개수의 군집이 찾아지는 eps 값과 minPts 값을 찾아보시오.

[Q10] [Q9]에서 찾은 군집화 결과물에서 Noise로 판별된 객체의 수가 몇 개인지 확인해 보시오.

[종합]

[Q11] 이 데이터셋에 가장 적합한 군집화 알고리즘은 무엇이라고 생각하는지 본인이 생각한 근거를 이용하여 서술하시오.