Statistics II

for Machine Learning

Chap. 7: Cluster Analysis

Oh, Hyung Sool

Cluster Analysis(군집분석)

- Cluster analysis is a statistical technique in which algorithms are used to group a set of objects or data points into groups based on their **similarity**.

- The result of cluster analysis is a set of clusters, where each cluster is **distinct** from one another, and the objects or data points within each cluster are **largely similar** to each other.

- The purpose of cluster analysis is to help reveal patterns and structures within a dataset that may provide insights into underlying relationships and associations.

- Applications of cluster analysis
  - **Market Segmentation**
    : to segment customers into groups based on their buying behavior, demographics, or other characteristics.
  - **Image Processing**
    : to group pixels with similar properties together, allowing for the identification of objects and patterns in images.
  - **Biology and Medicine**
    : to identify genes associated with specific diseases or to group patients with similar clinical characteristics together.
  - **Social Network Analysis**
    : to group individuals with similar social connections and characteristics together.
  - **Anomaly Detection**
    : to detect anomalies in data, such as fraudulent financial transactions, unusual patterns in network traffic, or outliers in medical data.

Cluster Analysis(군집분석)

Cluster analysis (clustering, segmentation, quantization, ...) is the data mining core task to find clusters.

But what is a cluster?

- cannot be precisely defined
- many different principles and models have been defined
- even more algorithms, with very different results
- when is a result "valid"?
- results are subjective "in the eye of the beholder"
- no specific definition seems "best" in the general case

Common themes found in definition attempts:

- more homogeneous
- more similar
- cohesive

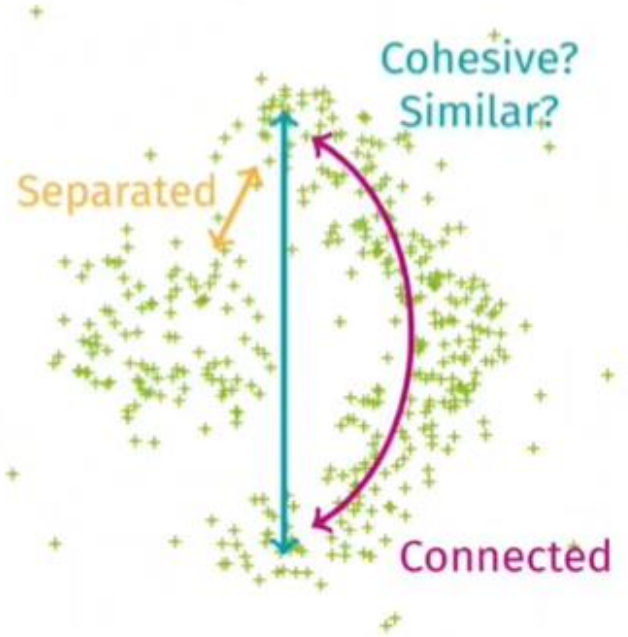
Cluster Analysis(군집분석)

Cluster analysis (clustering, segmentation, quantization, …)
is the data mining core task to divide the data into clusters such that:

- *similar* (related) objects should be in the *same* cluster
- *dissimilar* (unrelated) objects should be in *different* clusters
- clusters are *not* defined beforehand (otherwise: use classification)
- clusters have (statistical, geometric, …) properties such as:
  - connectivity
  - separation
  - least squared deviation
  - density

Clustering algorithms have different

- cluster models ("what is a cluster for this algorithm?")
- induction principles ("how does the algorithm find clusters?")



Cluster Analysis(군집분석)

Basic Steps for Clustering

Feature selection

- select information (about objects) concerning the task of interest
- aim at minimal information redundancy
- weighting of information

Clustering algorithm and parameters

- distance and similarity measure suitable for the problem
- cluster quality criterion / cost function / objective
- algorithms to use with this distance and quality criterion

Validation and interpretation of the results

- validation test
- integration with applications

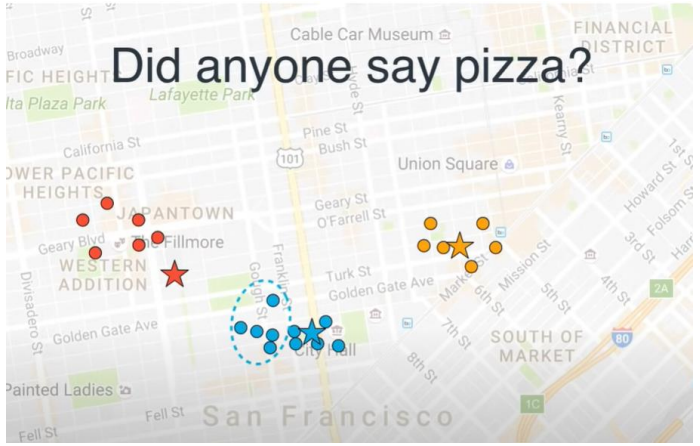
Cluster Analysis(군집분석)

- Types of Cluster Analysis
  - Hierarchical clustering(계층적 군집분석)
  - K-means clustering

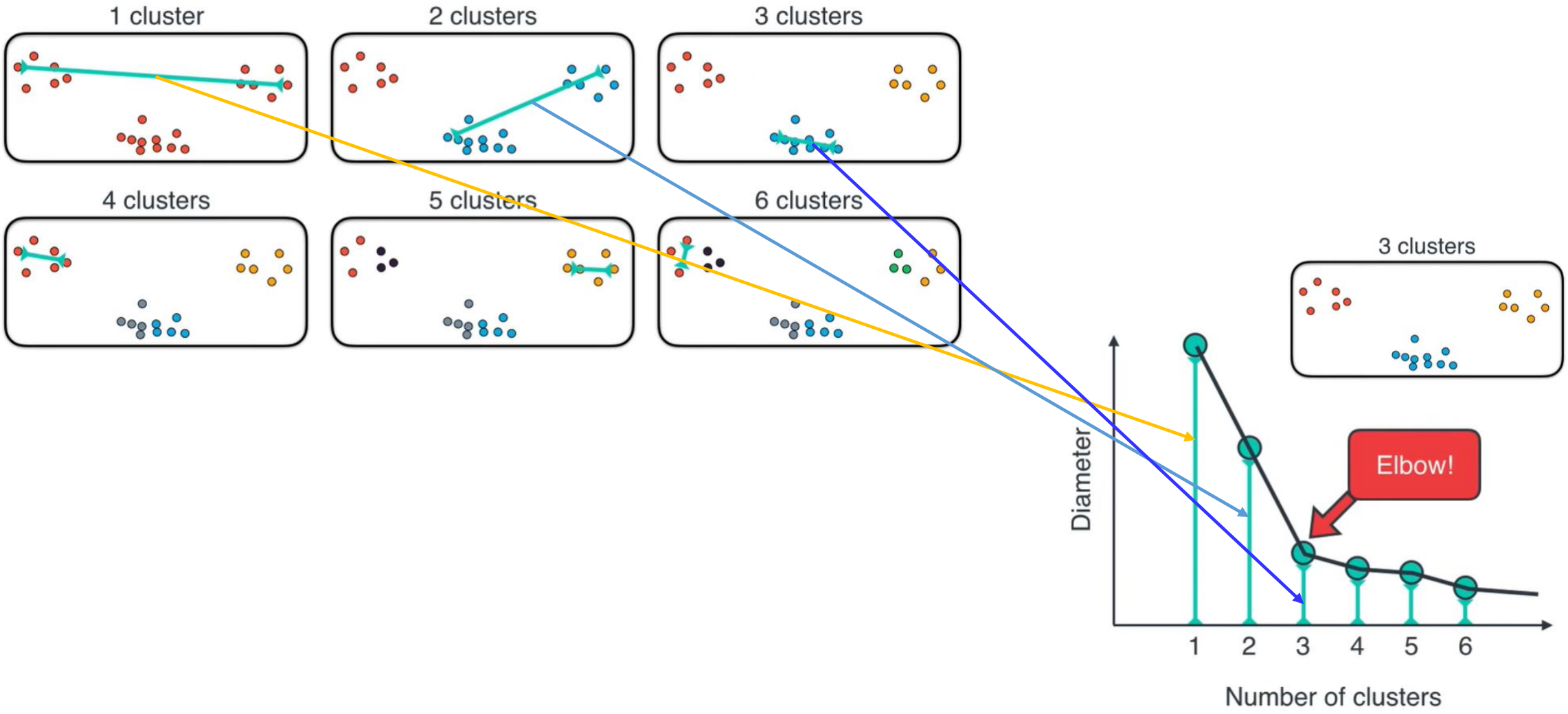
- **K-means clustering**
  - 개별 사례(Case) 또는 데이터와 K개 중심점과의 거리를 측정하여 이 가운데 가장 가까운 거리를 갖는 군집에 개별 사례나 데이터를 할당하는 방식으로 군집화를 수행
  - 수행 절차
    - ✓ 시작 단계에서 중심점의 수 K를 정해야 한다.
    - ✓ 중심점과 개별 케이스 간의 **거리**를 바탕으로 케이스를 분류
    - ✓ 각 군집에 할당된 케이스들을 이용하여 새로운 군집 평균을 계산하고, 이를 토대로 모든 케이스를 다시 분류
    - ✓ 군집 평균에 더 이상의 큰 변화가 없을 때 까지 이 과정을 수행
    - ✓ K-means clustering은 계층적 군집분석에서 다루기 어려운 큰 규모의 데이터 셋을 다룰 수 있다.
    - ✓ K-means clustering은 군집 중심점을 계산할 때 평균을 사용하기 때문에 모든 변수가 **연속형**이어야 하며, 분석 결과가 이상점에 의해 영향을 받을 수 있다.

K-means Clustering

K-means Clustering



Elbow method



Cluster Analysis(군집분석)

Hierarchical clustering

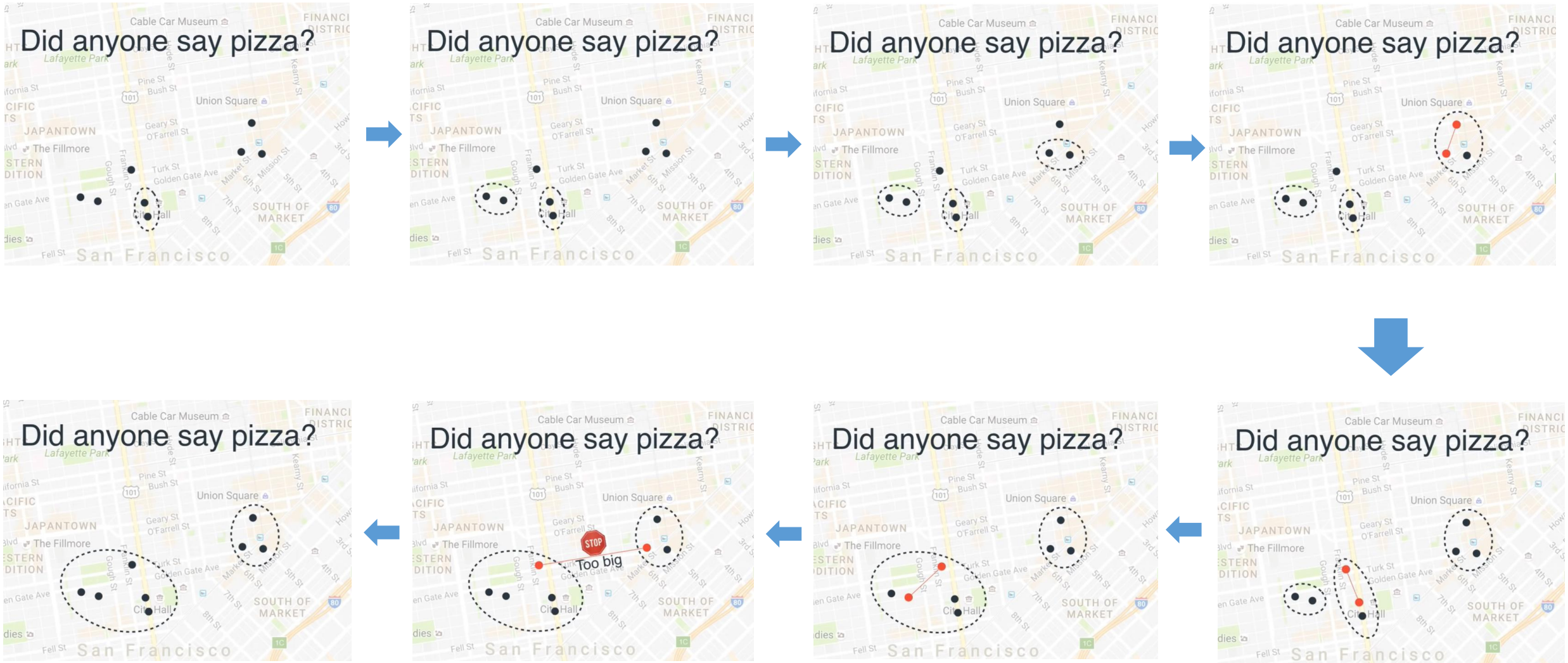
- 모든 사례(Case) 또는 데이터가 각자 하나의 군집을 형성하면서 시작하여 이후 군집 단계가 진행되면서 단계별로 유사한 군집끼리 서로 합쳐지며 **최종단계에서는 하나의 군집만** 남는다.

- 각 단계가 진행될 때마다 **군집 간의 최소 거리를 기준**으로 개별 케이스들이 기존의 군집에 흡수되거나 두 개의 케이스가 결합하여 새로운 군집을 만들거나, 기존의 두 군집이 결합하여 새로운 군집을 형성하는 등의 방식으로 군집의 개수가 줄어들게됨

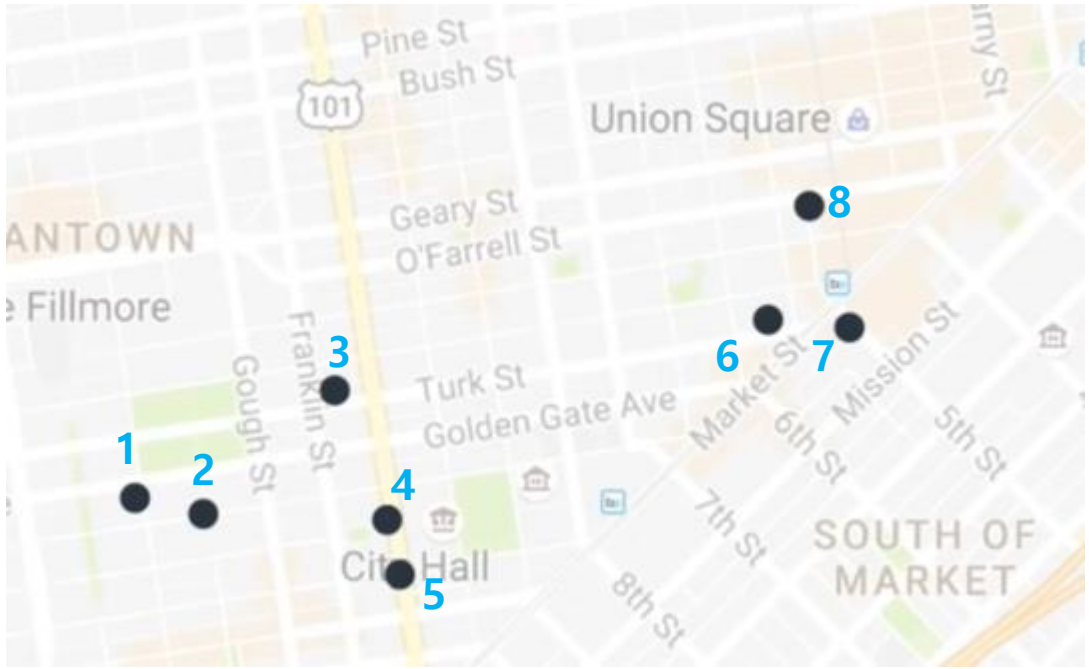
- 군집 간의 거리 측정
  - ✓ 단일 연결법(single linkage): 모든 케이스 쌍의 거리 중 가장 가까운 거리를 사용
  - ✓ 완전 연결법(complete linkage): 모든 케이스 쌍의 거리 중 가장 먼 거리를 사용
  - ✓ 평균 연결법(average linkage): 모든 케이스 쌍의 거리를 평균한 거리를 사용
  - ✓ 중심 연결법(centroid linkage): 중심 간의 거리를 사용
  - ✓ 최소 분산 연결법(minimum variance linkage): 모든 케이스 간의 총분산을 거리로 사용

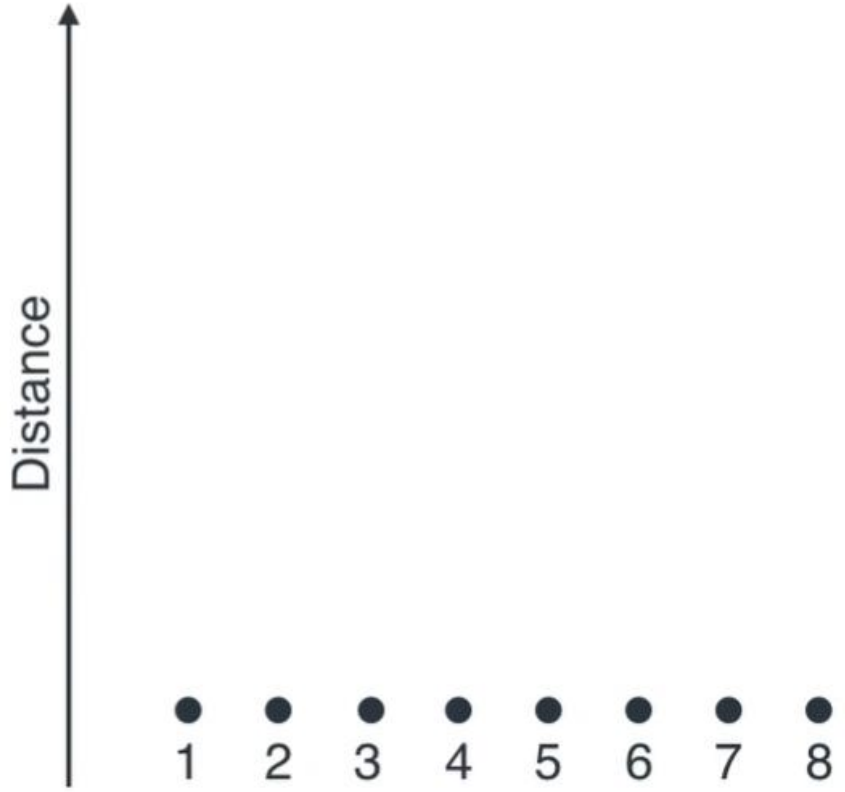
Hierarchical Clustering

Hierarchical Clustering

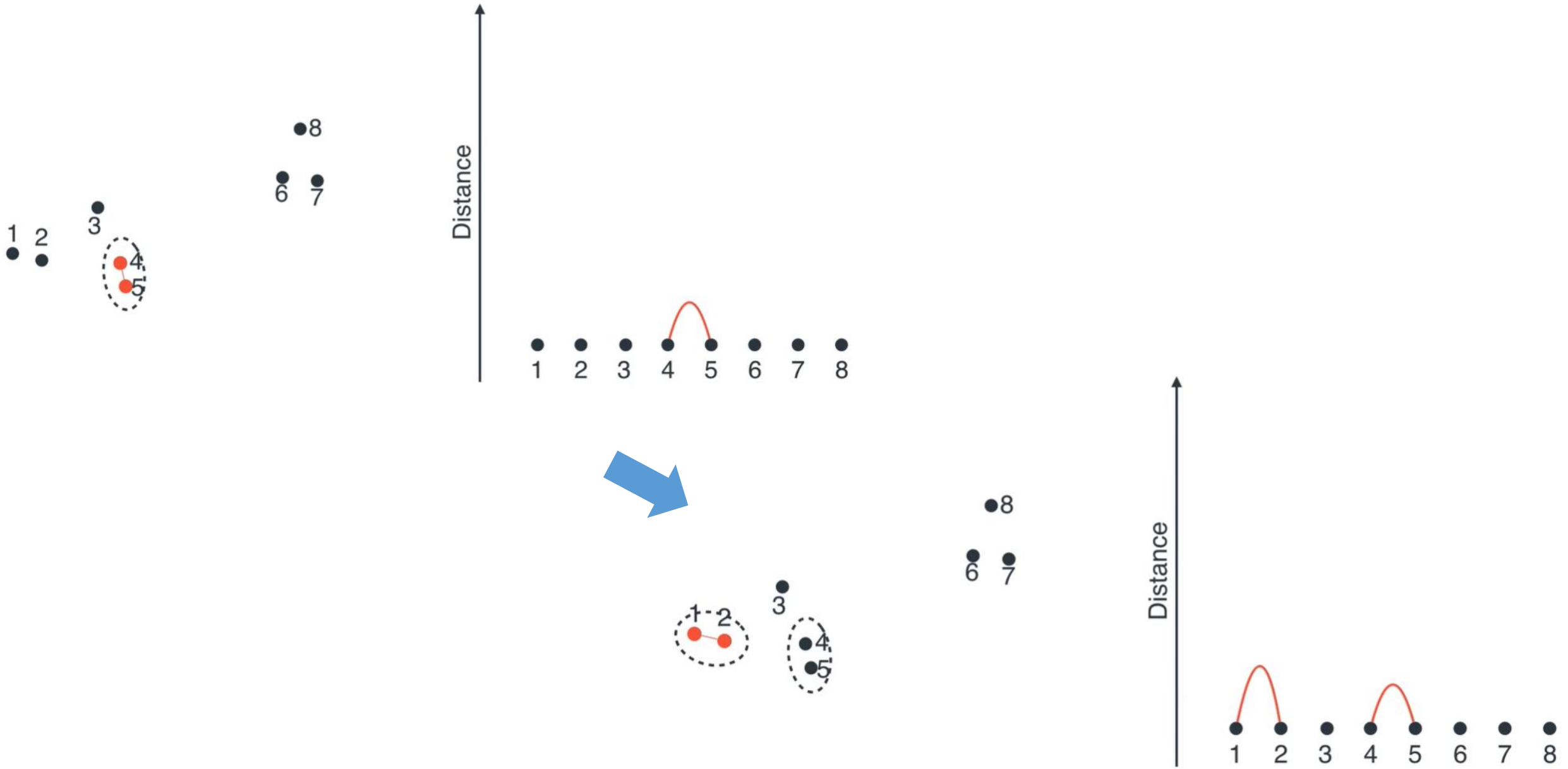


Dendrogram

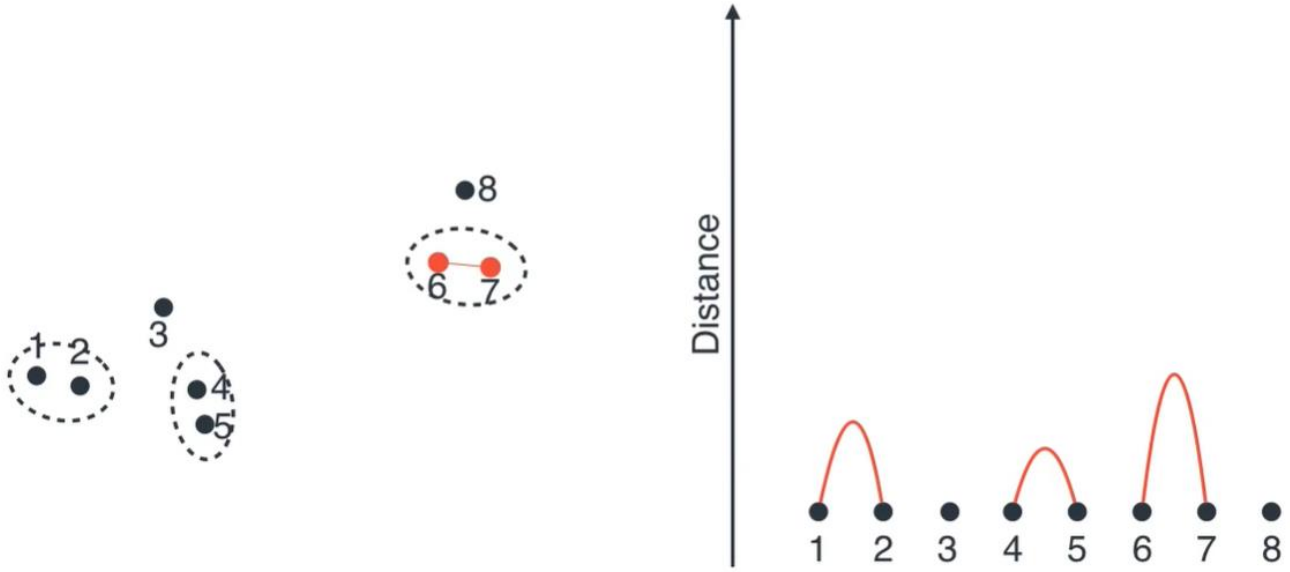

Pine St
Bush St
101
Union Square
Geary St
O'Farrell St
ANTOWN
Fillmore
Franklin St
Gough St
Turk St
Golden Gate Ave
Market St
6th St
Mission St
5th St
7th St
8th St
SOUTH OF MARKET
Hall
1
2
3
4
5
6
7
8

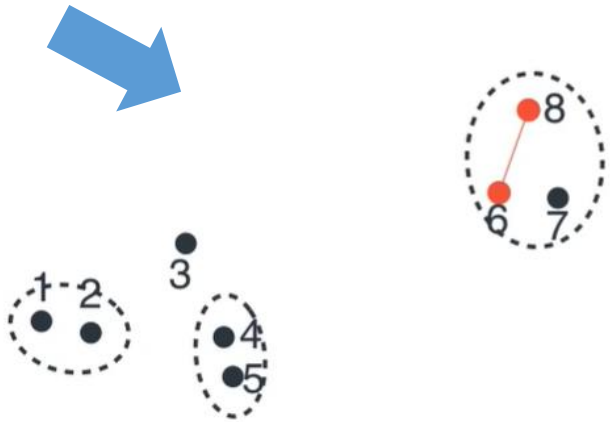

Distance
1
2
3
4
5
6
7
8

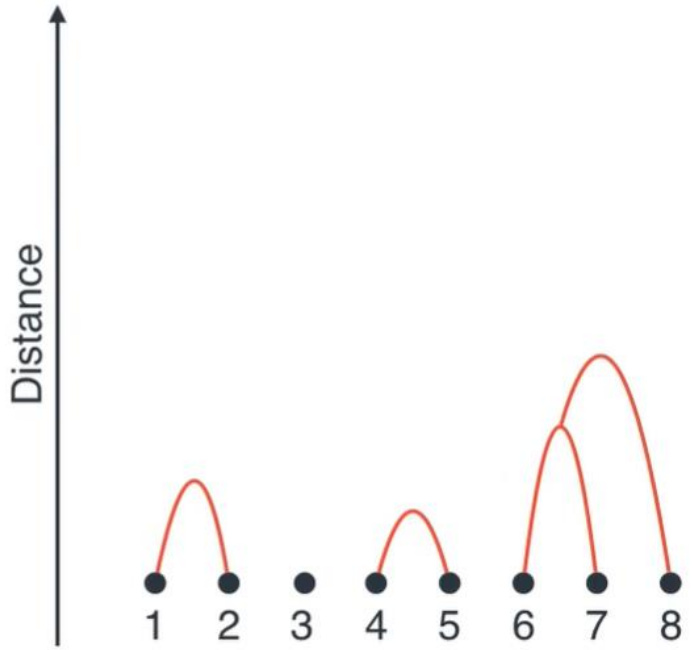
Dendrogram



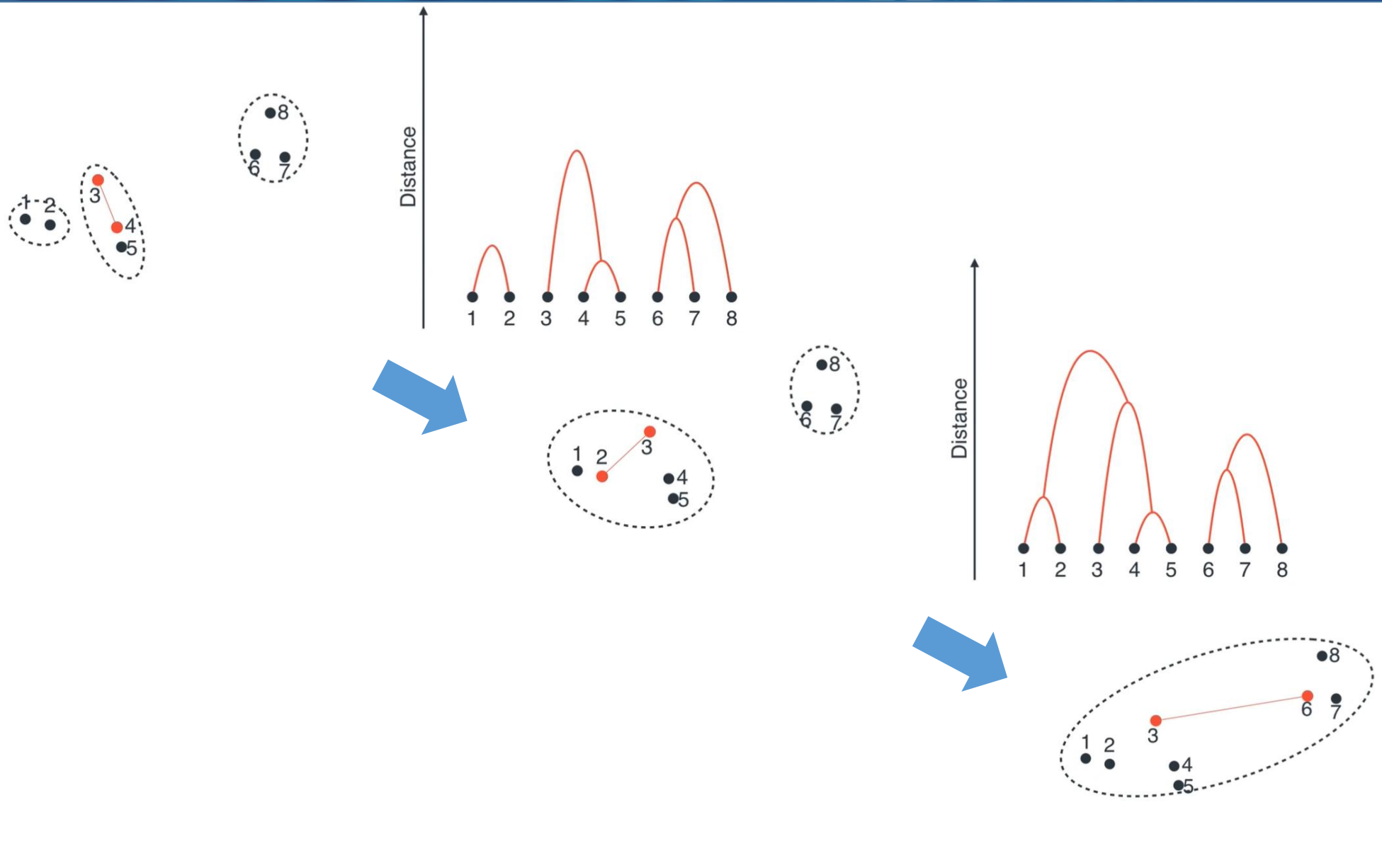
Dendrogram



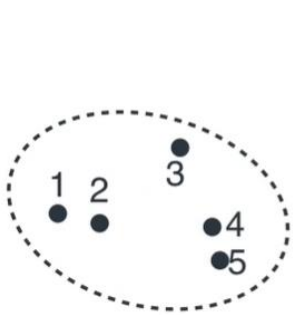


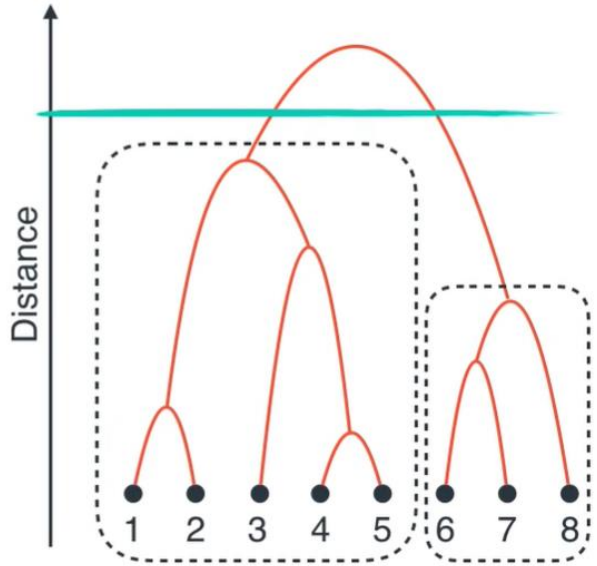


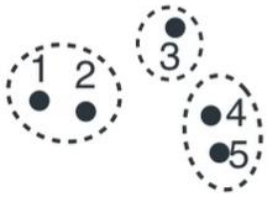
Dendrogram

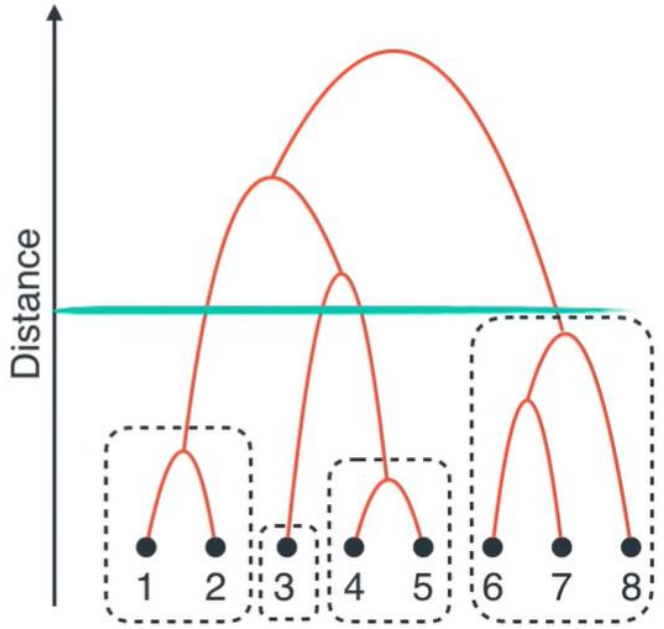


Dendrogram



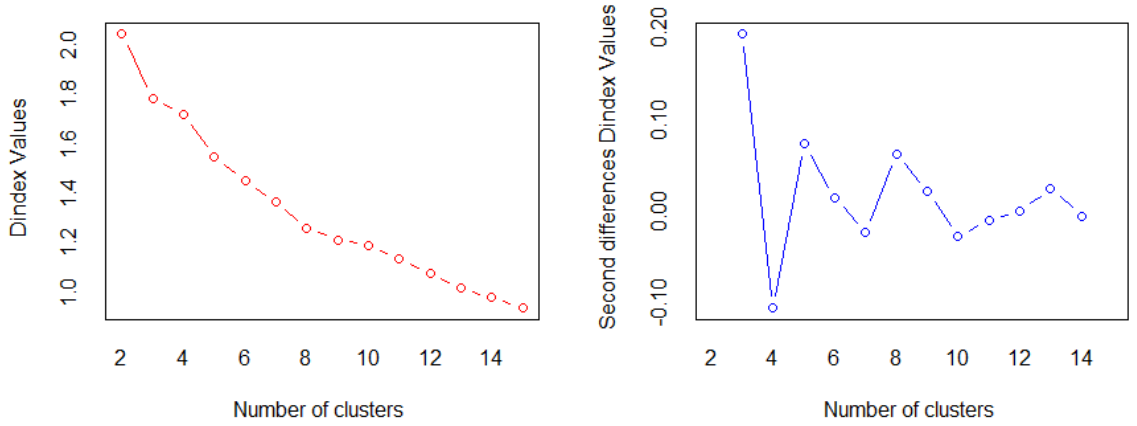




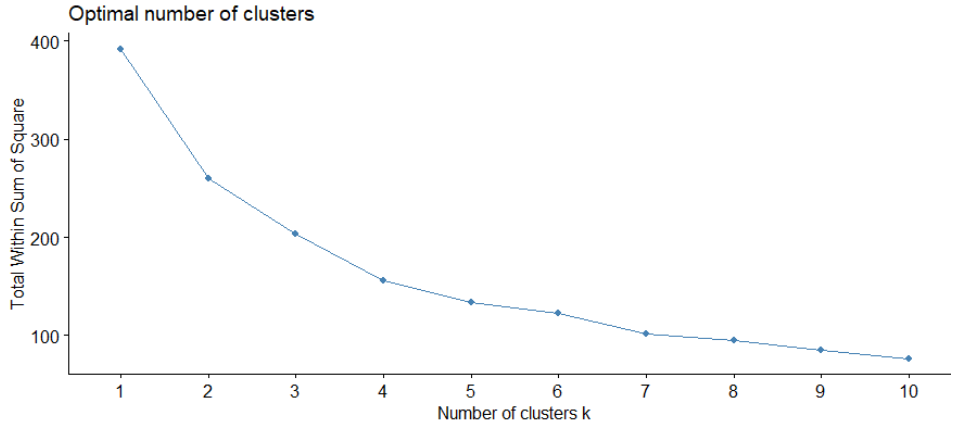


K-means Clustering: 예제

```
#### 데이터의 단위와 크기가 다르기 때문에 표준화하여야 한다.
state.scaled <- scale(state.x77)


#### clustering 학습을 위한 패키지 설치
install.packages("NbClust")
library(NbClust)

set.seed(123)  ### clustering 학습을 위한 seed No. 를 지정해주는 것
nc <- NbClust(state.scaled, distance = "euclidean",
              method = "kmeans",
              min.nc = 2, max.nc = 15 )

str(nc)
nc$Best.nc   ### 각 지표별로(KL, CH, Hartigan,...) 지지하는 군집의 갯수를 나타냄

table(nc$Best.nc[1,]) ### table()를 이용하여 지지한 군의 개수별 몇개의 지표가 지지하는지 확인


##### "fviz"함수를 사용하기 위해 필요한 패키지 설치
install.packages("factoextra")
install.packages("ggplot2")
library(factoextra)
library(ggplot2)


fviz_nbclust(state.scaled, kmeans, method = "wss") ### "wss": within sum of square
```




```
> table(nc$Best.nc[1,]) ### table()를 이용하여 지지한 군의 개수별 몇개의 지표가 지지하는지 확인

 0  1  2  3  4  5  6  8  9 15
 2  1  6  6  2  1  1  2  1  4
```



K-means Clustering: 예제

```
set.seed(123)
clustering.km <- kmeans(state.scaled, centers = 3, nstart = 25)

str(clustering.km)

clustering.km$cluster    ### Clustering 결과
clustering.km$size       ### 3개 Clustering 별로 포함된 갯수
clustering.km$centers    ### 3개 Clustering 각각의 특성을 확인할 수 있다.
```

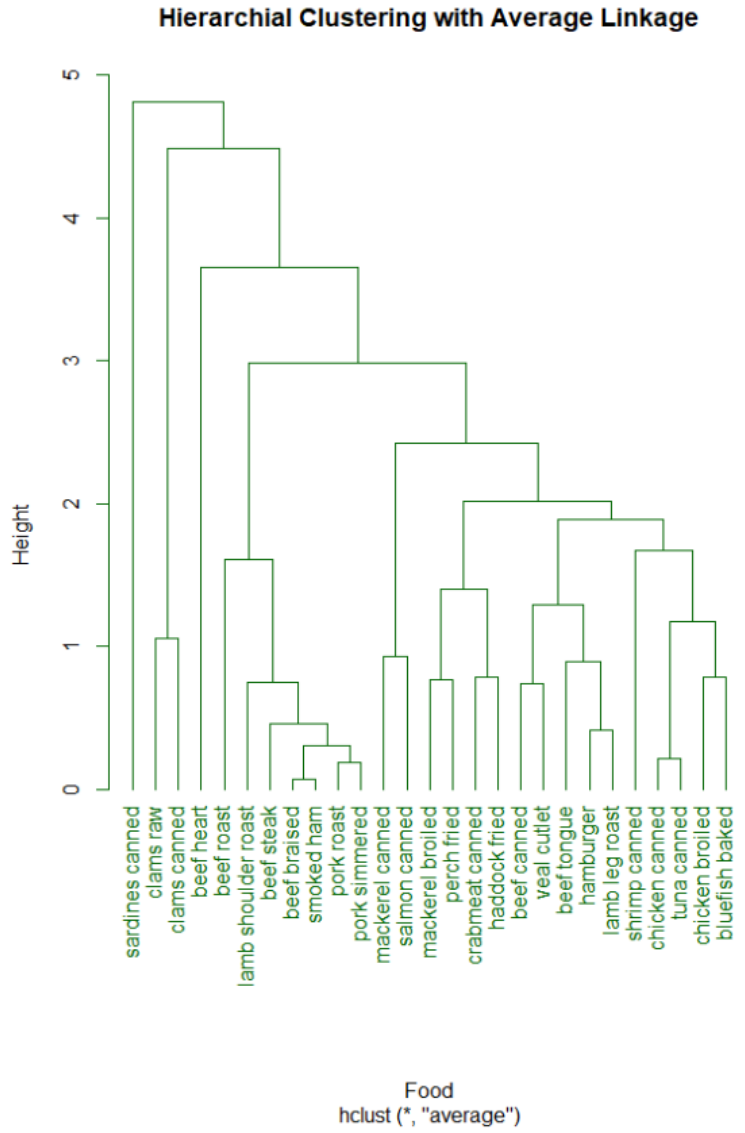
kmeans() 함수의 입력인자	
x	학습할 데이터(수치형 matrix)
centers	중심(k)의 수
iter.max	최대반복 산정수
nstart	초기값
algorithm	수행할 알고리즘 ("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen") default = "Hartigan-Wong"
trace	산정과정 출력 여부

```
> clustering.km$size       ### 3개 Clustering 별로 포함된 갯수
[1] 11 15 24
> clustering.km$centers    ### 3개 Clustering 각각의 특성을 확인할 수 있다.
  Population     Income   Illiteracy   Life Exp     Murder    HS Grad      Frost       Area
1 -0.2269956 -1.3014617  1.391527063 -1.1773136  1.0919809 -1.4157826 -0.7206500 -0.2340290
2  0.9462026  0.7416690  0.005468667 -0.3242467  0.5676042  0.1558335 -0.1960979  0.4483198
3 -0.4873370  0.1329601 -0.641201154  0.7422562 -0.8552439  0.5515044  0.4528591 -0.1729366
```

kmeans() 함수의 출력인자	
cluster	학습데이터의 모델에 의한 군집 분류 결과
centers	각 군집의 중심점
totss	전체변동
withinss	각 군집내 변동
tot.withinss	군집내 변동의 합
betweenss	군집간 변동
size	각 군집내 데이터 수
iter	반복횟수

Hierarchical Clustering: 예제

```
install.packages("flexclust")   ### 데이터 셋 "nutrient"를 불러오기 위해 필요한 패키지의 설치
library(flexclust)

data(nutrient)
str(nutrient)
head(nutrient)


nutrition <- nutrient

### 음식이름 대문자를 소문자로 변경하는 경우, 음식이름만을 변경하는 것이기 때문에 해당되는 행(row)에만
### 그래서, 음식이름을 소문자로 변경한 결과를 row.names(nutrition) 로 저장하여야만 한다.

row.names(nutrition) <- tolower(row.names(nutrition))
head(nutrition)


#### 데이터의 단위와 크기가 다르기 때문에 표준화하여야 한다.
nutrition.scaled <- scale(nutrition)



#### hierarchical clustering을 위해서는 가장 먼저 거리를 계산하여야 한다.
d <- dist(nutrition.scaled)   ### default method는 유클리드 거리

clustering.average <- hclust(d, method = "average")  ### 평균거리를 이용한 군집 간의 clustering 실시

plot(clustering.average, hang = -1,
     col="darkgreen", xlab = "Food",
     main = "Hierarchial Clustering with Average Linkage")
```



dist() 함수의 입력인자	
x	데이터
method	거리 산정법 ("euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski")
diag	대각행렬 출력여부
upper	행렬의 위부분 출력여부
p	minkowski 거리의 산정계수

hclust() 함수의 입력인자	
d	거리 매트릭스
method	분석방법 ("ward", "single", "complete", "average", "mcquitty", "median", "centroid" 등)

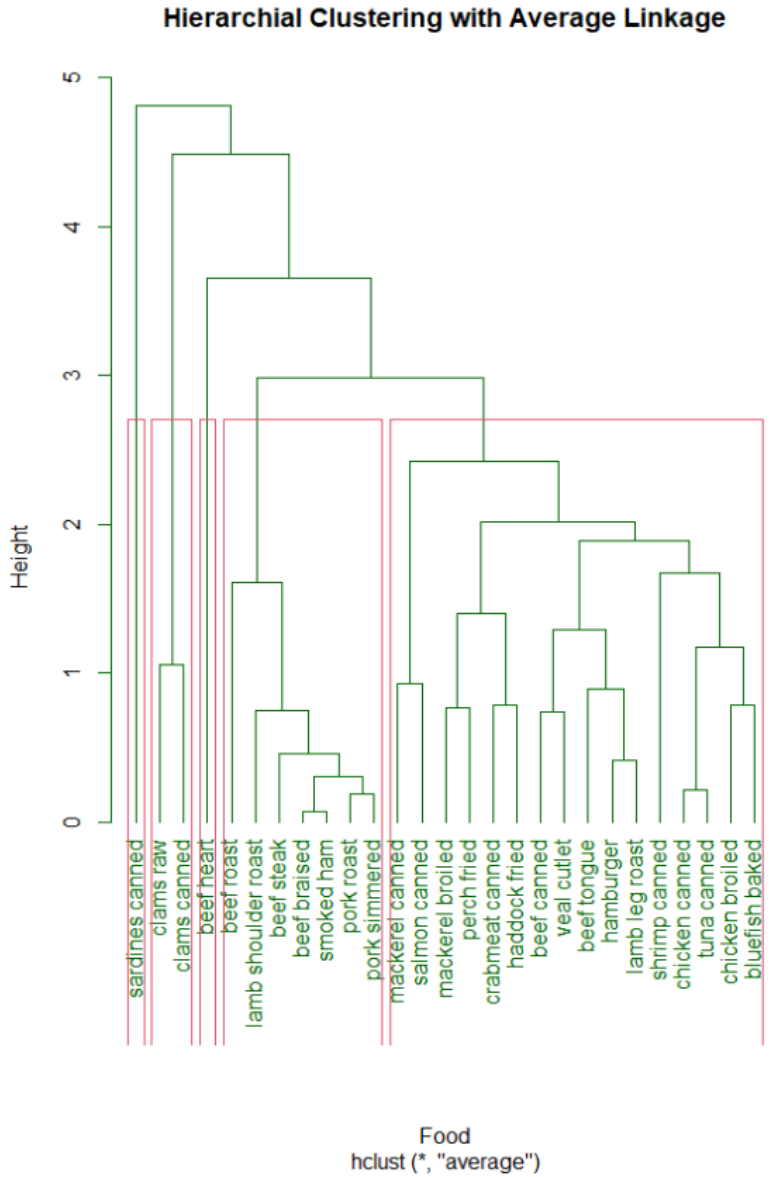
- 단일 연결법(single linkage): 가장 가까운 거리를 사용
- 완전 연결법(complete linkage): 가장 먼 거리를 사용
- 평균 연결법(average linkage): 평균한 거리를 사용
- 중심 연결법(centroid linkage): 중심 간의 거리를 사용

Hierarchical Clustering: 예제

```
set.seed(100)  ### clustering 학습을 위한 seed No. 를 지정해주는 것
nc <- NbClust(nutrition.scaled, distance = "euclidean",
              method = "average",
              min.nc = 3, max.nc = 15 )

str(nc)
nc$Best.nc   ### 각 지표별로(KL, CH, Hartigan,...) 지지하는 군집의 갯수를 나타냄

table(nc$Best.nc[1,]) ### table()를 이용하여 지지한 군의 개수별 몇개의 지표가 지


clusters <- cutree(clustering.average, k=5)
clusters
table(clusters)



plot(clustering.average, hang = -1,
     col="darkgreen", xlab = "Food",
     main = "Hierarchial Clustering with Average Linkage")

rect.hclust(clustering.average, k=5)  ### 클러스터를 덴드로그램에 표시
```

- cutree() 함수는 계층적 군집분석(hierarchical clustering)의 결과를 원하는 수의 군집으로 나누는 데 사용되는 함수

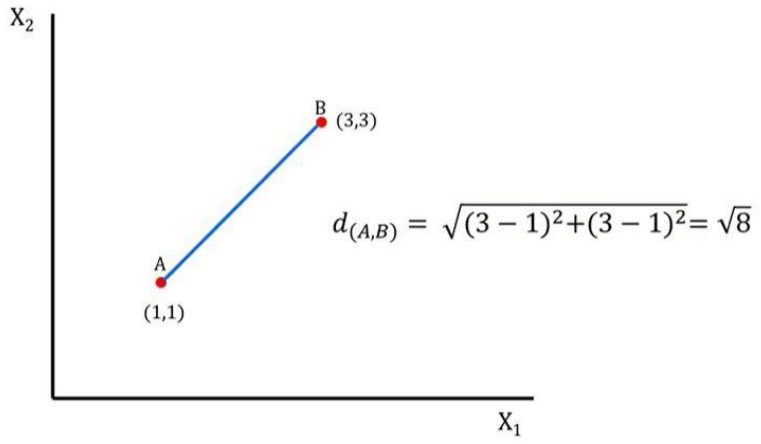
```
> table(nc$Best.nc[1,]) ### tabl
 0  3  4  5  9 10 13 14 15
 2  5  3  7  1  1  2  1  4
```



Clustering에 사용되는 거리 종류

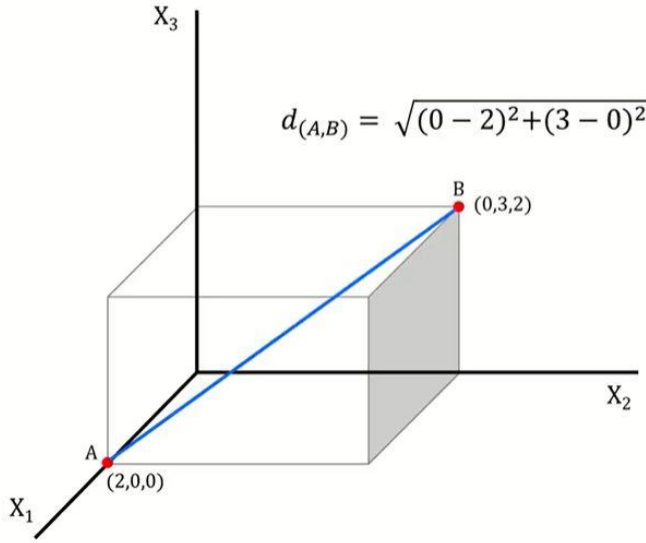
- **Euclidean Distance**

Euclidean Distance



$$d_{(A,B)} = \sqrt{(3-1)^2+(3-1)^2} = \sqrt{8}$$

Euclidean Distance

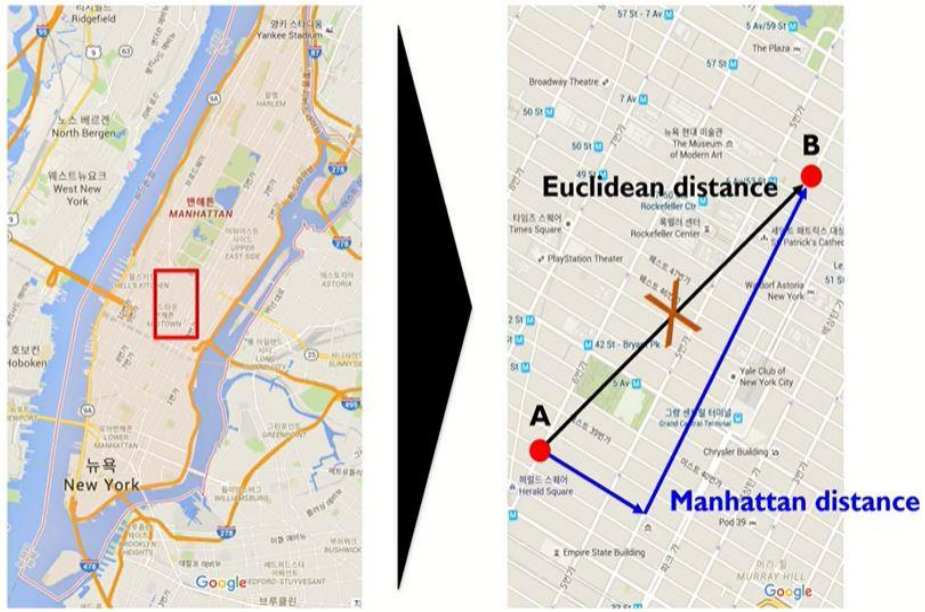


$$d_{(A,B)} = \sqrt{(0-2)^2+(3-0)^2+(2-0)^2} = \sqrt{17}$$

Clustering에 사용되는 거리 종류

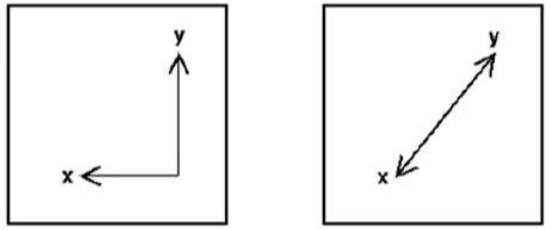
- **Manhattan Distance**

Manhattan Distance



Manhattan Distance

$$d_{Manhattan\,(X,Y)} = \sum_{i=1}^{n} |x_i - y_i|$$



• X에서 Y로 이동 시 각 좌표축 방향으로만 이동할 경우에 계산되는 거리

Clustering에 사용되는 거리 종류

● Mahalanobis Distance

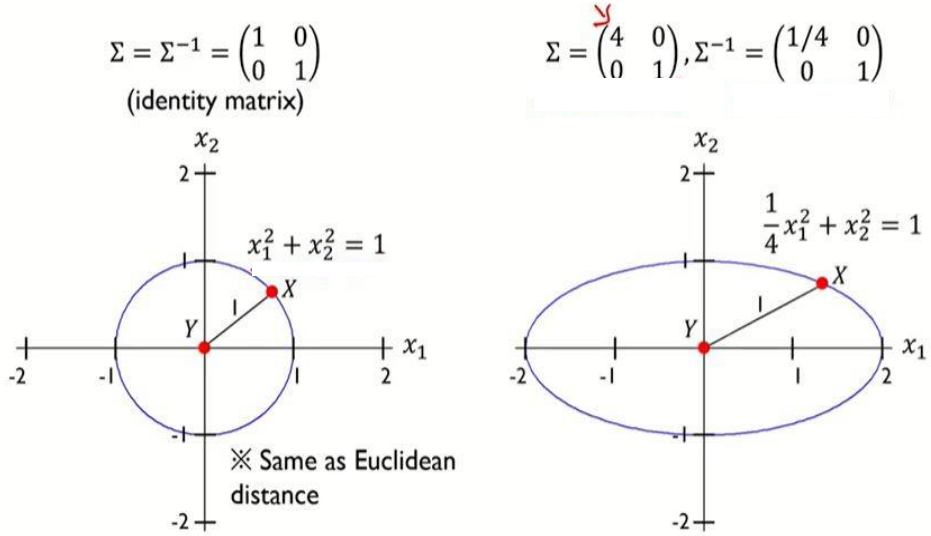
Mahalanobis Distance

$$d_{Mahalanobis\,(X,Y)} = \sqrt{(X-Y)^T\Sigma^{-1}(X-Y)},$$

Σ^{-1}: inverse of covariance matrix

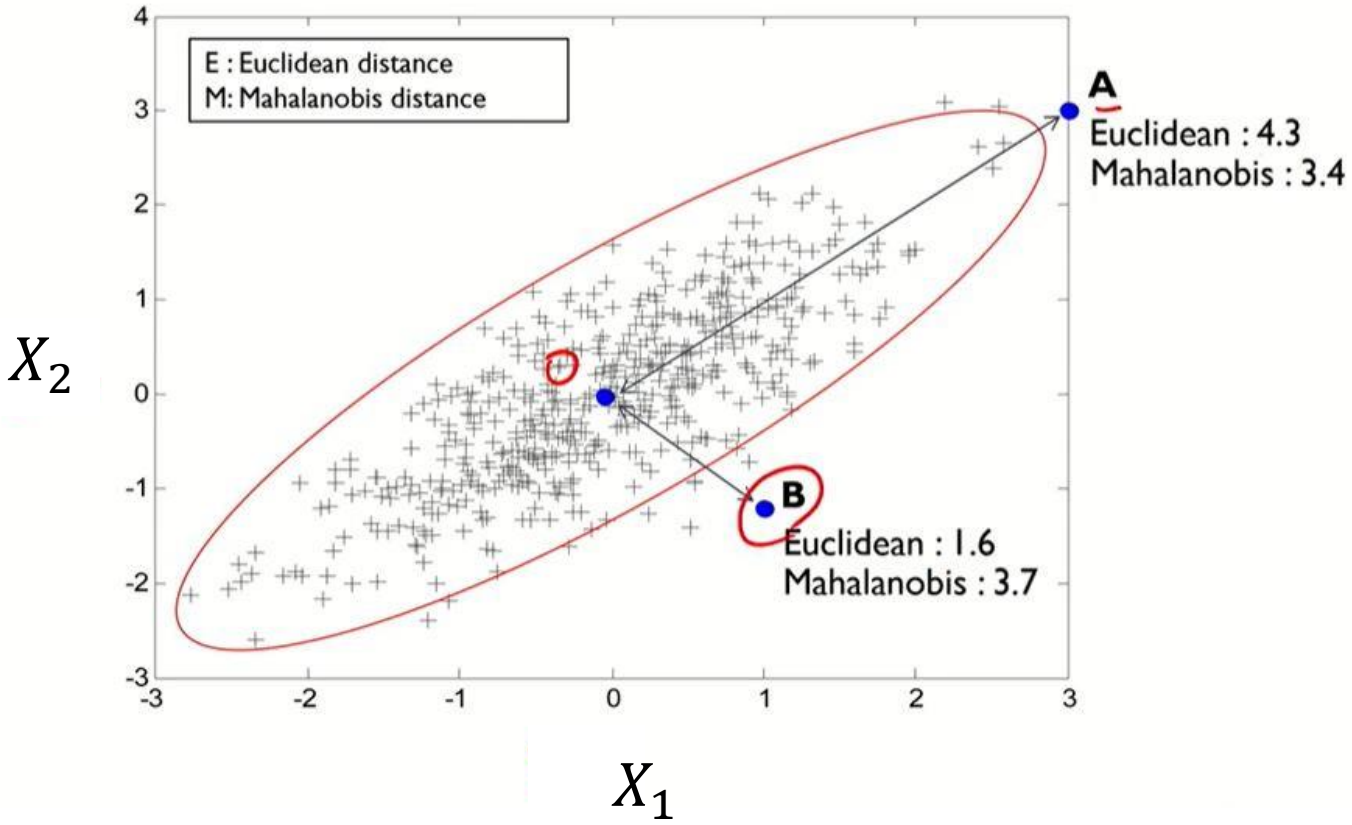
- 변수 내 분산, 변수 간 공분산을 모두 반영하여 X,Y 간 거리를 계산하는 방식
- 데이터의 covariance matrix가 identity matrix인 경우는 Euclidean distance와 동일함

Mahalanobis Distance



Clustering에 사용되는 거리 종류

Mahalanobis Distance



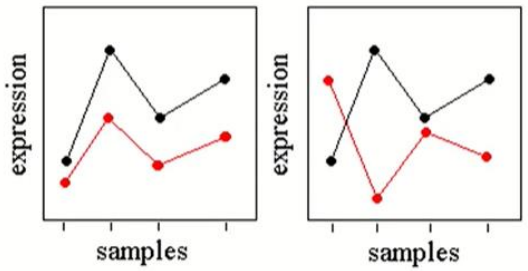
Clustering에 사용되는 거리 종류

● Spearman Rank Correlation Distance

Correlation Distance

$$d_{Corr(X,Y)} = 1 - r$$

$$where\ r = \sigma_{XY}$$



• 데이터 간 Pearson correlation을 거리측도로 사용하는 방식으로, 데이터 패턴의 유사도를 반영할 수 있음

Spearman Rank Correlation Distance

$$d_{Spearman(X,Y)} = 1 - \rho,$$

$$where\ \rho = 1 - \frac{6\sum_{i=1}^{n}(rank(x_i) - rank(y_i))^2}{n(n^2-1)}$$

• ρ를 Spearman correlation이라 하며, 이는 데이터의 rank를 이용하여 correlation distance를 계산하는 방식임

• ρ의 범위는 -1 부터 1로, Pearson correlation과 동일

Clustering에 사용되는 거리 종류

Spearman Rank Correlation Distance

계절 평균 낮 최고 기온

지역	봄	여름	가을	겨울
서울	17.06	28.43	19.07	3.50
뉴욕	16.32	28.22	18.37	5.43
시드니	22.23	17.03	21.90	25.63

지역 별 계절 기온 순위

지역	봄	여름	가을	겨울
서울	3	1	2	4
뉴욕	3	1	2	4
시드니	2	4	3	1

서울 - 뉴욕 간 Spearman correlation distance:

$$\rho = 1 - \frac{6\{(3-3)^2 + (1-1)^2 + (2-2)^2 + (4-4)^2\}}{4(4^2-1)} = 1 \rightarrow d_{(서울,뉴욕)} = 1 - 1 = 0$$

서울 - 시드니 간 Spearman correlation distance:

$$\rho = 1 - \frac{6\{(3-2)^2 + (1-4)^2 + (2-3)^2 + (4-1)^2\}}{4(4^2-1)} = -1 \rightarrow d_{(서울,시드니)} = 1 - (-1) = 2$$

PVA vs. Cluster Analysis

- 주성분 분석은: **변수 관점**에서의 처리 방식
  - ✓ 데이터의 분산을 최대한 보존하면서 차원 축소를 수행
  - ✓ 즉, 데이터의 분산이 가장 큰 방향으로 **새 변수를 생성**하고, **생성된 변수를 주성분**이라고 한다.
  - ✓ **데이터의 변동성**을 가장 잘 설명하는 **변수(:주성분)을 추출**하여 데이터의 구조를 이해하고 분석하는 데 유용
  - ✓ 예를 들어, 제품의 품질을 평가하기 위해 제품의 여러 특성을 측정하였다고 가정하자. 이 경우 주성분 분석을 사용하여 *제품의 품질을 가장 잘 설명하는 특성을 추출*할 수 있습니다.

- 군집분석은: **데이터 관점**에서의 처리 방식
  - ✓ **유사한 데이터**를 하나의 군집으로 묶는 방법
  - ✓ 데이터의 특성을 기반으로 **데이터를 군집으로 분류**하여, 데이터의 구조를 이해하고 분석하기 쉽게 한다.
  - ✓ 예를 들어, 고객의 구매 데이터를 분석하여 고객을 유사한 특성을 가진 군집으로 분류한다고 가정하자. 이 경우 군집분석을 사용하여 *고객의 성향이나 구매 패턴*을 이해할 수 있다.

- 요인분석은: **데이터 관점**에서의 처리 방식
  - ✓ 데이터의 내재된 구조를 설명하는 요인을 추출하는 방법
  - ✓ **데이터의 상관관계를 기반**으로 요인을 추출하여 데이터의 구조를 이해하고 분석하는데 유용
  - ✓ 예를 들어, 제품의 품질을 평가하기 위해 제품의 여러 특성을 측정하였다고 가정하자. 이 경우 요인분석을 사용하여 제품의 품질을 결정하는 내재된 요인을 추출할 수 있다.