




# Hierarchical Clustering

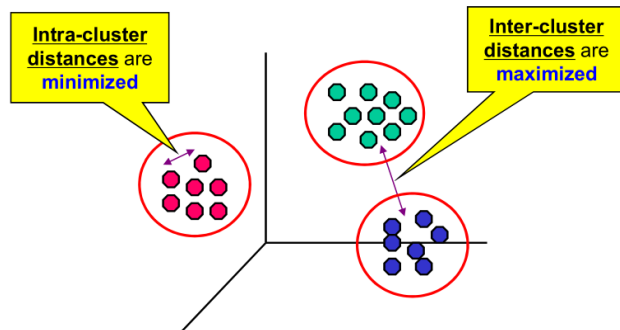
유승완



# Clustering

# Clustering ( 군집화 )

- 비슷한 개체끼리 한 그룹으로, 다른 개체는 다른 그룹으로 묶는 것
- 대표적인 비지도학습(unsupervised learning)
- 군집 간 분산(inter-cluster variance) 최대화  
군집 내 분산(inner-cluster variance) 최소화



# Application

---

## Understanding

- Browsing 시에 관련되어 있는 문서를 Grouping
- 비슷한 기능을 가지고 있는 유전자나 단백질들을 Grouping
- 비슷한 추세를 가진 주식들을 Grouping

## Summarization

- 큰 사이즈의 데이터셋의 크기를 줄여줌

# Hierarchical Clustering

# Partitional Clustering vs Hierarchical Clustering

---

- Partitional Clustering

각 데이터가 겹치지 않는 exclusive한 clustering

- Hierarchical Clustering

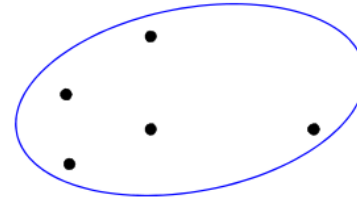
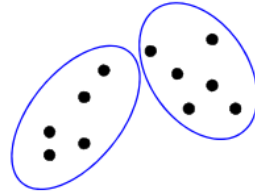
데이터들이 hierarchical 하게 묶이며 중첩되는 clustering

# Partitional Clustering

---



Original Points

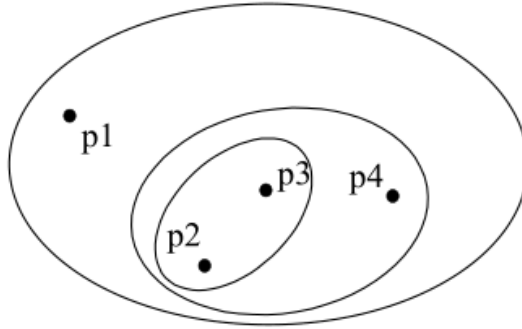


A Partitional Clustering

non-overlapping subsets  
(exclusive clusters)

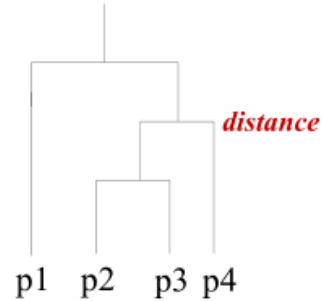
# Hierarchical Clustering

---



Hierarchical Clustering

**A set of nested clusters**



Dendrogram



# Type of Hierarchical Clustering

---

## ◆ Agglomerative : bottom up

1. 각각의 Point 를 cluster라고 생각하고 시작
2. 모든 데이터가 모인 1개의 클러스터가 만들어질 때까지 merge

## ◆ Divisive : top down

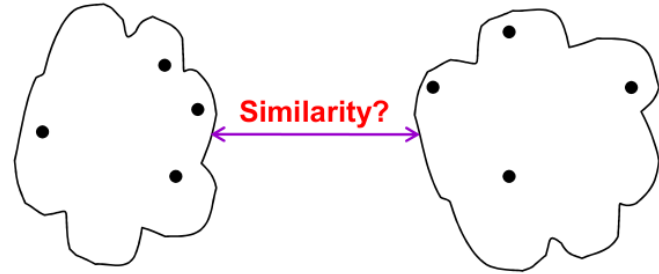
1. 모든 데이터를 포함한 1개의 Cluster로 시작
2. 각각의 point가 cluster가 될 때까지 Cluster를 split

▶ Cluster들을 split 하거나 merge 하는 measure로 similarity나 distance matrix를 활용

# How to Define Inter-Cluster Similarity

---

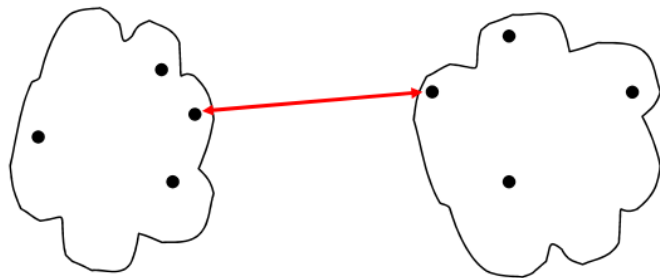
- MIN
- MAX
- Group Average
- Ward's Method



# MIN ( Single Link )

---

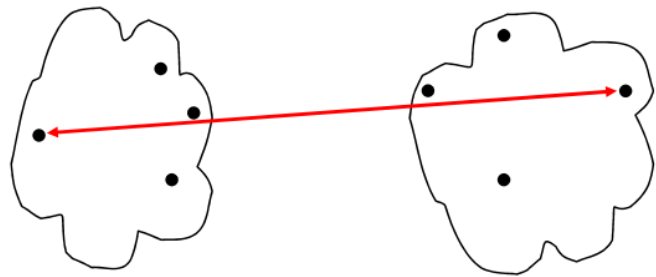
- 각각의 클러스터 내에 가장 가까운 점들의 거리를 기준
- Cluster가 원형이 아닌 다양한 형태로 생길 수 있음
- Noise와 Outlier에 Sensitive



# MAX ( Complete Link)

---

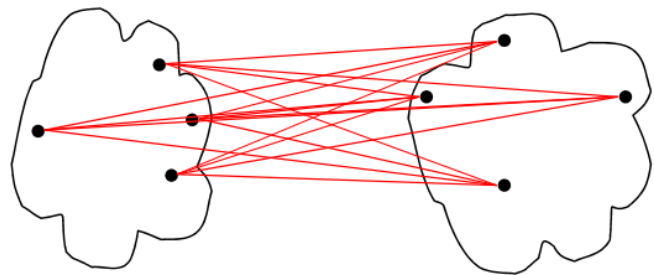
- 각각의 클러스터 내에 가장 먼 점들의 거리를 기준
- 큰 Cluster들이 쪼개지는 경향이 있음
- 원형 모양의 Cluster로 biased 되는 경향이 있음
- Noise와 Outlier에 less susceptible



# Group Average

---

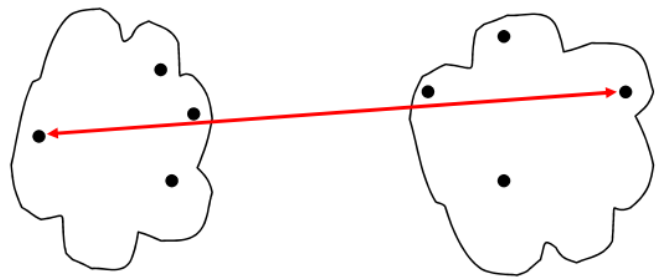
- 두 클러스터 내에 점들의 거리 평균을 기준
- MIN 과 MAX Measure들의 사이 효과
- MIN Measure 보다는 noise와 outlier에 덜 민감함
- 원형 모양의 Cluster로 biased 되는 경향이 있음



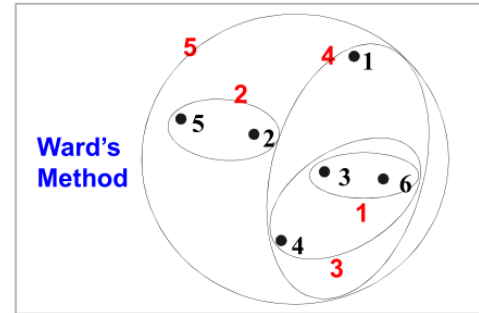
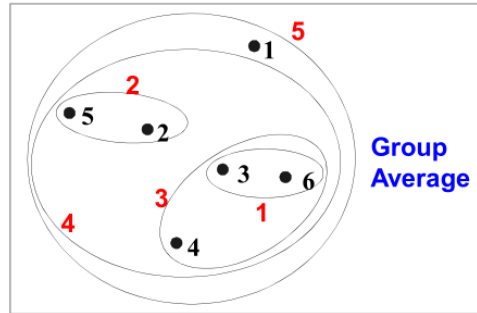
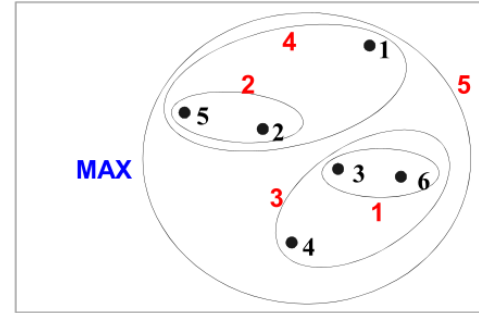
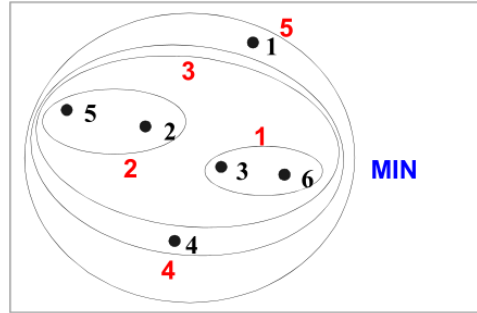
# Ward's Method

---

- 두 클러스터가 합쳐질 때 squared error의 증가량을 기준
- Noise와 Outlier에 Sensitive
- 원형 모양의 Cluster로 biased 되는 경향이 있음



# Comparsion



# Example

---

	1	2	3	4	5
1	0.00	0.10	0.90	0.35	0.80
2	0.10	0.00	0.30	0.40	0.50
3	0.90	0.30	0.00	0.60	0.70
4	0.35	0.40	0.60	0.00	0.20
5	0.80	0.50	0.70	0.20	0.00



# Problem and Limitations

---

- 한번 클러스터가 합쳐지면 Undo가 불가능함
- Local minimum에 빠지기 쉽다.
- Distance Measure 마다 단점이 있다.

MIN : outlier와 noise에 민감함

MAX / Average / Ward : 다양한 형태와 사이즈의 Cluster를 찾지 못함.

MAX : 큰 Cluster는 쪼개게 된다.



EOD



# Reference

---

고려대학교 산업경영공학부 백준걸 교수님 - 데이터마이닝 수업자료 Chapter 7 Clustering