Clustering

Clustering

- Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set.
- We seek a partition of the data into distinct groups so that the observations within each group are quite similar to each other,
- It make this concrete, we must define what it means for two or more observations to be similar or different.
- Indeed, this is often a domain-specific consideration that must be made based on knowledge of the data being studied.

유형

- 계층적 방법
  - 개별 대상 간의 거리에 의하여 가장 가까이에 있는 대상들로부터 시작하여 결합해 감으로써 나무 모양의 계층구조를 형성해가는 방법.
  - 자료의 크기가 크면 분석하기 어려움
  - 최단 연결법, 최장 연결법, 중심 연결법, 평균 연결법 등
- 분할 방법
  - 구하고자 하는 군집의 수를 정한 상태에서 설정된 군집의 중심에 가장 가까운 개체를 하나씩 포함해 가는 방식으로 군집을 형성해가는 방법
  - 많은 자료를 빠르고 쉽게 분류 할 수 있음
  - K-means 군집 등

유사도

- 두 관측값 : $x = (x_1, \ldots, x_p), y = (y_1, \ldots, y_p)$
- 유클리드(Euclid) 거리

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_p - y_p)^2}$$

- 맨하탄 (Manhattan) 거리

$$d(x, y) = |x_1 - y_1| + \cdots + |x_p - y_p|$$

- 민코우스키 (Minkowski) 거리

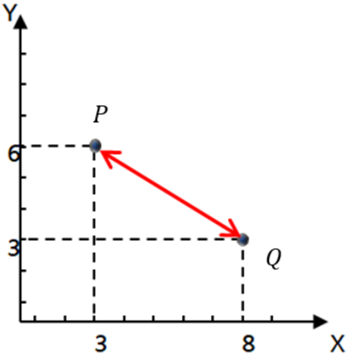
$$d(x, y) = \{(x_1 - y_1)^p + \cdots + (x_p - y_p)^p\}$$

- 마할라노비스 (Mahalanobis) 거리 (Σ:공분산)

$$d(x, y) = (x - y)^T \Sigma (x - y)$$

거리 - 예

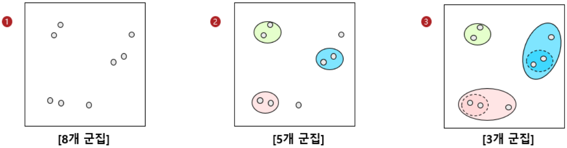
- 두 관측값 : $P = (3, 6), Q = (8, 3), \Sigma = \begin{pmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{pmatrix}$



데이터 준비 - 표준화

- 변수들의 값을 동일한 기준의 값으로 변환하여 단위에 대한 영향도를 제거하는 방법
- 데이터 : 자녀의 수, 신용카드의 수, 가구 소득, 자동차 보유수 등과 같이 서로 다른 단위들을 사용
- 표준화 : $Z = \dfrac{X - \bar{X}}{s}$

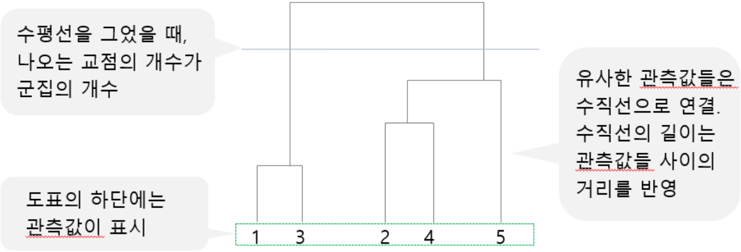
계층적군집분석 - 병합 알고리즘



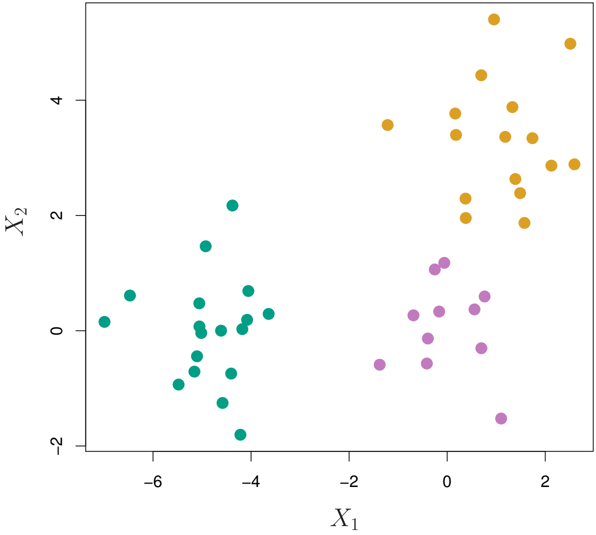
1. 각 관측값을 하나의 군집으로 간주하여 가장 거리가 가까운 두 군집을 병합하여 $n-1$개의 군집을 형성
2. 군집들 중에서 거리가 가장 가까운 두 군집을 병합
3. 2번 과정마다 군집의 개수가 하나씩 줄어드는데 모든 자료가 하나의 군집에 속할 때까지 2 반복

계층적군집분석 - 덴드로그램(Dendrogram)

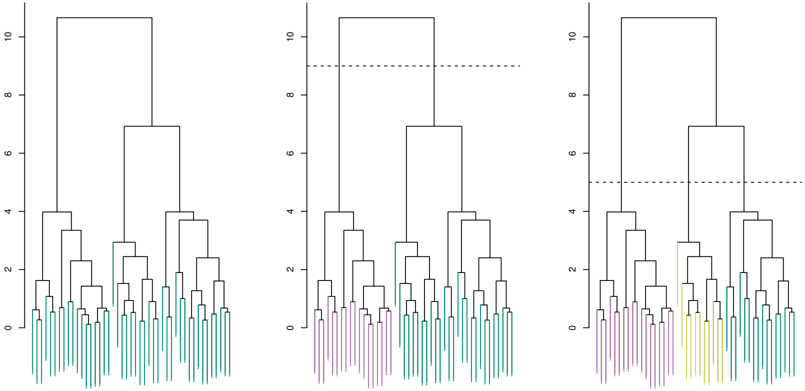
- 나무 형태의 도표를 활용하여 군집화 과정 및 결과를 시각적으로 간단하게 요약



계층적군집분석 - 예제 : 산점도



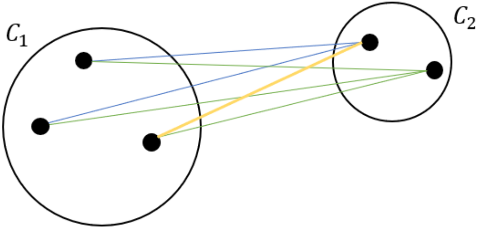
계층적군집분석 - 예제 : 덴드로그램



계층적군집분석 - 군집간의 거리 정의

1. 최단연결법

- Single Linkage Method
- 최인접 이웃 클러스터링 (nearest neighbor clustering)
- 두 군집 C_1, C_2 사이의 거리는 두 군집간의 최단 거리로 정의
- $d(C_1, C_2) = min\{d(x, y) : x \in C_1, y \in C_2\}$

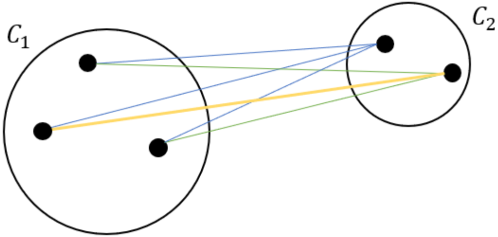


계층적군집분석 - 군집간의 거리 정의

- 최단연결의 특징
  - 같은 군집 내 속하는 관측값은 다른 군집에 속하는 관측값에 비하여 거리가 가까운 변수를 적어도 하나는 갖고 있음
  - 데이터들의 개형이 고리 모양으로 형성되어 있는 경우에는 군집이 부적절한 결과를 나타낼 수 있음
  - 새로운 관측값이 추가되는 경우, 기존 군집 내 관측값들 중 하나에 근접하기만 하면 됨

2. 최장연결법

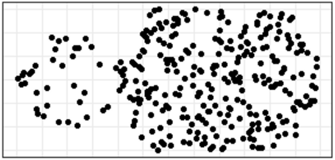
- Complete Linkage Method
- 가장 먼 이웃 클러스터링 (farthest-neighbor clustering)
- 두 군집 C_1, C_2 사이의 거리는 두 군집간의 최장 거리로 정의
- $d(C_1, C_2) = max\{d(x, y) : x \in C_1, y \in C_2\}$



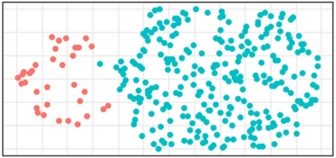
계층적군집분석 - 예제

	1	2	3	4	5
1	0				
2	7	0			
3	1	6	0		
4	9	3	8	0	
5	8	5	7	4	0

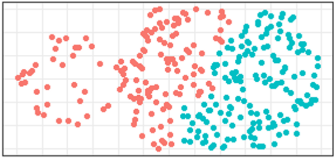
계층적군집분석 - 최장 vs. 최단



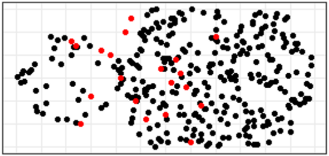
최단연결법



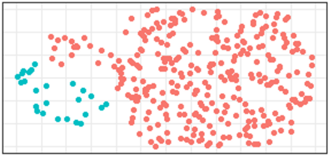
최장연결법



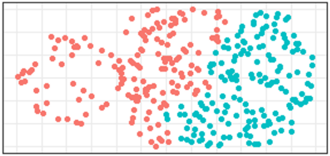
계층적군집분석 - 최장 vs. 최단



최단연결법



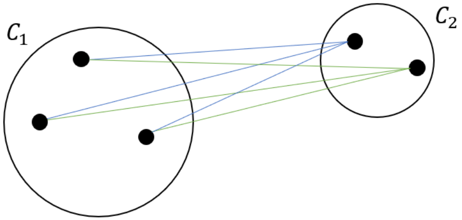
최장연결법



계층적군집분석 - 군집간의 거리 정의

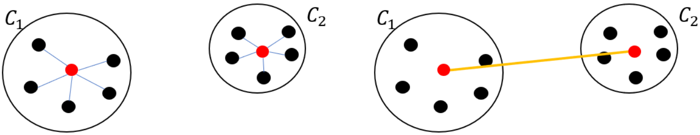
3. 거리평균 (Average distance)

- 최단/최장 연결법은 이상값이나 노이즈에 과하게 민감
- 이상값이나 노이즈에 대한 민감성 문제 해소
- $d(C_1, C_2) = average\{d(x, y) : x \in C_1, y \in C_2\}$



계층적군집분석 - 군집간의 거리 정의

4. Ward's Method
   - 각 집단 내 분산을 최소화
   - 이상치에 민감하나 각 집단간의 개체 수를 비슷하게 군집화
5. Centroid linkage
   - 각 집단 내 평균점 기준
   - 이상치의 영향을 적게 받음



계층적군집분석 - 단점

- 계산 속도 : $n \times n$ 거리행렬을 계산하고 저장해야 하므로 데이터가 큰 경우 횟수가 많아지고 계산 속도가 느려짐
- 안정성 : 데이터를 재정렬하거나 몇 개의 관측값을 제외시킬 경우, 전혀 다른 군집 결과가 나타날 수 있음
- 거리 선택 :군집간의 거리를 선택할 때 연결법에 따라 완전 다른 군집들이 형성될 수 있음
- 이상값에 민감

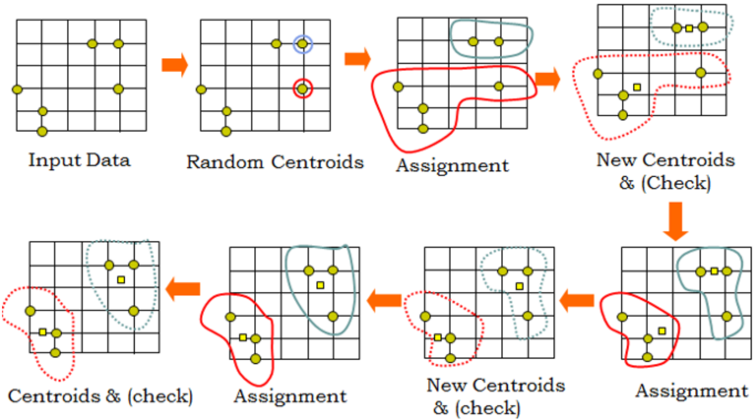
K-평균 군집분석 (K-means Clustering)

- 데이터 셋 D에 대해, 분할 $C_1, \ldots, C_K$는 다음의 성질을 만족
  - $C_1 \cup \cdots \cup C_K = D$: 하나의 관측값은 적어도 하나의 군집에 포함
  - $C_k \cap C_{k'} = \phi, \;\; k \neq k'$: 두 개 이상의 군집에 동시에 포함되는 관측값은 없음
- 군집내 응집도(within cluster variation : WCV)를 가장 작게 하는 군집으로 할당

$$\underset{C_1,\ldots,C_K}{\operatorname{argmin}} \left\{ \sum_{k=1}^{K} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

K-평균 군집분석

Simple example



K-평균 군집분석 - 알고리즘

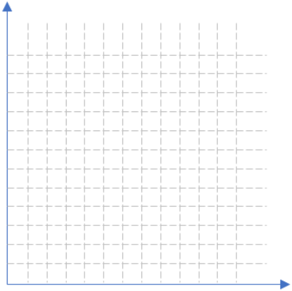
[Step 0] 초기 군집개수인 K를 선택

[Step 1] 데이터에서 K개의 중심을 랜덤하게 선택하여 거리를 계산한 후, 가장 가까운 것끼리 군집을 생성 (Assignment)

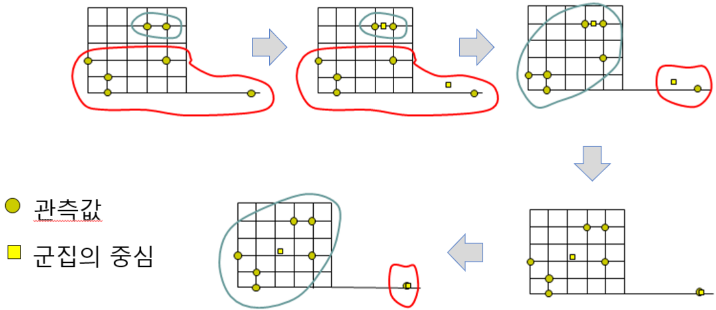
[Step 2] 각 군집에서 새로운 중심을 설정 (New Centroids)

[Step 3] 중심이 변하지 않을 때까지, Step 1부터 다시 반복 (Converage)

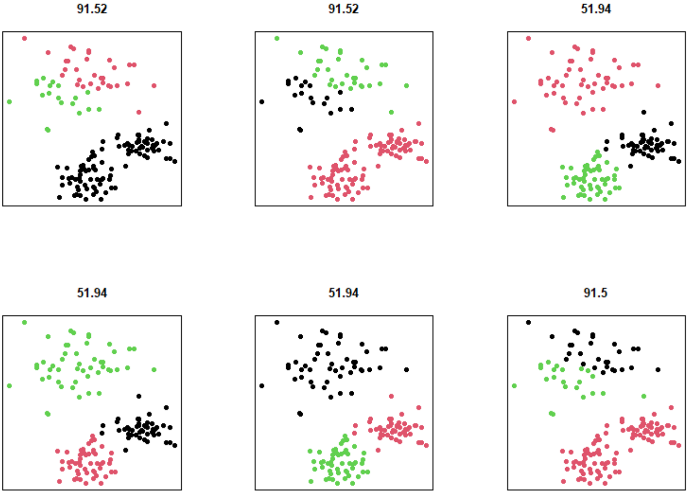
K-평균 군집분석 - 예제



K-평균 군집분석 - 이상값



K-평균 군집분석 - 초기 중심

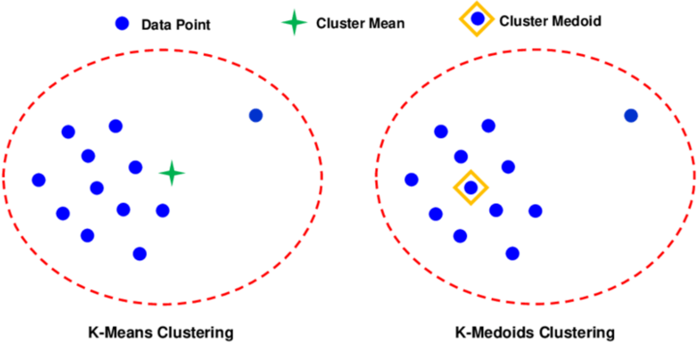


K-평균 군집분석 - 초기 군집 수의 결정

- K-평균 군집 방법의 결과는 초기 군집 수 K의 결정에 민감하게 반응함
- 여러 가지 K값에 대한 군집분석을 수행
- 관측값 간의 평균거리(Within Sum of Squares)와 군집간의 평균거리 (Between Sum of Squares) 비교를 통해 평가
- 자료의 시각화를 통한 최적 군집수의 결정 - 차원의 축소(PCA 등)

K-Medoids Clustering

- K-means 군집분석에서 평균 대신 중앙값을 중심위치로 선정
- $S = \{x_1, \ldots, x_n\}$: $x_{medoid} = \text{argmin}_{y \in S} \sum_{i=1}^{n} d(y, x_i)$



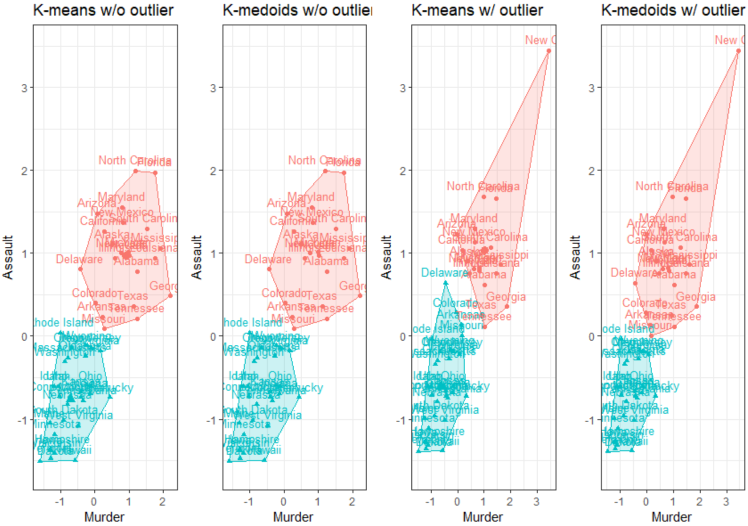
K-Means vs. K-Medoids

- 두 방법 모두 관측 자료 사이의 거리를 사용하여 군집 구성
- 이상치의 영향을 받음
- 군집 내 응집도를 최대로 하는 최적의 군집 구조를 찾을 수 없음 : 주어진 초기 군집 중심 위치로 부터 군집 내 응집도를 최대로 하는 군집 구조를 순차적으로 구성
- 사전에 군집 수에 대한 예상이 필요
- 초기 중심에 따라 군집 결과가 달라질 수 있음

K-Means vs. K-Medoids

- 일반적으로 K-means에 비해 K-Medoids가 더 많은 계산을 요구 : 자료 위치의 평균을 찾는 것보다 중심이 되는 자료를 찾는 것이 더 복잡
- K-Medoids가 K-Means보다 outlier의 영향을 상대적으로 덜 받음
- 자료에 Label이 있는 경우 K-Medoids가 해석하기 더 유리함

K-Means vs. K-Medoids - 예제



K-Medoids Clustering - Voronoi iteration

- K-means 알고리즘과 동일한 원리를 갖는 알고리즘

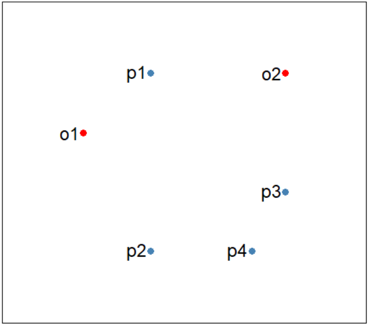
Step 1 K개의 초기 medoids를 임의로 결정

Step 2 각 K개의 medoids로부터 가까운 점들을 군집으로 할당

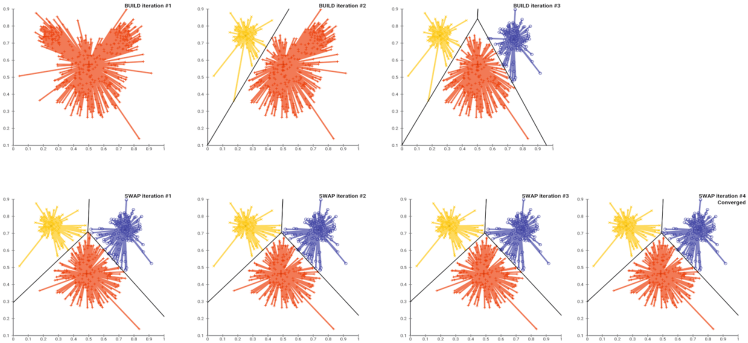
Step 3 결정된 군집에서 군집 내 응집도가 최개다 되도록 새로운 medoids 결정

Step 4 군집의 변화가 없을 때까지 위 과정 반복

K-Medoids Clustering - PAM



K-Medoids Clustering - PAM



K-Medoids Clustering - CLARA

- Clustering Large Applications
- Kaufmann and Rousseeuw, 1990
- 분석 대상 자료가 대용량 자료인 경우 자료로부터 적당한 랜덤 샘플을 추출하여 PAM 알고리즘을 적용

Step 1 자료로부터 M개의랜덤 샘플 추출

Step 2 추출한 자료로부터 PAM 알고리즘을 적용하여 K개의 medoids 선정

Step 3 선정된 medoids로부터 전체 자료의 군집 구성

K-Medoids Clustering - CLARANS

- Clustering Large Applications based upon Randomized Search
- Ng and Han, 1994
- CLARA 알고리즘에서 랜덤 샘플을 추출하는 과정을 여러 번 반복하는 알고리즘

Step 1 자료로부터 M개의랜덤 샘플 추출

Step 2 추출한 자료로부터 PAM 알고리즘을 적용하여 K개의 medoids 선정

Step 3 위 과정을 주어진 횟수만큼 반복하여 전체 자료의 군집을 구성하고 군집내 응집도가 가장 좋은 medoids를 선정

거리기반 군집분석의 평가 (1/3)

- Elbow Method
  - 군집의 수 K를 $1, 2, 3, \ldots$으로 변화시켜 가면서
  - 군집내 응집도 W_K를 계산하고
  - K를 가로축, W_K를 세로축으로 두어 그림을 그린다
  - 그림에서 W_K의 변화가 가장 작은 지점을 최적의 K값으로 결정

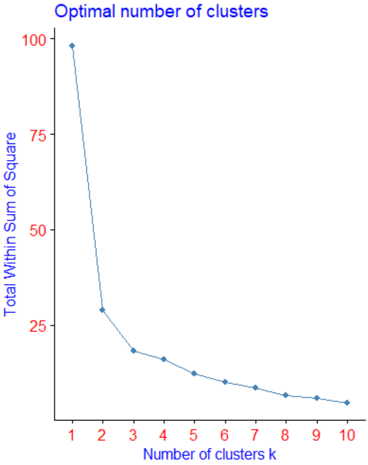
거리기반 군집분석의 평가 (2/3)

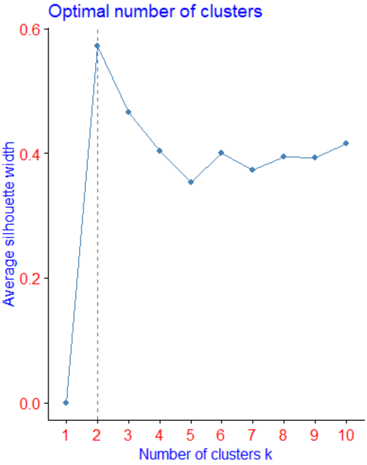
- Silhouette Method
  - 군집의 수 K를 $1, 2, 3, \ldots$으로 변화시켜 가면서
  - A : 각 자료에서 해당 자료가 속한 군집 내 모든 자료와의 거리의 평균
  - B : 각 자료에서 해당 자료가 속하지 않은 다른 군집 내 모든 자료와의 거리의 평균
  - 각 자료의 실루엣(Silhouette) 계산

$$S = \frac{B - A}{max(A, B)}$$

  - 모든 자료의 실루엣 값의 평균이 가장 큰 K를 최적의 K값으로 결정.

거리기반 군집분석의 평가 (3/3)

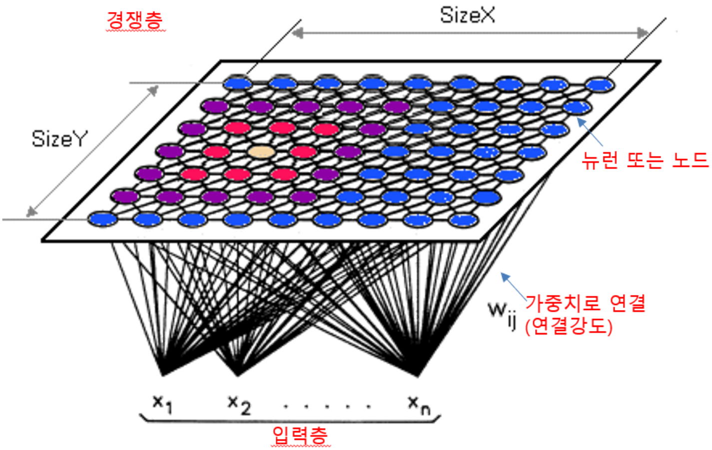




자기조직화지도 : self-organizing map (SOM)

- 비지도학습 인공신경망의 한 종류로써 입력데이터를 저차원 (일반적으로 2차원)의 공간으로 축소시키는 알고리즘
- 입력층(Input layer) : 입력변수의 개수와 동일한 뉴런수 존재
- 경쟁층(Competitive layer) : 2차원 격자(grid)로 구성되며, 사용자가 미리 정해 놓은 군집의 수만큼 뉴런 수 존재
- 입력층에 있는 자료는 학습을 통해 경쟁층에 정렬 (map)
- 입력층에 있는 각각의 뉴런은 경쟁층에 있는 각각의 뉴런과 연결되어 있음

자기조직화지도 : self-organizing map (SOM)



SOM : 알고리즘

[Step 1] SOM 맵 뉴런에 대한 가중 벡터(weight vector) 초기화

$$\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m), m : \text{노드 수}$$

[Step 2] 학습

[Step 2-1] 새로운 입력 벡터 제시 ($\boldsymbol{x}$)

[Step 2-2] 입력벡터와 각 뉴런의 가중 벡터와의 유사도 계산

$$D_i = \|\boldsymbol{x} - \boldsymbol{w}_i\|, \ \ i = 1, 2, \ldots, m$$

[Step 2-3] 최소 거리에 있는 출력 뉴런 선택 (BMU)

$$c = \underset{i=1,2,\ldots,m}{\operatorname{argmin}} \ D_i$$

[Step 2-4] 선택된 뉴런의 가중 벡터 재조정

$$\boldsymbol{w}_c^{new} = \boldsymbol{w}_c + \alpha(t)(\boldsymbol{x} - \boldsymbol{w}_c)$$

[Step 3] Step2 반복

SOM - 특징

- 경쟁 학습으로 각각의 뉴런이 입력 벡터와 얼마나 가까운지 계산하여 가중 벡터를 반복적으로 재조정하여 학습
- 가중 벡터 및 입력 패턴과 가장 유사한 경쟁층 뉴런이 승자
- 승자 독식 구조로 인해 경쟁층에는 승자 뉴런만이 나타나며, 승자와 유사한 연결 강도를 갖는 입력 패턴이 동일한 경쟁 뉴런으로 배열됨
- 수행속도가 매우 빠름
- 패턴 발견, 이미지 분석 등에서 뛰어난 성능을 보임
- 잠재적으로 실시간 학습 처리 가능

SOM - 예제

- 예제 : 4개의 입력벡터를 2개의 그룹으로 클러스터링
  - 입력벡터 : (1,0,0,1), (0,0,1,1), (0,1,0,1), (0,0,1,0)
  - 경재층의 노드 수 : 2개
  - SOM의 최대반복 횟수 설정
  - 유사도 : 유클리드 거리 $D_i = \|\boldsymbol{x} - \boldsymbol{w}_i\|^2 = \sum_{k=1}^{n}(x_k - w_{ik})$
  - 초기학습률 : $\alpha(1) = 0.6$, 한번의 연산이 끝나면 학습률 조정

$$\alpha(t+1) = 0.4 * \alpha(t)$$

밀도기반 군집분석

- DB scan (Density-based spatial clustering of applications with noise)

- 군집간의 거리를 활용하는 K-means 나 계층적 군집분석과는 달리 조밀하게 몰려 있어 밀도가 높은 부분을 군집화 하는 방식
- 한 점을 기준으로 ε 내에 점이 m개 이상 있으면 하나의 군집으로 인식
- 필요 인자 :
  - 기준점으로부터의 거리 : ε
  - 군집화를 위한 반경 내의 최소 데이터 수 : m (minPts)

밀도기반 군집분석 - 용어

- 중심점(core point) : ε 반경 내에 m개 이상의 점을 가지고 있는 점
- 경계점(boder point) : ε 반경 내에 m개 미만의 점을 가지고 있지만 다른 군집에 속한 점
- noise point : 중심점도 아니고, 어느 군집에도 속하지 않는 점 (outlier)

밀도기반 군집분석 - 예제 ($m = 4$)

- P_1 : ε 반경 내에 5개의 점이 있으므로 중심점 $\Rightarrow$ 군집화
- P_2 : ε 반경 내에 3개의 점이 있지만, P_1이 중심인 군집에 포함 $\Rightarrow$ 경계점
- P_3 : ε 반경 내에 4개의 점이 있으므로 중심점 $\Rightarrow$ 반경 내에 다른 중심점 P_1 이 포함되어 있는데, 이 경우 중심점 P_1과 P_3가 연결되어 있다고 하고 하나의 군집으로 묶임
- P_4 : noise point

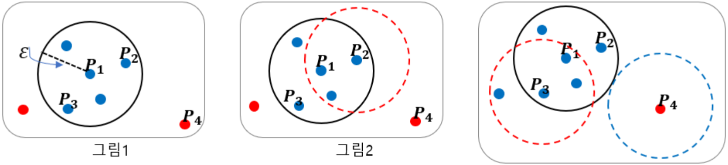


그림1 그림2 그림3

밀도기반 군집분석 - 알고리즘

[Step1] 모든 점의 ε 반경에 있는 점을 찾고 중심점 식별

[Step2] 중심점 중 연결된 구성요소 확인

[Step3] 중심점이 아닌 점들의 경우, 중심점의 ε 반경에 인접한 군집에 할당 그렇지 않은 경우 이를 noise point로 식별

밀도기반 군집분석 - 장단점

- 장점

- 모든 형태로 군집 가능
- 노이즈 조절 가능
- 사전 군집 수 설정 불필요

- 단점

- 군집 경계를 찾기 위해서는 밀도가 낮아지는 지점이 필요
- 실제 세계에서 정확한 군집 구조를 찾기가 어려움