데이터 전처리
Data Preprocessing

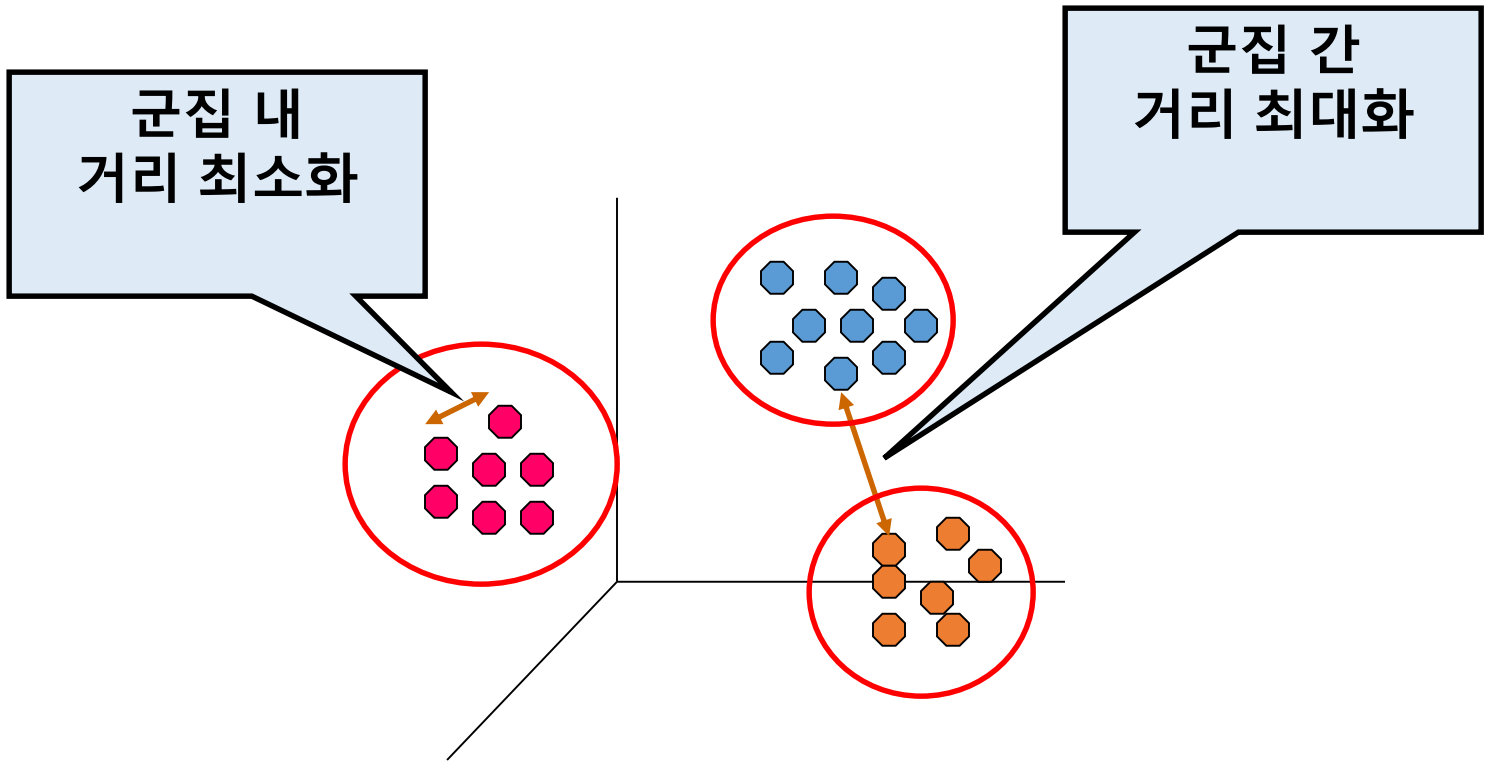
10
군집 분석(Cluster Analysis)

목차

1. 군집 분석

군집 분석Cluster Analysis

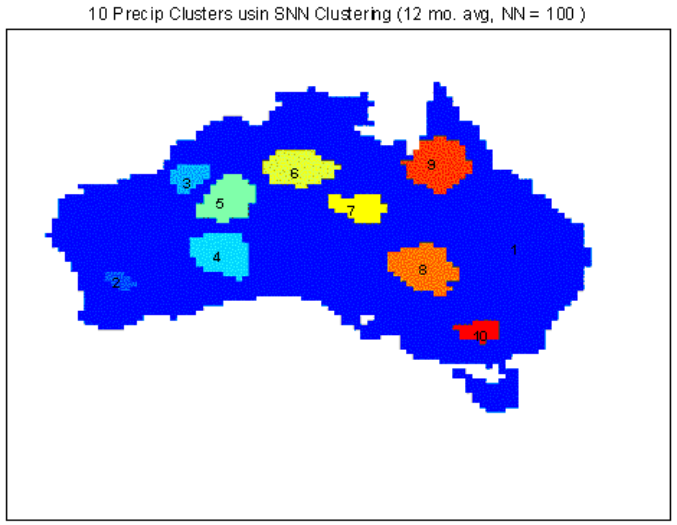
- 그룹의 객체가 서로 유사하거나(또는 관련된) 다른 그룹의 객체와 다르거나(또는 관련 없는) 객체의 그룹 찾기



군집 분석 응용

- 이해Understanding
  - 탐색을 위한 관련된 문서 그룹
  - 유사한 기능을 갖는 유전자 및 단백질 그룹
  - 유사한 가격 변동을 가진 주식 그룹
- 요약Summarization
  - 대규모 데이터 집합의 크기 줄이기

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



호주의 강수량 군집화

군집 분석이 아닌것은?

- 단순한 세분화Simple segmentation
  - 학생을 알파벳 순으로, 성별로 다른 등록 그룹으로 나누기

- 질의 결과Results of a query
  - 그룹화는 외부 사양specification의 결과
  - 군집화는 데이터를 기반으로 하는 객체의 그룹화

- 지도 분류Supervised classification
  - 클래스 라벨 정보가 필요

- 연관 분석Association Analysis
  - 로컬Local vs. 글로벌 연결global connections

군집 개념이 모호할 수 있음

얼마나 많은 군집이 있는가?

6개 군집

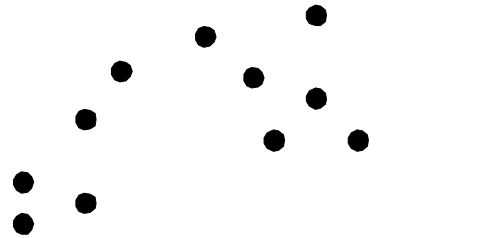
2개 군집

4개 군집

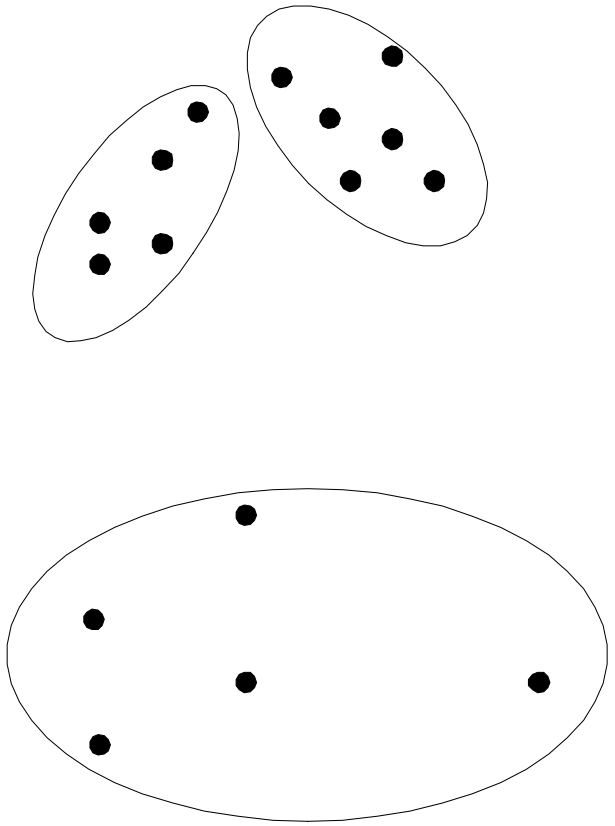
군집화 유형

- 군집화는 군집 집합
- 군집의 계층 집합과 분할 집합 간의 중요한 차이점
- 분할 군집화Partitional Clustering
  - 각 데이터 객체가 정확히 하나의 부분집합에 있도록 비중복non-overlapping 부분 집합(군집)으로 데이터 객체를 나눔
- 계층 군집화Hierarchical clustering
  - 계층 트리로 구성된 중첩된 군집 집합

분할 군집화Partitional Clustering

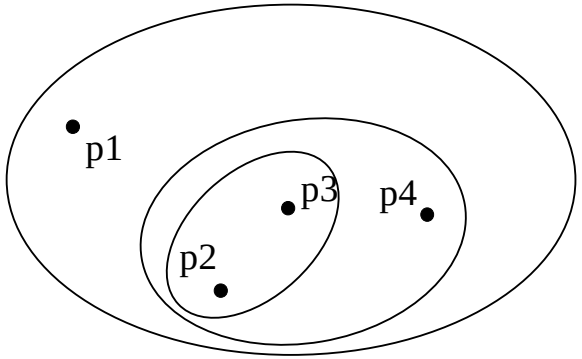


원본 포인트

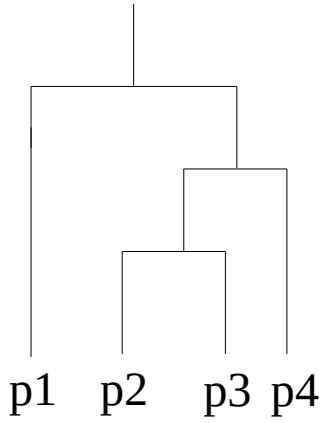


분할 군집화

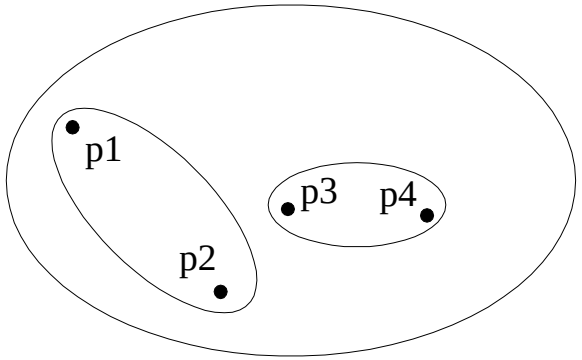
계층 군집화Hierarchical Clustering



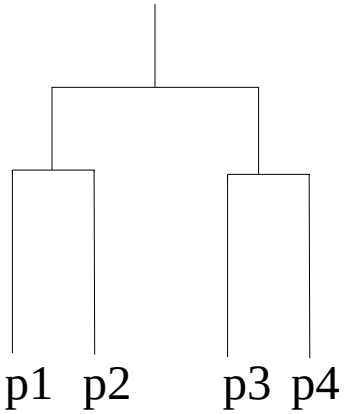
전통적인 계층 군집화



전통적인 덴드로그램Dendrogram



비전통적인 계층 군집화



비전통적인 덴드로그램Dendrogram

군집 집합 간의 다른 구별Distinctions

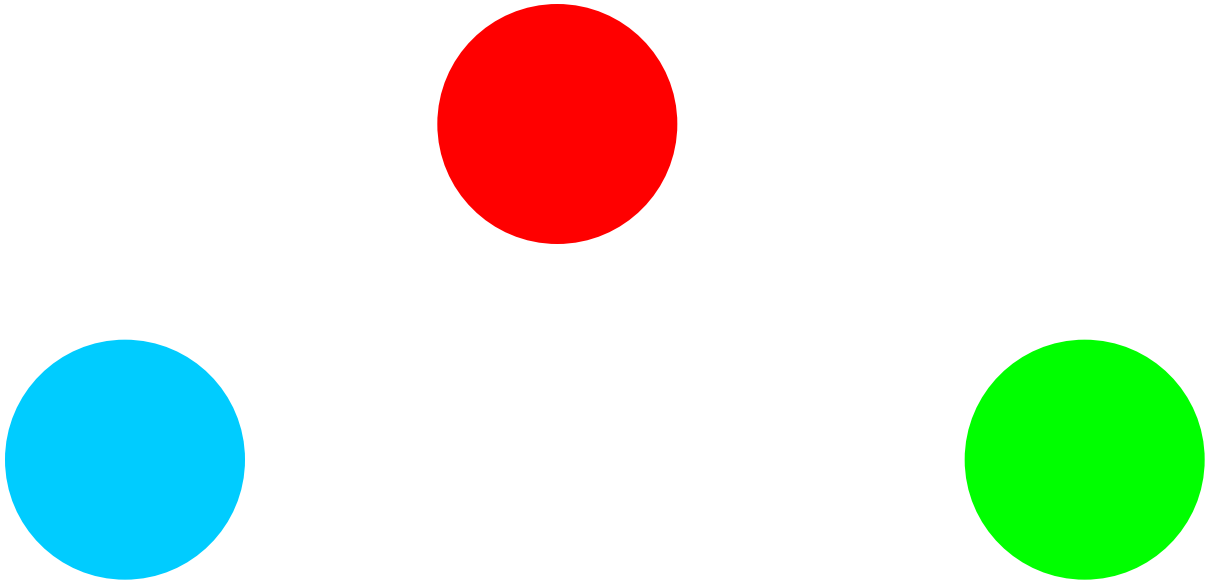
- 독점exclusive vs. 비독점non- exclusive
  - 비베타적 군집화에서 포인트는 여러 군집에 속할 수 있음
  - 여러 클래스 또는 '경계' 포인트를 나타낼 수 있음
- 퍼지fuzzy vs. 비퍼지non-fuzzy
  - 퍼지 군집화에서, 포인트는 0과 1 사이의 가중치를 가진 모든 군집에 속함
  - 가중치는 1로 합쳐야 함
  - 확률적 군집화는 유사한 특성을 가짐
- 부분partial vs. 완료complete
- 어떤 경우에는 일부 데이터만 군집화하기 원함
- 이종의Heterogeneous vs. 동종의homogeneous
  - 넓게 다른 크기, 모양, 그리고 밀도의 군집

군집 유형

- 잘 분리된 군집 Well-separated clusters
- 중심 기반 군집 Center-based clusters
- 인접 군집 Contiguous clusters
- 밀도 기반 군집 Density-based clusters
- 소유Property 또는 개념Conceptual
- 목적 함수에 의해 기술됨

군집 유형: 잘 분리된 군집Well-separated

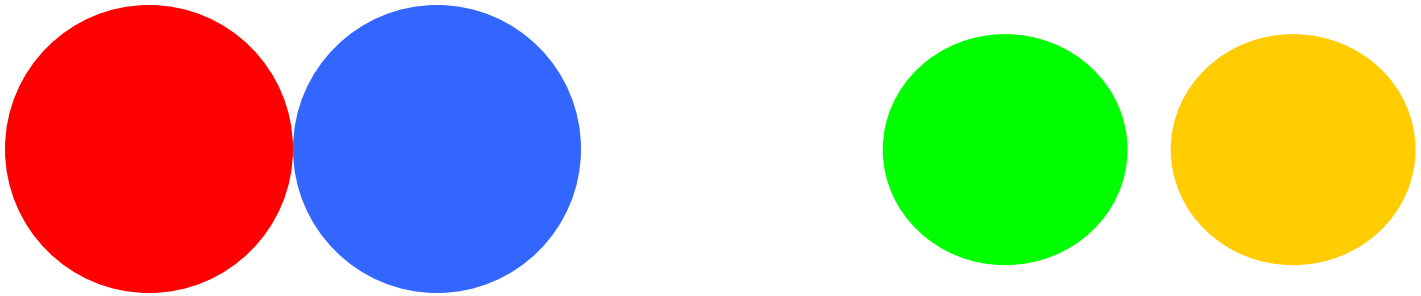
- 잘 분리된 군집 Well-Separated Clusters:
  - 군집은 군집에 없는 어느 포인트보다 군집에 있는 다른 모든 포인트에 가까운(또는 더 유사한) 군집에 있는 군집의 어느 포인트의 집합



3개의 잘 분리된 군집

군집 유형: 중심 기반Center-Based

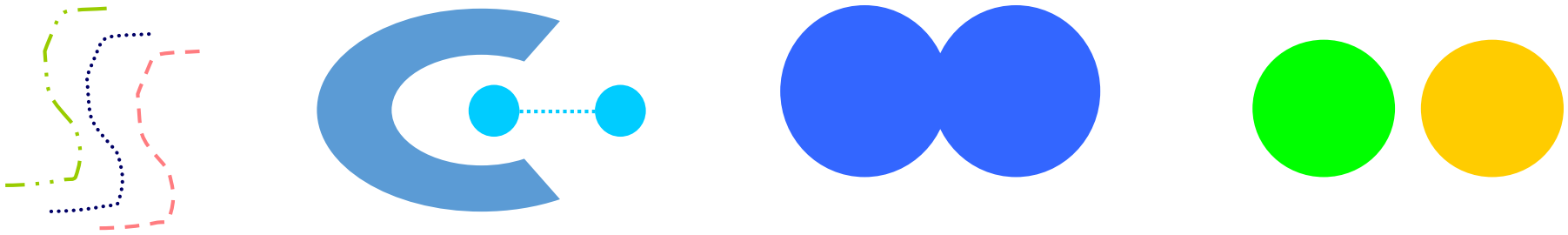
- 중심 기반Center-based
  - 군집은 군집의 객체가 다른 군집의 중심보다 군집의 "중심"에 더 가깝게(더 유사한)있는 객체 집합
  - 군집의 중심은 종종 중심점centroid, 군집의 모든 포인트의 평균, 또는 medoid, 이는 군집의 가장 대표적인 포인트



4개의 중심 기반 군집

군집 유형: 인접 기반Contiguous-Based

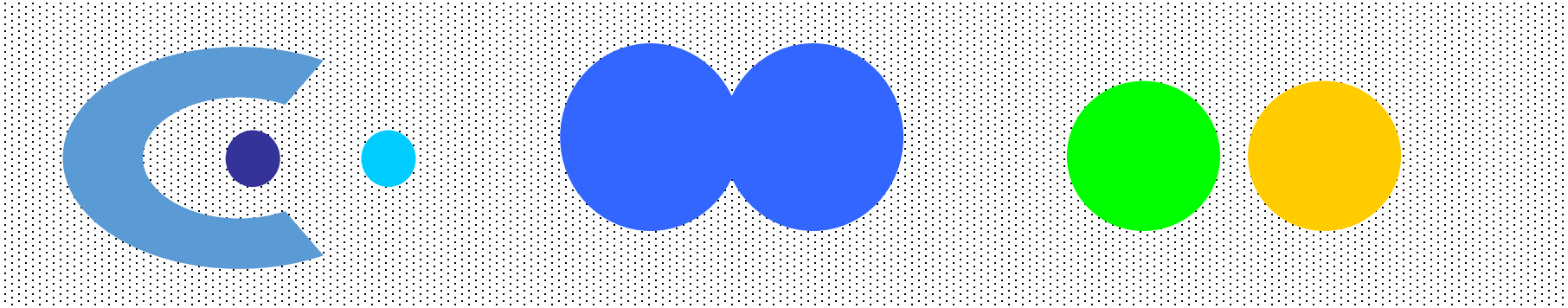
- 인접 군집Contiguous Cluster (인접 노드 또는 이행 노드)
  - 군집은 군집의 한 포인트가 군집에 없던 포인트보다 군집의 하나 이상의 다른 지점에 더 가깝거나 더 비슷한 포인트 집합



8 contiguous clusters

군집 유형: 밀도 기반Density-based

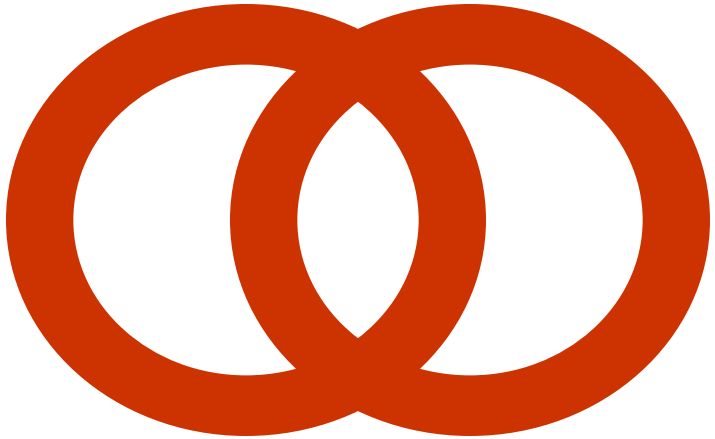
- 밀도 기반Density-based
  - 군집은 고밀도high density의 다른 범위에서 저밀도low-density 범위에 의해 분리되는 밀도가 높은 포인트 범위
  - 군집이 불규칙하거나 얽힌 경우, 그리고 노이즈 및 이상치가 있는 경우에 사용



6 density-based clusters

군집 유형: 개념적 군집Conceptual Clusters

- 공유 속성Shared Property 또는 개념적 군집Conceptual Clusters
  - 공통 속성을 공유하거나 특정 개념을 나타내는 군집을 찾음



2개의 겹치는 원형

군집 유형: 목적 함수Objective Function

- 목적 함수에 의해 정의된 군집
  - 목적 함수에 최소화하거나 최대화하는 군집을 찾음
  - 포인트를 군집으로 나눌 수 있는 가능한 모든 방법을 열거하고, 주어진 목적 함수를 사용하여 각 잠재적 군집 집합의 '우수성'을 평가 (NP Hard)
  - 글로벌 또는 로컬 목적objectives 을 가질 수 있음
    - 계층 군집화 알고리즘은 일반적으로 로컬 목적을 가지고 있음
    - 분할 알고리즘은 일반적으로 글로벌 목적을 가지고 있음
  - 글로벌 목적 함수 접근의 변형은 데이터를 매개 변수화된 모델에 맞추는 것
    - 모델의 매개 변수는 데이터로부터 결정
    - 혼합 모델Mixture models은 데이터가 여러 통계 분포의 '혼합 ' 이라고 가정

군집화 문제를 다른 문제에 매핑

- 군집화 문제를 다른 도메인에 매핑하고 해당 도메인의 관련 문제를 해결
  - 근접 행렬은 가중 그래프를 정의, 여기서 노드는 군집화 된 포인트이고, 가중치가 있는 에지edges는 포인트 사이의 인접성을 나타냄
  - 군집화는 각 군집마다 하나씩 연결된 구성 요소로 그래프를 분리하는 것과 같음
  - 군집간의 에지 가중치를 최소화하고 군집 내의 에지 가중치를 최대화하고 싶음

입력 데이터의 특성이 중요

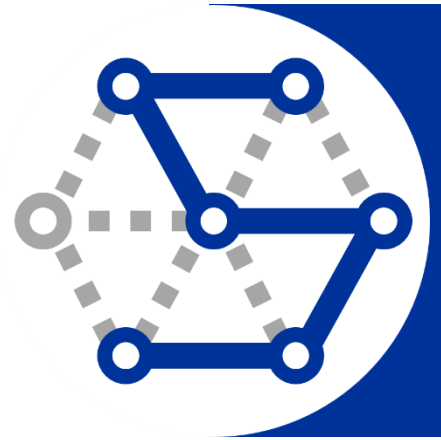
- 근접proximity 또는 밀도density 측정 유형
  - 군집화의 중심
  - 데이터 및 애플리케이션에 의존

- 근접 또는 밀도에 영향을 미치는 데이터 특성Data characteristics
  - 차원Dimensionality
    - 희박성Sparseness
  - 속성 유형Attribute type
  - 데이터의 특수 관계Special relationships
    - 예를 들어, 자기 상관autocorrelation
  - 데이터의 분포Distribution

- 노이즈Noise 및 이상치Outliers
  - 종종 군집화 알고리즘의 작동을 방해

군집화 알고리즘Clustering Algorithms

- K-means 및 다양한 변형
- 계층 군집화Hierarchical clustering
- 밀도 기반 군집화Density-based clustering



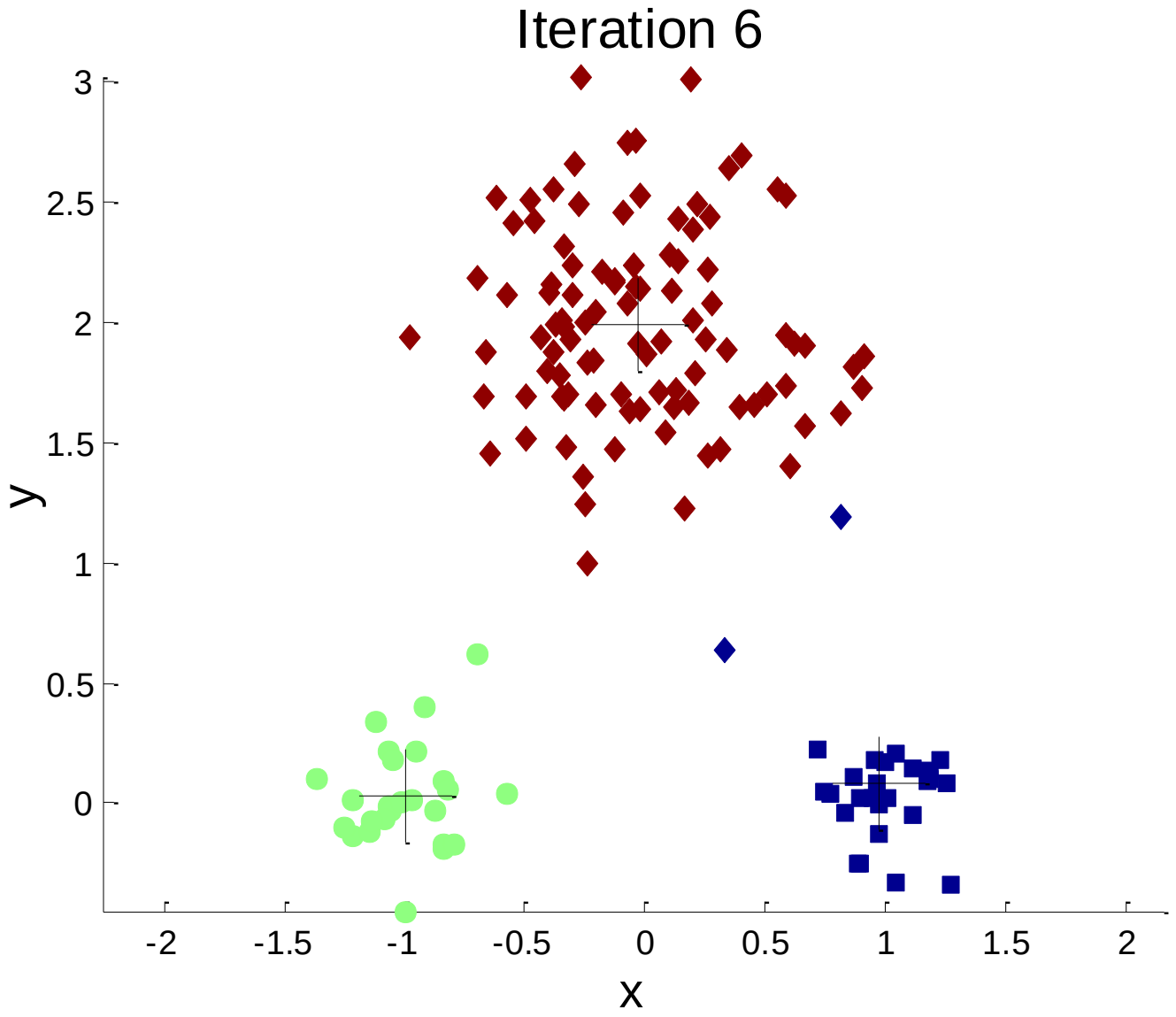
2. K-means

K-means 군집화

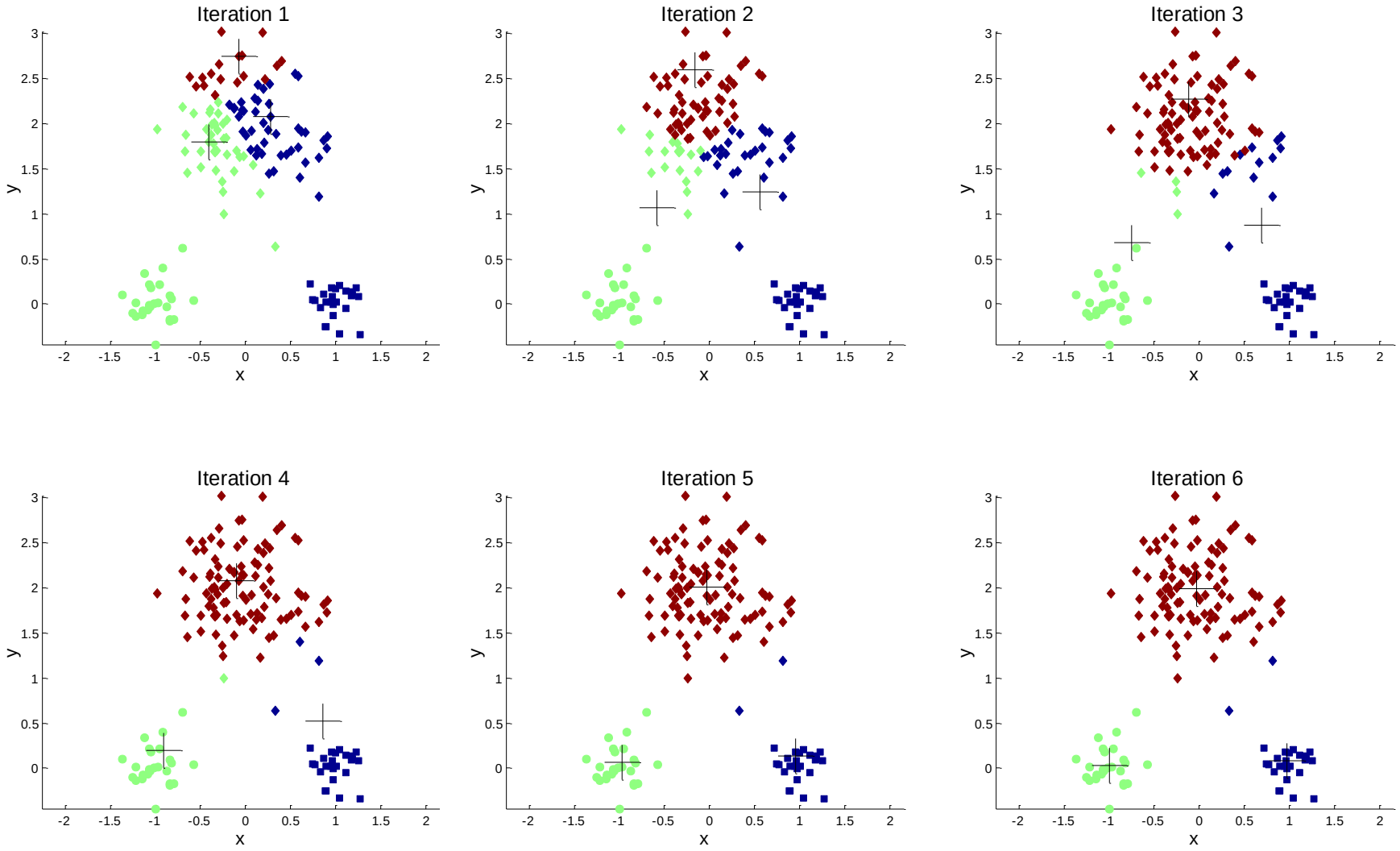
- 분할 군집화 접근
- 군집의 수 K를 지정
- 각 군집은 centroid(중심점)와 연합됨
- 각 포인트는 가장 가까운 중심을 가진 군집에 할당
- 기본 알고리즘은 매우 간단함

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-means 군집화 예제



K-means 군집화 예제



K-means 군집화 상세한 설명

- 초기 중심점centroids는 종종 무작위로 선택
  - 생산되는 군집은 실행마다 다름
- 중심centroid은 일반적으로 군집에 포인트들의 평균
- 근접Closeness은 유클리드 거리Euclidean distance, 코사인 유사성cosine similarity, 상관관계correlation 등으로 측정됨
- K-means는 위에서 언급한 공통 유사성 측정에 대해 수렴
- 대부분의 수렴은 처음 몇 번의 반복에서 발생
  - 종종 멈추는 조건stopping condition은 "비교적 적은 수의 점이 군집을 바꿀때까지"로 변경됨
- 복잡성Complexity은 $O(n * K * I * d)$
  - n = 포인트의 수, K = 군집의 수, I = 반복 횟수, d = 속성의 수

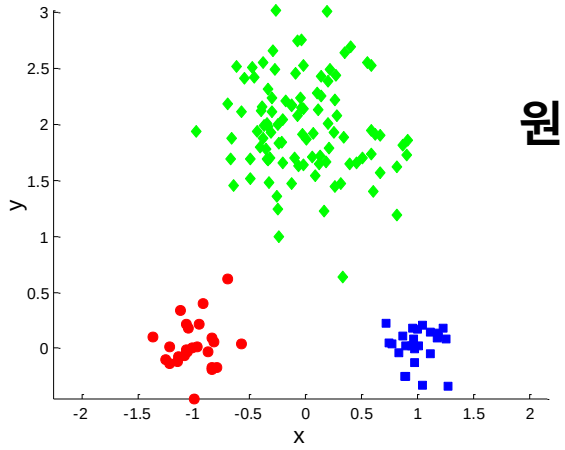
K-means Clusters 평가

- 가장 일반적인 측정값은 오차제곱합 SSE(Sum of Squared Error)
  - 각 포인트에 대해, 오차error는 가장 가까운 군집까지의 거리
  - SSE를 얻으려면 오차를 제곱하고 합해야 함

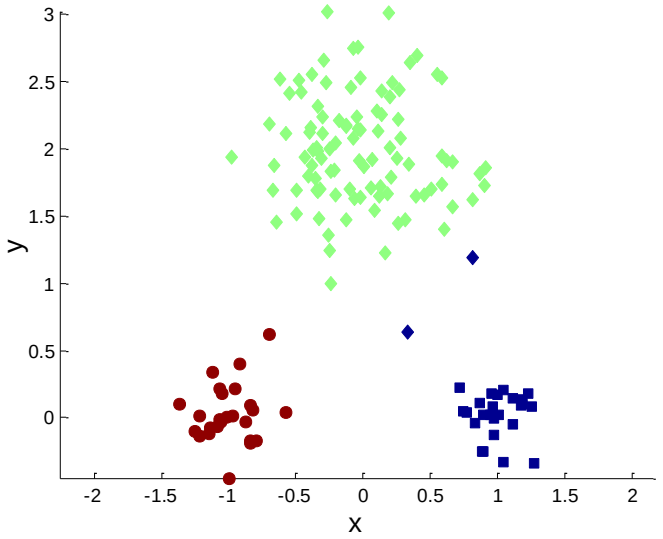
$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

  - x는 군집 C_i에 있는 데이터 포인트이고, m_i는 군집 C_i의 대표 포인트
    - m_i가 군집의 중심(평균)에 해당함을 나타낼 수 있음
  - 주어진 두개의 군집 집합에서 우리는 가장 작은 오차를 갖는 군집을 선호
  - SSE를 줄이는 쉬운 방법 중 하나는 군집 수를 K를 증가시켜 늘리는 것
    - 더 작은 K를 가진 좋은 군집화는 더 높은 K를 가지는 빈약한 군집화보다 작은 SSE를 가짐

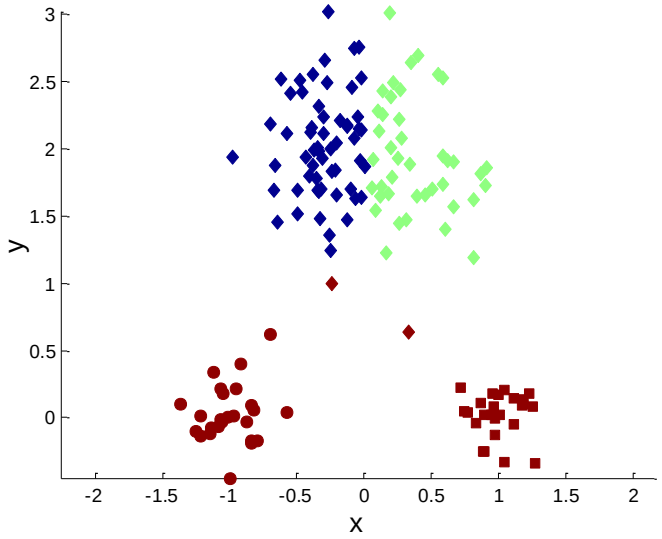
두개의 다른 K-means 군집화



원본 포인트



최적의 군집화



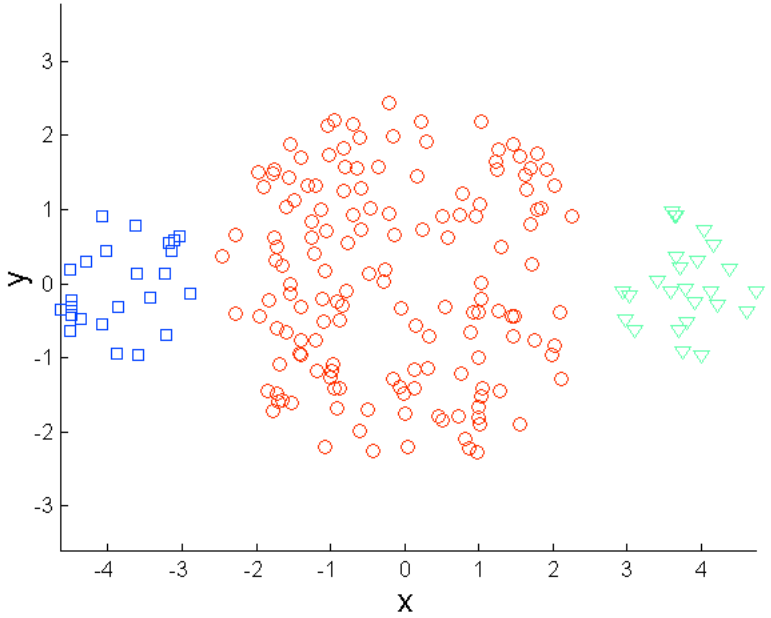
차선의 군집화

K-means의 한계

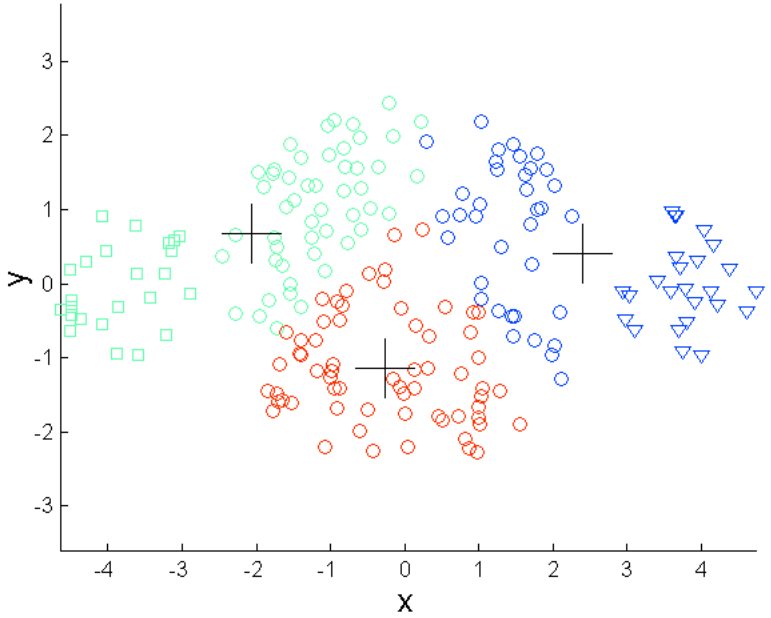
- K-means는 군집이 다음 조건이 다른 경우 문제
  - 크기Sizes
  - 밀도Densities
  - 비구 모양Non-globular shapes

- K-means는 데이터에 이상치가 포함되어 있을 때 문제

K-means의 한계: 다른 크기 Differing Sizes

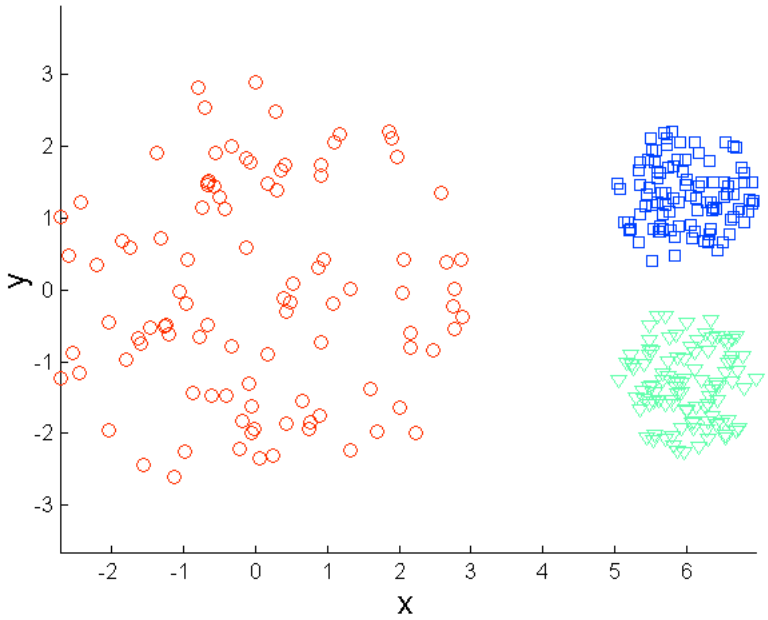


원본 포인트

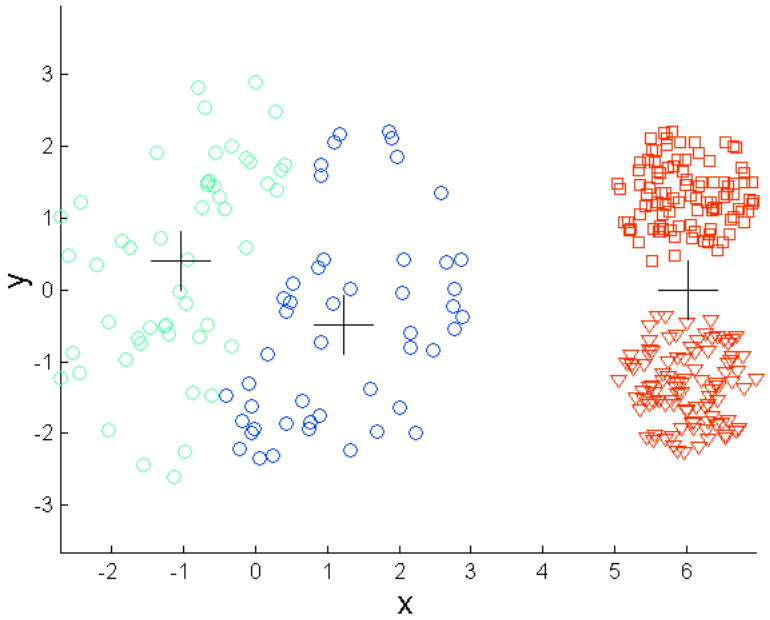


K-means (3개의 군집)

K-means의 한계: 다른 밀도 Differing Density

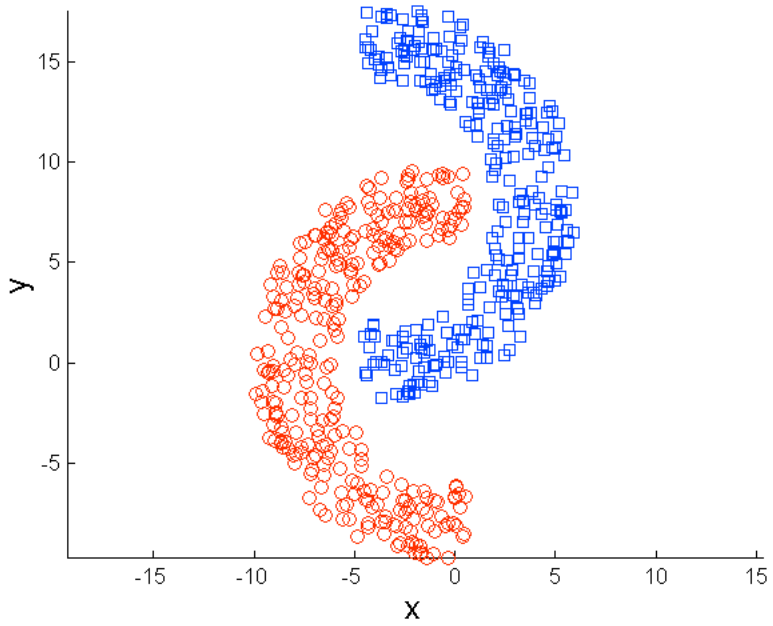


원본 포인트

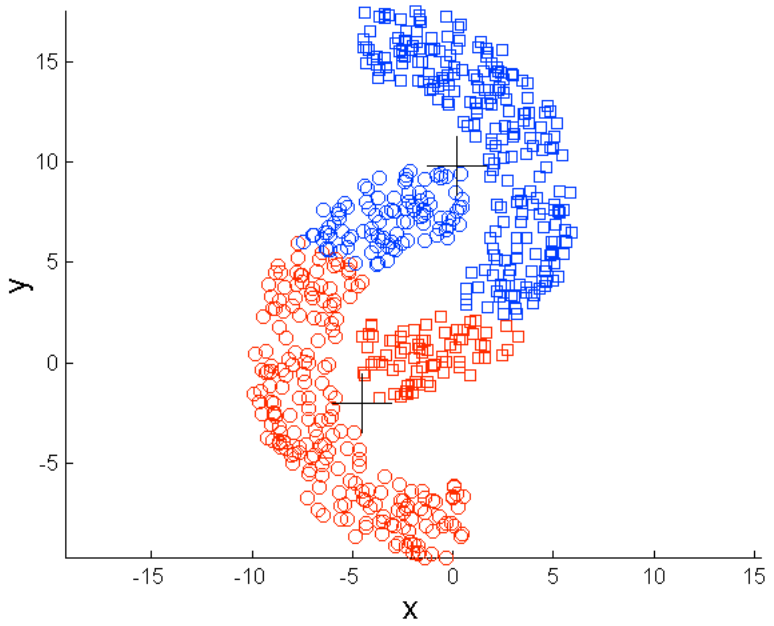


K-means (3개의 군집)

K-means의 한계: 비구 모양Non-globular Shapes

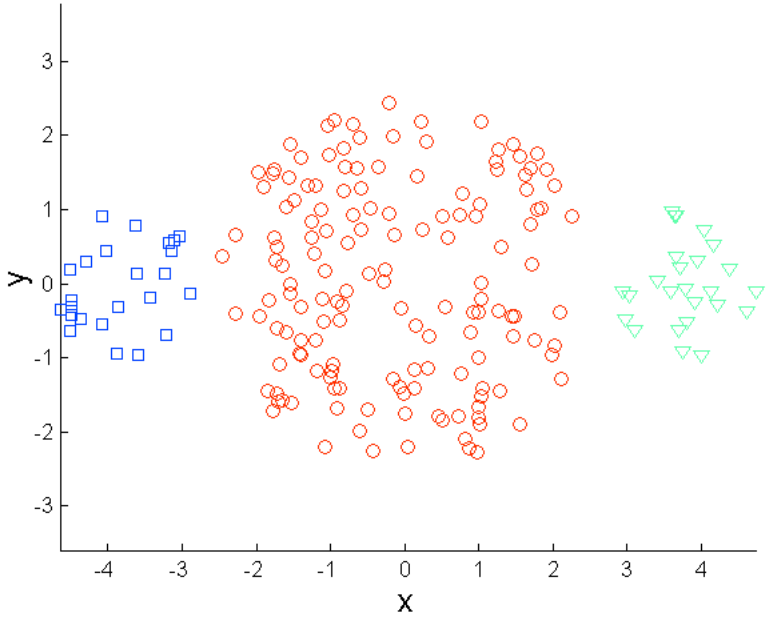


원본 포인트

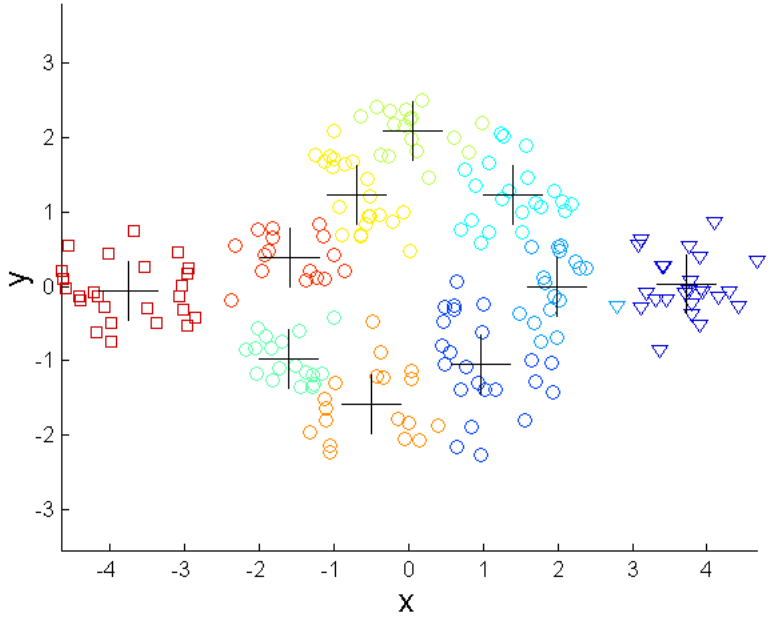


K-means (2개의 군집)

K-means 한계 극복



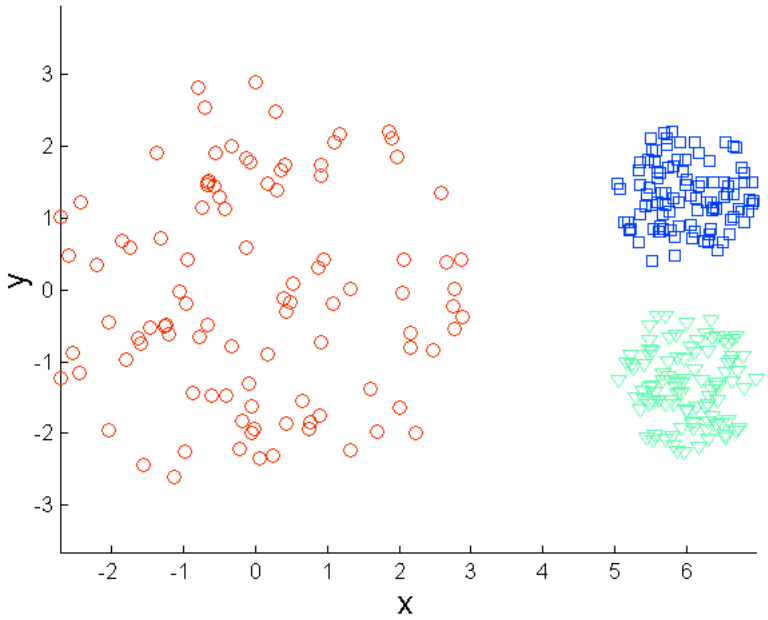
원본 포인트



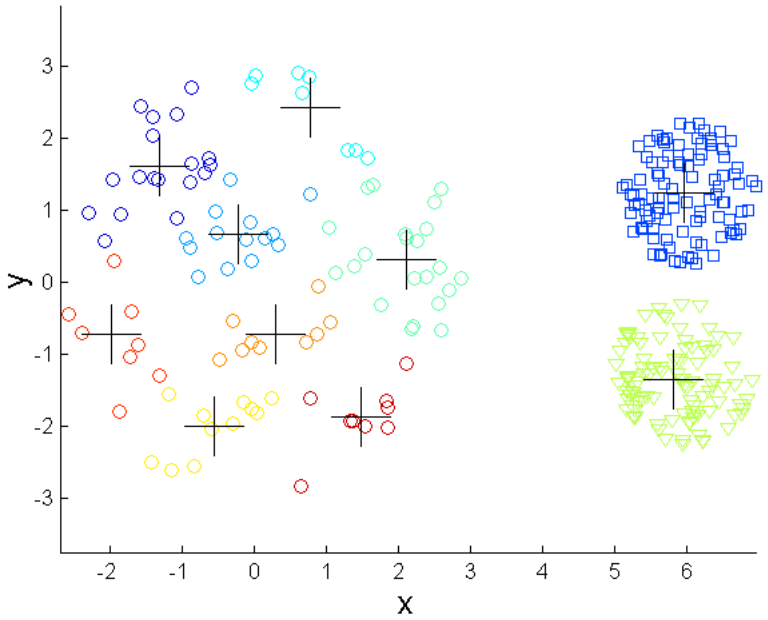
K-means 군집화

한 가지 해결책은 많은 군집을 사용하는 것
군집의 일부는 찾았지만 함께 사용해야 함

K-means 한계 극복

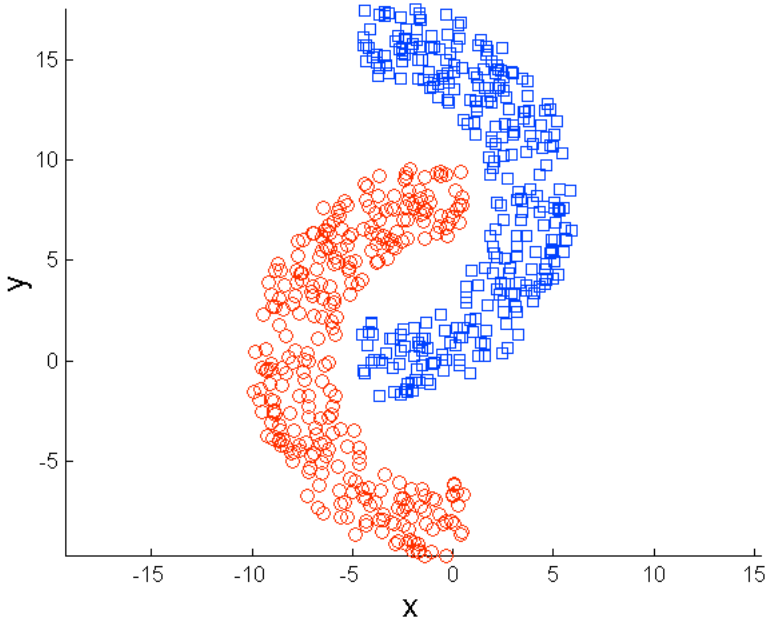


원본 포인트

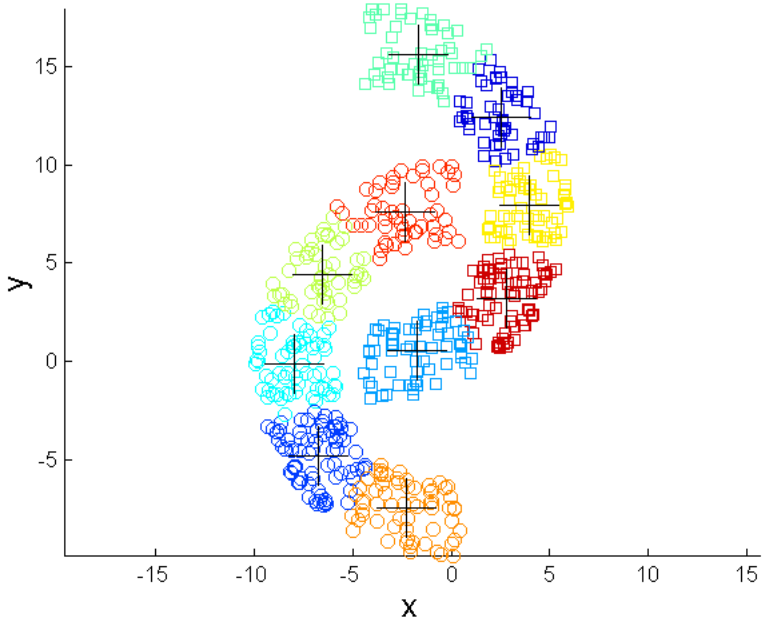


K-means 군집화

K-means 한계 극복

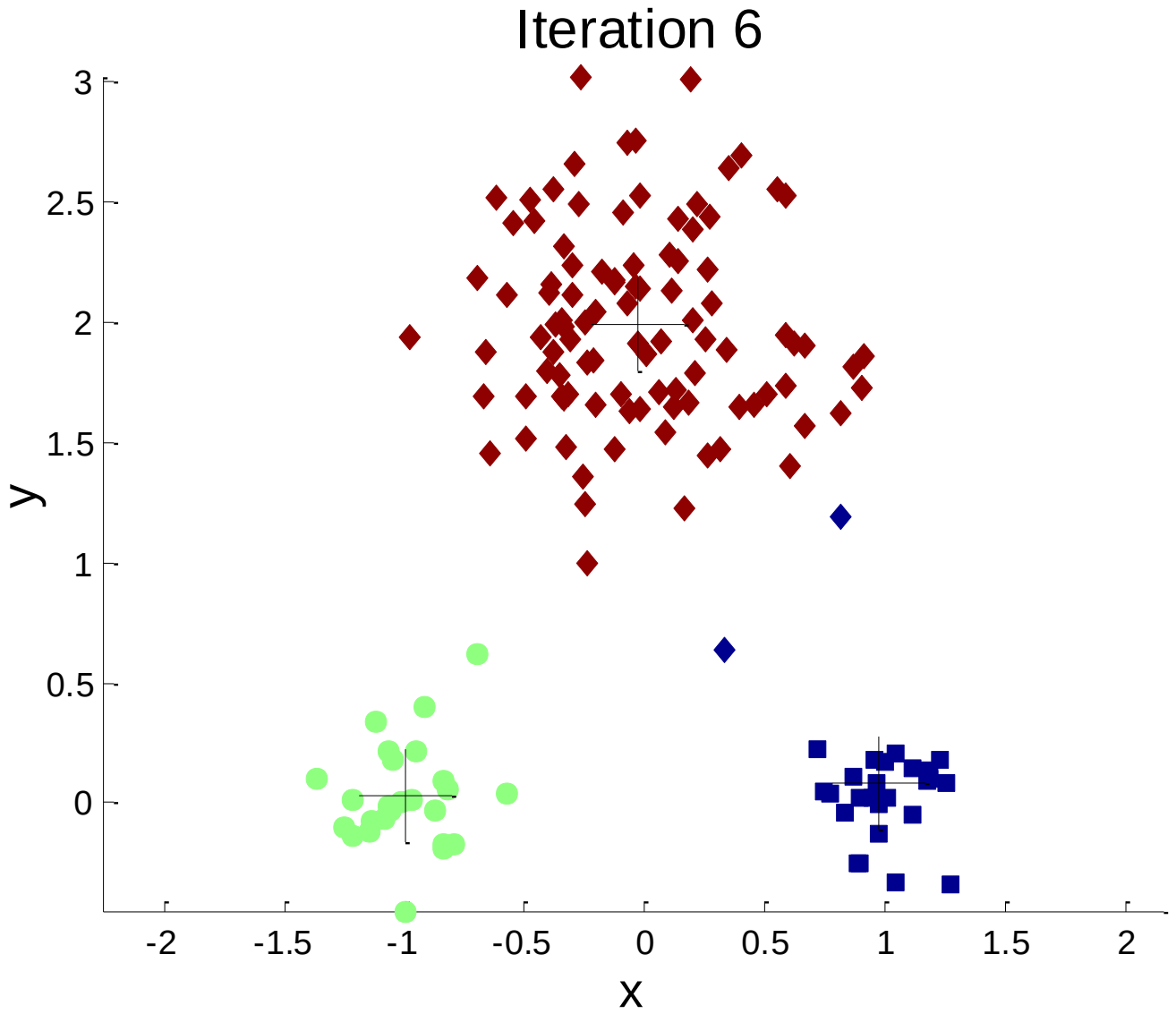


원본 포인트

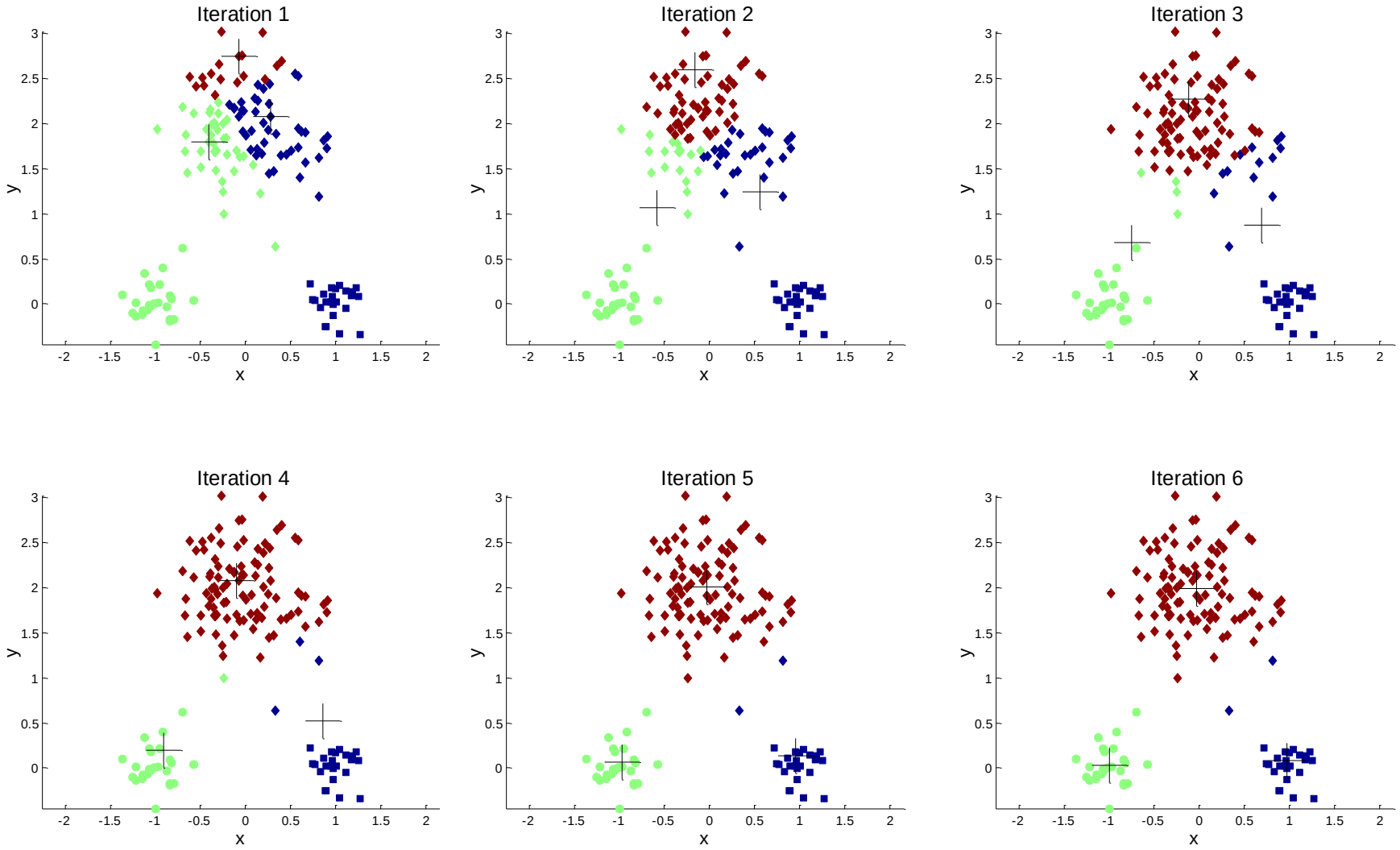


K-means 군집화

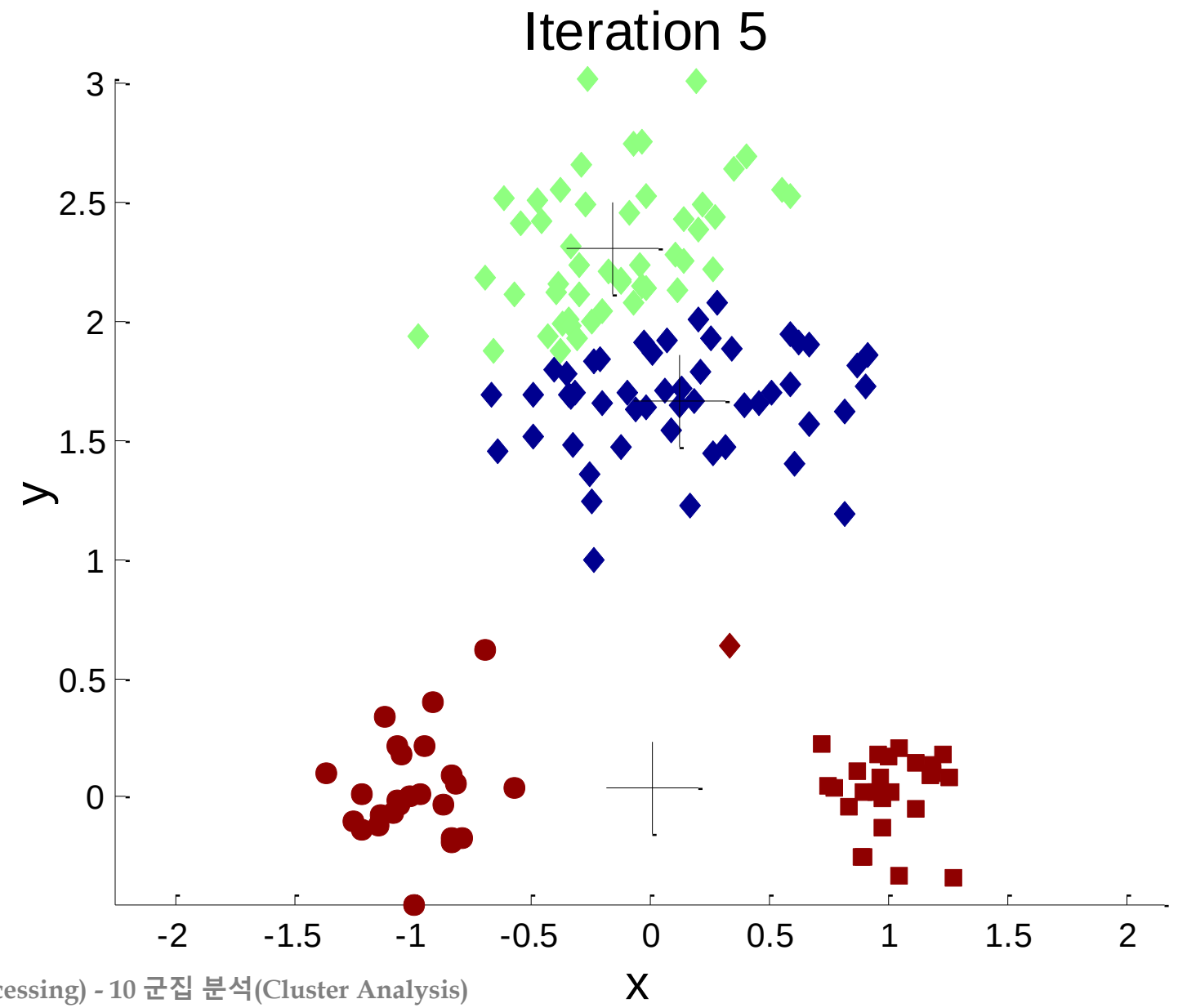
초기 중심점 선택의 중요성



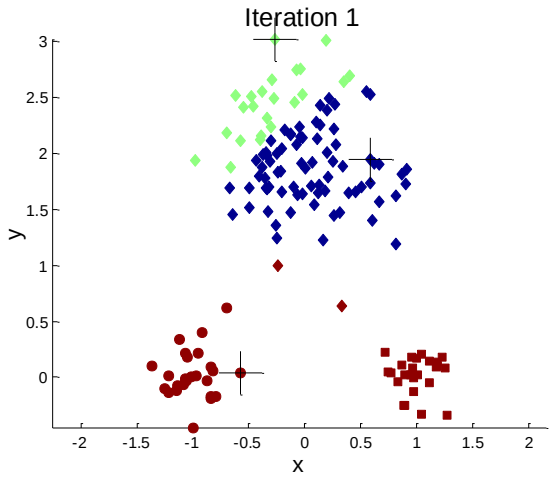
초기 중심점 선택의 중요성

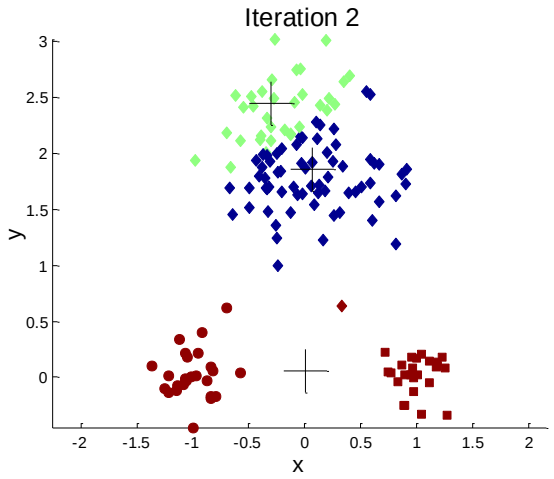


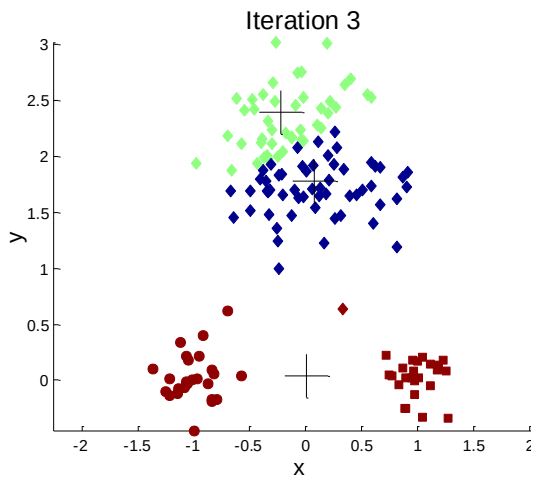
초기 중심점 선택의 중요성

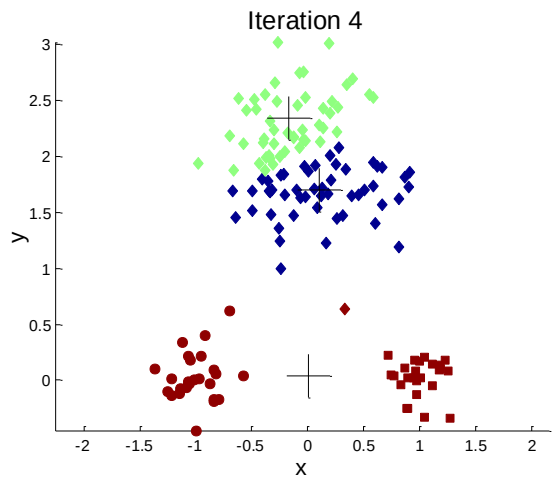


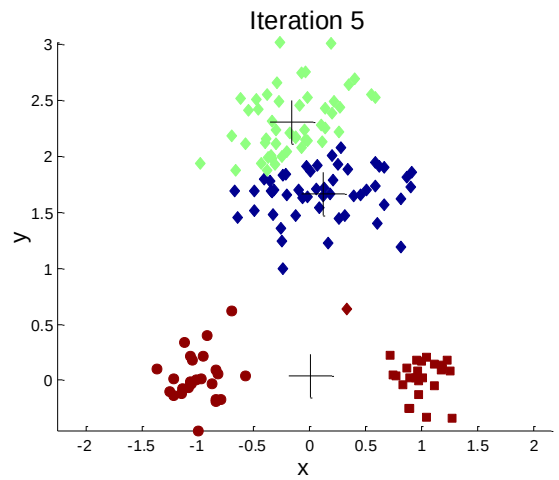
초기 중심점 선택의 중요성











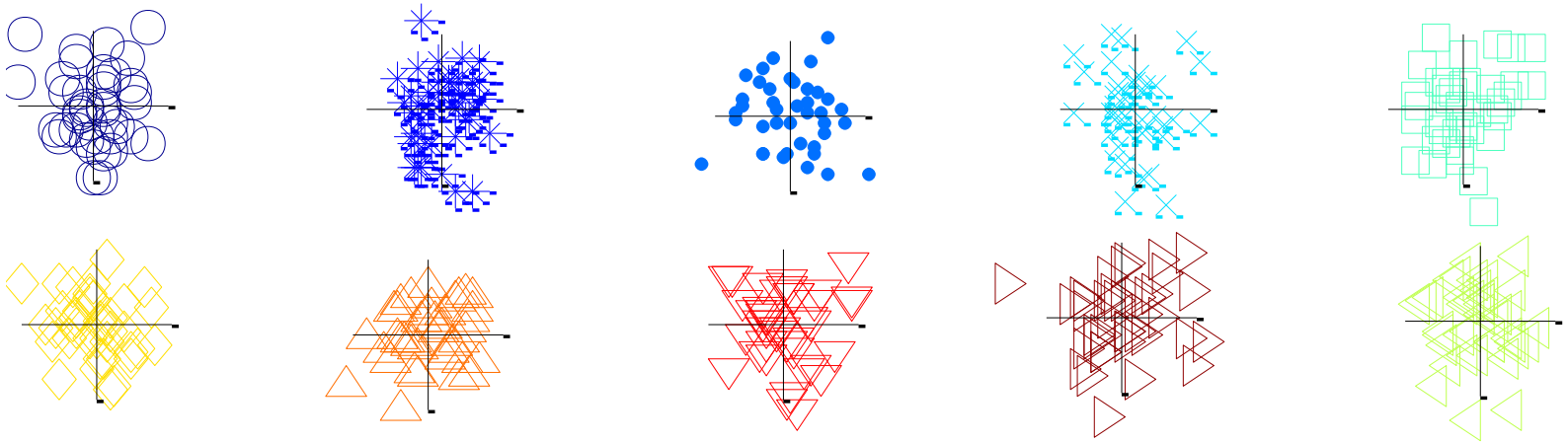
초기 포인트 선택시 문제

- K개의 '실제' 군집이 있다면 각 군집에서 하나의 중심점을 선택할 확률은 작음
  - K가 크면 확률은 상대적으로 작음
  - 군집이 같은 크기 n인 경우

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

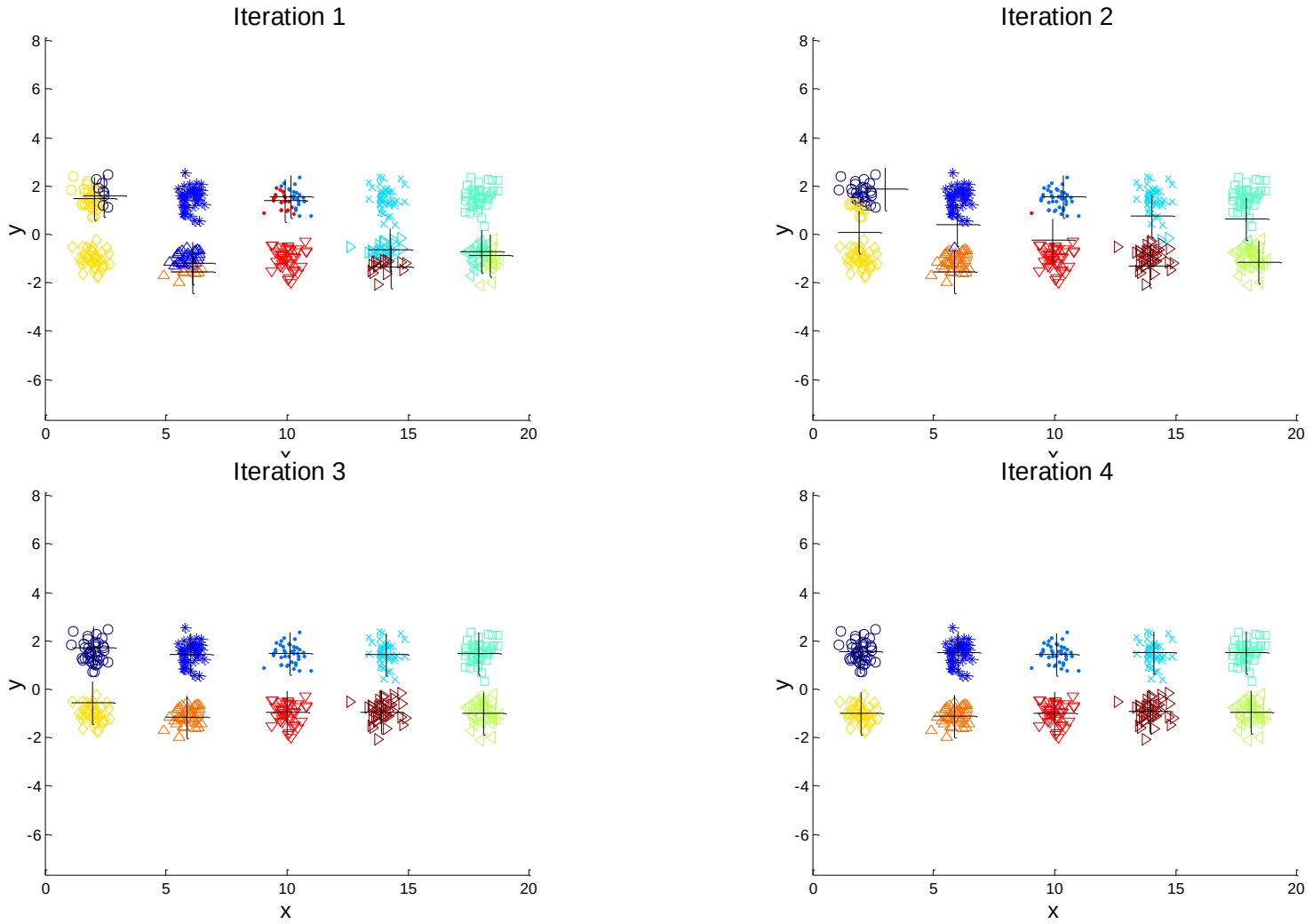
  - 예를 들어, K=10이면 확률 = $10!/10^{10} = 0.00036$
  - 초기 중심점이 때로는 옳은 방향으로 재조정 될 때가 있고, 때로는 그렇지 못할 때가 있음
  - 다섯 쌍의 군집 예제를 고려해보자

10개의 군집 예제



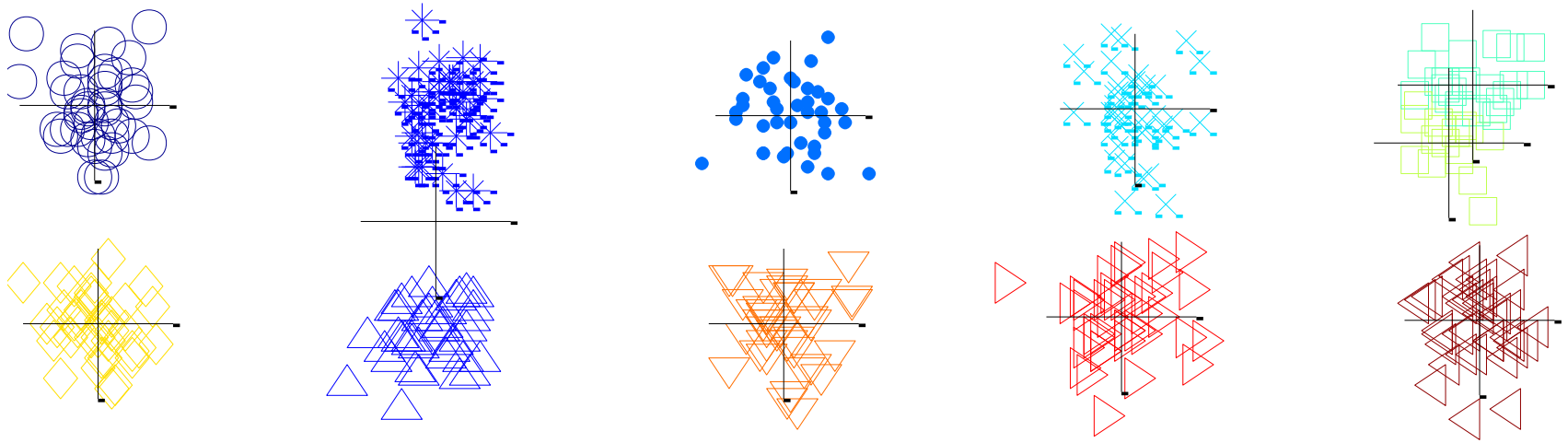
각 군집 쌍의 한 군집에서 두 개의 초기 중심점으로 시작

10개의 군집 예제



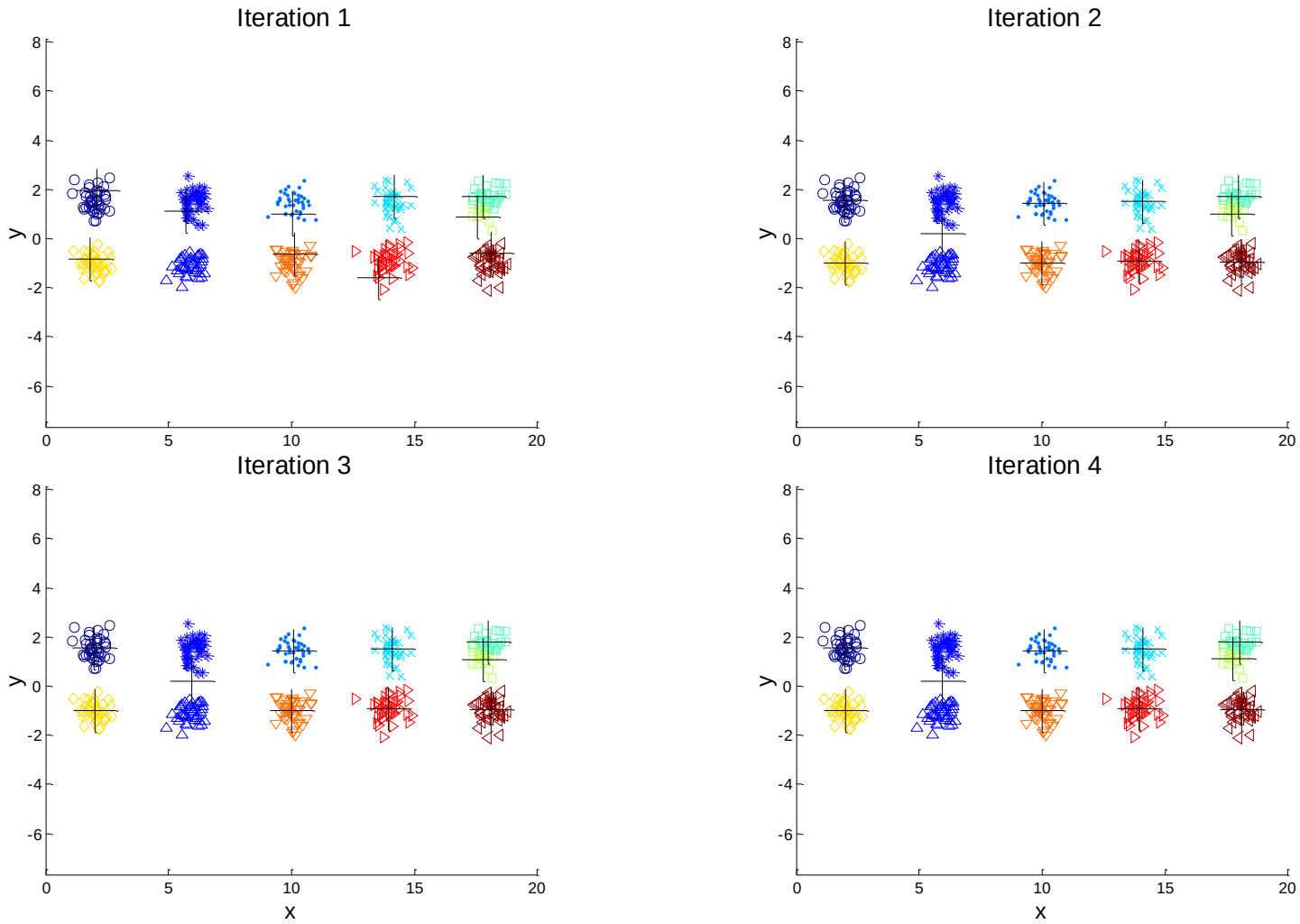
각 군집 쌍의 한 군집에서 두 개의 초기 중심점으로 시작

10개의 군집 예제



세 개의 초기 중심을 갖는 일부 군집 쌍으로 시작하는 반변 다른 군집은 하나만 포함

10개의 군집 예제



세 개의 초기 중심을 갖는 일부 군집 쌍으로 시작하는 반면 다른 군집은 하나만 포함

초기 중심점 문제 해결 방법

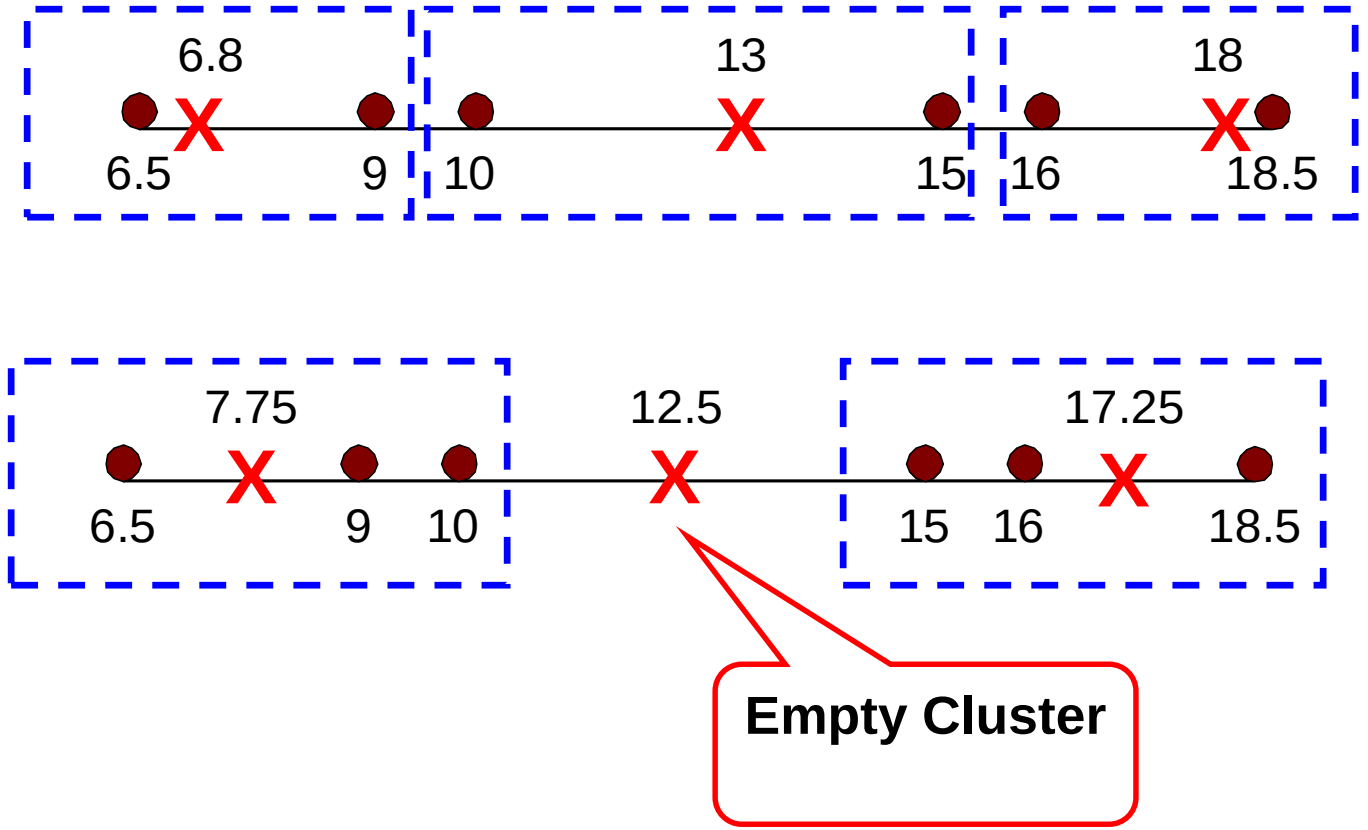
- 다중 실행Multiple runs
  - 도움이 되지만 확률은 당신 편이 아님
- 초기 중심점 결정을 위해 샘플링과 계층 군집화 사용
- 초기 중심점 K를 좀 더 선택하고, 이 초기 중심점 중 하나를 선택
  - 가장 넓게 분리된 것을 선택
- 후처리Postprocessing
- 더 많은 수의 군집을 생성한 다음 계층 군집화를 수행
- 이분법Bisecting K-means
  - 초기화 문제에 취약하지 않음

K-means++

- 이 방법은 무작위 초기화보다 느리지만, SSE 측면에서 일관되게 더 나은 결과를 생성
  - k-means++ 알고리즘은 기대한 $O(\log k)$에 근사한 비율을 보장, 여기서 k는 중심의 수
- 초기 중심 집합 C를 선택하려면 다음을 수행
- 첫 번째 중심점을 무작위로 선택
- For k – 1 steps
  - N개의 포인트 각각에 대해, x_i, $1 \le i \le N$, 현재 선택된 중심점들에 대한 최소 제곱 거리를 찾음, $C_1, \ldots, C_j$, $1 \le j < k$, 즉, $\min_j d^2(C_j, x_i)$
  - $\frac{\min_j d^2(C_j, x_i)}{\sum_i \min_j d^2(C_j, x_i)}$에 비례하는 확률을 가진 포인트를 선택하여 새로운 중심을 무작위로 선택
- End For

빈 군집 Empty Clusters

- K-means는 빈 군집을 산출할 수 있음



빈 군집 처리

- 기본적인 K-means 알고리즘은 빈 군집 생성 가능

- 몇 가지 전략
  - SSE에 가장 많이 기여한 부분을 선택
  - SSE가 가장 높은 지점을 군집에서 선택
  - 빈 군집이 여러 개 있는 경우 위의 작업을 여러 번 반복 가능

점진적으로 중심 업데이트

- 기본적인 K-means 알고리즘에서 중심점들은 모든 점들이 중심점에 할당 된 후에 업데이트

- 대안은 각 할당(점진적 접근) 후에 중심을 업데이트 하는 것
  - 각 할당은 0 또는 2개의 중심을 업데이트
  - 더 비쌈
  - 순서 의존성을 접함
  - 빈 군집을 얻지 말기
  - 가중치를 사용하여 영향을 변경할 수 있음

전처리Pre-processing와 후처리Post-processing

- 전처리Pre-processing
  - 데이터 표준화
  - 이상치 제거
- 후처리Post-processing
  - 이상치를 나타내는 작은 군집 제거
  - 느슨한loose 군집 분리, 즉 상대적으로 높은 SSE를 갖는 군집
  - 상대적으로 낮은 SSE를 가지고 가까운 군집 병합
  - 군집화 처리 중에 이러한 단계를 사용 가능
    - ISODATA(Iterative Self-Organizing Data Analysis Technique yAy!)

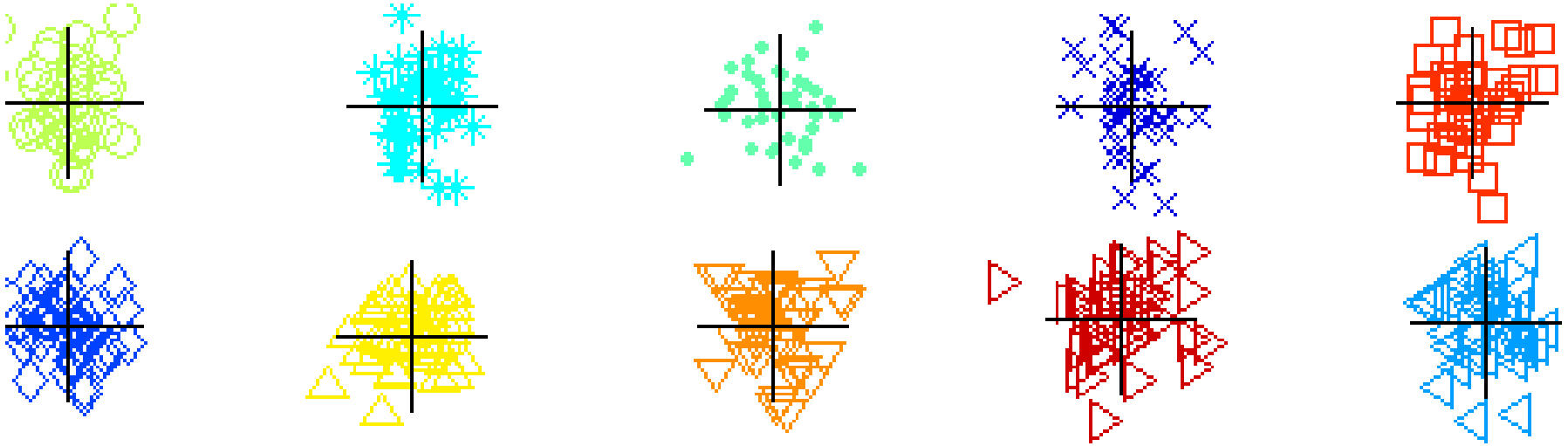
이분법 K-means

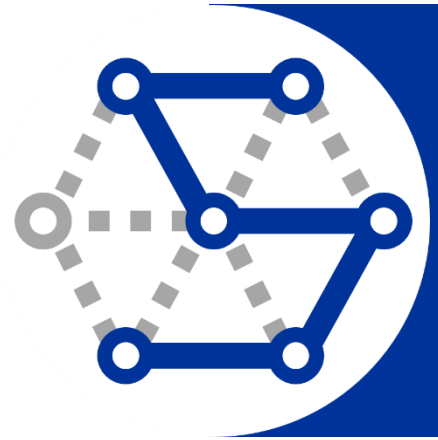
- 이분법 K-means 알고리즘
  - 분할 또는 계층 군집화를 생성할 수 있는 K-means의 변형

1: Initialize the list of clusters to contain the cluster containing all points.
2: **repeat**
3: Select a cluster from the list of clusters
4: **for** $i = 1$ to *number_of_iterations* **do**
5: Bisect the selected cluster using basic K-means
6: **end for**
7: Add the two clusters from the bisection with the lowest SSE to the list of clusters.
8: **until** Until the list of clusters contains K clusters

CLUTO: http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

이분법 K-means 예제

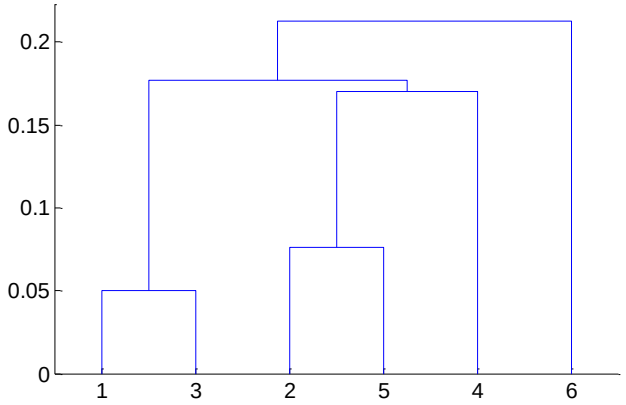


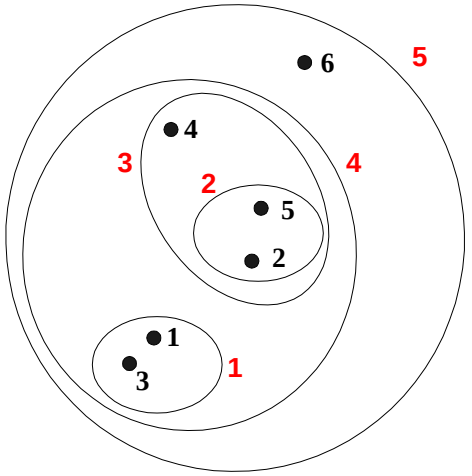


3. 계층 군집화

계층 군집화Hierarchical Clustering

- 계층 트리로 구성된 중첩된 군집 집합을 생성
- 덴드로그램dendrogram으로 시각화 할 수 있음
  - 병합 또는 분할 시퀀스를 기록하는 다이어그램과 같은 트리





계층 군집화의 장점

- 특정 군집의 수를 가정할 필요는 없음
  - 임의의 원하는 군집 수는 적절한 수준에서 덴드로그램 '절단cutting'에 의해 얻을 수 있음

- 의미있는 텍소노미taxonomies와 일치할 수 있음
  - 생물 과학 분야의 사례 (예: 동물 왕국, 계통 발생 재구성, …)

계층 군집화Hierarchical Clustering

- 계층 군집화의 두 가지 주요 유형
  - 응집성Agglomerative:
    - 포인트를 개별 군집으로 시작
    - 각 단계에서 하나의 군집(또는 k개의 군집)가 남을 때까지 가장 가까운 군집 쌍을 병합
  - 분열성Divisive:
    - 모든 것을 포함하는 하나의 군집에서 시작
    - 각 단계에서 각 군집이 개별 포인트(또는 k개의 군집)가 포함될 때까지 군집 분열
- 전통적인 계층 알고리즘은 유사도similarity 또는 거리 행렬distance matrix을 사용
  - 한 번에 하나의 군집 병합 또는 분할

응집성 군집화 알고리즘Agglomerative Clustering Algorithm

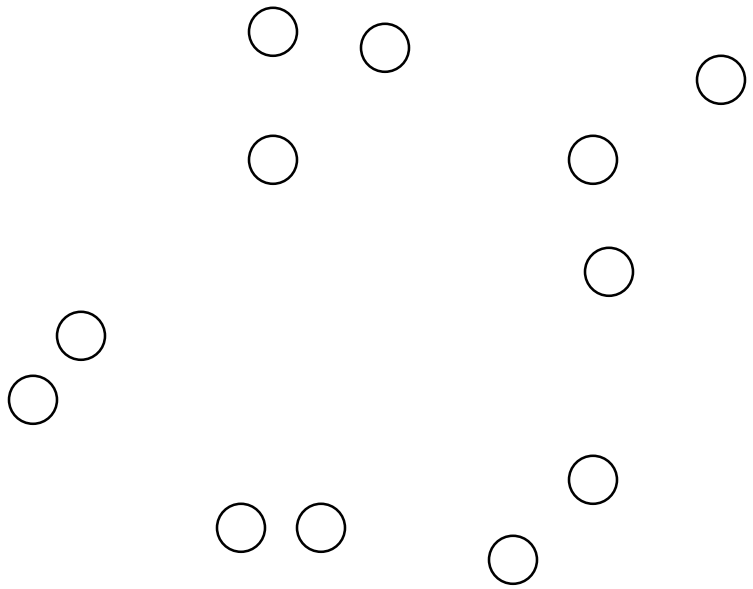
- 가장 널리 사용되는 계층 군집화 기법

- 기본 알고리즘은 복잡하지 않음straightforward
  - 근접 행렬proximity matrix 계산
  - 각 데이터 포인트가 군집이 되도록 함
  - 단일 군집이 남아 있을 때까지 반복Repeat
    - 가장 가까운 두 군집을 병합
    - 근접 행렬 업데이트

- 키 작동은 두 군집의 근접 계산
  - 군집 간의 거리를 정의하는 다양한 접근 방식은 다른 알고리즘으로 구분

시작 상황

- 개별 점과 근접 행렬의 군집으로 시작

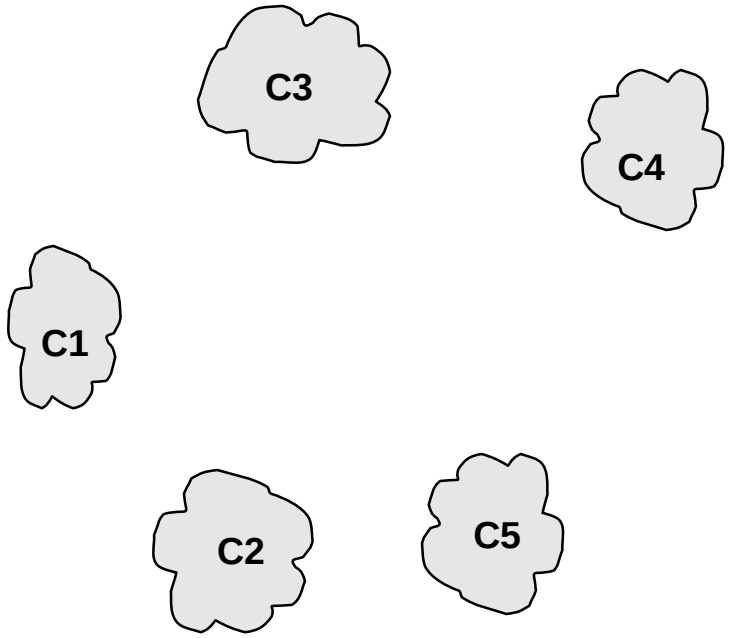


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

근접 행렬

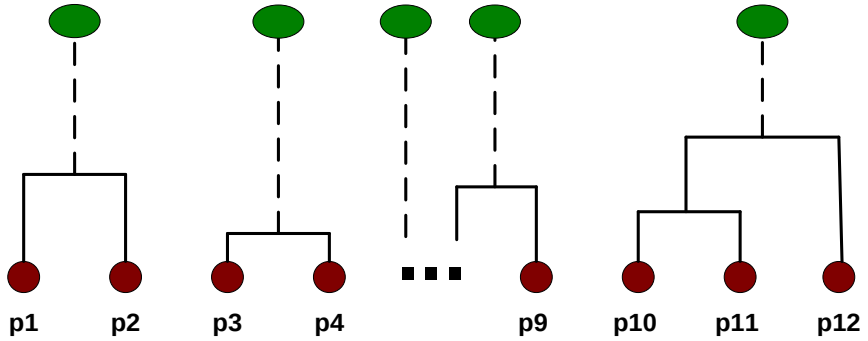
중간 상황

- 병합 단계가 끝나면 군집이 생김



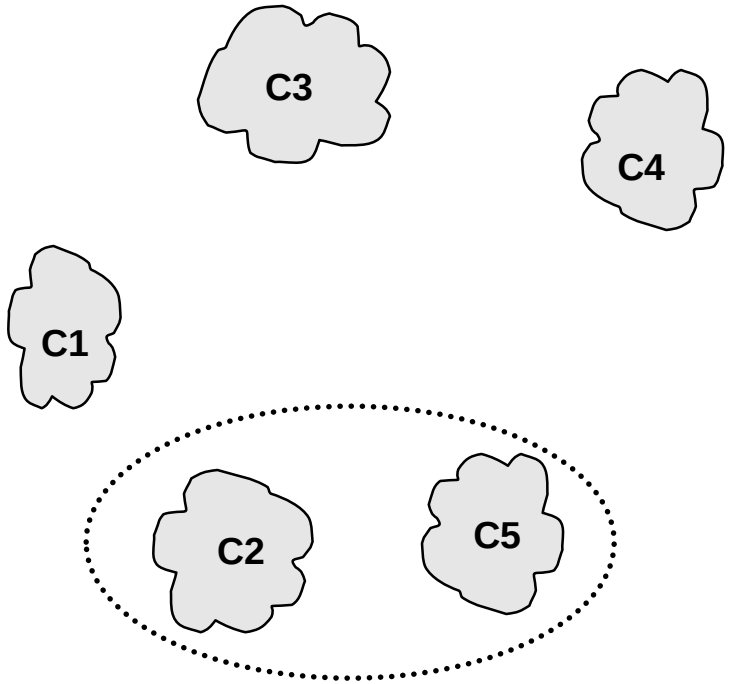
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

근접 행렬



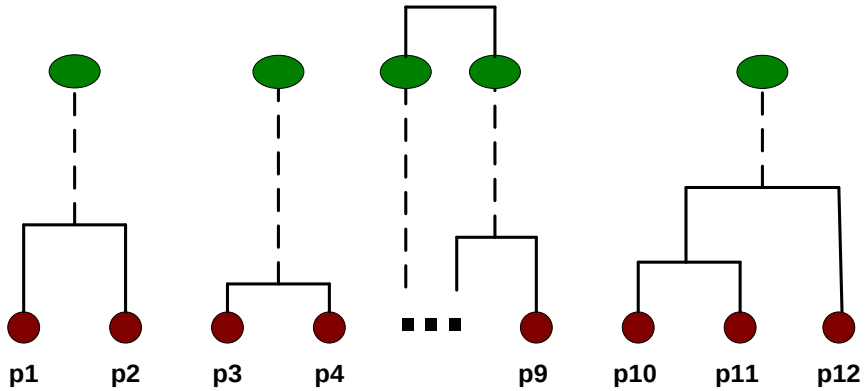
중간 상황

- 가장 가까운 두 군집 (C2와 C5)를 병합하고 근접 행렬을 업데이트



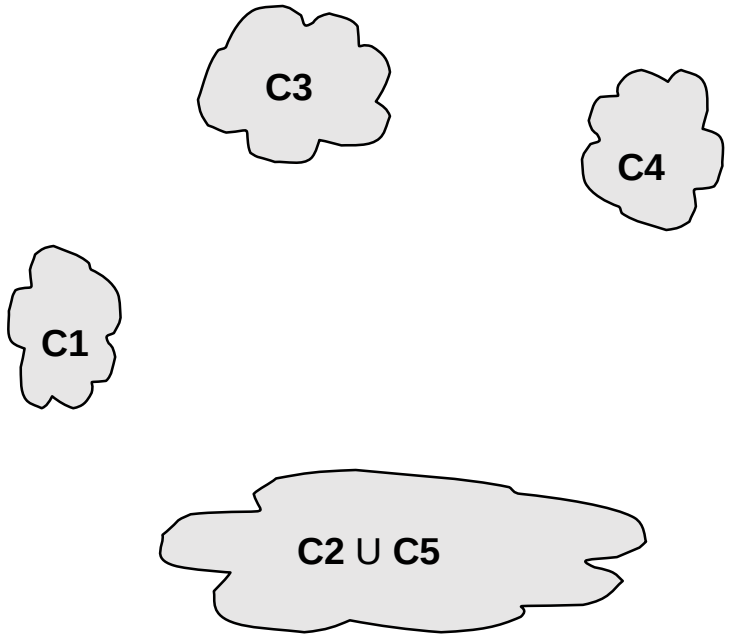
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

근접 행렬



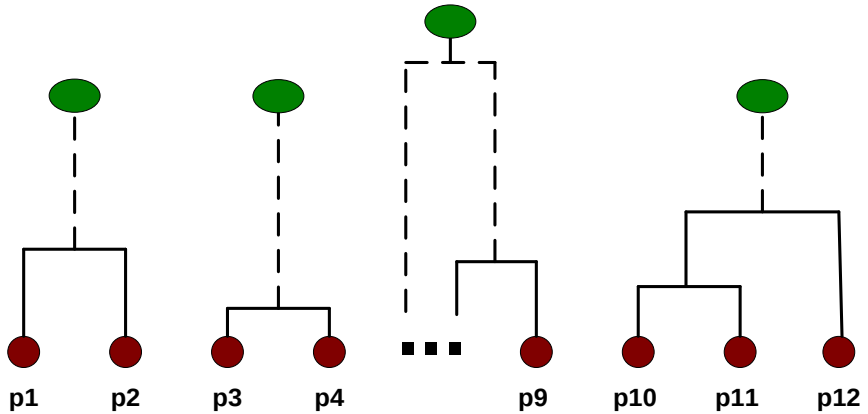
병합 후

- 문제는 "우리가 근접 행렬을 어떻게 업데이트 합니까?"

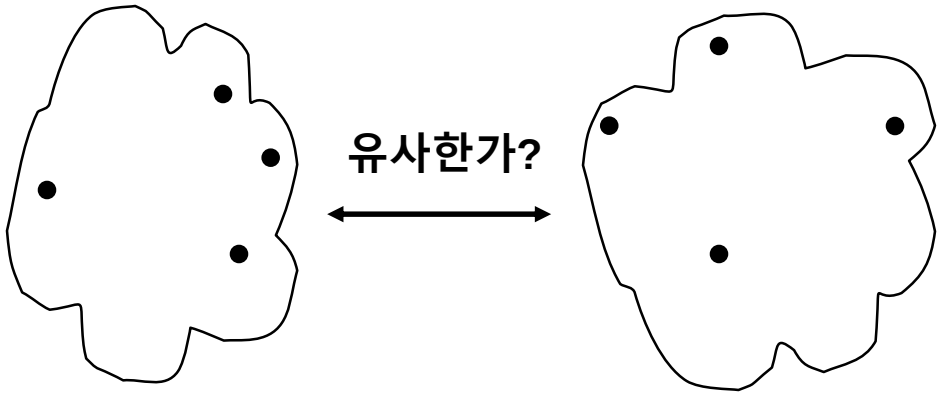


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

근접 행렬



군집 간 거리를 정의하는 방법

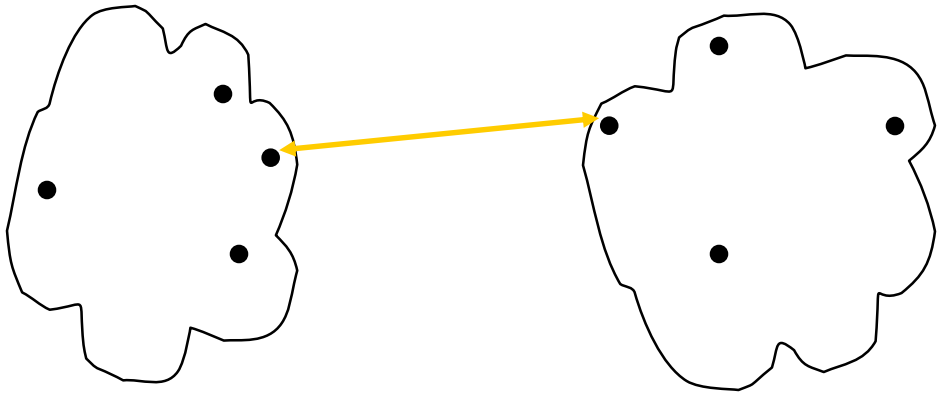


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

근접 행렬

- 최소값MIN
- 최대값MAX
- 그룹 평균Group Average
- Centroids 간의 거리Distance Between Centroids
- 목적 함수에 의해 구동되는 다른 방법들
  - Ward의 방법은 제곱 오류를 사용

군집 간 유사성을 정의하는 방법

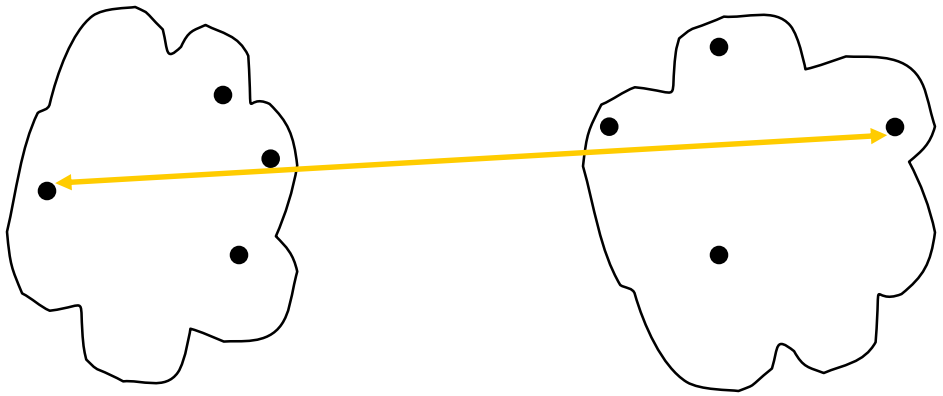


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

근접 행렬

- 최소값MIN
- 최대값MAX
- 그룹 평균Group Average
- Centroids 간의 거리Distance Between Centroids
- 목적 함수에 의해 구동되는 다른 방법들
  - Ward의 방법은 제곱 오류를 사용

군집 간 거리를 정의하는 방법

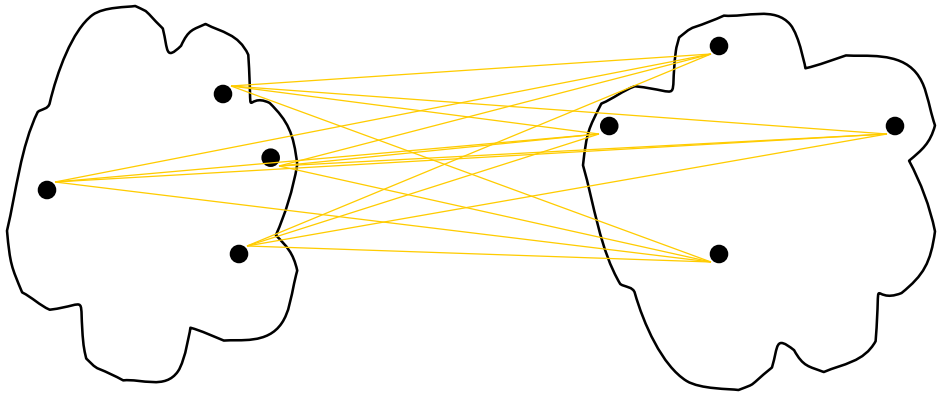


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

근접 행렬

- 최소값MIN
- 최대값MAX
- 그룹 평균Group Average
- Centroids 간의 거리Distance Between Centroids
- 목적 함수에 의해 구동되는 다른 방법들
  - Ward의 방법은 제곱 오류를 사용

군집 간 거리를 정의하는 방법

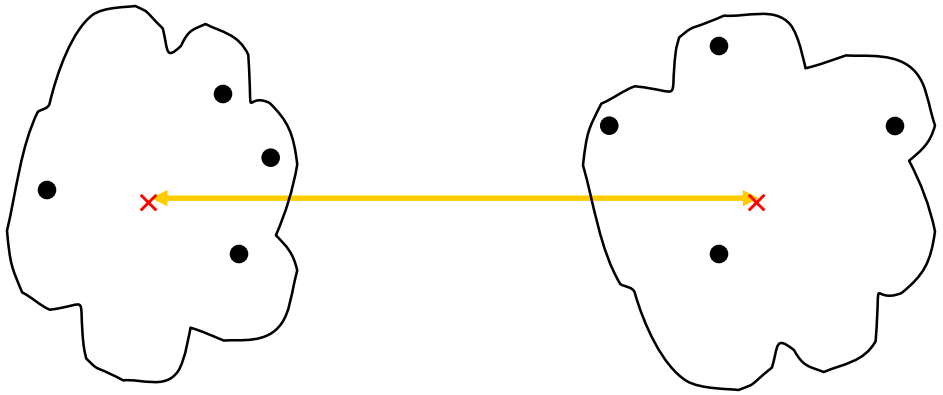


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

근접 행렬

- 최소값MIN
- 최대값MAX
- 그룹 평균Group Average
- Centroids 간의 거리Distance Between Centroids
- 목적 함수에 의해 구동되는 다른 방법들
  - Ward의 방법은 제곱 오류를 사용

군집 간 거리를 정의하는 방법



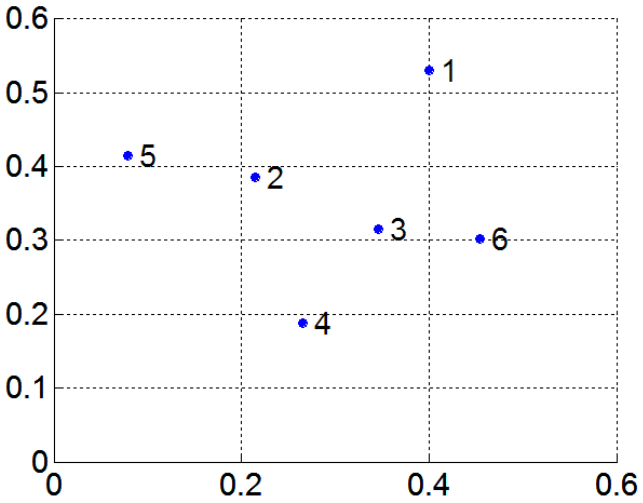
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

근접 행렬

- 최소값MIN
- 최대값MAX
- 그룹 평균Group Average
- Centroids 간의 거리Distance Between Centroids
- 목적 함수에 의해 구동되는 다른 방법들
  - Ward의 방법은 제곱 오류를 사용

최소값MIN 또는 단일 링크Single Link

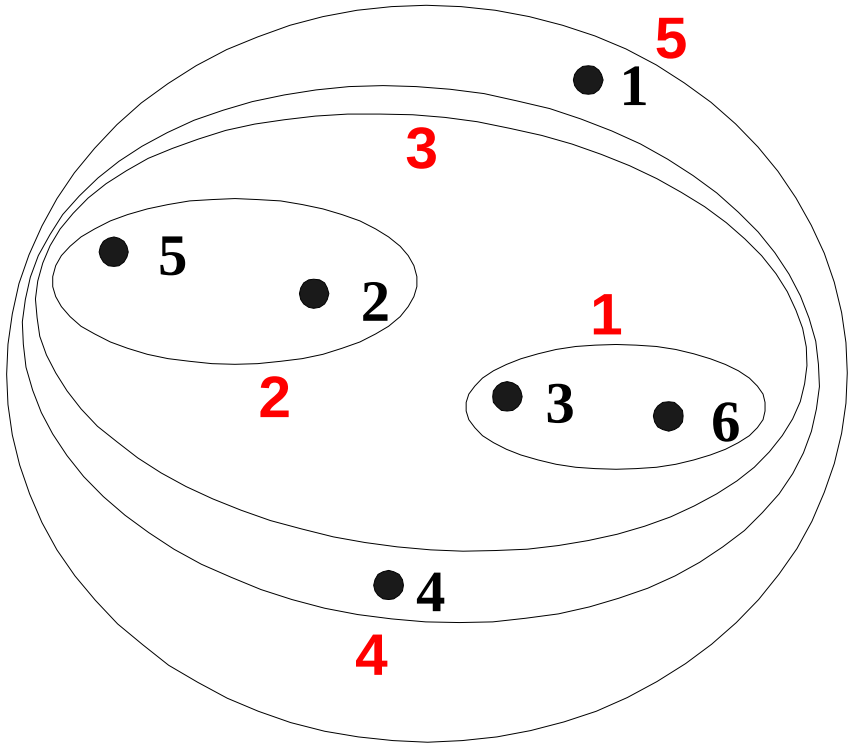
- 두 군집의 근접성은 다른 군집의 두 개의 가장 가까운 포인트를 기반으로 함
  - 한 쌍의 점, 즉 근접 그래프의 한 링크에 의해 결정
- 예제:



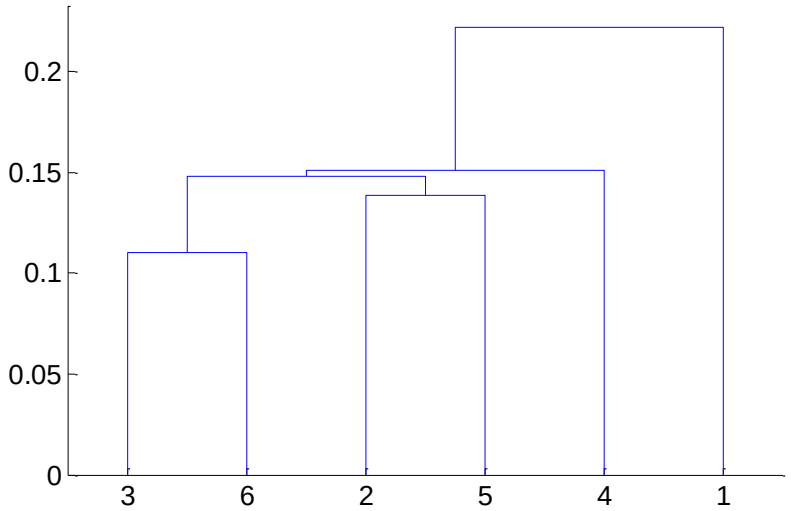
거리 행렬

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

계층 군집화Hierarchical Clustering: 최소값MIN

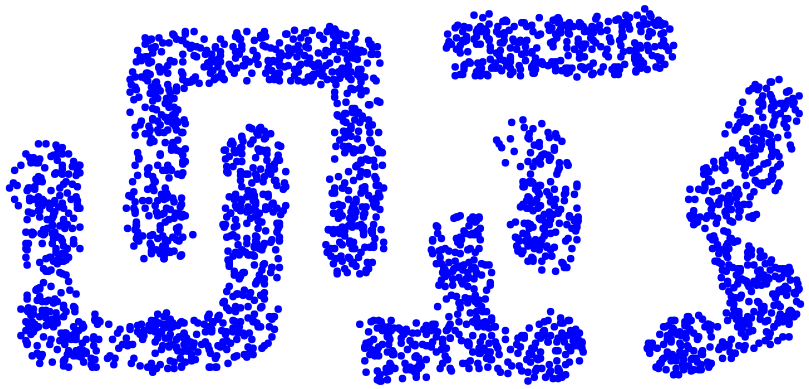


중첩된 군집

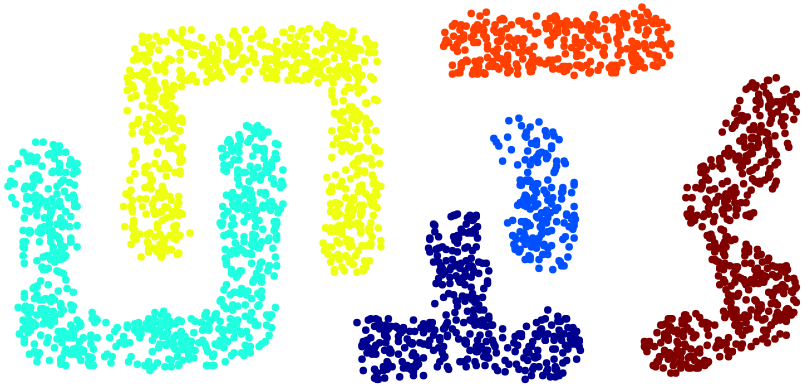


덴드로그램

최소값MIN의 장점



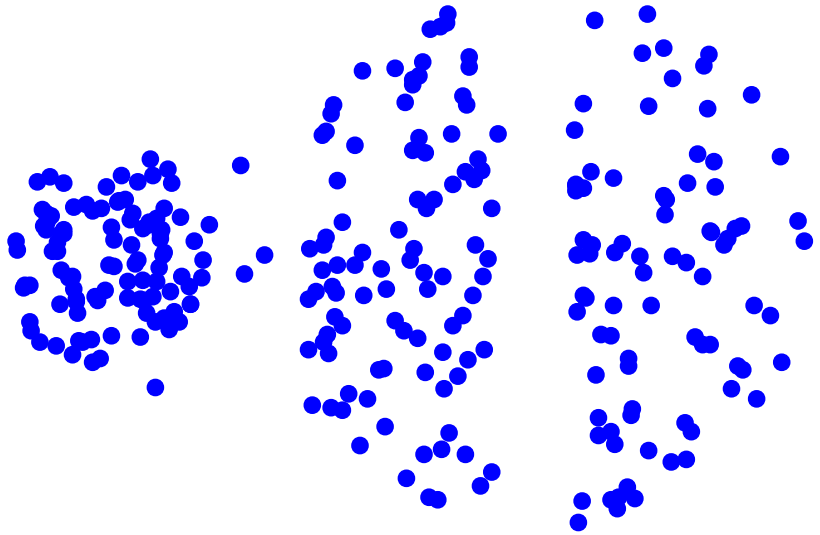
Original Points



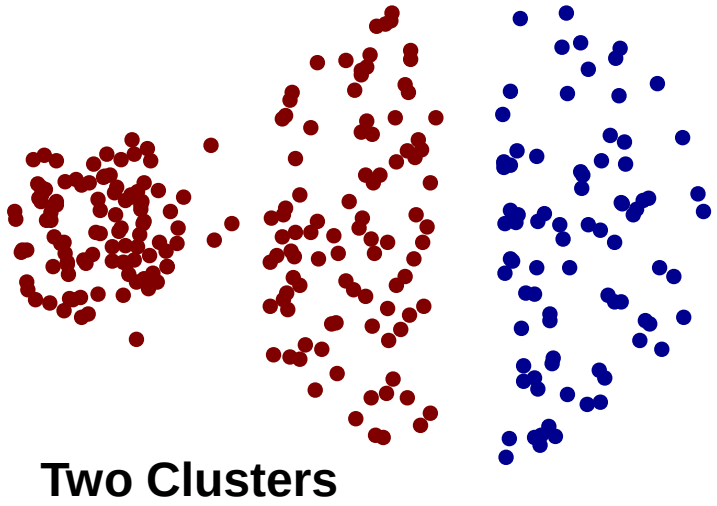
Six Clusters

- 타원형이 아닌 도형 처리 가능

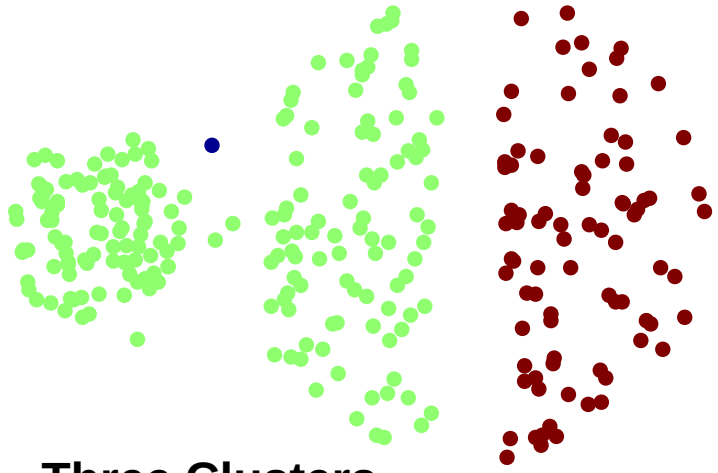
최소값MIN의 한계



Original Points



Two Clusters

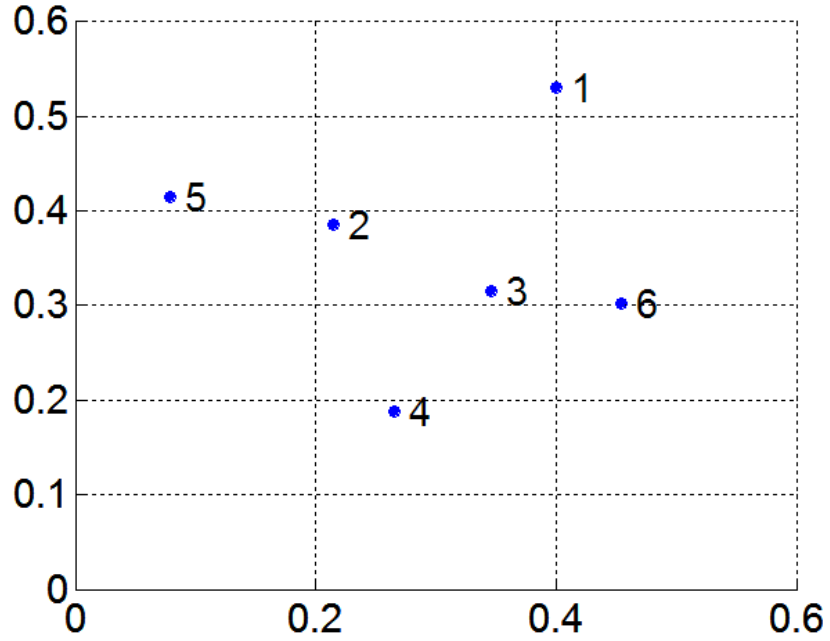


Three Clusters

- 노이즈 및 이상치에 민감

최대값MAX 또는 완전한 연계Complete Linkage

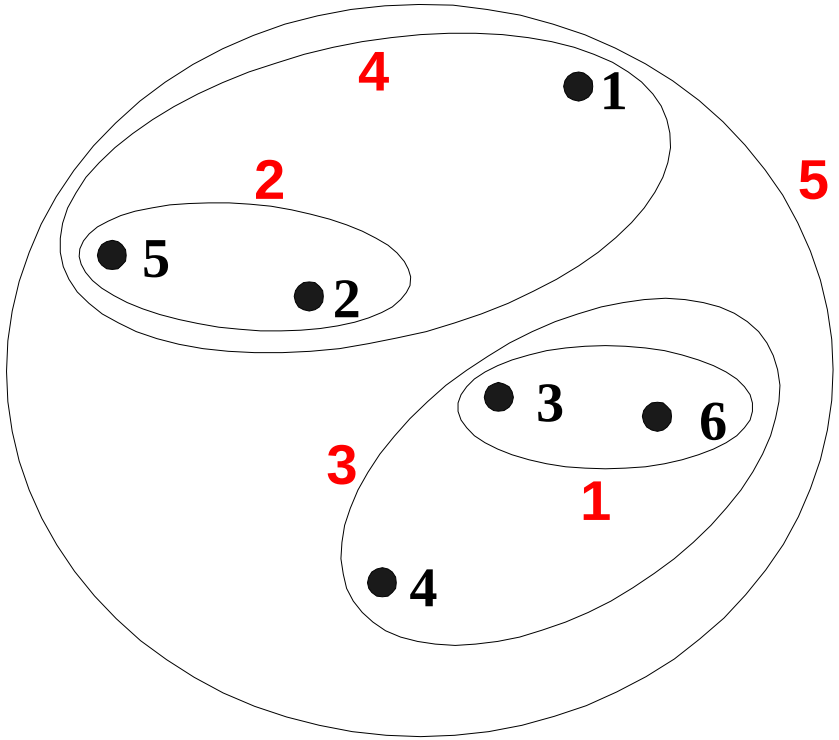
- 두 군집의 근접성은 다른 군집의 두 개의 가장 먼 포인트를 기반으로 함
  - 두 군집의 모든 포인트 쌍에 의해 결정



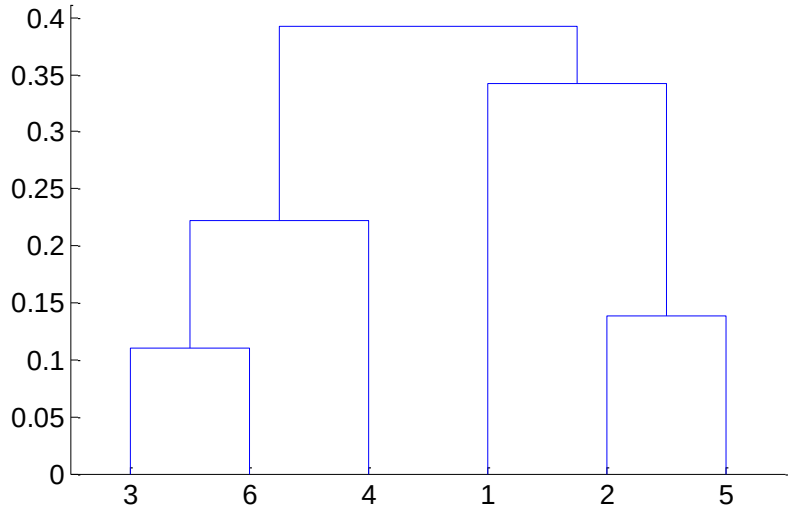
거리 행렬

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

계층 군집화Hierarchical Clustering: 최대값MAX

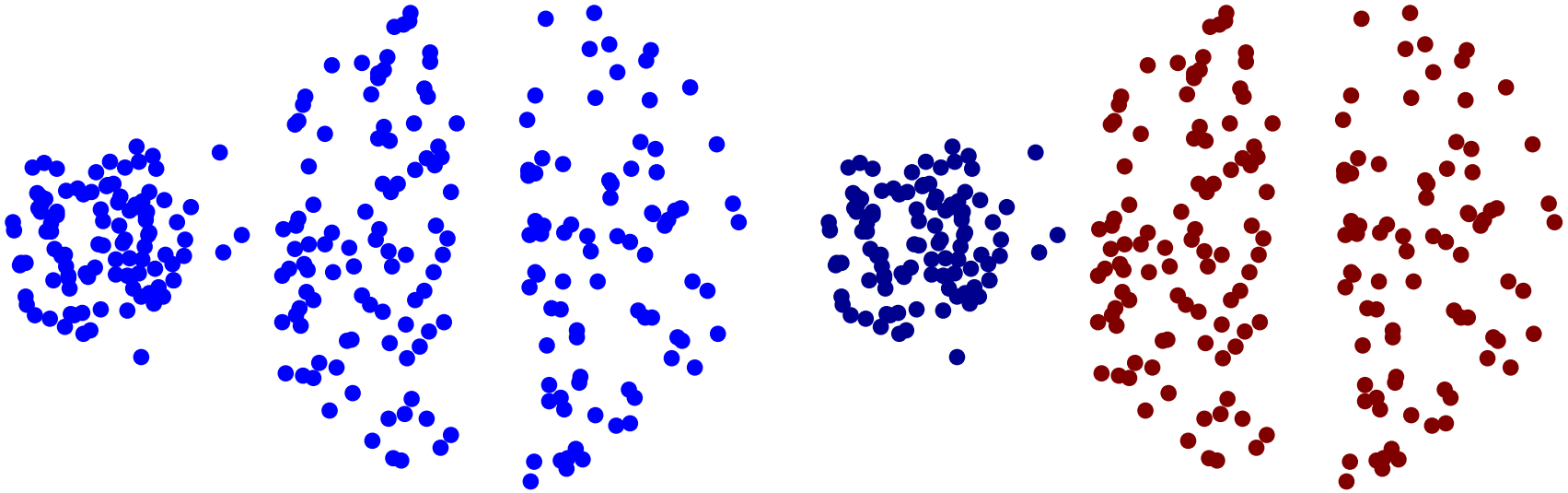


중첩된 군집



덴드로그램

최대값MAX의 장점

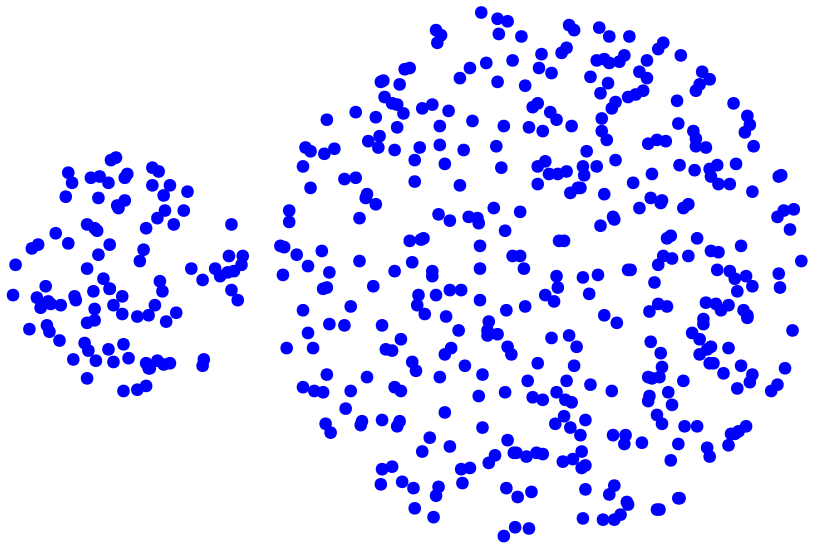


Original Points

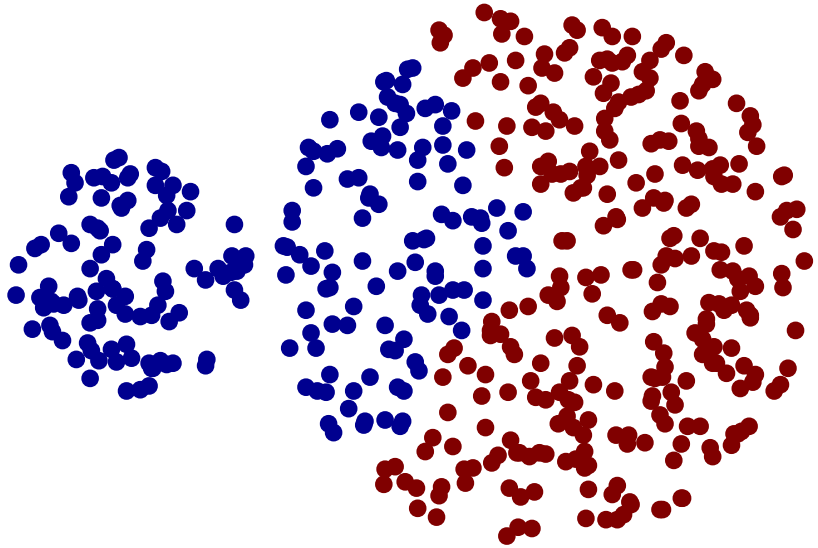
Two Clusters

- 노이즈 및 이상치에 영향을 덜 받음

최대값MAX의 한계



Original Points



Two Clusters

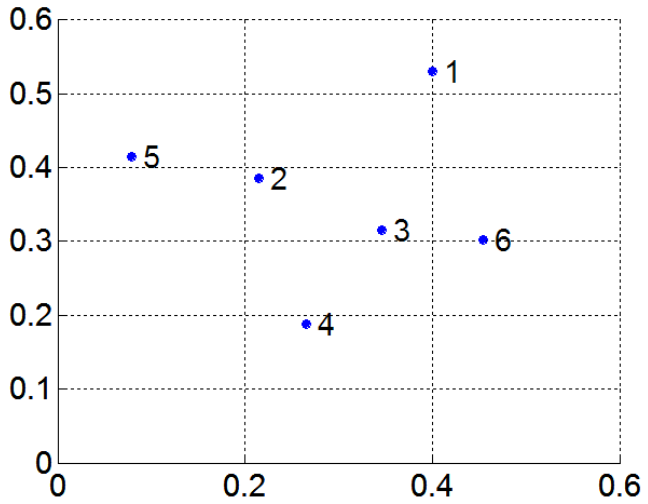
- 큰 군집을 깨부술 수 있음
- 공 모양의globular 군집으로 편향

그룹 평균Group Average

- 두 군집의 근접성은 두 군집의 포인트 사이의 쌍방향 근접성의 평균

$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}$$

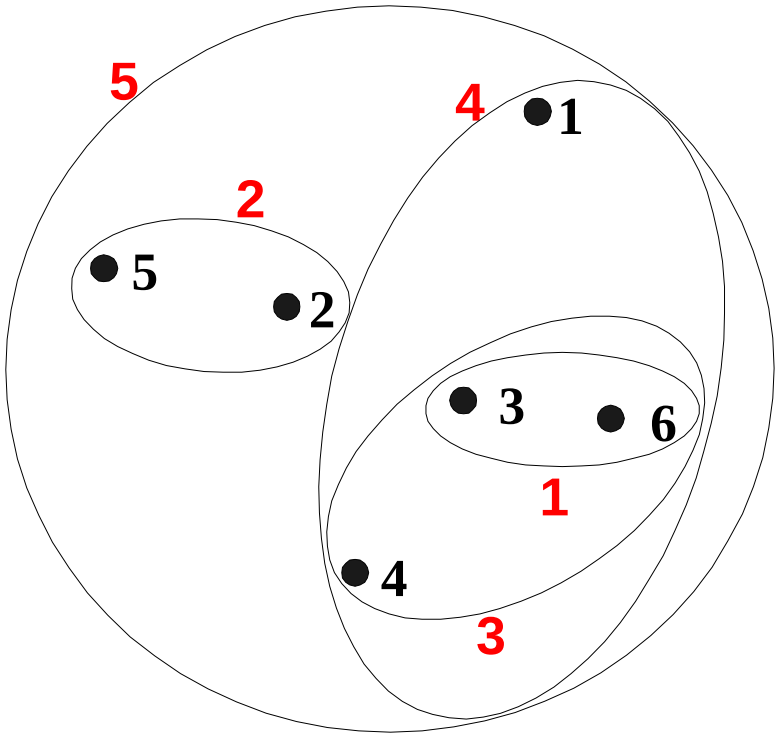
- 총 근접성이 큰 군집을 선호하기 때문에 확장성에 대한 평균 연결성을 사용해야 함



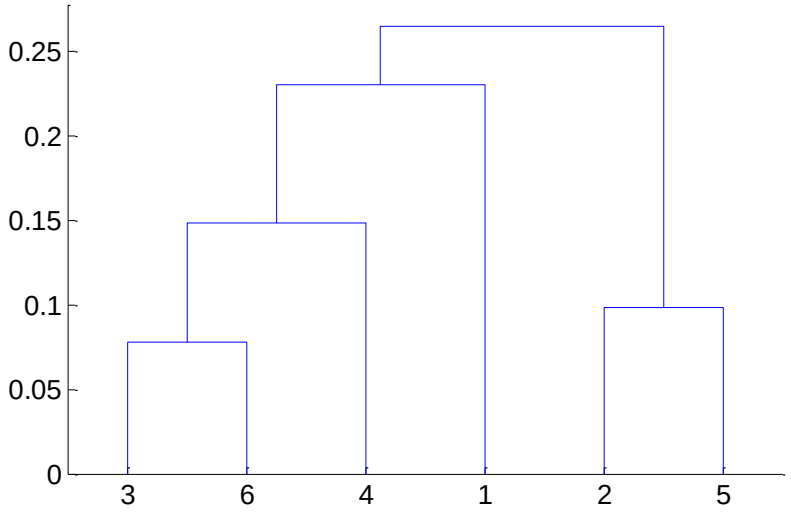
Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

계층 군집화Hierarchical Clustering: 그룹 평균Group Average



중첩된 군집



덴드로그램

계층 군집화Hierarchical Clustering: 그룹 평균Group Average

- 단일 링크Single Link와 전체 링크Complete Link의 타협

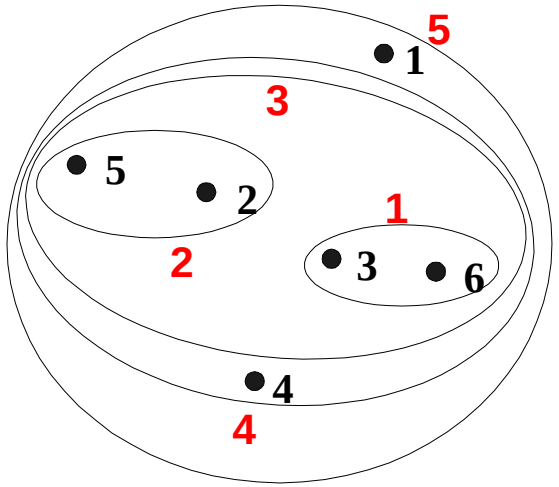
- 강점Strengths
  - 노이즈 및 이상치의 영향을 덜 받음

- 한계Limitations
  - 공 모양의 군집으로 편향

군집 유사성Cluster Similarity: Ward 방법Method

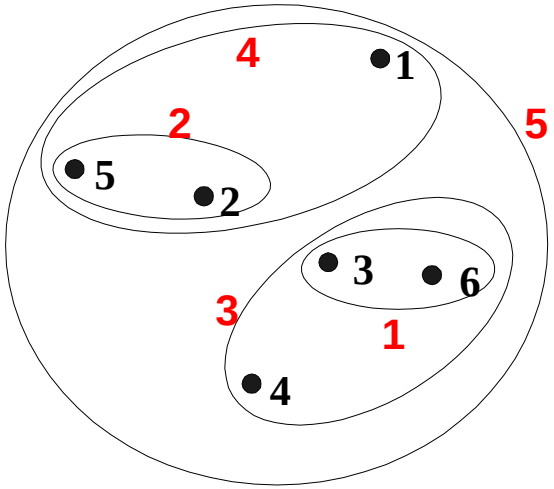
- 두 군집의 유사성은 두 군집이 병합될 때, 제곱 오류의 증가를 기반으로 함
  - 포인트 간의 거리가 거리 제곱인 경우 그룹 평균과 유사
- 노이즈 및 이상치의 영향을 덜 받음
- 공 모양의 군집으로 편향
- K-means의 계층 아날로그
  - K-means 초기화에 사용 가능

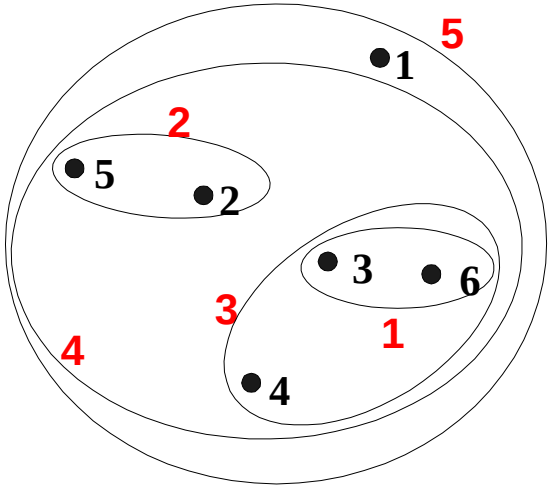
계층 군집화Hierarchical Clustering: 비교Comparison



최소값MIN

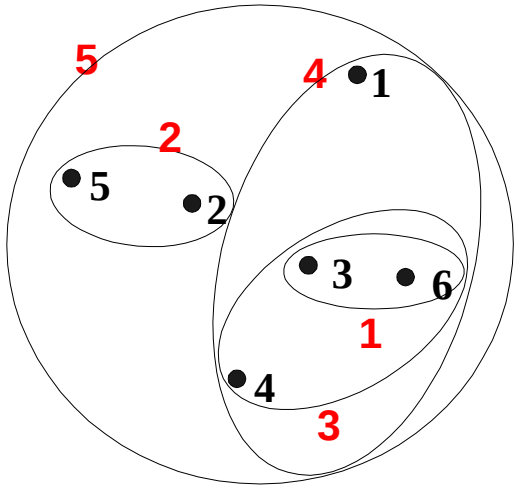
최대값MAX





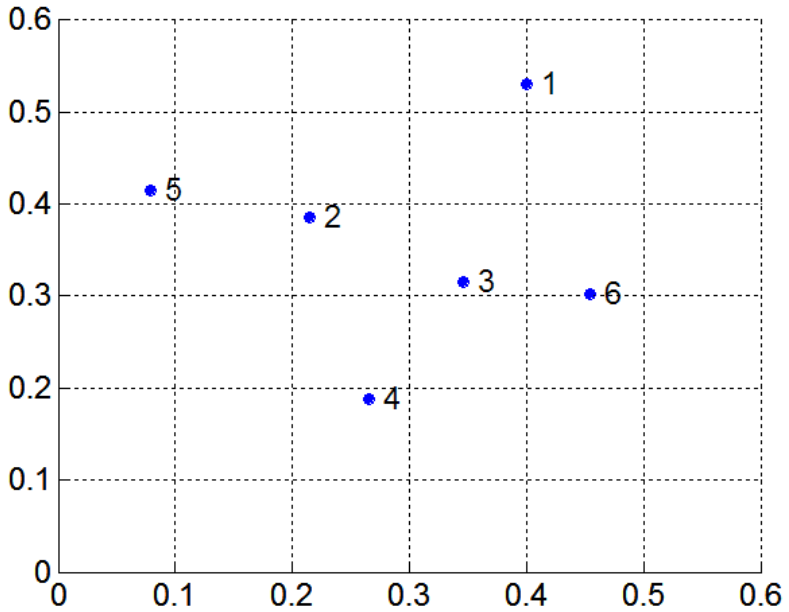
그룹 평균Group Average

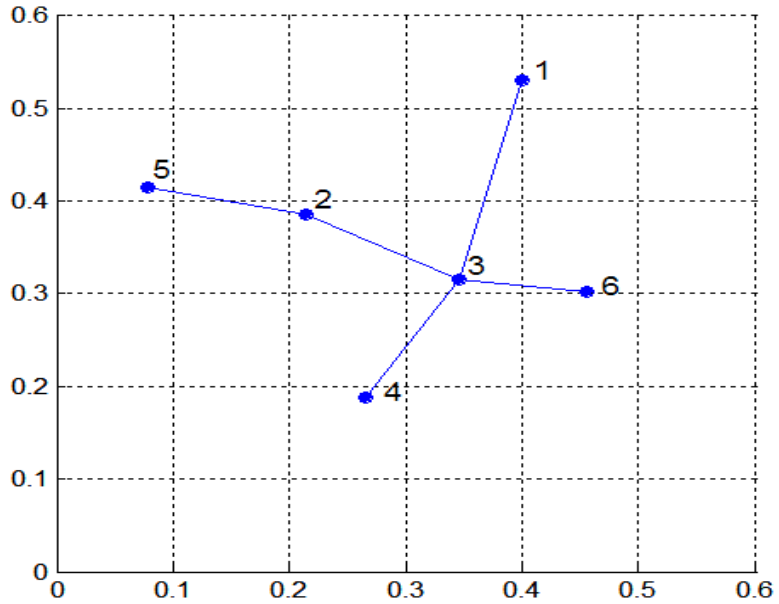
Ward 방법Method



MST: 구분하는 계층 군집화 Divisive Hierarchical Clustering

- MST(Minimum Spanning Tree) 빌드
  - 어떤 포인트로 구성된 트리로 시작
  - 연속적인 단계에서 한 포인트(p)가 현재 트리에 있지만 다른 점(q)은 현재 트리에 없는 점 (p, q) 중 가장 가까운 포인트를 찾음
  - q를 트리에 추가하고 p와 q 사이에 모서리를 놓음





MST: 구분하는 계층 군집화Divisive Hierarchical Clustering

- MST를 사용하여 군집 계층 구성

Algorithm 7.5 MST Divisive Hierarchical Clustering Algorithm

1: Compute a minimum spanning tree for the proximity graph.
2: **repeat**
3: Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).
4: **until** Only singleton clusters remain

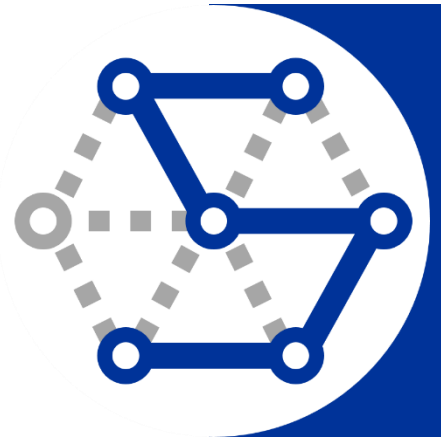
계층 군집화Hierarchical Clustering: 시간 및 공간 요구사항

- $O(N^2)$ 공간을 근접 행렬에 사용
  - N은 포인트 수

- 많은 경우 $O(N^3)$ 공간 사용
  - N개의 단계가 있으며 각 단계에서 크기, N^2, 근접 행렬을 업데이트하고 검색해야 함
  - 복잡성은 약간은 교묘하게 $O(N^2 \log(N))$ 시간으로 줄일 수 있음

계층 군집화Hierarchical Clustering: 문제점 및 한계

- 두 개의 군집을 결합하는 결정이 내려지면 취소할 수 없음
- 글로벌 목적 함수가 직접 최소화되지 않음
- 다른 계획에는 다음 중 하나 이상의 문제가 있음:
  - 노이즈 및 이상치에 대한 민감도
  - 다른 크기와 공 모양의 군집을 다루기 어려움
  - 큰 군집이 깨짐

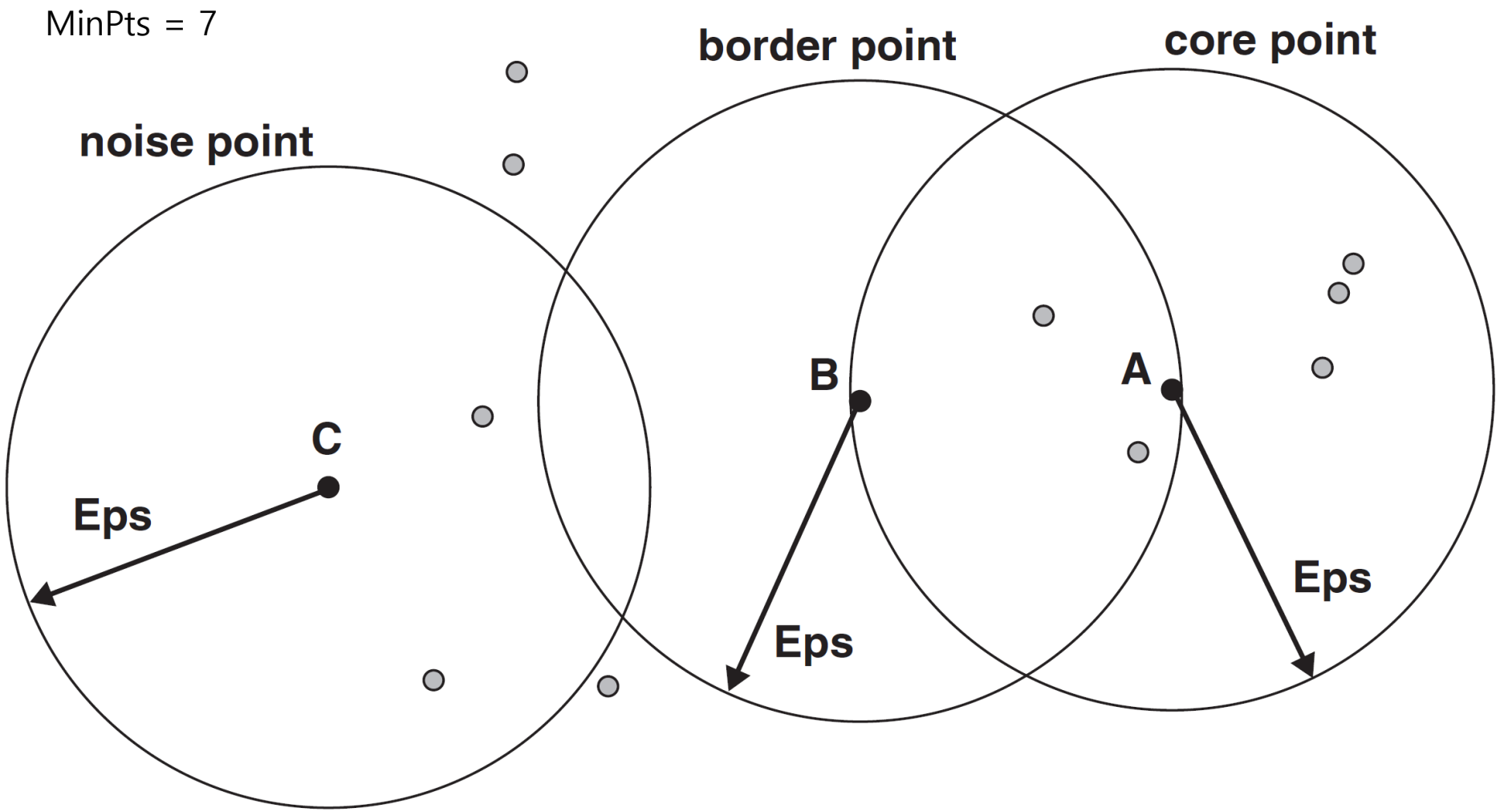


4. DBSCAN

DBSCAN

- DBSCAN은 밀도 기반 알고리즘density-based algorithm
  - 밀도Density = 지정된 반경 내 포인트 수(Eps)
  - 포인트는 지정된 최소 포인트(MinPts)가 있는 경우, 핵심 포인트core point
    - 군집의 내부에 있는 포인트
    - 포인트 자체를 계산
  - 경계 포인트border point은 핵심 포인트가 아니지만 핵심 포인트 근처에 있음
  - 노이즈 포인트noise point 는 핵심 포인트 또는 경계 포인트가 아닌 모든 포인트

DBSCAN: 핵심, 경계, 및 노이즈 포인트



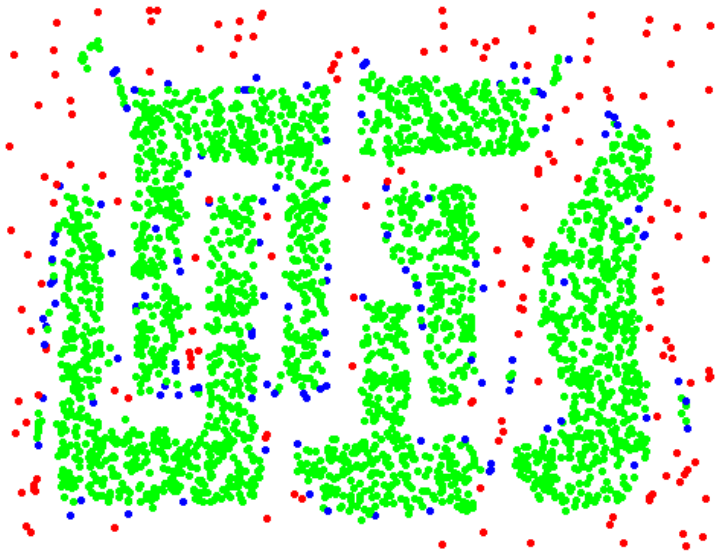
DBSCAN 알고리즘

- 노이즈 포인트 제거
- 나머지 포인트에서 군집화 수행

$current\_cluster\_label \leftarrow 1$
for all core points **do**
  if the core point has no cluster label **then**
    $current\_cluster\_label \leftarrow current\_cluster\_label + 1$
    Label the current core point with cluster label $current\_cluster\_label$
  end if
  for all points in the Eps-neighborhood, except i^{th} the point itself **do**
    if the point does not have a cluster label **then**
      Label the point with cluster label $current\_cluster\_label$
    end if
  end for
end for

DBSCAN: 핵심, 경계, 및 노이즈 포인트

원본 포인트

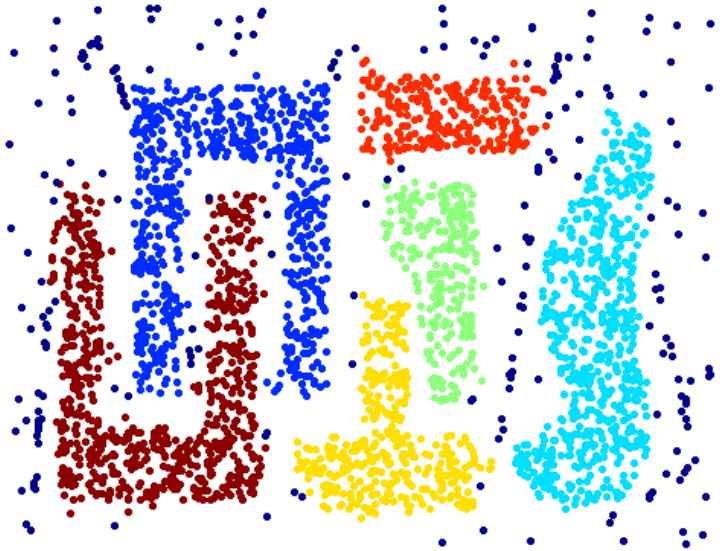


포인트 유형: 핵심, 경계 및 노이즈

Eps = 10, MinPts = 4

DBSCAN이 잘 작동할 때

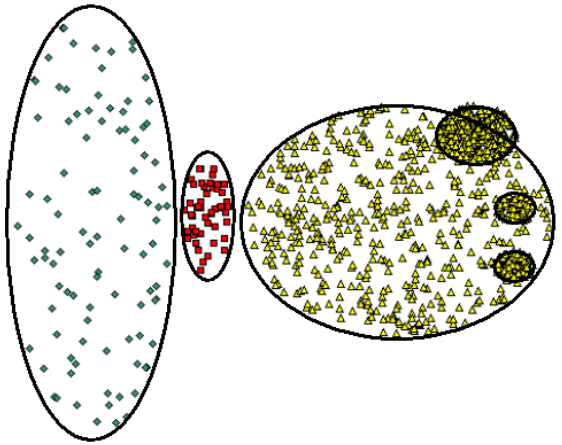
원본 포인트



군집

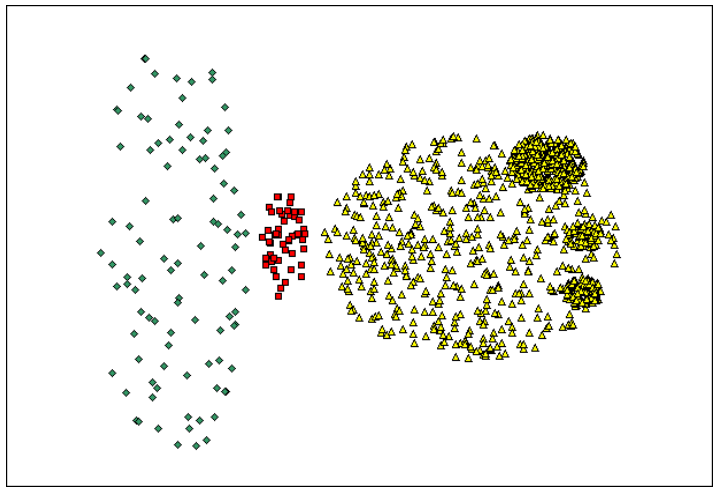
- 노이즈에 강함
- 다양한 모양과 크기의 군집 처리 가능

DBSCAN이 잘 동작하지 않을 때

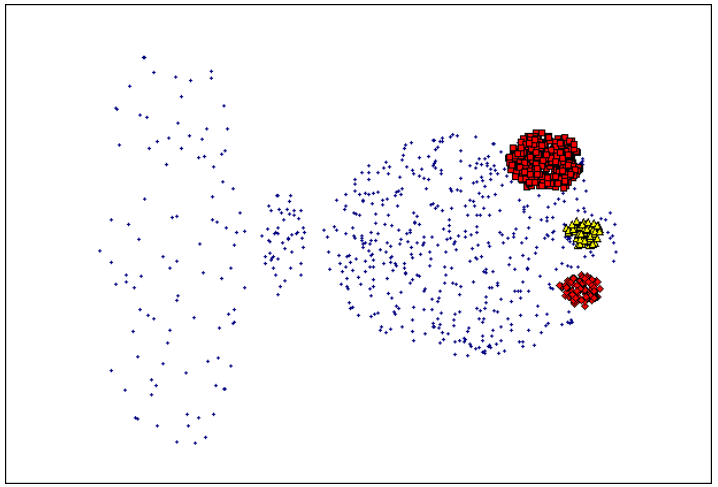


원본 포인트

- 다양한 밀도
- 고차원 데이터



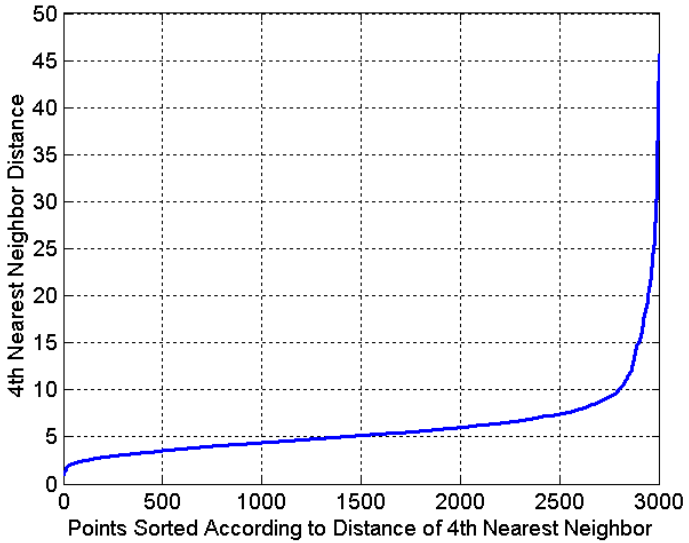
(MinPts=4, Eps=9.75).

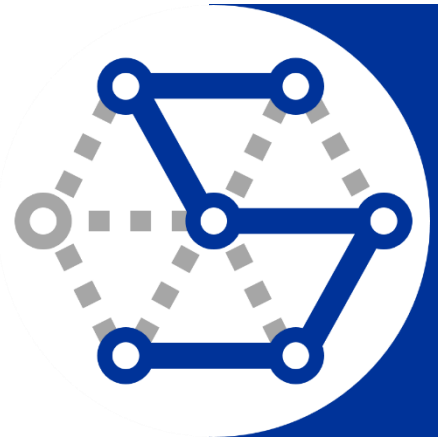


(MinPts=4, Eps=9.92)

DBSCAN: Eps 및 MinPts 결정

- 군집의 포인트에 대해 k번째 가장 가까운 이웃nearest neighbors 이 대략 동일한 거리에 있다는 아이디어
- 노이즈 포인트는 먼 거리에서 k번째 가장 가까운 이웃을 가짐
- 그래서 k번째 가장 가까운 이웃에 대한 모든 포인트의 정렬된 거리 구성





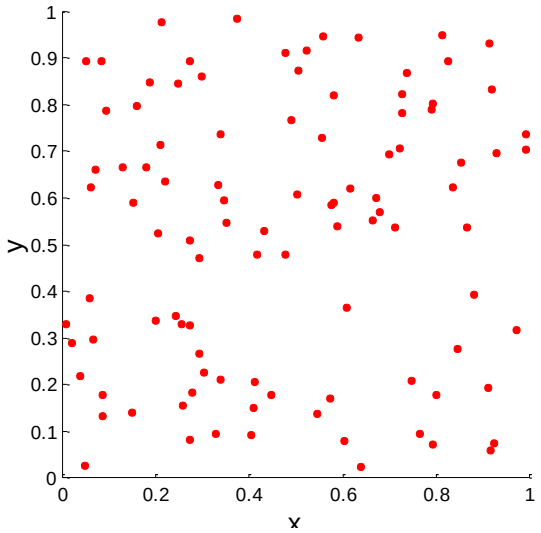
5. 군집 평가

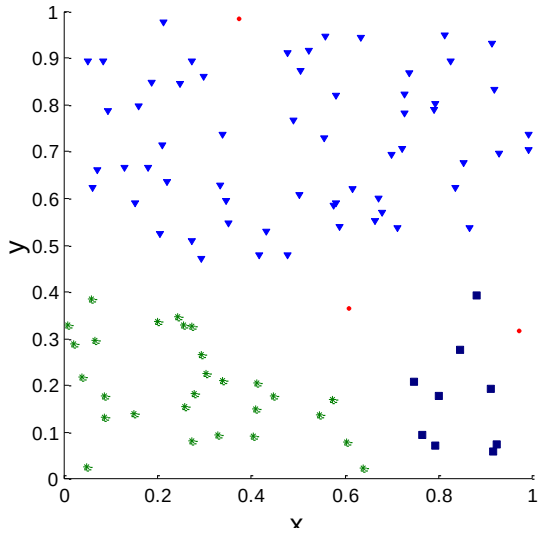
군집 유효성Cluster Validity

- 지도 분류supervised classification를 위해 우리 모델이 얼마나 훌륭한지 평가할 수 있는 다양한 측정 방법이 존재
  - 정확도accuracy, 정밀도precision, 리콜recall
- 군집 분석의 경우, 결과 군집의 "장점"을 평가하는 방법은?
- 그러나 "군집은 보는 사람의 눈에 있다!"
- 그렇다면 우리는 왜 그것을 평가하기 원하는가?
  - 노이즈 패턴을 찾는 것을 피하기 위해
  - 군집화 알고리즘 비교
  - 군집 의 두 세트를 비교
  - 두 군집을 비교

랜덤 데이터Random Data에서 발견된 군집

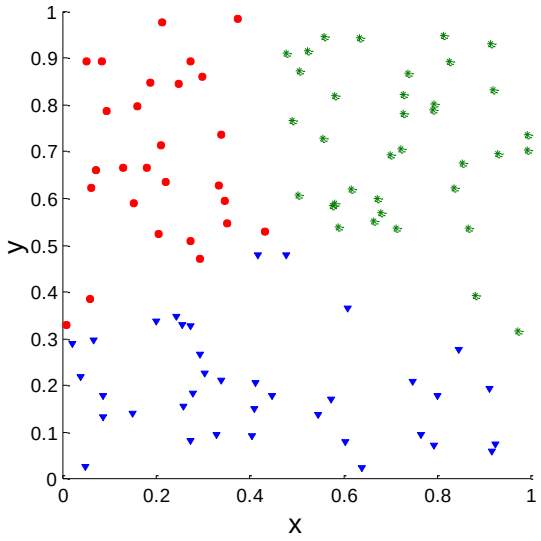
랜덤
포인트

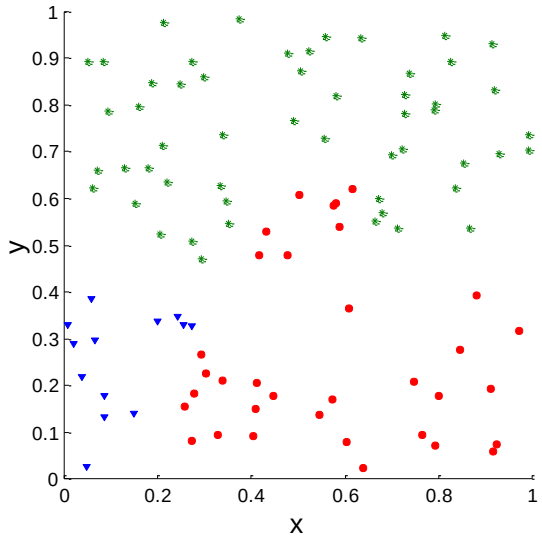




DBSCAN

K-means





Complete
Link

군집 유효성 검사의 다른 측면

1. 한 데이터 집합의 군집화 경향clustering tendency, 즉 비-랜덤 구조가 데이터에 실제로 존재하는지를 구별하는 것을 결정
2. 군집 분석 결과를 외부에서 알려진 결과(예: 외부적으로 주어진 클래스 레이블)와 비교
3. 군집 분석 결과가 외부 정보를 참조하지 않고 데이터에 얼마나 잘 맞는지 평가
   - 데이터만 사용
4. 군집 분석의 서로 다른 두 집합의 결과를 비교하여 어느 것이 더 나은지 결정
5. '정확한'군집 수 결정

- 2, 3, 4의 경우 전체 군집화를 평가할 것인지 아니면 개별 군집만 평가할 것인지를 더 구분할 수 있음

군집 유효성 측정

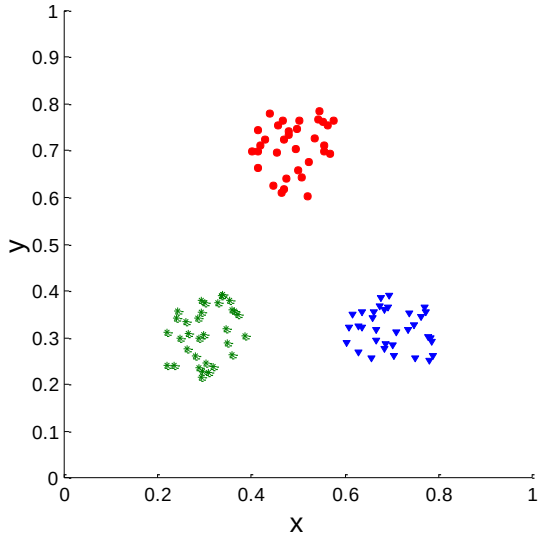
- 군집 유효성의 다양한 측면을 판단하기 위해 적용되는 수치적 수단은 다음 세가지 유형으로 분류
  - 외부 색인External Index: 군집 레이블이 외부에서 제공되는 클래스 레이블과 일치하는 정도를 측정하는데 사용
    - Entropy
  - 내부 색인Internal Index: 외부 정보를 고려하지 않고 군집화 구조의 장점을 측정하는데 사용
    - Sum of Squared Error(SSE)
  - 상대 색인Relative Index: 두 개의 다른 군집화 또는 군집을 비교하는데 사용
    - 외부 또는 내부 색인(예: SSE 또는 엔트로피)이 이 함수에 사용되는 경우가 많음
- 때로는 이것들을 지표indices 대신에 기준criteria 이라고 부르기도 함
  - 그러나 때로는 기준이 일반적인 전략이며 지수는 기준을 구현하는 수치 척도

상관관계Correlation를 통한 군집 유효성 측정

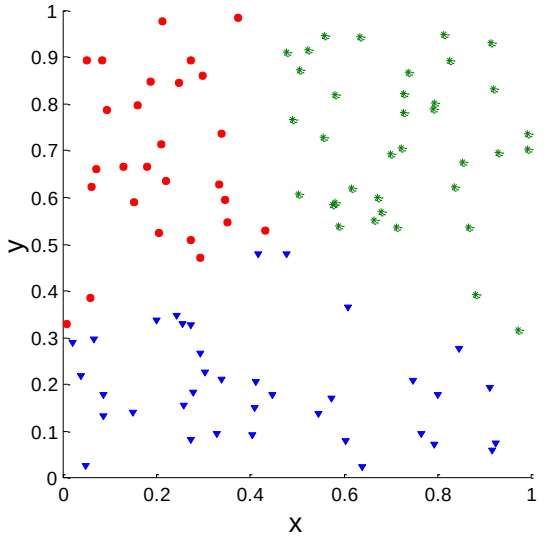
- 2개의 행렬
  - 근접 행렬Proximity Matrix
  - 이상적인 유사성 행렬Similarity Matrix
    - 각 데이터 요소에 대해 하나의 행과 하나의 열
    - 연관된 포인트 쌍이 동일한 군집에 속하면 항목은 1
    - 연관된 포인트 쌍이 다른 군집에 속하면 항목은 0
- 두 행렬 간의 상관관계 계산
  - 행렬은 대칭이므로 n(n-1)/2개 항목 간의 상관관계만 계산해야 함
- 상관관계가 높으면 동일한 군집에 속한 지점이 서로 가깝다는 것을 나타냄
- 일부 밀도 또는 인접성 기반 군집에 대해서는 적절한 측정 방법이 아님

상관관계Correlation를 통한 군집 유효성 측정

- k에 대한 이상적인 유사성과 인접성 행렬의 상관관계는 다음 두 데이터 집합의 군집화를 의미



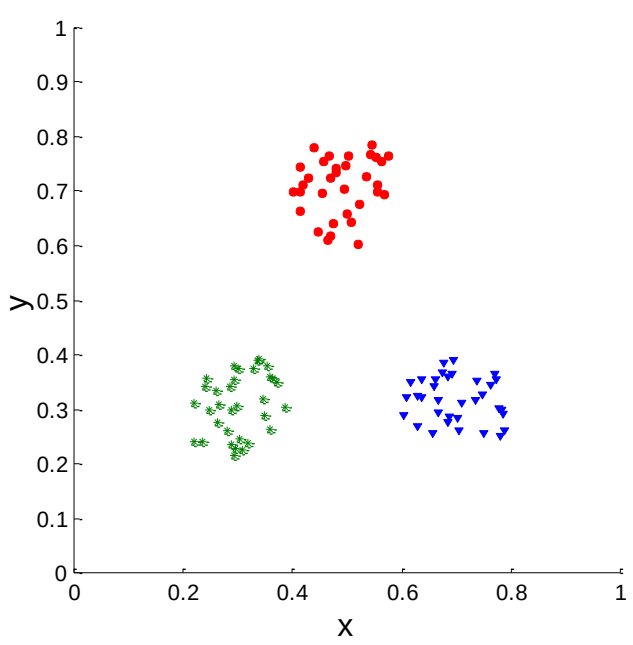
Corr = -0.9235

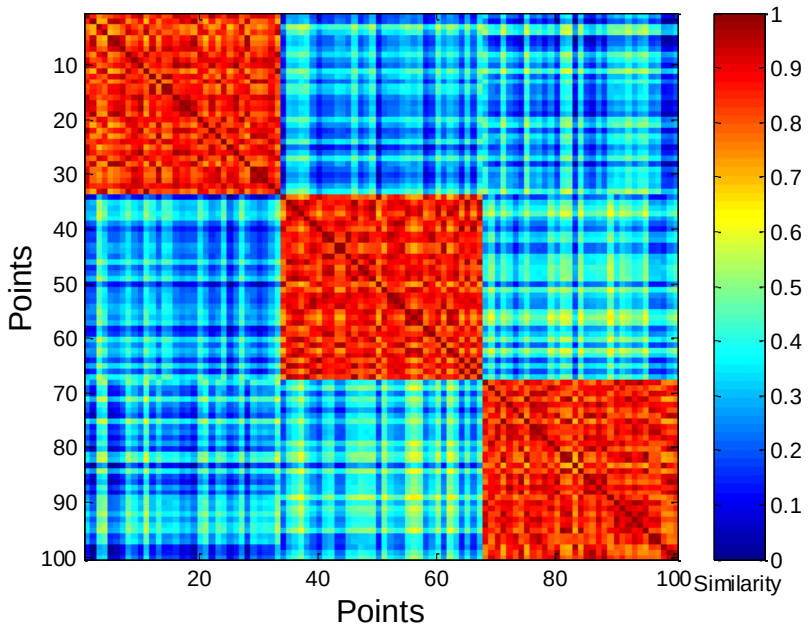


Corr = -0.5810

군집 유효성 검사에 유사성 행렬 사용

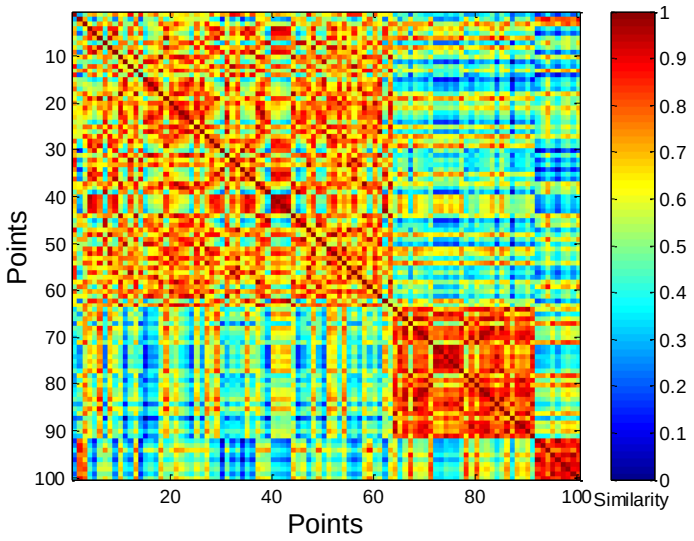
- 군집 레이블과 관련하여 유사성 행렬을 정렬하고 시각적으로 검사

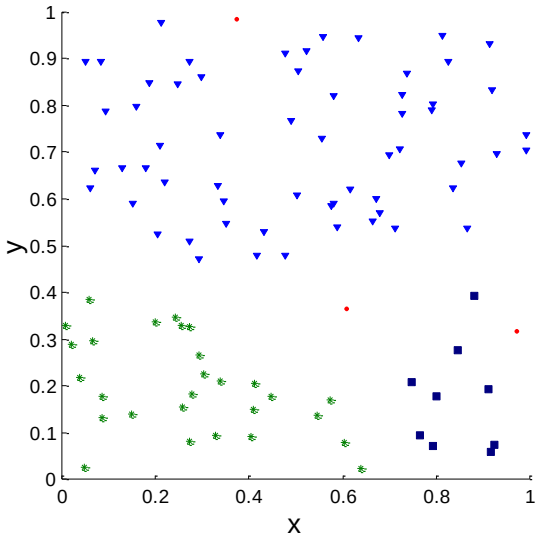




군집 유효성 검사에 유사성 행렬 사용

- 랜덤 데이터의 군집은 너무 선명하지 않음

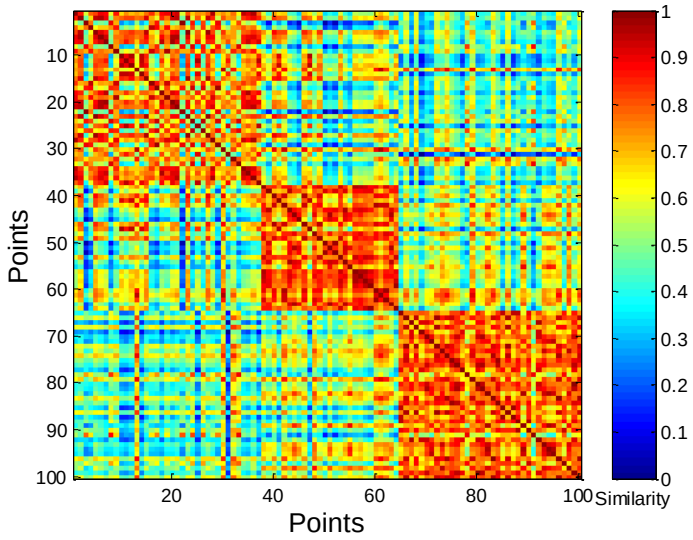


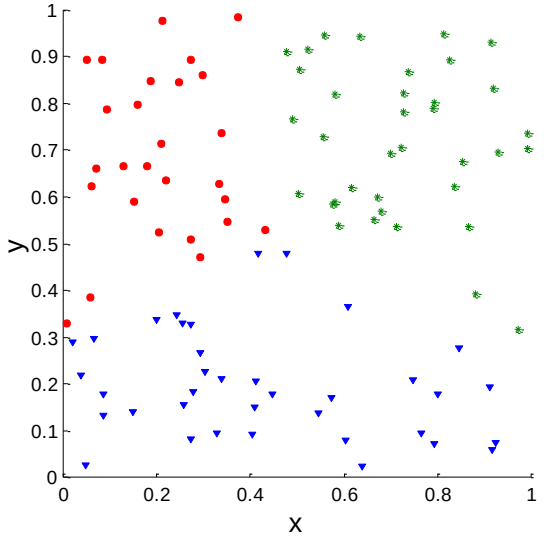


DBSCAN

군집 유효성 검사에 유사성 행렬 사용

- 랜덤 데이터의 군집은 너무 선명하지 않음

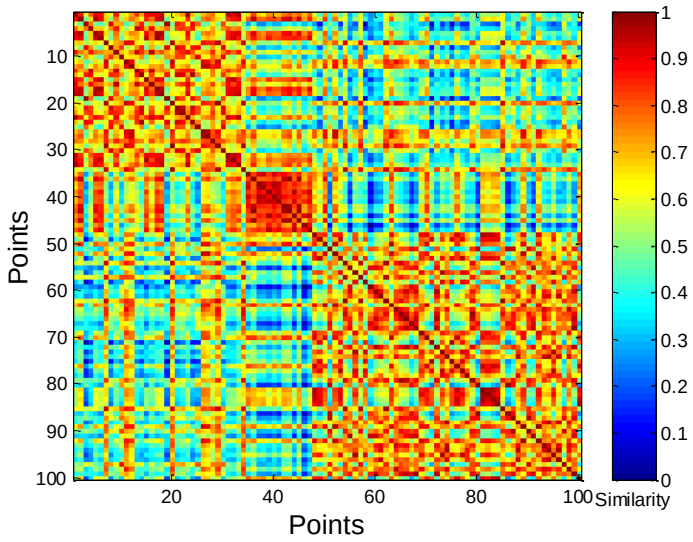


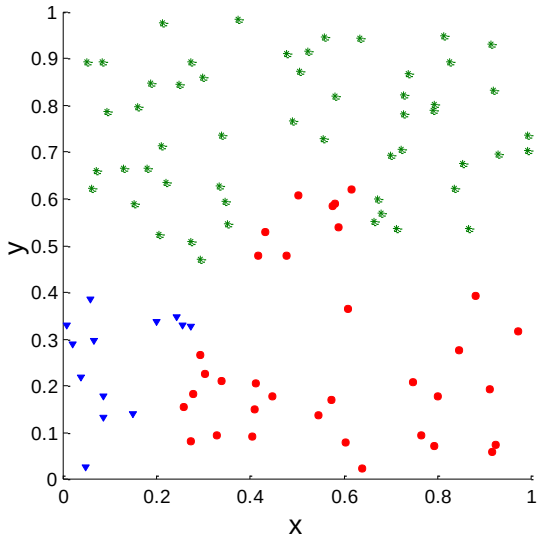


K-means

군집 유효성 검사에 유사성 행렬 사용

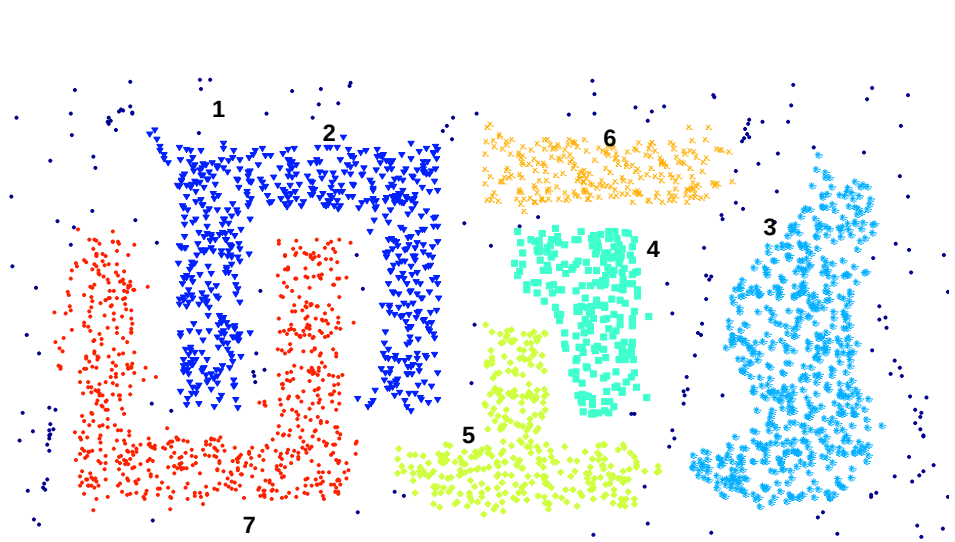
- 랜덤 데이터의 군집은 너무 선명하지 않음

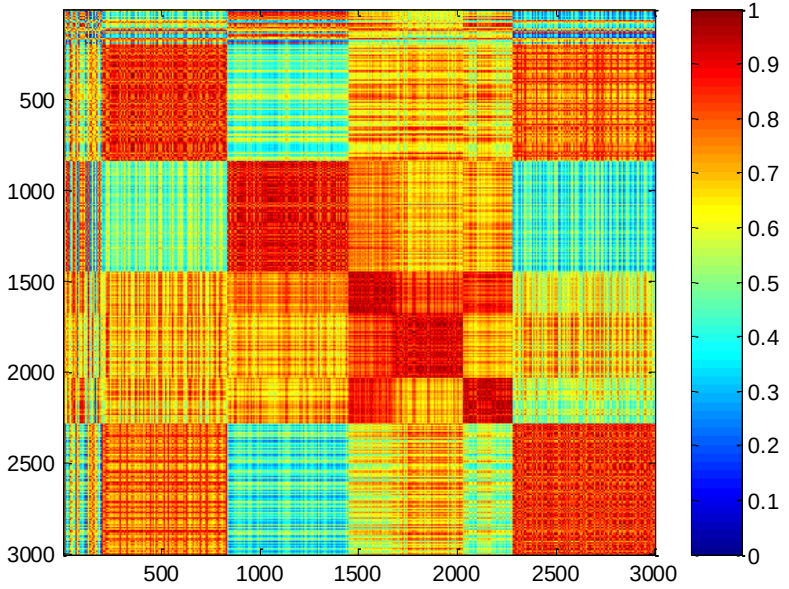




Complete Link

군집 유효성 검사에 유사성 행렬 사용

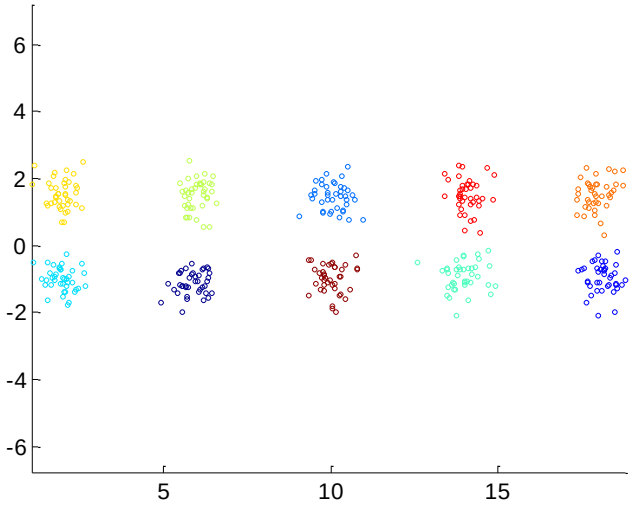


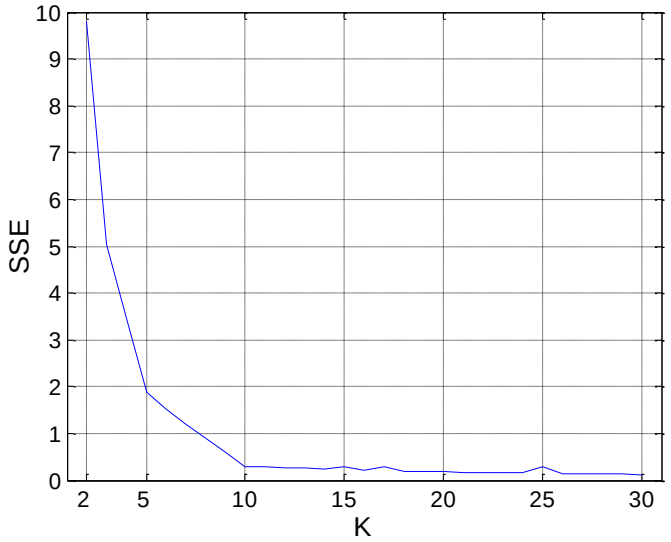


DBSCAN

내부 측정Internal Measures: SSE

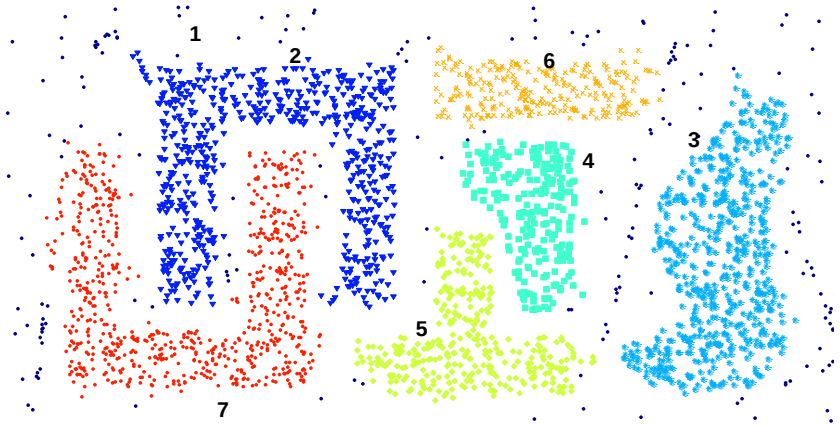
- 더 복잡한 그림의 군집은 잘 분리되어 있지 않음
- 내부 색인Internal Index: 외부 정보를 고려하지 않고 군집화 구조의 장점을 측정하는데 사용
  - SSE
- SSE는 두 개의 군집화 또는 두 개의 군집(평균 SSE)를 비교하는데 적합
- 군집 수를 추정하는 데에도 사용할 수 있음

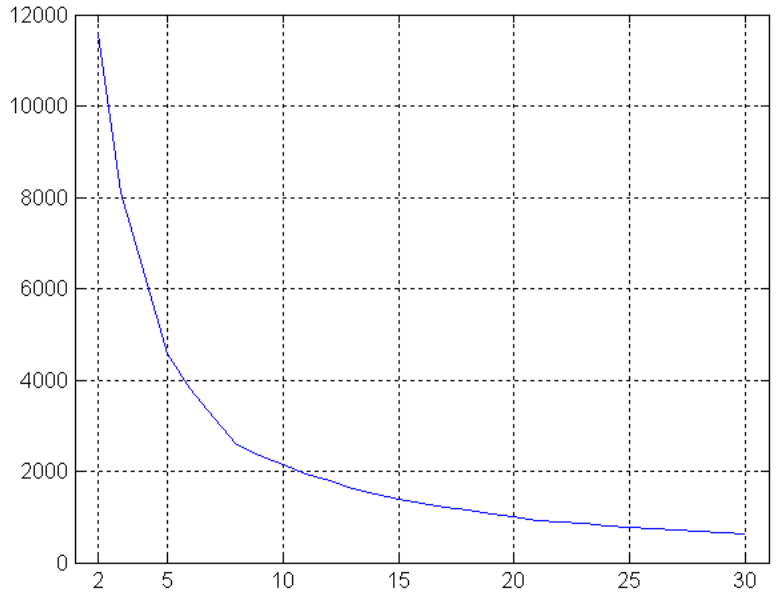




내부 측정Internal Measures: SSE

- 보다 복잡한 데이터 집합의 SSE 곡선





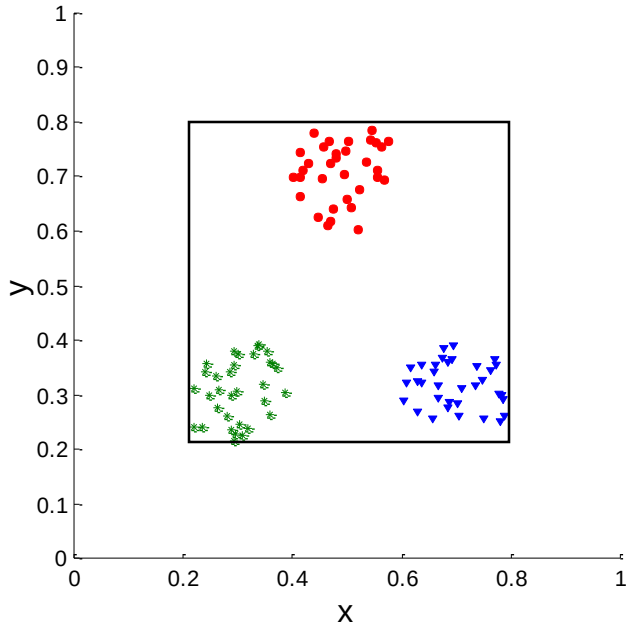
K-means를 사용하여 발견된 군집의 SSE

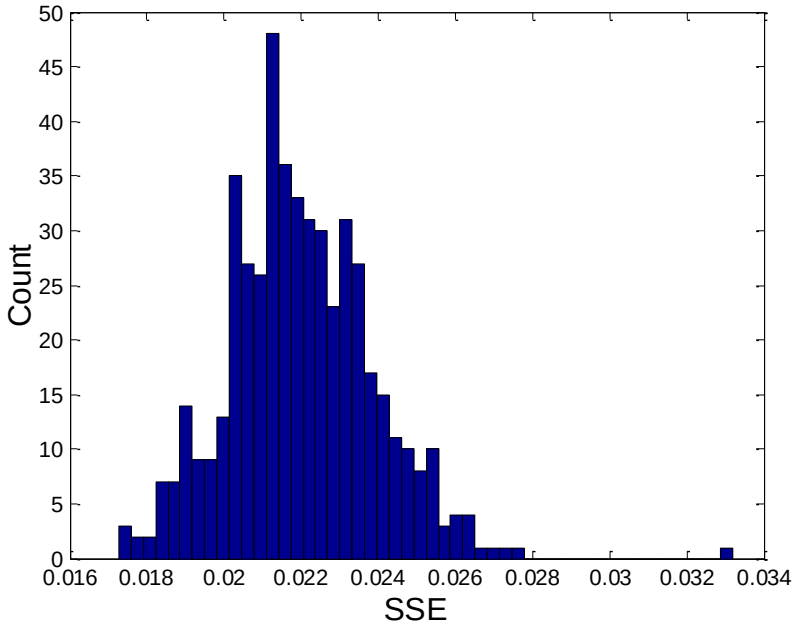
군집 유효성을 위한 프레임워크

- 모든 조치를 해석할 수 있는 프레임워크 필요
  - 예를 들어, 평가 척도가 10이라는 값이 좋거나, 공평하거나, 나쁘거나?
- 통계는 군집 유효성에 대한 프레임워크를 제공
  - 군집화 결과가 "이례적"일수록 데이터에서 유효한 구조를 나타낼 확률이 높음
  - 랜덤 데이터 또는 군집화의 결과인 인덱스 값을 군집화 결과값과 비교할 수 있음
    - 인덱스 값이 거의 없는 경우, 군집 결과는 유효
  - 이러한 접근법은 더 복잡하고 이해하기가 더 어려움
- 두 가지 다른 군집 분석 집합의 결과를 비교하기 위해 프레임워크가 덜 필요
  - 그러나 두 인덱스 값의 차이가 중요한지에 대한 질문이 있음

SSE를 위한 통계적 프레임워크Statistical Framework

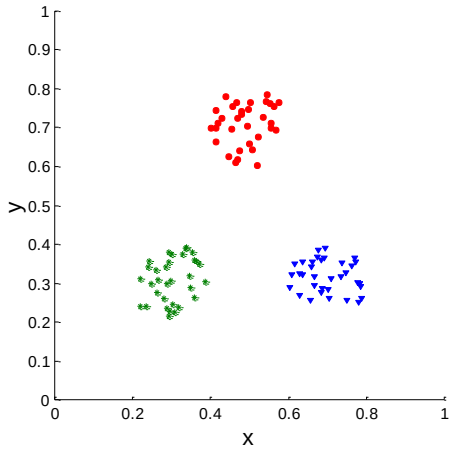
- 예제
  - 랜덤 데이터의 3개 군집에 대해 0.005의 SSE를 비교
  - 히스토그램은 x와 y값에 대해 0.2 – 0.8의 범위에 분산된 100개의 랜덤 데이터 포인트의 500 집합에 있는 3개의 군집 SSE를 보여줌



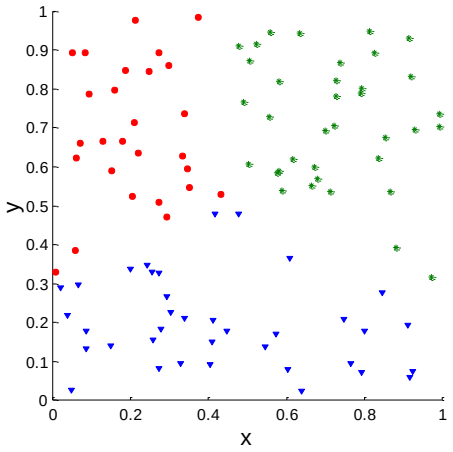


상관관계에 대한 통계적 프레임워크Statistical Framework

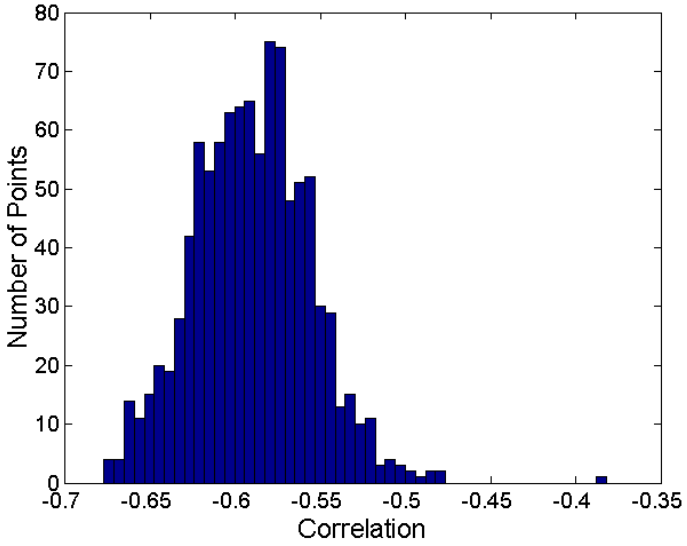
- k에 대한 이상적인 유사성과 인접성 행렬의 상관관계는 다음 두 데이터 집합의 군집화를 의미



Corr = -0.9235



Corr = -0.5810



내부 측정Internal Measures: 응집력Cohesion과 분리력Separation

- 군집 응집력Cluster Cohesion: 군집 의 객체와 얼마나 밀접한 관련이 있는지 측정
  - 예: SSE
- 군집 분리력Cluster Separation: 군집이 다른 군집과 얼마나 다른지 또는 잘 분리되어 있는지 측정
  - 예: Squared Error
  - 응집력은 군집 내 제곱합(SSE)

$$SSE = WSS = \sum_{i} \sum_{x \in C_i} (x - m_i)^2$$

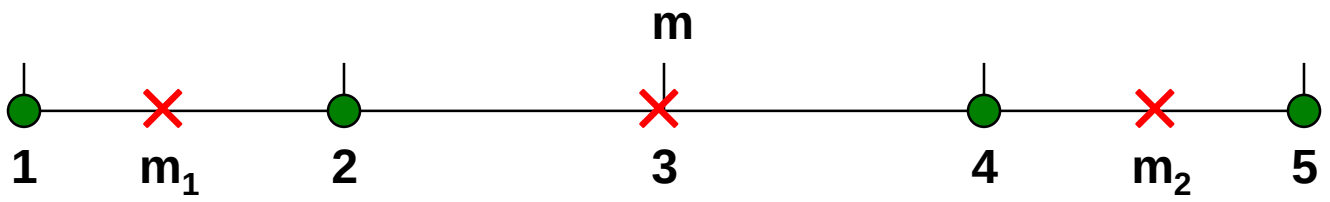
  - 분리는 군집 제곱합 사이에서 측정

$$BSS = \sum_{i} |C_i| (m - m_i)^2$$

    - 여기서 |Ci|는 군집 i의 크기

내부 측정Internal Measures: 응집력Cohesion과 분리력Separation

- 예: SSE
  - BSS + WSS = 상수constant



K=1 cluster:

$$SSE = WSS = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$BSS = 4 \times (3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

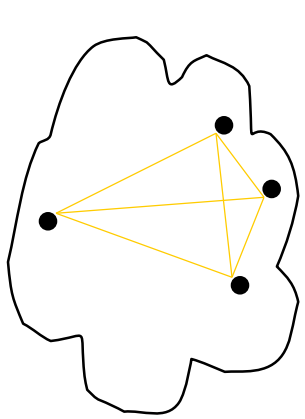
$$SSE = WSS = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$BSS = 2 \times (3-1.5)^2 + 2 \times (4.5-3)^2 = 9$$

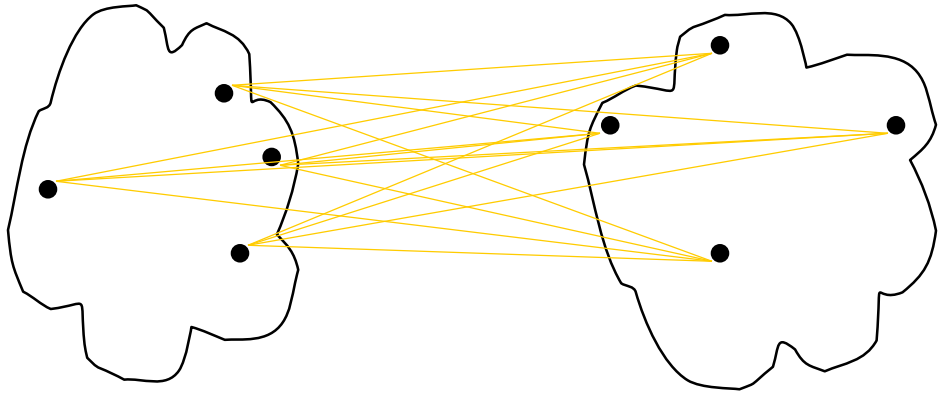
$$Total = 1 + 9 = 10$$

내부 측정Internal Measures: 응집력Cohesion과 분리력Separation

- 근접 그래프 기반 접근법은 응집력과 분리력에 사용될 수 있음
  - 군집 응집력은 군집 내의 모든 링크의 가중치 합계
  - 군집 분리력은 군집 노드와 군집 외부의 노드 사이의 가중치 합



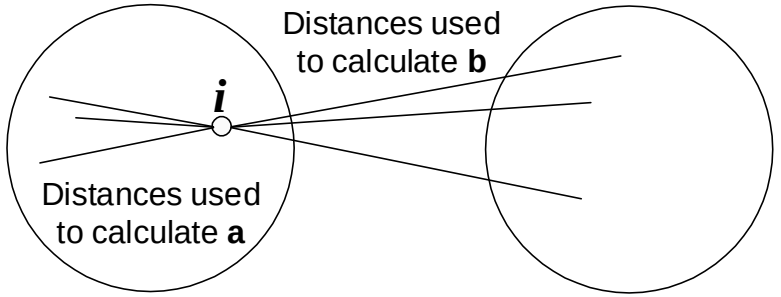
응집력



분리력

내부 측정Internal Measures: 실루엣 계수Silhouette Coefficient

- 실루엣 계수 아이디어는 응집력과 분리력의 개념을 결합하지만, 군집 및 군집화 뿐만 아니라 개별 포인트에 대한 것
- 개별 포인트 i에 대해서
  - a = 군집의 포인트에 대한 i의 평균 거리를 계산
  - b = min(다른 군집의 포인트까지의 평균 거리) 계산
  - 포인트에 대한 실루엣 계수는 $s = (b - a)/\max(a, b)$로 주어짐



  - 일반적으로 0에서 1 사이
  - 1에 가까울수록 좋음
- 군집 또는 군집화의 평균 실루엣 계수를 계산할 수 있음

군집 유효성의 외부 측정: 엔트로피 및 순도

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

Q & A