Ch 9. Clustering

9.5 학습 벡터 양자화

9.6 가우시안 혼합 클러스터링

9.7 밀도 클러스터링

9.8 계층 클러스터링

21.04.07 이재학

9.5.1 학습 벡터 양자화(LVQ)

- **LVQ(Learning Vector Quantization)**

  원형 벡터를 찾는 방식은 K 평균 클러스터링과 유사

  - 각 데이터에 클러스터 라벨을 설정
  - **학습율** 원형 벡터를 구하는 과정에서 필요한 스칼라 값

$(x_1, x_2, x_3, \ldots xn)$

K 평균 클러스터링 Set

$((x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots (xn, y_n))$

LVQ Set

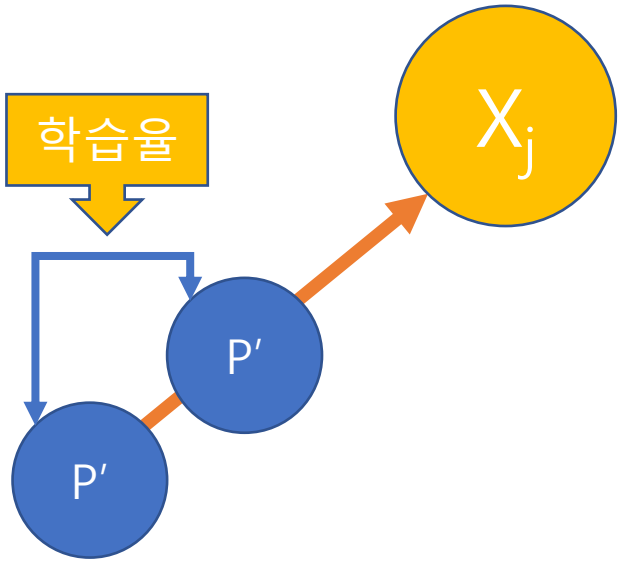
Y는 클러스터 라벨

9.5.2 학습 벡터 양자화(LVQ) 원형 벡터 갱신

- **LVQ 원형 벡터 갱신**

$P' = P_{i*} + n*(X_j - P_{i*}) \cdots\cdots\cdots$ (레이블이 같으면)

$\|P' - X_j\|_2 = (1-n)\|P_{i*} - X_j\|_2 \cdots$ (레이블이 다르면)



- P' : 갱신할 원형 데이터
- X_j 임의로 선택한 데이터 샘플
- N: 학습율(0~1 사이의 값을 가짐)
- P_{i*} : 랜덤으로 샘플한 벡터에서 가장 가까운 원형 벡터

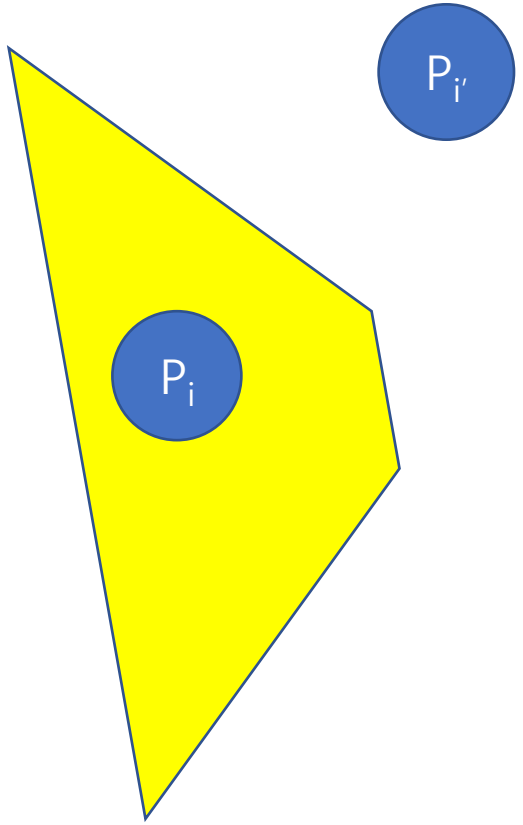
9.5.3 보로노이 분할

- **Voronoi 분할**

  샘플 Xi에 대해서 가장 가까운 원형 벡터 P_i와
  기타(다른 레이블의)$P_{i'}$ 가라고 한다면 다음 식이 성립

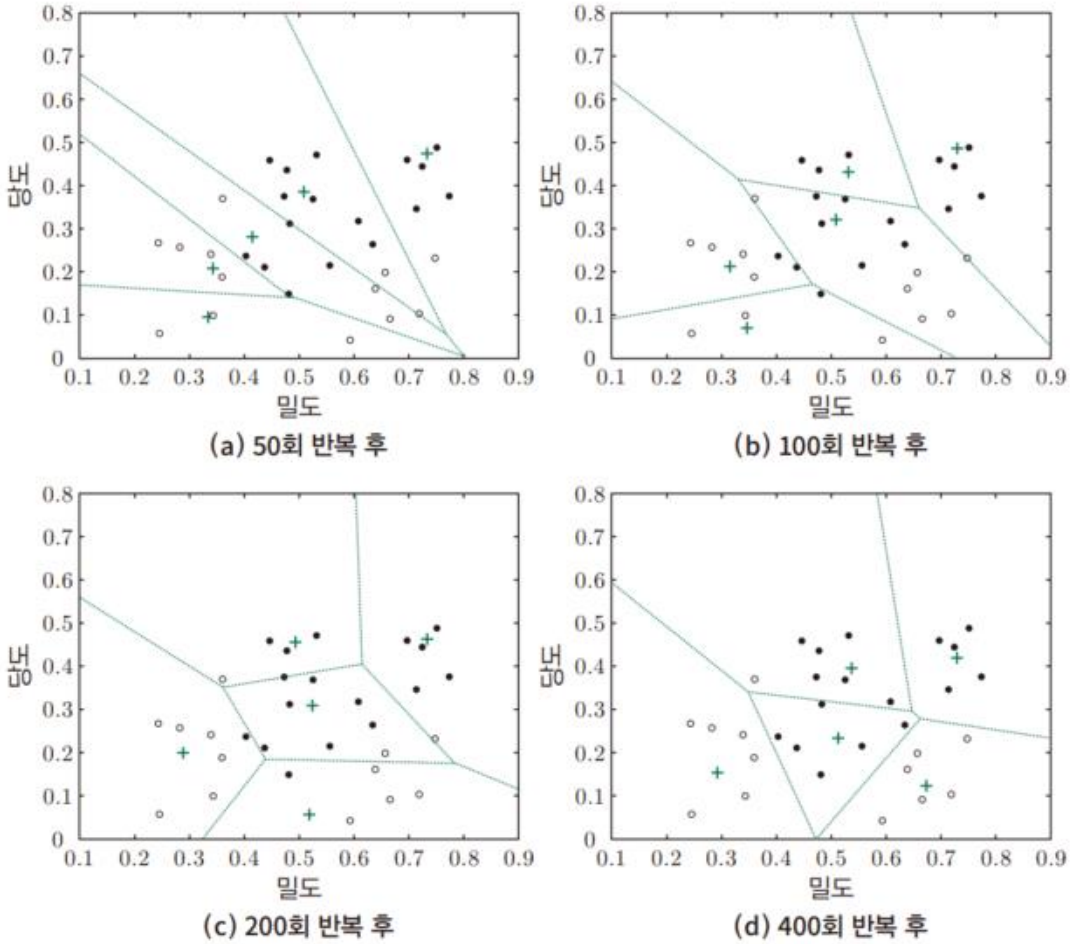
$$R_i = \{\boldsymbol{x} \in \mathcal{X} \mid \|\boldsymbol{x} - \boldsymbol{p}_i\|_2 \leqslant \|\boldsymbol{x} - \boldsymbol{p}_{i'}\|_2, i' \neq i\} .$$

  - R_i(레이블이 i인 영역) 안의 모든 샘플은 레이블이 다른 원형 벡터와의 거리는 Pi의 거리보다 항상 **같거나 작다**. 이 점들의 영역을 보로노이 분할이라고 한다.



9.5.4 예제

- 예제



(a) 50회 반복 후 (b) 100회 반복 후

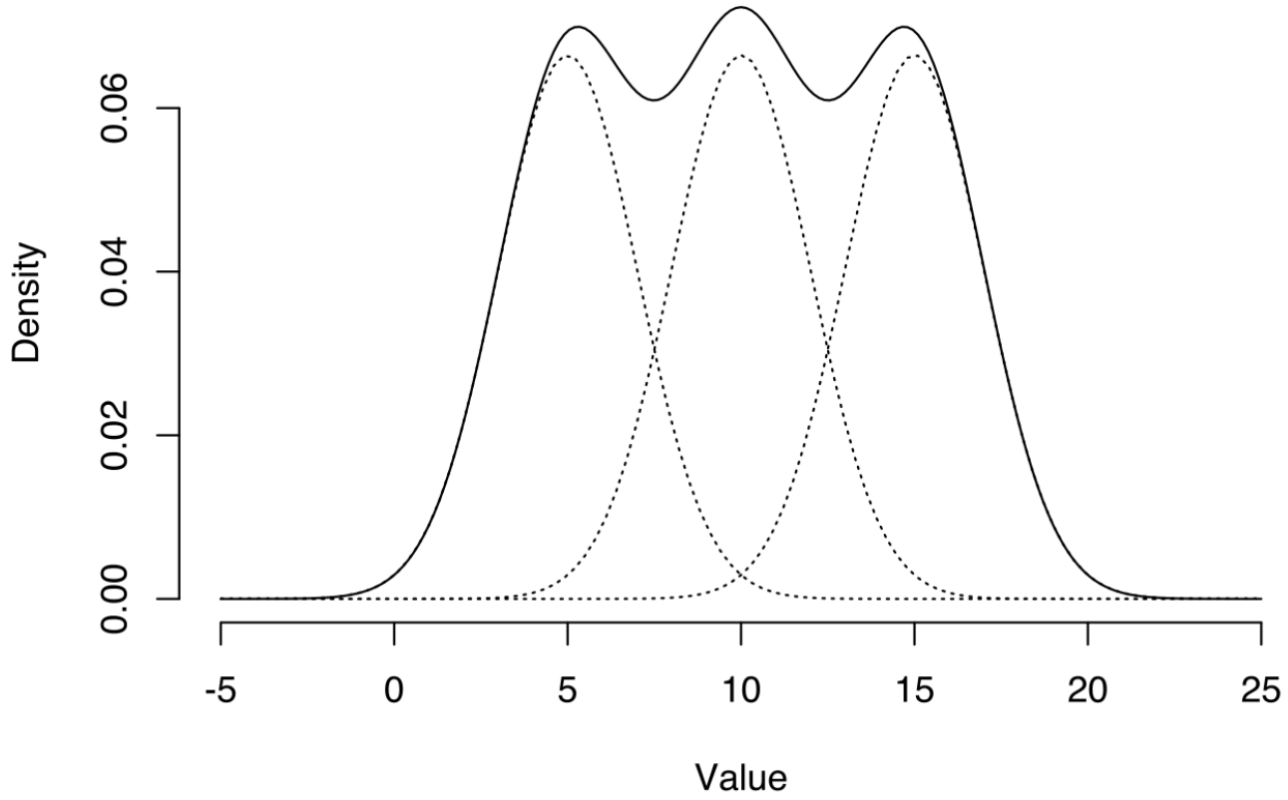
(c) 200회 반복 후 (d) 400회 반복 후

그림 9.5 ＼ 수박 데이터 세트 4.0에서 LVQ 알고리즘($q = 5$)을 실행했을 때, 다른 반복 횟수에 따른 클러스터링 결과

9.6.1 가우시안 혼합 클러스터링

▪ 비지도 학습 모델

- 가우시안 혼합 모델그래프



9.6.2 가우시안 혼합 클러스터링 과정

▪ 가우시안 혼합 클러스터링을 구하는 과정

- N 차원에서의 확률 벡터를 구함

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})},$$

- 이렇게 구한 $PM(Z_i = I \mid X_j)$는 혼합 성분으로 생성될 사후 확률을 나타냅니다.

$$p_{\mathcal{M}}(z_j = i \mid \boldsymbol{x}_j) = \frac{P(z_j = i) \cdot p_{\mathcal{M}}(\boldsymbol{x}_j \mid z_j = i)}{p_{\mathcal{M}}(\boldsymbol{x}_j)} = \frac{\alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{\sum_{l=1}^{k} \alpha_l \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}.$$

- 모델 파라미터 $a_i, u_i, \sum_l$ 의 값을 최대 우도 측정법을 통해 구함

$$LL(D) = \ln\left(\prod_{j=1}^{m} p_{\mathcal{M}}(\boldsymbol{x}_j)\right) = \sum_{j=1}^{m} \ln\left(\sum_{i=1}^{k} \alpha_i \cdot p(\boldsymbol{x}_j \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\right)$$

.....???

9.6.4 확률 용어 정리

- **우도**

  분자는 고정된 양을 가지고
  분모의 값을 조정하는 것

$$\{\frac{K_1}{X_1}, \frac{K_2}{X_2}, \dots, K_n/X_n\}$$

- **공분산**

  확률 변수가 2개 이상 있을
  때의 각 **확률변수들의 평균**

$$Corr(X, Y) = \rho = \frac{Cov(X, Y)}{S(X)S(Y)} = \frac{E[(X-\mu_x)(Y-\mu_y)]}{\sqrt{E(X-\mu_x)^2 \cdot E(Y-\mu_y)^2}}$$

- **라그랑주승수**

  제약에 대한 최적 환경 값을 구하는 공식

$$J(X;\theta,\ \lambda) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k N(x_n|\mu_k,\ \Sigma_k) + \lambda\left(1 - \sum_{k=1}^{K} \pi_k\right)$$

$$\frac{\partial J(X;\theta,\ \lambda)}{\partial \pi_k} = \sum_{n=1}^{N} \frac{N(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\ \Sigma_j)} - \lambda = 0$$

$$\Leftrightarrow \sum_{k=1}^{K}\sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\ \Sigma_j)} - \lambda \sum_{k=1}^{K} \pi_k = 0$$

$$\Leftrightarrow \sum_{k=1}^{K}\sum_{n=1}^{N} \gamma(z_{nk}) - \lambda = 0 \quad \left(\because \sum_{k=1}^{K} \pi_k = 1\right)$$

$$\therefore \lambda = N \quad \left(\because \sum_{k=1}^{K} \gamma(z_{nk}) = 1\right)$$

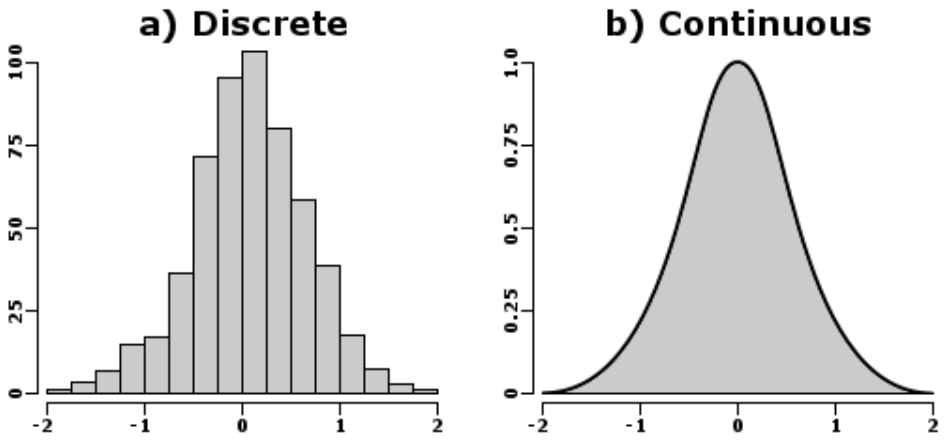
$$\frac{\partial J(X;\theta,\ \lambda)}{\partial \pi_k} = \sum_{n=1}^{N} \frac{N(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\ \Sigma_j)} - N = 0$$

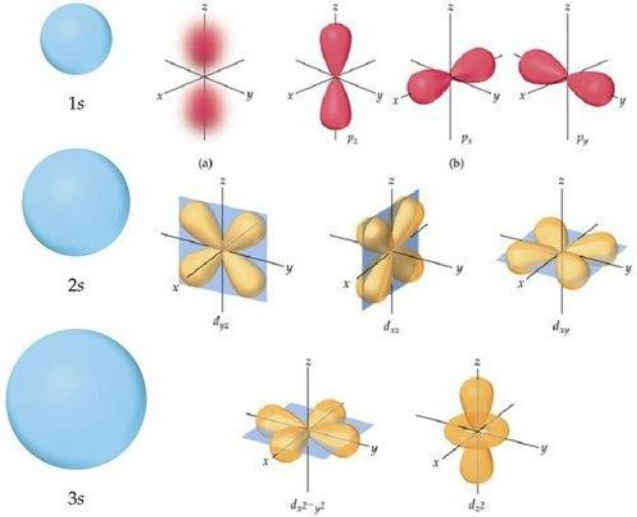
$$\Leftrightarrow \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k,\ \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x_n|\mu_j,\ \Sigma_j)} - N\pi_k = 0$$

$$\therefore \pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk})$$

9.6.3 가우시안 혼합 클러스터링 예제

- **보어의 양자역학**
  - 확률분포 그래프



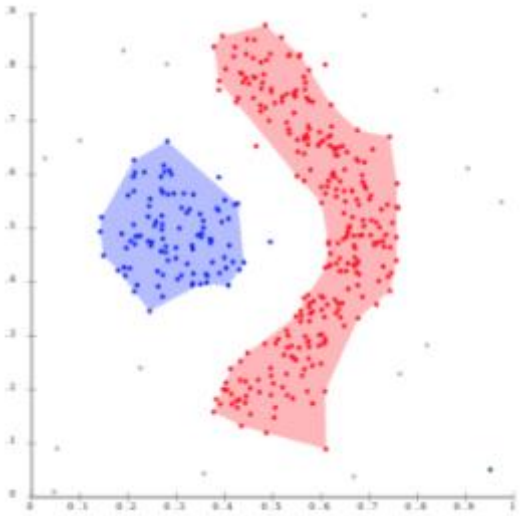


9.7.1 밀도 클러스터링

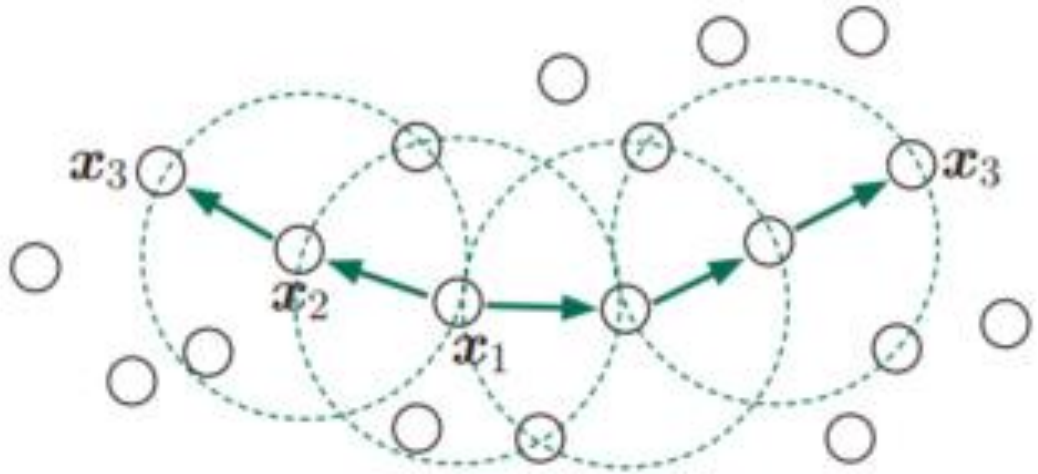
- **DBSCAN(Density-based spatial clustering of application with noise)**

  K 평균이나 **계층 구조**의 클러스터링 같은 경우 **군집간의 거리**를 사용한 클러스터링

  밀도 클러스터링은 **밀도가 높은 부분**을 클러스터링



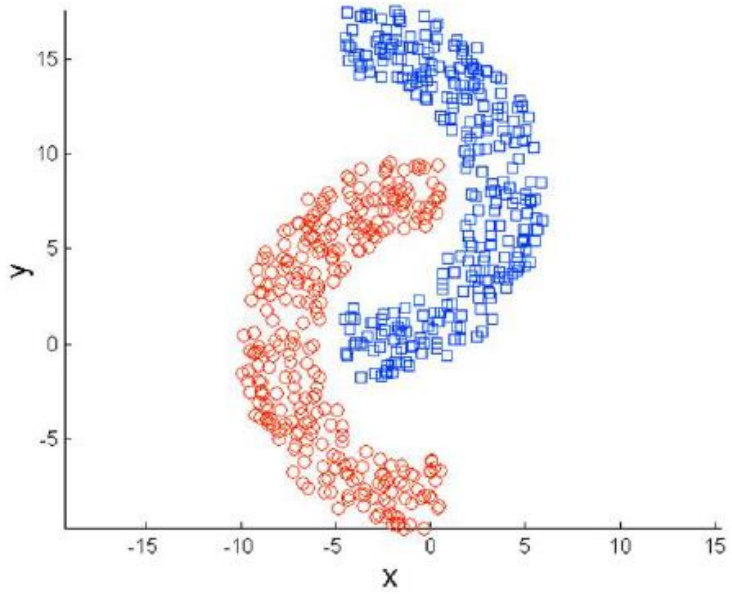
9.3.2 밀도 클러스터링 원리

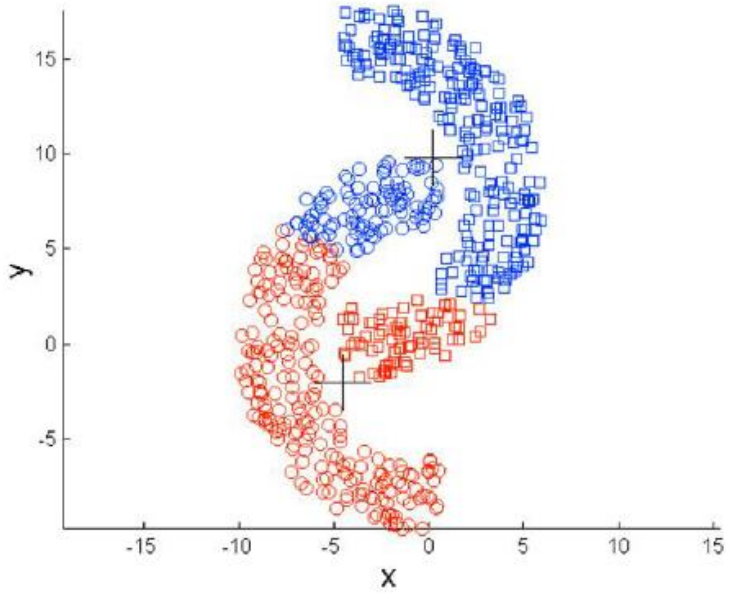


- **∈(앱실론)- 이웃 지역**: 한 데이터에서 앱실론 반경 안에 있는 샘플들
- **핵심 대상**: ∈-이웃 지역 안에 미리 지정한 minPts 만큼의 샘플이 포함되어 있는 샘플
- **직접 접근 가능한**(directly density-reachable): ∈-이웃 지역 안에 미리 지정한 minPts 만큼의 샘플이 포함되어 있는 샘플
- **접근 가능한**(density-reachable): $p_1=x_i$이고 $p_n=x_j$ 일 때 p_{i+1}은 P_i의 직접 접근 가능한 밀도라면 x_j는 X_i의 접근 가능한 밀도

9.7.3 밀도 클러스터링의 장점

- **K 평균 클러스터링의 한계를 보완**
  - 클러스터 수를 정하지 않아도 됨
  - 기학학적인 분포에 특화됨



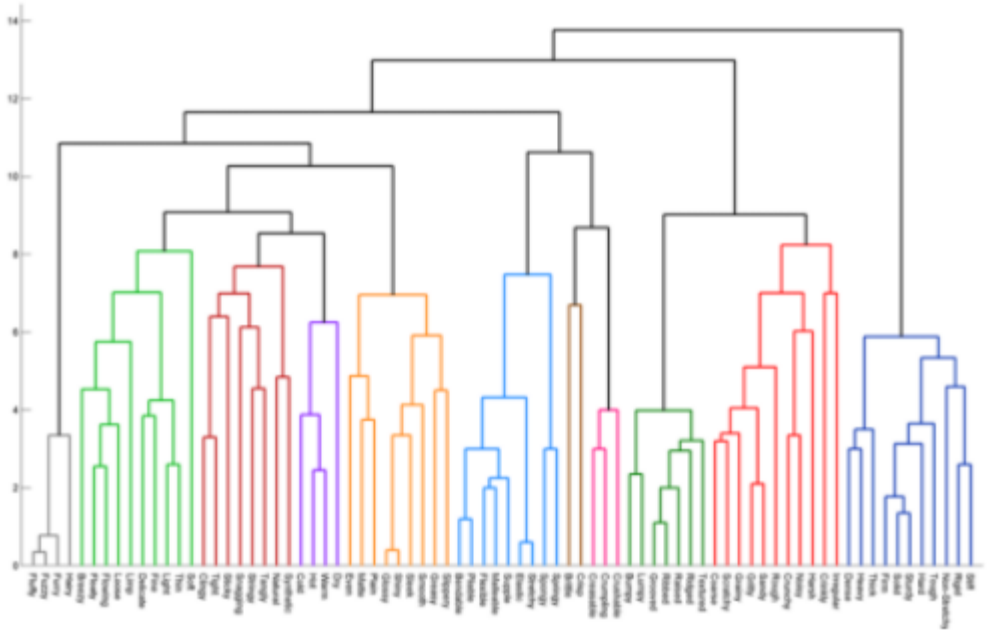


9.7.4 밀도 클러스터링의 단점

- **조건이 잘 갖춰져야 함**
  - **데이터가 입력되는 순서**
  - **거리 측정 방법**
  - **데이터의 특성** 이 다 갖춰져야 함

9.8.1 계층 클러스터링 Hierarchical clustering

- 트리 구조를 이용



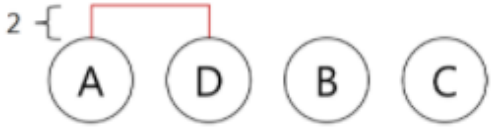
- 덴드로그램

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

9.8.2 덴드로그램 Dendrogram

- **Step 1. 미리 구해 놓은 거리나 유사도를 토대로 덴드로그램을 구성**

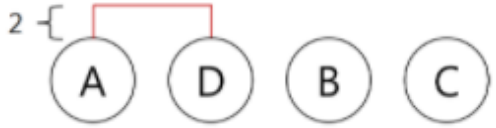
	A	B	C	D
A		20	7	2
B			10	25
C				3
D				



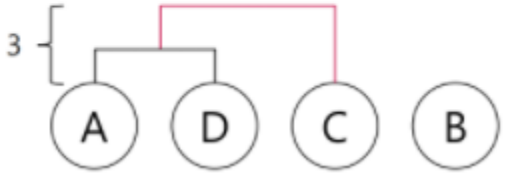
	A	B	C	D
A		20	7	2
B			10	25
C				3
D				

9.8.3 덴드로그램-2 Dendrogram

- **Step 2. 군집-> 하나의 개체로 선언**



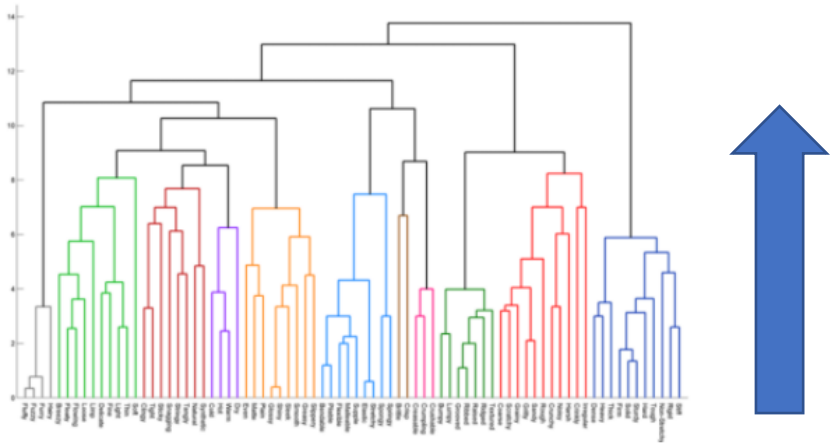
	A	B	C	D
A		20	7	2
B			10	25
C				3
D				



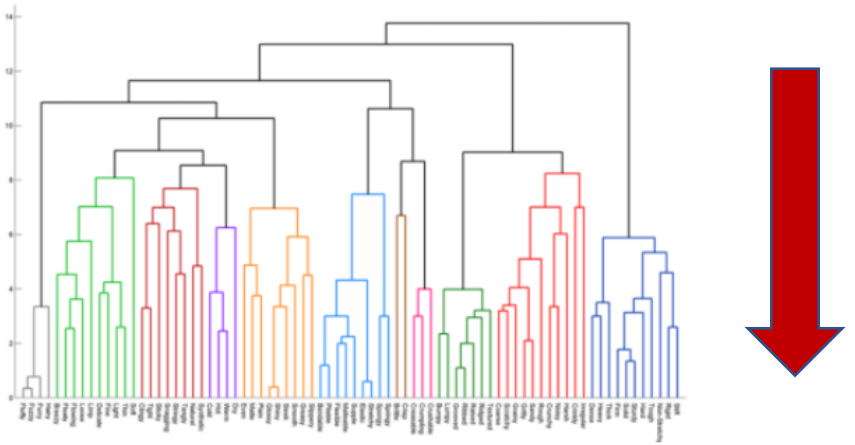
	AD	B	C	
AD		20	3	
B			10	
C				

9.8.4 계층 클러스터링 종류

- 상향식 클러스터링(AGNES)



- 하향식 클러스터링(DIANA)



9.8.5 계층 클러스터의 장단점

- **장점**
  - 군집수를 필요에 따라 **조정 가능**

- **단점**
  - 연산량이 K-평균 군집화보다 **무거운 편**

Etc. 참고자료

- **고려대학교 DSBA - Multivariate Data Analysis 강의**
  - Ch 9. Clustering

- **참고 블로그**
  - ratsgo's blog
  - https://untitledtblog.tistory.com/146
  - http://matrix.skku.ac.kr/math4ai-intro/W11/
  - https://ratsgo.github.io/machine%20learning/2017/04/18/HC/