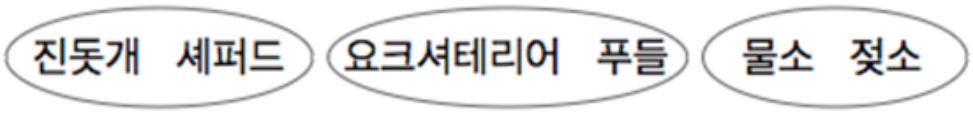
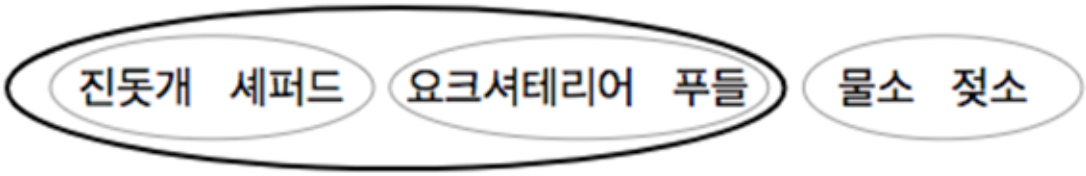
계층적 군집화 (Hierachial Clustering)

1. Hierarchial Clustering (계층 분석)

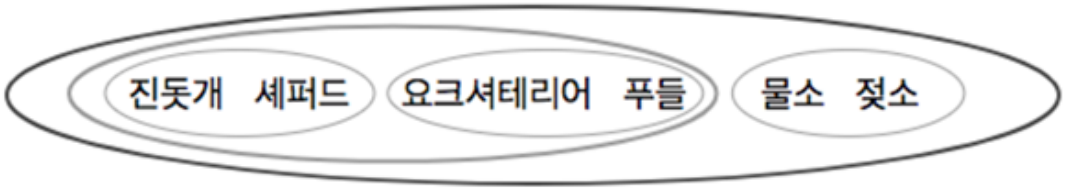
-최종적으로 하나의 군집이 될 때까지 비슷한 군집끼리 계속 묶는 클러스터링 방법



1) 중형견, 소형견, 소와 같이 세개의 군집으로 클러스터링



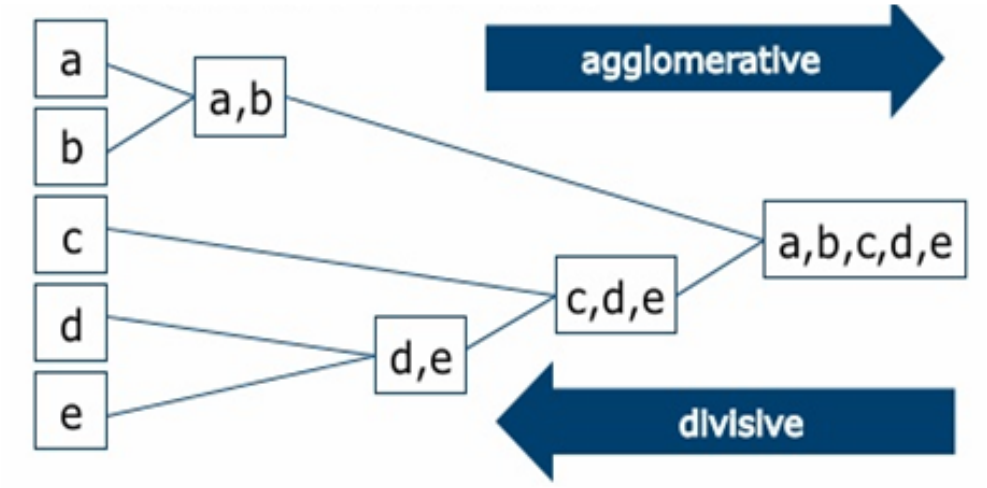
2) 중형견과 소형견을 개라는 하나의 군집으로 묶음



3) 개와 소를 동물이라는 하나의 군집으로 다시 묶음

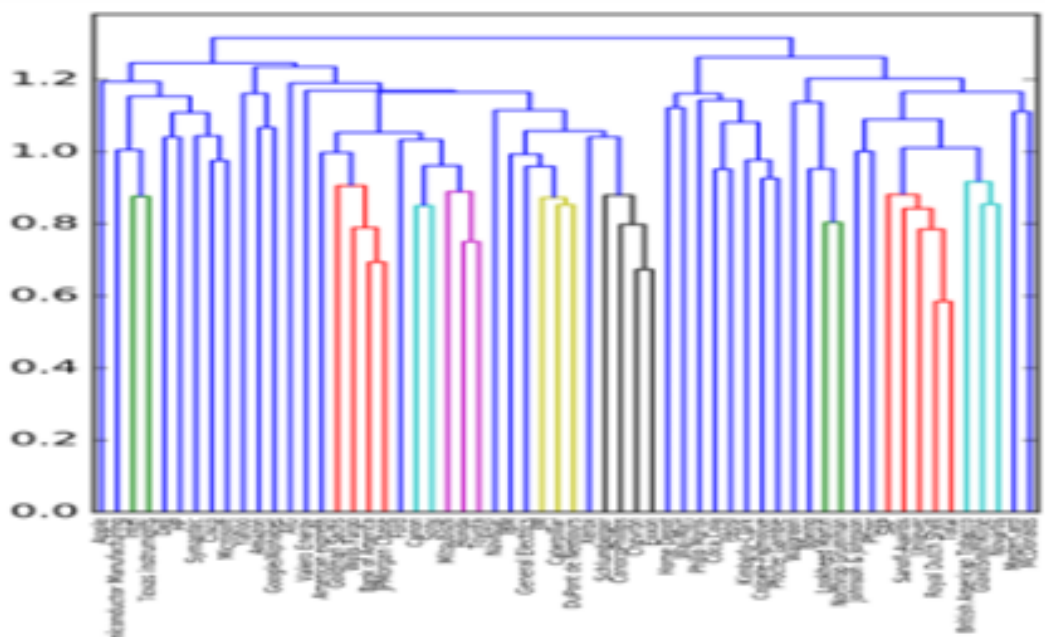
-각 클러스터가 다른 클러스터에 속해 있기(nested)때문에 계층적(hierarchial)이라 불림

-가장 큰 군집에서 시작해서 n개의 개체, 혹은 n개의 클러스터로 쪼개는 방식도 가능



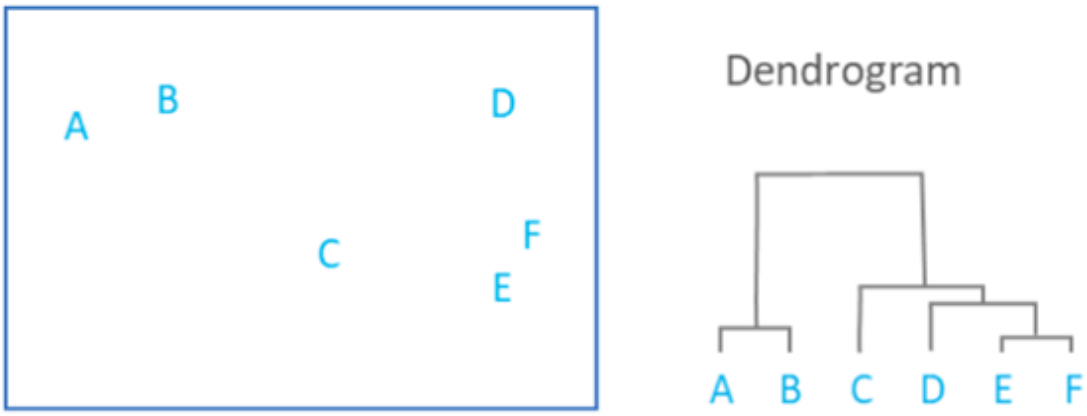
-코딩을 할 때 몇 개의 클러스터로 데이터를 정리할 것인지 정해주지 않아도 됨.

Dendrogram을 보고 '가장 적절한' 수준에서 그래프를 잘라 클러스터 수를 정함.



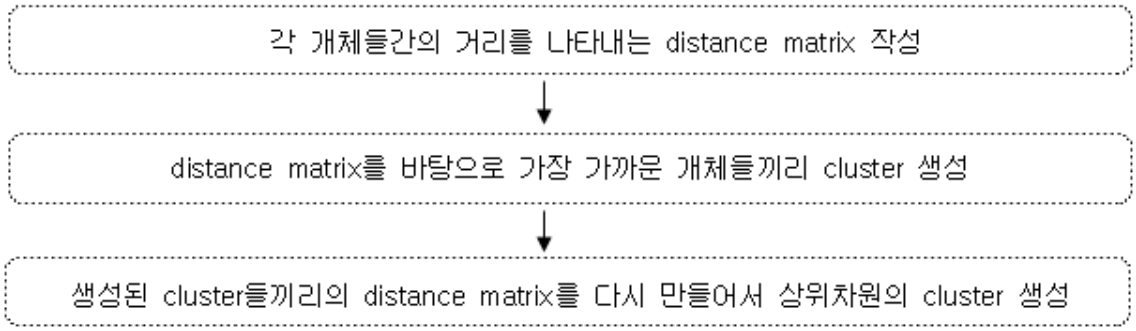
▲Dendrogram, 세로축은 distance를 나타낸다.

-Dendrogram은 distance matrix를 시각적으로 보여주는 것으로, 막대기의 세로 길이가 길수록 그 군집/개체들의 거리가 멀다고 생각할 수 있다.

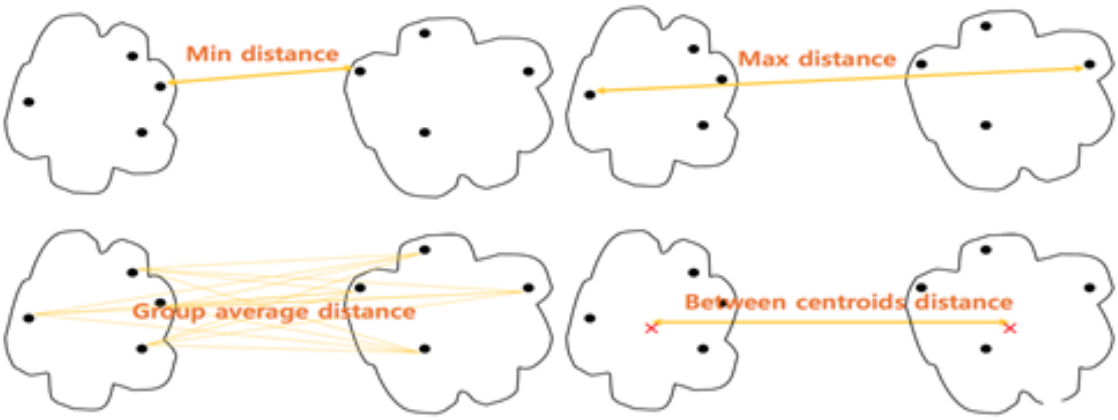


▲E는 F와, A는 B와 가까이 있어서 서로를 묶는 막대기의 길이가 짧은 반면에 (A,B)그리고 (C,D,E,F)는 서로 멀리 떨어져 있어 막대기의 길이가 더 길다.

-클러스터링이 실제로 진행되는 알고리즘은 다음과 같다:



-한편, 군집-군집 간 거리를 계산할 수 있는 방법들은 다음과 같다.



1) Min distance / Single link

클러스터 u의 모든 데이터 i 와 클러스터 v의 모든 데이터 j 의 모든 조합에 대해 측정된 거리 중 최솟값

2) Max distance / Complete link

클러스터 u의 모든 데이터 i와 클러스터 v의 모든 데이터 j의 모든 조합에 대해 측정된 거리 중 최댓값

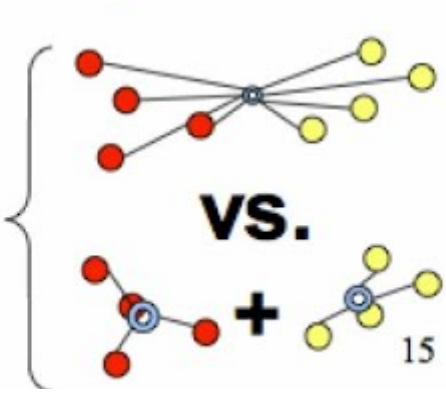
3) Group Average distance

클러스터 u의 모든 데이터 i와 클러스터 v의 모든 데이터 j의 모든 조합에 대해 측정한 거리들의 평균값

4) Between centroid distance

각 클러스터의 중심점(centroid)를 정의한 다음, 두 중심점의 거리를 클러스터 간의 거리로 정의

5) Ward's Method



위 4개의 거리와는 살짝 다른 개념으로, 서로 다른 두 군집 간의 유사성을, 두 군집이 만만 합쳐졌을 때 그 오차 제곱합(ESS)의 증가분에 기반해서 측정함. 즉, 거리 행렬(distance matrix)를 구할 때 오차 제곱합의 증분(increase of ESS)을 두 군집 사이의 거리로 측정하게 된다.

유사성 측도: ESS(error sum of squares)의 증분

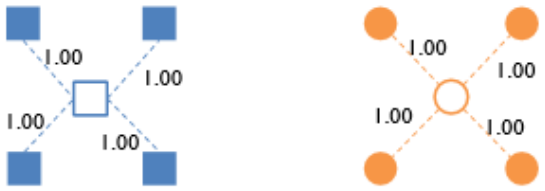
$$d(i+j,k)=\frac{\left\|\mu_{i+j}-\mu_{k}\right\|^{2}}{\frac{1}{n_1}+\frac{1}{n_2}}$$

,

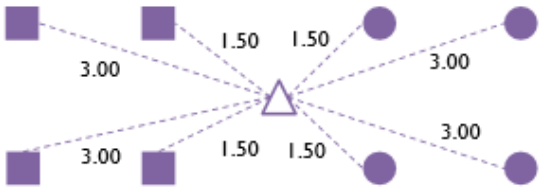
$$ESS=\sum_{k=1}^{K}\sum_{x_i\in C_k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_{kj})^2$$

* ESS : error sum of squares
** k : number of clusters (1 ~ K)
x_i : elements of cluster C_k
j : number of variables (1 ~ n)

- SSE before merge: $1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 = 8$



- SSE after merge: $4 \times 1.5^2 + 4 \times 3^2 = 45$



- Ward distance: **45-8 = 37**

단일 연결법이 noise나 outlier에 민감한 반면에, Ward 연결법은 노이즈나 이상치에 덜 민감한 장점이 있다. 그리고, Ward 연결법은 비슷한 크기의 군집끼리 묶어주는 경향이 있다.

중심연결법과 Ward 연결법의 유사성 측정 수식은 비슷하지만, 중심 연결법의 유사성 측도 대비 Ward 연결법에는 가중값이 추가되었다는 점이 다르다.

2. K 정하기

위에서 언급한 것처럼, hierarchial clustering에서는 사용자가 나중에 클러스터의 개수를 정하게 됨. 그렇다면 적절한 클러스터의 개수(k)를 어떻게 정할 수 있을까?

-기본적으로 클러스터링이 잘 되었다면, (Ward's Method를 제외하면)

i) 같은 클러스터 안의 data끼리는 거리가 짧아야 하며

ii) 다른 클러스터 안의 data끼리는 거리가 멀어야 한다.

따라서 우리는 Within-cluster sum of squares(WWS)와 Between-cluster sum of squares (BSS)를 측정해야 하며, 계산 방법은 아래에 나와있다.

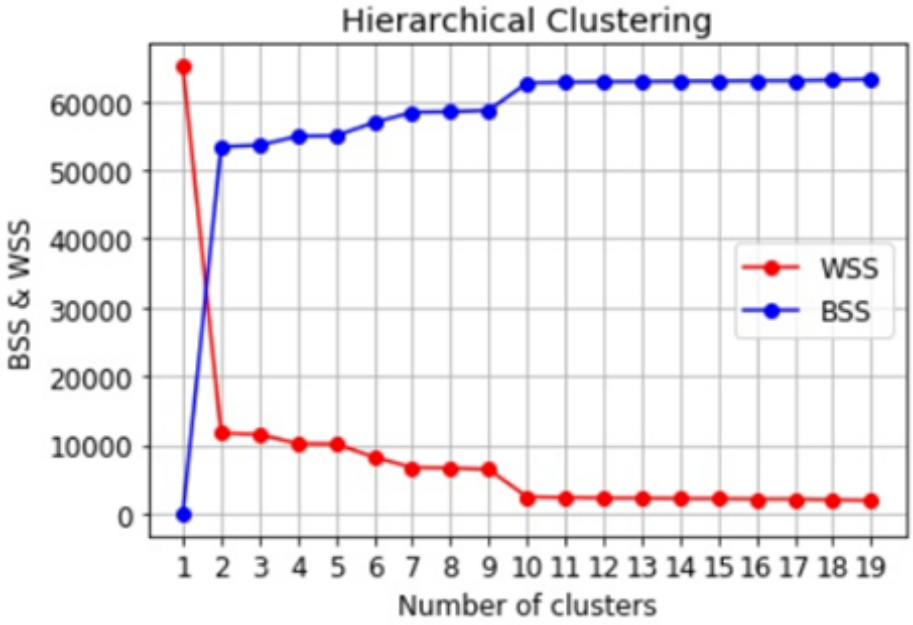
$$\mathrm{WSS}(C) = \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, \mu_i)^2$$

$$\mathrm{BSS}(C) = \sum_{i=1}^{k} |C_i| d(\mu, \mu_i)^2$$

▲μ_i는 cluster C_i의 centroid를 나타내며, μ는 전체 data set의 centroid이다.

위에서는 Euclidean 방식을 이용해 거리를 계산했다.

-Cluster의 개수에 따른 WSS와 BSS의 값을 모두 계산해 표로 살펴본 뒤, 가장 적합한 k를 구하면 된다.



▲k=2일 때나, 그 이상일 때나 WSS와 BSS 값이 크게 변하지 않는다.

따라서 k=2로 클러스터 개수를 정할 수 있다.

-단점 : i) 각 군집들이 어떤 특성을 가지고 있는지에 대해서는 알려주지 못함.

ii) 거리계산 방법이 달라질 때마다 추천 군집의 수가 달라짐.

iii) 군집으로 한 번 설정하면 군집 이동이 불가능함.

iv) 관측치가 150개 이하일 때 주로 사용됨. (계산시간이 오래 걸리기 때문)