**군집분석(Clustering)**

**군집분석(Clustering)**

**Theory**

군집 분석(Clustering Analysis)은 각 개체의 유사성을 측정하여 유사성이 높은 대상 집단끼리 분류하고, 군집에 속한 객체들의 유사성과 서로 다른 군집에 속한 개체간의 상이성을 규명하는 다변량 분석 기법이다. 군집 분석에 이용되는 자료는 **별도의 반응변수가 요구되지 않으며**, 오로지 개체간의 유사성(Similarity)에만 기초하여 군집을 형성한다.

**거리(Distance)**

군집분석에서 관측 데이터 간 유사성이나 근접성을 측정할 때 거리를 사용하는데, 사용되는 거리의 종류는 아래와 같다.

* 유클리디안 거리(Euclidean distance)
* 맨하탄 거리(Manhattan Distance)
* 민코우스키 거리(Minkowski Distance)
* 코사인 유사도(Cosine Similarity)

**계층적 군집분석(Hierarchical Clustering)**

**Theory**

계층적 군집분석은 n 개의 군집으로 시작해 점차 군집의 개수를 줄여 나가는 방법이다. 계층적 군집을 형성하는 방법에는 **합병형 방법(Agglomerative)과 분리형 방법(Divisive)**이 있다.

 Hierarchical Clustering은 Tree기반의 모델이다. 2차원의 데이터의 경우를 생각해보자. 2차원 데이터는 좌표로 가시적으로 군집을 시각화시킬수 있지만, 3차원은 보기가 힘들어진다. 그리고 4차원이 넘어서면, 시각화가 거의 불가능해진다. Hierarchical clustering은 이러한 3차원 이상의 군집에서도 dendogram을 통해 직관적인 cluster 형성이 가능하다는 장점이 있다. 가장 기본적인 Hierarchical Clustering을 먼저 확인해보자.

**분리형 계측적 군집분석**

* 최단연결법(Single linkage, nearest neighbor) : method = ‘single’

N\*n거리행렬에서 거리가 가장 가까운 데이터를 묶어서 군집을 형성한다.

군집과 군집, 데이터와의 거리 계산 시 최단거리를 계산하여 거리행렬을 수정한다.

수정된 거리행렬에서 거리가 가장 가까운 데이터 혹은 군집을 형성한다.

* 최장연결법(Complete linkage, farthest neighbor) : method=’complete’

거리 계산 시 최장거리를 계산하여 거리행렬을 수정하는 방법.

* 평균연결법(Average linkage) : method= ‘average’

거리 계산 시 평균을 거리로 계산하여 거리행렬을 수정하는 방법.

* 와드연결법(Ward linkage) : ‘ward’

군집 내 편차들의 제곱합을 고려한 방법으로 군집 간 정보 손실을 최소화하는 방법

* 그 외 : method = ‘centroid’, ‘median’

**합병형 계층적 군집분석**

 병합 군집은 계층적 군집을 만들어서 군집이 반복하여 진행되면서 모든 포인트들이 하나의 포인트를 가진 클러스터에서 마지막 클러스터까지 이동하여 최종 군집이 형성되는 것이다. 여기서 주의할 점은 병합군집(agglomerative clustering)은 군집이 한 번 형성되면 군집에 속한 개체가 다른 군집으로 이동할 수 없는 특성을 가지고있다는 점이다.

**군집화**

거리행렬을 통해 가장 가까운 거리의 객체들 간의 관계를 규명하고, 덴드로그램을 그린다.

군집의 수는 목적에 따라 다를 수 있찌만 대부분 5개 이상의 군집은 잘 활용하지 않는다.

**거리행렬 기준 덴드로그램 그리기 -> 가로선을 그어 군집의 개수 선택 -> 적절한 군집수 선정**

**비계층적 군집분석(Non-Hierarchical Clustering)**

**Theory**

비계층적 군집분석은 n개의 개체를 k개의 군집으로 나눌 수 있는 모든 가능한 방법을 점검해 최적화한 군집을 형성하는 것이다. 대표적으로 K-Means Clustering 군집분석이 있다.

**K-평균 군집분석 (Kmeans Clustering)**

주어진 데이터를 k개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차의 분산을 최소화하는 방식으로 동작한다.

1. 원하는 군집의 개수오 초기 값(seed)들을 정해 seed 중심으로 군집을 형성한 후
2. 각 데이터를 거리가 가장 가까운 seed가 있는 군집으로 분류한다.
3. 각 군집의 seed 값을 다시 계산한다.

**K-평균 군집분석의 특징**

* 거리 계산을 통해 군집화가 이루어지므로 연속형 변수에 활용이 가능하다.
* K개의 초기 중심값은 임의로 선택이 가능하며 가급적이면 멀리 떨어지는 것이 좋다.
* 초기 중심값의 선정에 따라 결과가 많이 달라진다.
* 초기 중심값으로부터 오차 제곱합을 최소화하는 방향으로 군집이 형성되는 탐욕적 알고리즘으로 안정된 군집은 보장하나, 최적이라는 보장은 없다.

**K-평균 군집분석의 장단점**

* **장점** : 알고리즘이 단순하며, 빠르게 수행되어 분석 방법 적용이 용의함,

계층적 군집분석에 비해 많은 양의 데이터를 다룰 수 있음.

내부 구조에 대한 사전정보가 없어도 의미있는 자료를 만들 수 있음.

* **단점** : 군집의 수, 가중치와 거리 정의가 어렵다.

사전에 주어진 목적이 없으므로 결과 해석이 어렵다.

잡음이나 이상치의 영향을 많이 받는다.

초기 군집 수 결정에 어려움이 있다.

**K-평균 군집분석의 클러스터 수를 정하는 법.**

1. 이니셔를 계산하여 더 이상 급격하게 이니셔가 낮아지지 않는 지점.
2. 실루엣점수를 계산했을 때, 모든 클러스터가 평균 실루엣 점수를 넘으며, 비교적 실루엣 점수가 비슷한 지점.

**혼합분포군집(Mixture Distribution Clustering)**

**Theory**

혼합 분포 군집은 모형 기반의 군집 방법이며, 데이터가 k개의 모수적 모형의 가중합으로 표현되는 모집단 모형으로부터 나왔다는 가정하에서 모수와 함께 가중치를 자료로부터 추정하는 방법을 사용한다.

K개의 각 모형은 군집을 의미하며, 각 데이터는 추정된 k개의 모형 중 어느 모형으로부터 나왔을 확률이 높은지에 따라 군집의 분류가 이루어진다.

흔히 혼합모형에서의 모수와 가중치의 추정은 EM 알고리즘이 사용된다.

**혼합 분포 군집모형의 특징**

* K-평균 군집의 절차와 유사하지만, 확률분포를 도입하여 군집을 수행한다.
* 군집을 몇 개의 모수로 표현할 수 있으며, 서로 다른 크기나 모양까지 잡아낼 수 있다.
* EM 알고리즘을 이용한 모수 추정에서 데이터가 커지면 수렴 시간이 조금 오래걸린다.
* 군집의 크기가 너무 작으면 추정의 정도가 떨어지거나 어려울 수 있다.
* 이상치에 민감하다.

**DBSCAN(밀도 기반 군집)**

**Theory**

밀도 기반 군집화의 대표적인 알고리즘인 DBSCAN은 간단하고 직관적인 알고리즘으로 되어 있음에도 데이터의 분포가 기하학적으로 복잡한 데이터 세트에도 효과적인 군집화가 가능합니다. DBSCAN은 특정 공간 내에 데이터 밀도 차이를 기반 알고리즘으로 하고 있어서 복잡한 기하학적 분포도를 가진 데이터 세트에도 군집화를 잘 수행합니다.

**주요 하이퍼 파라미터**

* EPS : 입실론 주변 영역의 반경을 의미합니다.
* Min\_samples : 핵심 포인트가 되기 위한 최소 데이터 개수