강원지역혁신플랫폼

기계학습

Machine Learning

계층적 군집 분석

학습목표

- 계층적 군집 분석 개념을 이해하고, 설명할 수 있습니다.

01 | 군집 분석

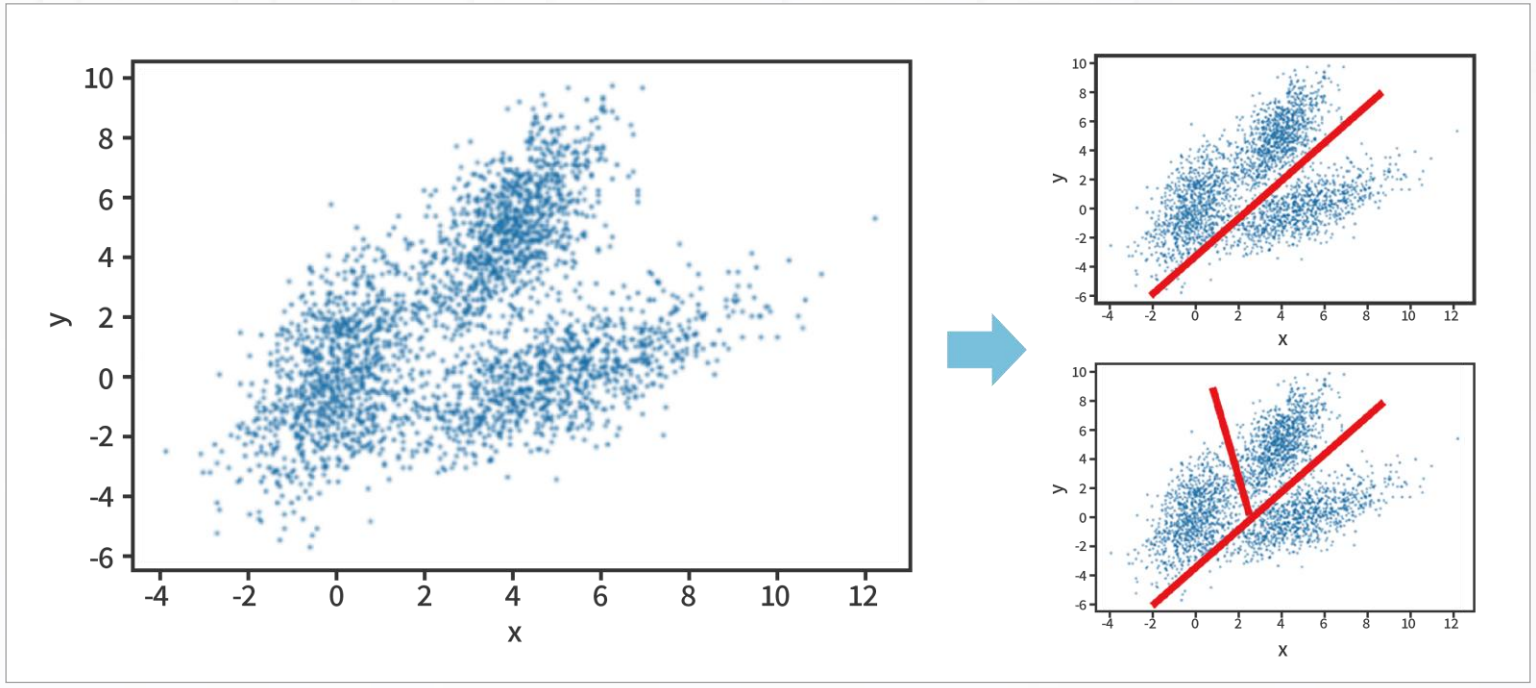
군집 분석 (Clustering)

- 군집 분석이란 개체들을 유사성에 기초하여 n개의 군집으로 집단화하여 집단의 특성을 분석하는 다변량 분석임
  - 즉, 군집 분석은 비슷한 특징을 가진 데이터들의 집합임
  - 군집 분석은 변수들이 속한 모집단 또는 범주에 대한 사전 정보(분류 기준)가 없는 경우 관측값들 사이의 유사성(거리)을 이용함
    - 개체들을 자연스럽게 몇 개의 그룹 또는 군집으로 나누어 그룹의 특성을 찾아내는 방법임

01 | 군집 분석

군집 분석은 라벨이 붙어져 있지 않은 데이터를 나누는 것이기 때문에
분석가는 몇 개의 군집으로 묶을지를 궁금해하지만 군집분석에는 정답이 없음

✦ 아래 그림처럼 하나의 데이터 셋을 두 개로 분류할 수도 있고 세 개로 분류할 수도 있기 때문임



01 | 군집 분석

- 군집 분석에는 P-값(p-value)처럼 의사 결정에 참고할 수 있는 검증 값도 없음
  - 어떤 변수와 분석법을 사용했는지에 따라 결과가 다르게 나타남
    - 그렇기 때문에 왜 이렇게 분류가 되었는지 이해할 수 있는 경험과 역량이 중요한 분석이기도 함

군집 분석에는 정답이 없음

01 | 군집 분석

군집 분석에서 **중요한 사항**은 아래와 같이 세 가지가 있음

1. **차원 축소**
2. **변수 종류** 및 이해
3. 사용하는 **방법론**과 **적용될 알고리즘**

01 | 군집 분석

1 차원 축소

✦ 유사한 변수들을 묶어서 차원을 축소함

› 예를 들어 변수들이 연봉, 나이, 연차, 직군 등 다양한 변수가 있는 경우에 나이와 연차는 비슷한 변수로 묶을 수 있음

- 이때는 나이를 제외하고 연차 하나만을 사용하는 것으로 차원을 축소할 수 있음
- 같은 군집 내에서는 차원이 동질적(homogeneous)이어야 함

01 | 군집 분석

2 변수 종류 및 이해

- **변수 종류**가 연속형 또는 명목형인지
  - 또한, **변수 개수**와 **특징**에 대한 **이해**가 **필요**함
    - **변수 종류**와 **특징**에 따라서 **사용되는 방법론**을 **고려**하게 됨

01 | 군집 분석

3 사용하는 방법론과 적용될 알고리즘

- **회귀 분석**에서는 **변수 자체**가 **중요**함
- **군집 분석**에서는 **거리**를 **어떻게 정의**하고 **측정할 것인지**가 **더 중요**함
  - 군집 분석의 **포인트**는 **이 거리**를 **측정하는 방법**이 **변수**의 **특성**과 **관계**가 **있기 때문**임
    - 만약에 사용할 **변수**의 **단위**가 **다를** 경우는 **표준화**가 **필요**함

01 | 군집 분석

- 군집 분석에서 거리를 계산하는 방법은 변수의 종류에 따라 다름
  - 물론 변수의 종류가 같은 경우에도 다양한 방법이 존재함
    - 연속형 변수(Continuous variable)의 경우 다음과 같은 거리 함수 등이 있음
      - 유클리드 거리 (Euclidean distance)
      - 맨허튼 거리 (Manhattan Distance)
      - 민코프스키 거리 (Minkowski Distance)
      - 통계적 거리 (Statistical Distance)
      - 마할라노비스 거리 (Mahalanobis distance)

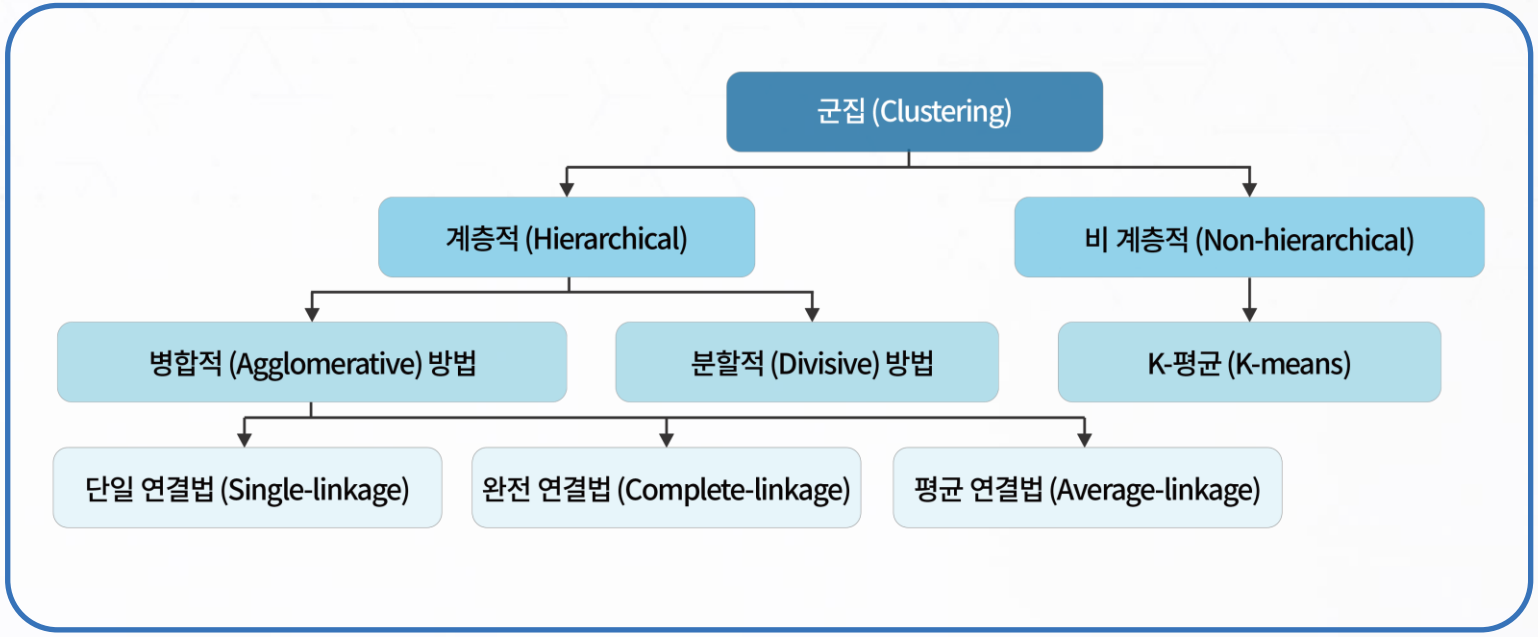
01 | 군집 분석

› **명목형 변수**(Categorical variable)의 경우 다음과 같은 거리 함수 등이 있음

- 자카드 거리 (Jaccard distance)
- 해밍 거리 (Hamming distance)
- Anderberg Coefficient
- Ochiai Coefficient
- Simple Matching Coefficient
- Rogers and Tanimoto Coefficient
- Sørensen–Dice Coefficient

01 | 군집 분석

› 다음은 군집 분석의 종류를 도식화한 것임



군집의 종류

01 | 군집 분석

- 계층적 군집에서 **군집 결합 기준**(Clustering linkage criteria)은 다음과 같음
  - **계층적 군집**에서 **군집 간**의 **결합 기준**으로 **다양한 방법**을 **사용**할 수 있음
    - Centroid
      - 각 군집의 **중심**을 **기준**으로 **결합**함
      - 이는 군집의 **평균 위치**를 **사용**함
    - Median
      - 각 군집의 **중간 값**을 **기준**으로 **결합**함
      - 이는 데이터의 **중앙 위치**를 **사용**함

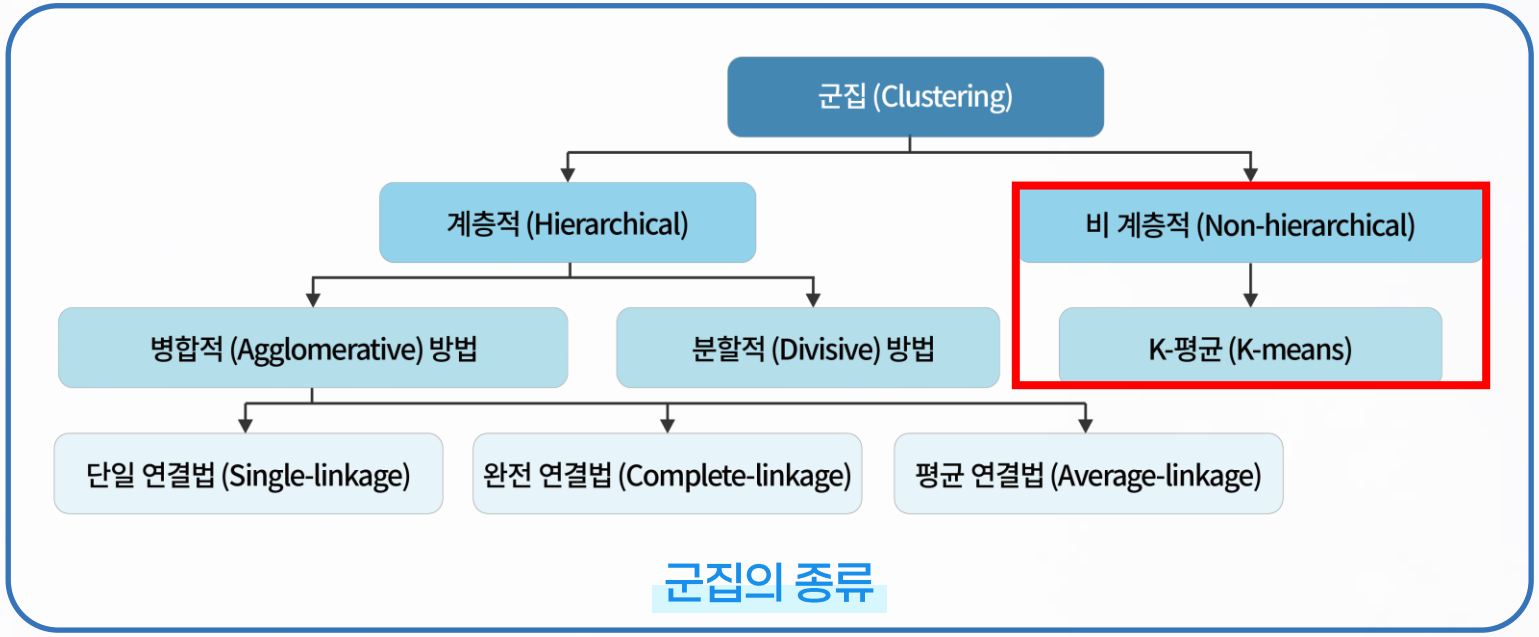
01 | 군집 분석

Ward

- 두 군집을 결합함으로써 증가하는 총 제곱 오차(Total Squared Error, TSE)를 최소화하는 방식으로 군집을 결합함
- 이는 데이터의 내부 분산을 최소화하는 데 중점을 둠

비계층적 군집 (Non-Hierarchical Clustering)

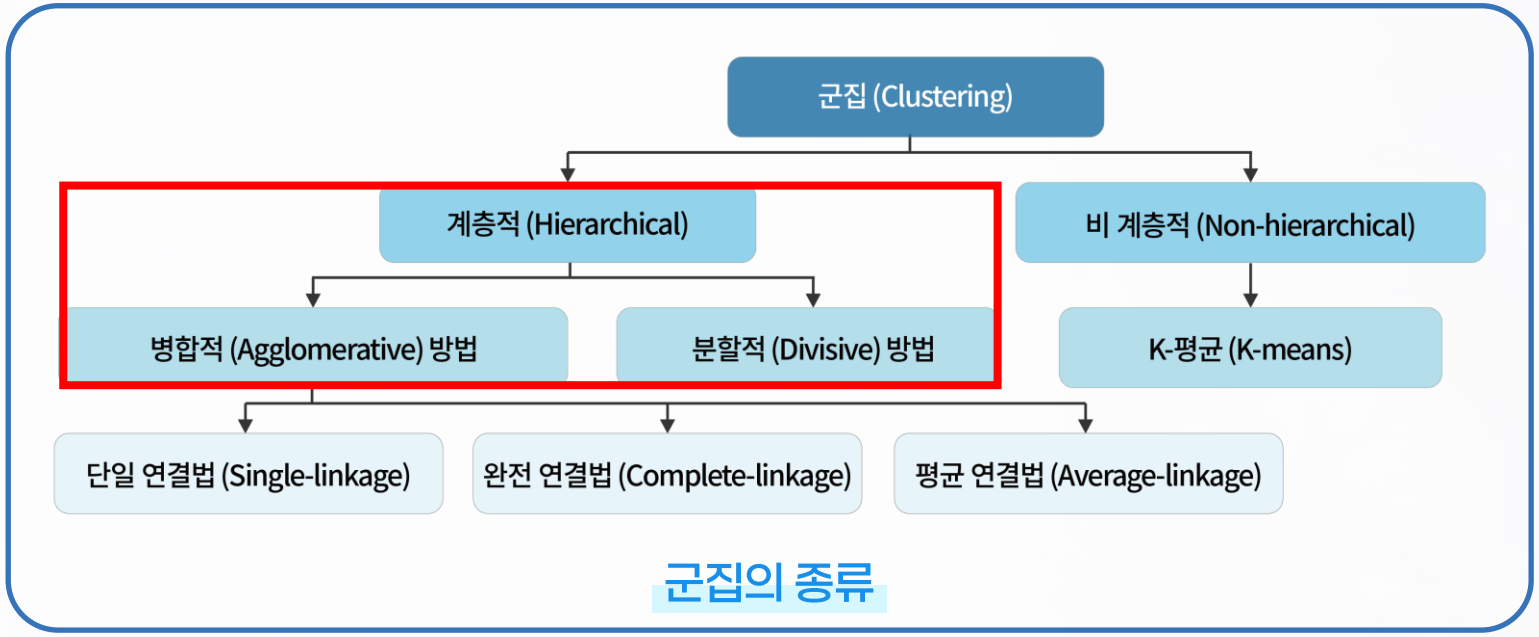
- 다변량 자료의 산포를 나타내는 여러 측도를 이용하여 판정 기준을 최적화시키는 방법으로 군집을 나누는 방법임
  - 한 번 분리된 개체도 반복적으로 시행하는 과정에서 재 분류될 수 있음
    - 사전에 군집의 개수가 정해져 있을 경우에 사용함
    - 대표적인 방법으로는 K-평균 군집(k-mean clustering)이 있음



군집의 종류

계층적 군집 (Hierarchical clustering)

- 고객 관리에서 사용자들의 특징에 따라 분류하는 경우를 생각해 보자.
  - 사용자들을 신규, 이탈, 복귀 고객처럼 접속한 시점과 마지막 접속에 따라 분류 기준을 정할 수도 있을 것임
    - 이 경우는 기준점이 명확한데 비해 어떤 경우에는 고객들을 분류하는 기준을 모르는 경우도 있음



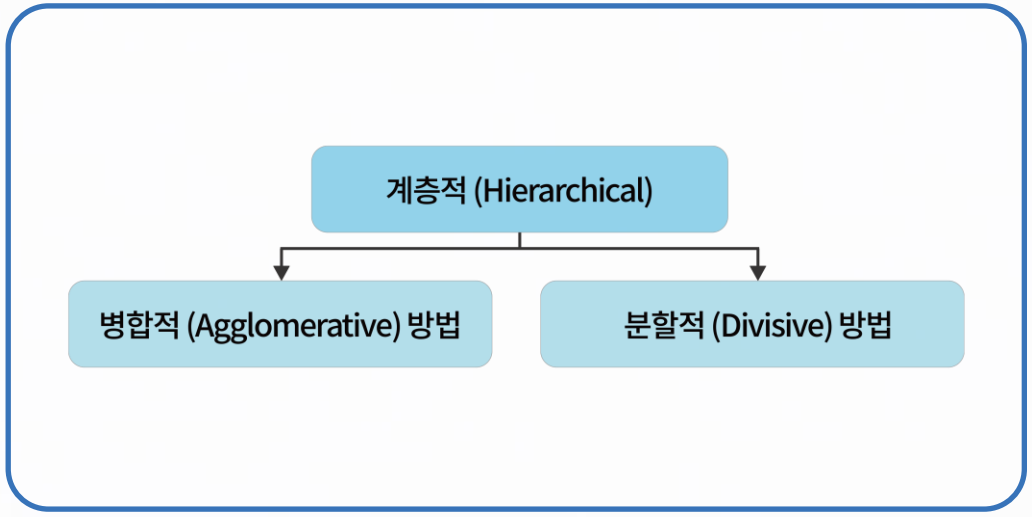
군집의 종류

03 | 계층적 군집

- 예를 들어 이미 **자사의 제품**을 **사용**하고 있는 경우 **충성**, **보통**, **불량** 고객으로 **나눈다**고 했을 때를 생각해 보자.
  - ✦ 이 **분류 기준**을 **어떻게 정해야 하는지**부터 **고민**이 될 수 있음
    - › **이러한 문제**를 **해결하기 위한 방법**으로 **계층적 군집**을 **활용**할 수도 있을 것임

03 | 계층적 군집

- 계층적 군집은 **군집**의 **개수**가 **정해지지 않았을 때** 사용함
  - 군집의 개수를 모를 때 사용하기 때문에 **몇 개**의 **군집**으로 **나누어야 하는지 결정**하기 **위해 사용**하기도 함
  - 계층적 군집은 아래 그림과 같이 **병합적 방법**과 **분할적 방법**이 있음
    - **병합적** 방법은 **가까운 개체**끼리 차례로 **묶어서** 차례로 **분리**하는 방법임
    - **분할적** 방법은 전체 데이터 집합을 **하나의 군집**에서 **시작**하여, **멀리 떨어진 개체**를 차례로 **분리**하는 방법임



03 | 계층적 군집

- 계층적 군집은 구현이 간단하고 이해하기 쉬움
  - 덴드로그램과 같은 그래프로 결과를 직관적으로 이해할 수 있음
  - 한 번 병합된 개체는 다시 분리되지 않는다는 특징을 가지고 있음

03 | 계층적 군집

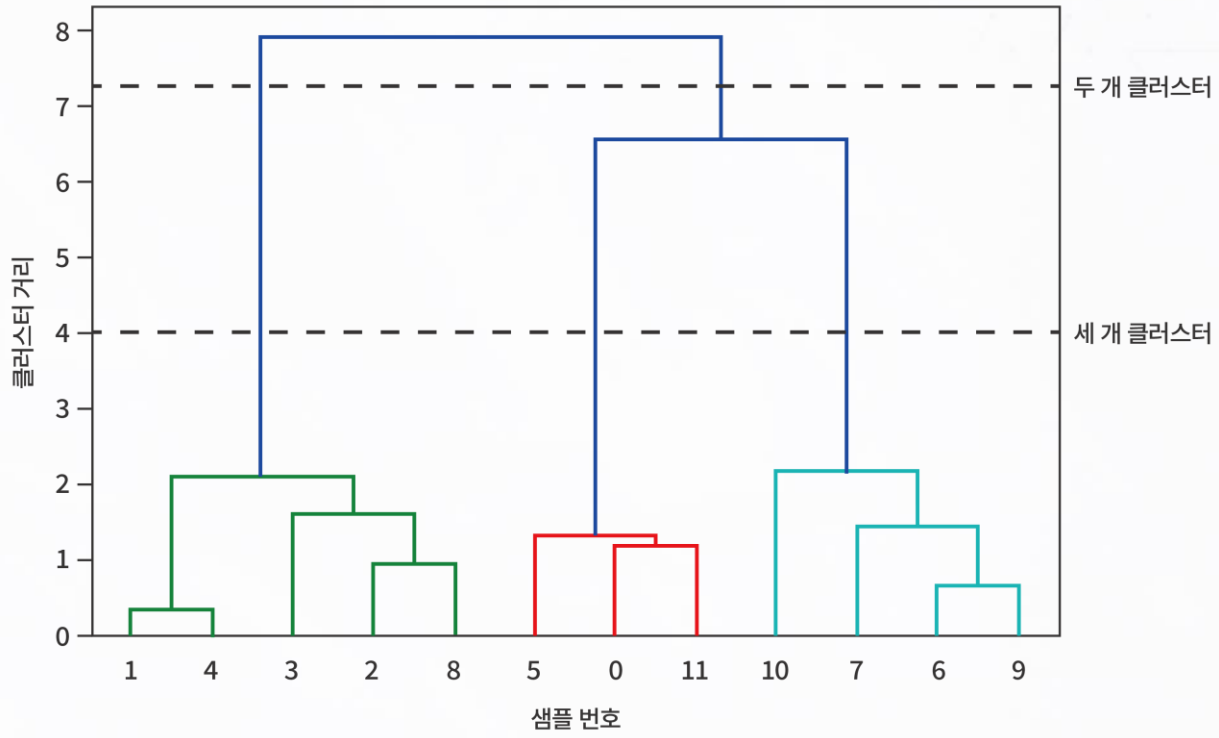
- 군집의 특성을 설명하기가 애매한 경우도 발생함
  - 사용된 거리측정 방법과 알고리즘에 따라 이상치에 민감도가 높을 수 있음
  - 큰 사이즈의 군집을 잘라버리는 경우도 발생하기 때문임

- 많은 변수를 투입하거나 데이터의 크기가 많은 경우 계산량이 많아 느려짐

03 | 계층적 군집: 덴드로그램

덴드로그램 (Dendrogram)

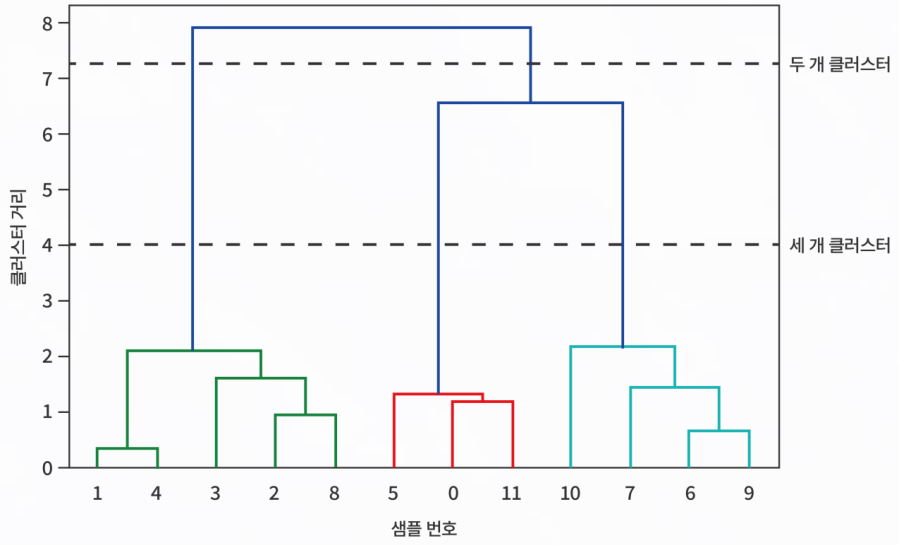
- 트리 구조를 갖는 다이어그램을 덴드로그램이라고 부름
  - y축은 군집과의 거리를 보여줌



03 | 계층적 군집: 덴드로그램

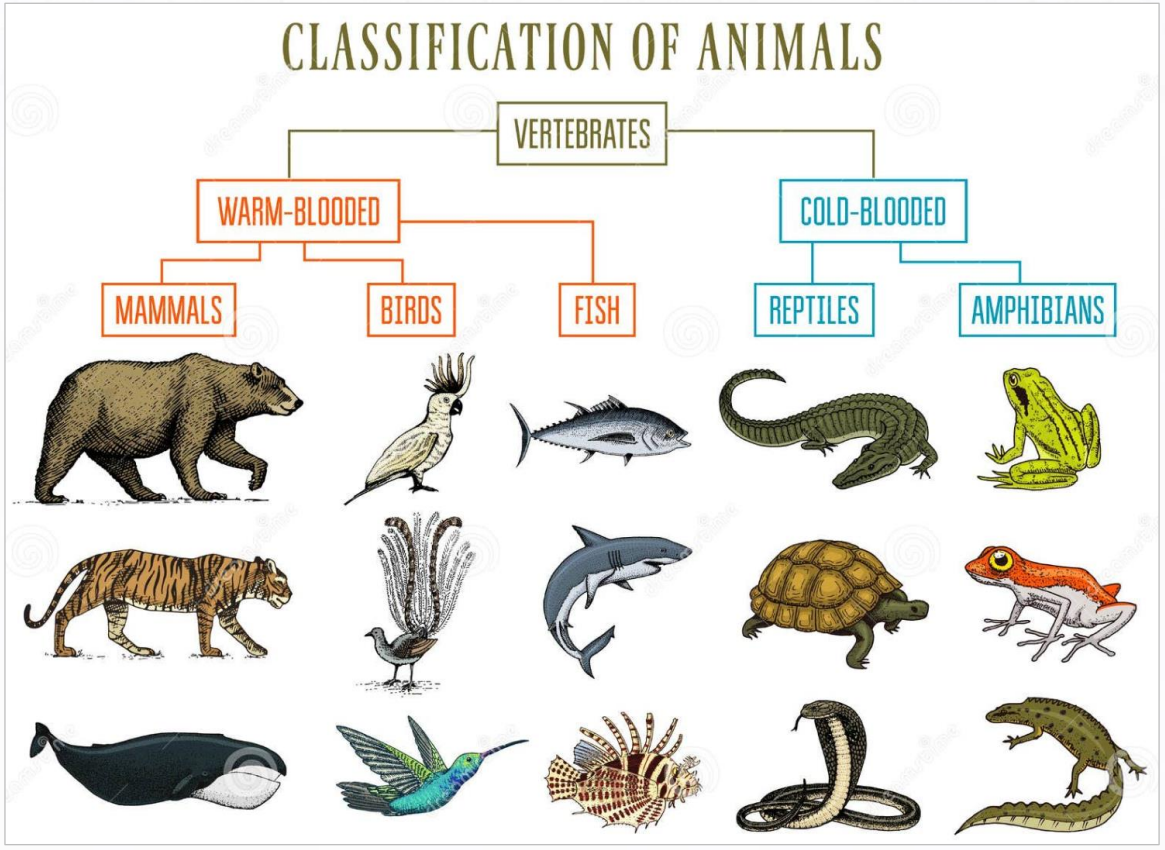
덴드로그램의 장점

- **군집 개수**를 **내 맘대로** 정할 수 있음
  - 덴드로그램의 장점은 **클러스터**의 **개수**를 **지정하지 않아도 된다**는 점임
  - 아래 그림과 같이 **덴드로그램**을 잘라서 **원하는 수준**의 **군집**을 **나눌 수** 있음
    - 아래 그림의 첫 번째 점선으로 두 개의 군집으로 나눔
    - 아래 그림의 두 번째 점선으로 세 개의 군집으로 나눔



03 | 계층적 군집: 덴드로그램

덴드로그램은 생물학의 분류나 고객 군의 분류에서 많이 사용함

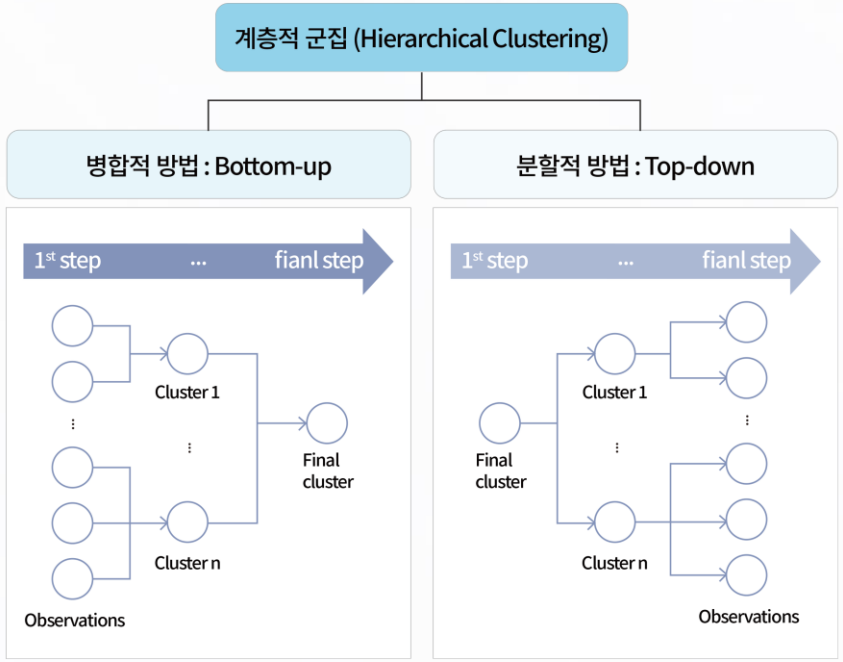


출처: https://www.dreamstime.com

03 | 계층적 군집: 덴드로그램

덴드로그램은 병합적 방법과 분할적 방법으로 표현할 수 있음

- 병합적 방법은 bottom-up 접근 방식임
  - 처음 시작할 때 모든 데이터가 각각의 군집으로 시작해서 주변과 병합해나가는 방식임
- 분리형 방법은 top-down 접근 방식임
  - 시작할 때 모든 데이터가 하나의 군집으로 시작해서 분리해나가는 방식임



03 | 계층적 군집: 덴드로그램

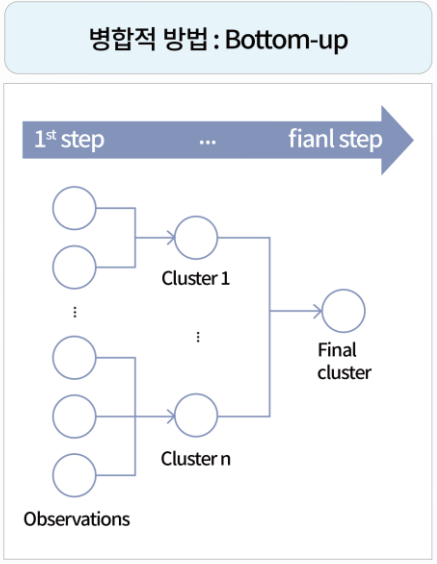
병합적(Agglomerative) 방법

✦ 병합적 방법의 계층적 군집 분석 정의

› 주어진 데이터에서 개별 데이터 하나 하나를 독립된 군집으로 가정함

- 이들을 특정 알고리즘에 의해 병합하여 상위단계 군집을 구성함
- 이렇게 구성된 상위단계 군집을 특정 알고리즘에 의해 또 다시 병합함

➔ 최종적으로 데이터 전체를 멤버로 하는 하나의 군집으로 구성하는 방법임

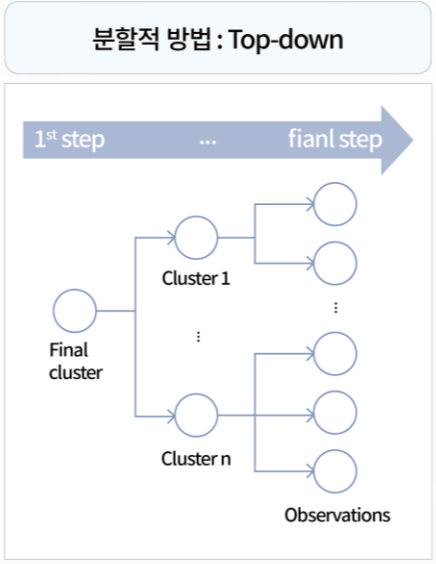


03 | 계층적 군집: 덴드로그램

분리형(Divisive) 방법

분리형 계층적 군집 분석의 정의

- **데이터 전체**를 **멤버**로 하는 **하나**의 **군집**에서 **시작**함
  - 이들을 **특정 알고리즘**에 의해 **분리**하여 **상위단계 군집**을 **구성**함
  - 이렇게 구성된 **상위단계 군집**을 **특정 알고리즘**에 의해 또 **다시 분리**함
    - **최종적**으로 **개별 데이터**로 **분리**해 나가는 식으로 **군집**을 **구성하는 방법**임



03 | 계층적 군집: 알고리즘

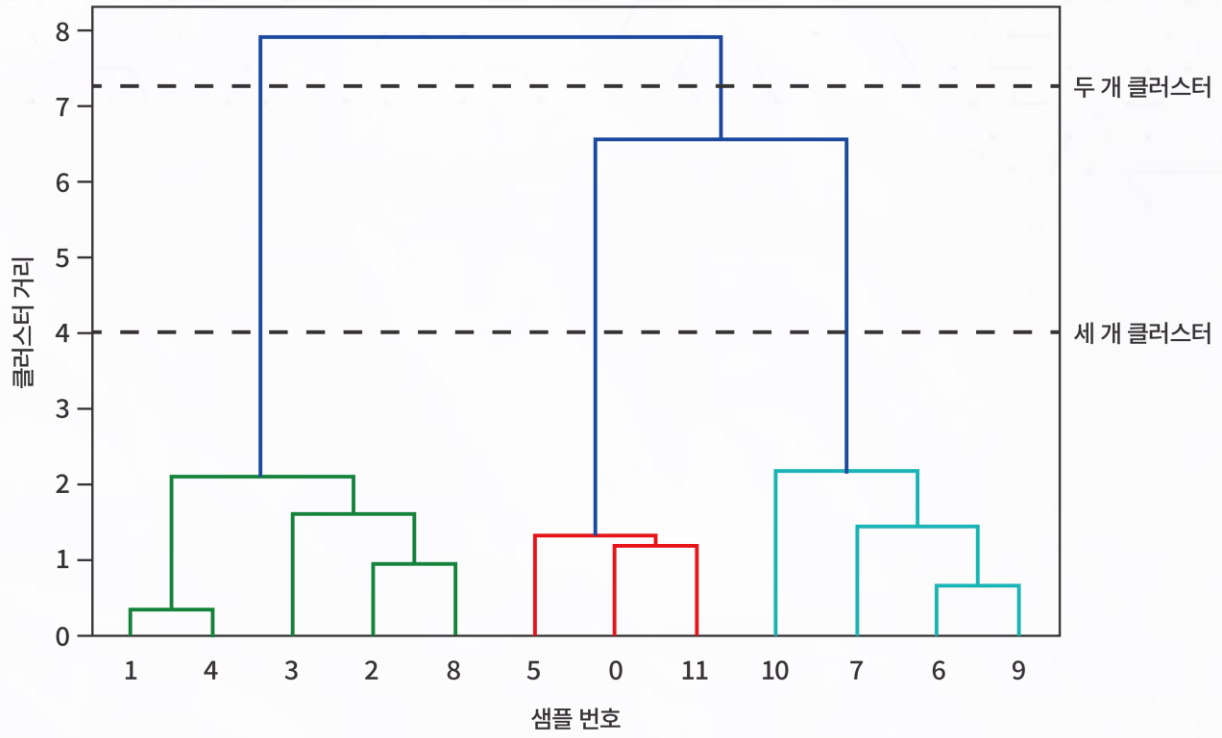
계층적 군집 알고리즘

- 계층적 군집이란 **계층적 트리 모형**을 이용해 **개별 개체**들을 **순차적**, **계층적**으로 **유사한 개체 내의 그룹**과 **통합**하여 **군집화**를 수행하는 **알고리즘**임
  - **개체**들이 **결합되는 순서**를 나타내는 **트리 형태**의 구조인 **덴드로그램** 덕분임

- K-평균 군집화(K-means Clustering)와 달리 **군집 수**를 **사전**에 **정하지 않아도 학습**을 **수행**할 수 있음

03 | 계층적 군집: 알고리즘

아래 그림과 같은 덴드로그램을 생성한 후 적절한 수준에서 트리를 자르면 전체 데이터를 몇 개 군집으로 나눌 수 있게 됨



03 | 계층적 군집: 알고리즘

계층적 군집 알고리즘의 학습과정은 다음과 같음

- 계층적 군집을 수행하려면 모든 개체들 간 거리(distance)나 유사도(similarity)가 이미 계산되어 있어야 함
  - 아래 그림과 같이 주어진 학습 데이터의 개체 수가 네 개이고 거리 행렬을 이미 구해놨다고 가정해 보자.
    - 거리가 가까운 관측치들끼리 차례대로 군집으로 묶어보자.

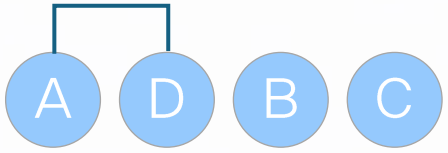
덴드로그램

	A	B	C	D
A	0	20	7	2
B		0	10	25
C			0	3
D				0

거리 행렬

03 | 계층적 군집: 알고리즘

- 아래 그림에서 거리가 가장 짧은 것이 2임
  - 이에 해당하는 개체는 A와 D이므로 먼저 A와 D를 하나의 군집으로 엮음
    - 아래 그림에서 덴드로그램의 높이는 관측치간 거리(2)가 됨
      - 여기서 A와 D를 한 군집으로 묶었으니 거리행렬을 바꿔주어야 함



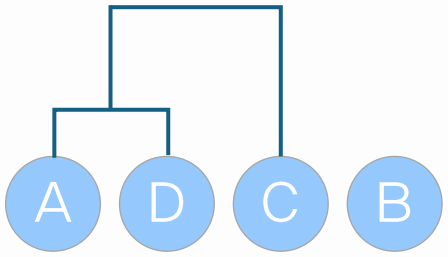
덴드로그램

	A	B	C	D
A	0	20	7	2
B		0	10	25
C			0	3
D				0

거리 행렬

03 | 계층적 군집: 알고리즘

- 아래 그림에서 개체-개체 거리를 군집-개체 거리로 계산해야 함
  - 아래 그림에서 AD와 B, AD와 C 이렇게 거리를 구해야 한다는 것임
    - 아래 그림에서 거리가 가장 짧은 것이 7(AD와 C)임
    - 이에 해당하는 개체는 AD와 C이므로 먼저 AD와 C를 하나의 군집으로 엮음
      - 여기서 AD와 C를 한 군집으로 묶었으니 거리행렬을 바꿔주어야 함



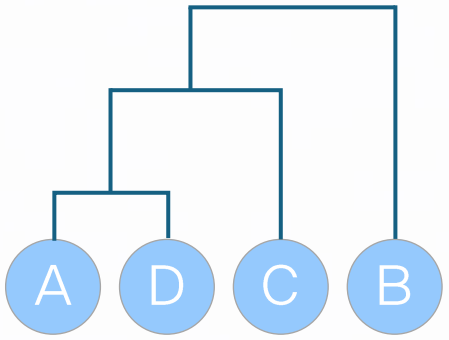
덴드로그램

	AD	B	C
AD	0	20	7
B		0	10
C			0
D			

거리 행렬

03 | 계층적 군집: 알고리즘

- 아래 그림과 같이 A, B, C, D 개체 **군집**을 **완성**하게 됨
  - 이렇게 각 군집의 거리를 **가장 가까운 원소의 거리 값**으로 **갱신**함
    - 이 **군집화하는 방법**이 **단일**(최단) **연결법**(single linkage method)임



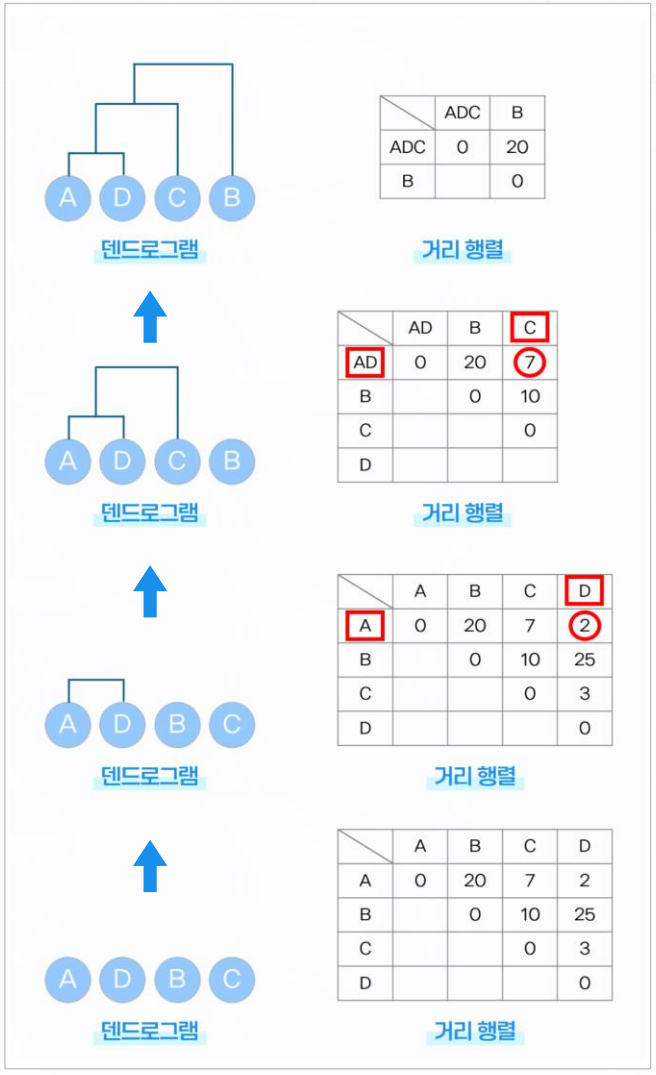
덴드로그램

	ADC	B
ADC	0	20
B		0

거리 행렬

03 | 계층적 군집: 알고리즘

아래의 그림은 **계층적 군집 알고리즘**(단일 연결법)의 **전체 학습 과정**을 나타냄



03 | 계층적 군집: 알고리즘

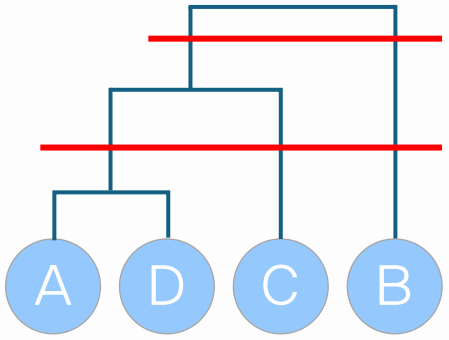
계층적 군집 알고리즘(단일 연결법)을 정리하면 다음과 같음

1. 1개의 관찰 값(개체/케이스)을 가지는 샘플 사이즈 n개의 군집으로 시각화함
2. 각각의 거리(유사성) 매트릭스를 계산함
3. 가장 유사한(거리가 가까운) 군집의 쌍(A, B)을 찾아서 하나의 군집(AB)으로 묶음
4. 3번이 진행됨에 따라 하나의 군집으로 묶이고 나서 생긴
   AB와 다른 개체의 거리 매트릭스를 업데이트함
5. 데이터의 모든 관찰 값(개체/케이스)이
   하나의 군집이 될 때까지 3번과 4번을 반복함

03 | 계층적 군집: 알고리즘

계층적 군집의 특징은 다음과 같음

- 계층적 군집은 K-평균 군집화와 달리 사전에 군집수 K를 설정할 필요가 없음
  - 아래의 그림처럼 덴드로그램의 최상층을 끊어주면 ADC와 B로 두 개 군집이 도출됨
  - 아래의 그림에서 두 번째 층을 끊으면 AD와 C, B 이렇게 세 개 군집이 나옴
  - 계층적 군집의 학습 결과물인 덴드로그램을 적절한 수준으로 잘라주면 된다는 것임
  - 반면, 계층적 군집의 계산복잡성은 $O(n^3)$로 K-평균 군집화보다는 무거운 편임



덴드로그램

	ADC	B
ADC	0	20
B		0

거리 행렬