




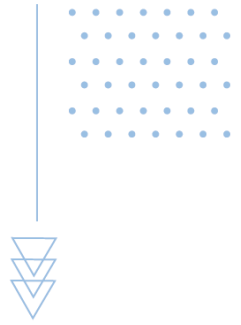
# IT와 비즈니스혁신

W11. 마이닝 기법 II: 군집 분석



# Contents

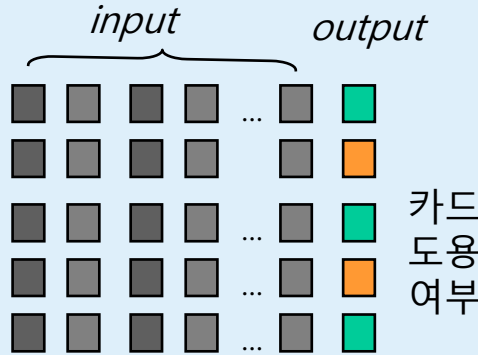
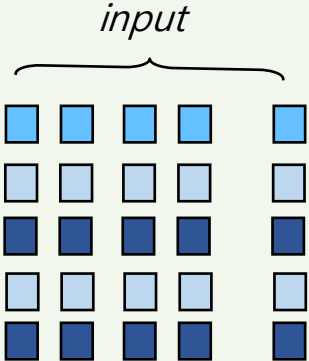
- I. 군집 분석 개요
  - II. 군집 분석 원리
  - III. 활용 사례
  - IV. 정리
- 
- 
- 



---

# 1. 군집 분석 개요

데이터 마이닝은 크게 **출력 변수**의 존재 여부에 따라 **지도학습과 자율학습**으로 나눌 수 있음

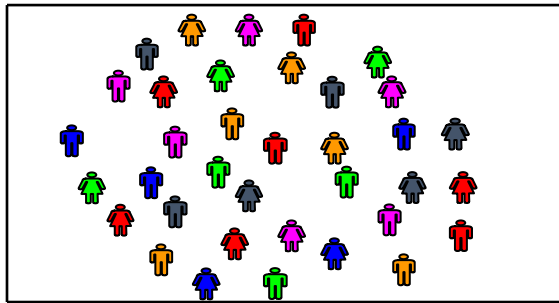
	지도학습 (supervised learning)	자율학습 (unsupervised learning)
의미	<ul style="list-style-type: none"> <li>입력 데이터와 정답(Label)을 제공 받아 이를 통해 입력(독립)과 출력 (Label, 종속,타겟) 으로 매칭할 수 있는 규칙 생성</li> </ul> <p>예. 카드번호, 성별, 나이,거래 내역 등 →  카드도용 여부</p>	<ul style="list-style-type: none"> <li>외부에서 정답(Label)이 주어지지 않음</li> <li>입력 데이터에서 패턴을 찾아내는 작업</li> </ul> <p>예. 군집화: 주어진 데이터를 3 개의 그룹으로 나눔 </p>
특징	출력 변수가 존재함	출력 변수가 존재하지 않음
분석 기법	의사결정나무, 회귀분석, 인공신경망, 판별분석 등	<b>군집분석</b> , 연관성 분석 등

## ■ 데이터 안에 존재하는 의미 있는 그룹 · 군집(cluster)을 찾아내는 작업

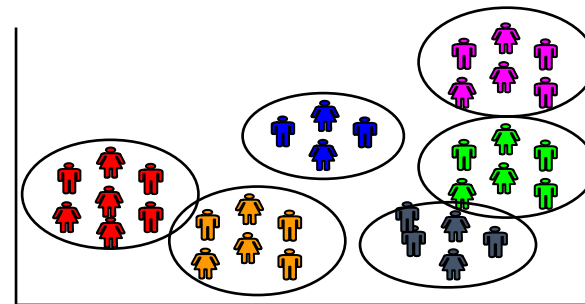
- 집단 또는 범주에 대한 사전 정보가 없는 데이터의 경우
- 주어진 관측 값을 사용하여 전체를 몇 개의 유사한 집단으로 그룹화하고 각 집단의 성격을 파악하기 위한 기법

## ■ 활용 사례

- 정치, 선거 등에서 유권자들을 특성에 따라 몇 개의 그룹으로 나누고 각 그룹 별로 다른 문구를 사용하여 선거 유세
- 고객을 몇 개의 그룹으로 나누어 그룹별로 다른 마케팅 전략 적용



군집화



### ☐ 데이터를 설명하기 위한 군집화

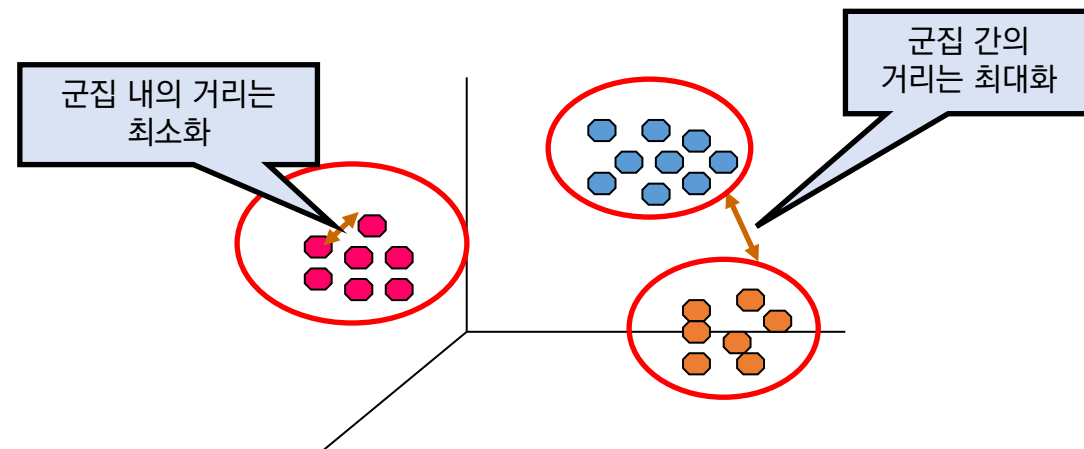
- 군집 분석의 결과로 나온 각 군집에 대한 특성을 설명
- 동일한 군집에 속하는 개체는 데이터 속성이 비슷하고, 서로 다른 군집에 속한 개체들과는 상이하도록 군집이 구성
- 응용 사례
  - 고객의 특성에 따라 고객을 군집화하고 각 군집에 적합한 마케팅 전략 수립
  - 유사 주제로 군집화된 문서 그룹에 대하여 문서 전체를 읽지 않고도 주요 주제를 이해, 요약할 수 있음

### ☐ 전처리 작업을 위한 군집화

- 데이터 축소가 필요한 경우 군집화 이용
  - 알고리즘의 계산 복잡도가 감소하나 정보 손실 불가피
- 군집 분석 이후 각 군집이 의미하는 것을 파악하기 위해 다른 데이터 마이닝 기법이나 통계적 분석 기법 적용

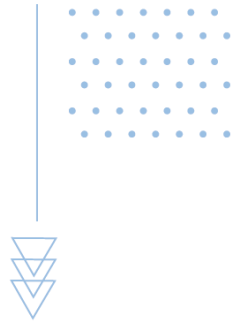
## 유사성 측정 방법

- 가장 보편적으로 사용되는 방법: 거리 측정
- 동일 군집 내 개체간의 거리는 최소화, 군집 간의 거리는 최대화



## 군집을 나누는 방법

- 비계층적 (non-hierarchical) 방법
  - 다변량 자료의 산포를 나타내는 여러 가지 측도를 이용하여 이들 판정기준을 최적화시키는 방법으로 군집을 나누는 방법
  - 한 번 분리된 개체도 반복적으로 시행하는 과정에서 재분류될 수 있음
  - 대표적인 방법: **k-means 군집 분석**
- 계층적 (hierarchical) 방법
  - 가까운 개체끼리 차례로 묶거나 멀리 떨어진 개체를 차례로 분리해 가는 군집방법
  - 한 번 병합된 개체는 다시 분리되지 않음
  - 응집형과 분리형



## II. 군집 분석 원리

$k$ -mean 군집 분석, 계층적 군집 분석



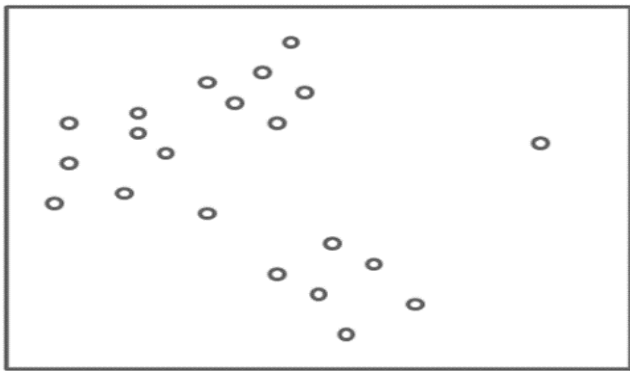
## 목표

- 데이터를 사전에 결정된  $k$ 개의 군집으로 할당

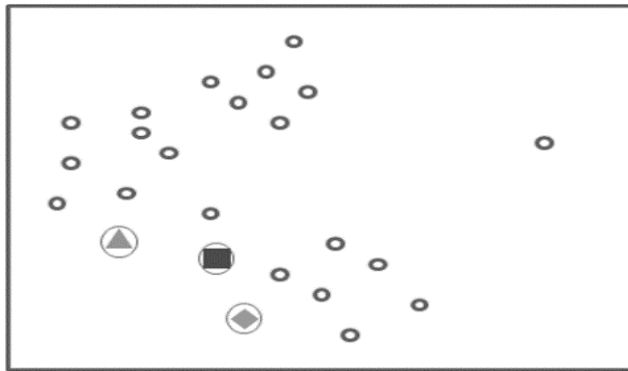
## 분석 단계

- 단계 1: 임의로  $k$ 개의 중심(centroid) 선택
- 단계 2: 각 데이터에 대하여 최근접 중심을 구하고 그 군집에 할당
- 단계 3: 단계 2의 결과에 따라 각 군집의 중심을 다시 계산
- 단계 2, 3을 반복
- 종료: 단계 2에서 모든 데이터의 군집이나 중심에 큰 변화가 없는 경우 또는 미리 정해진 반복횟수에 도달할 때

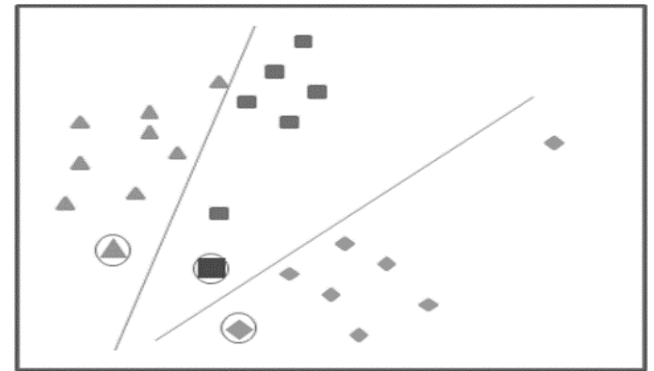
## $k$ -means 군집 분석 단계



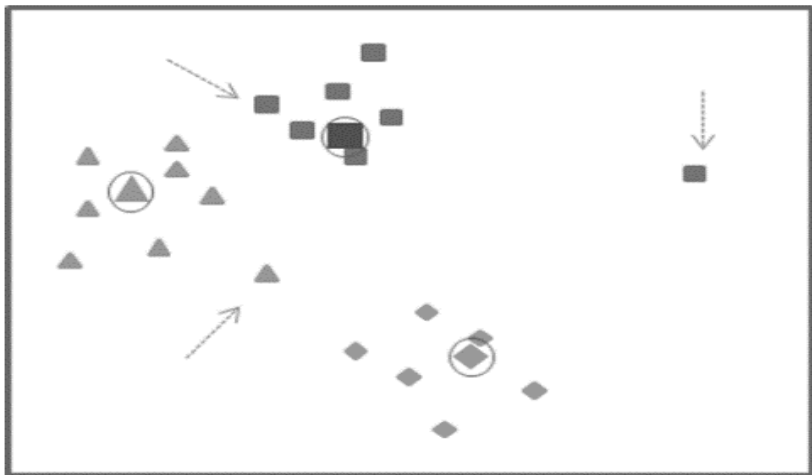
데이터



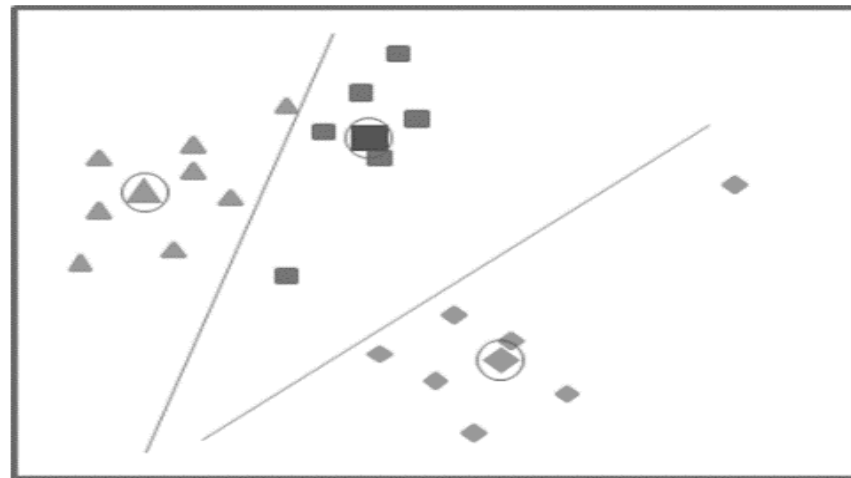
초기 임의로  $k$ 개의 중심 선택



최근접 중심의 군집으로 데이터를 할당



새로운 중심으로 데이터 재할당

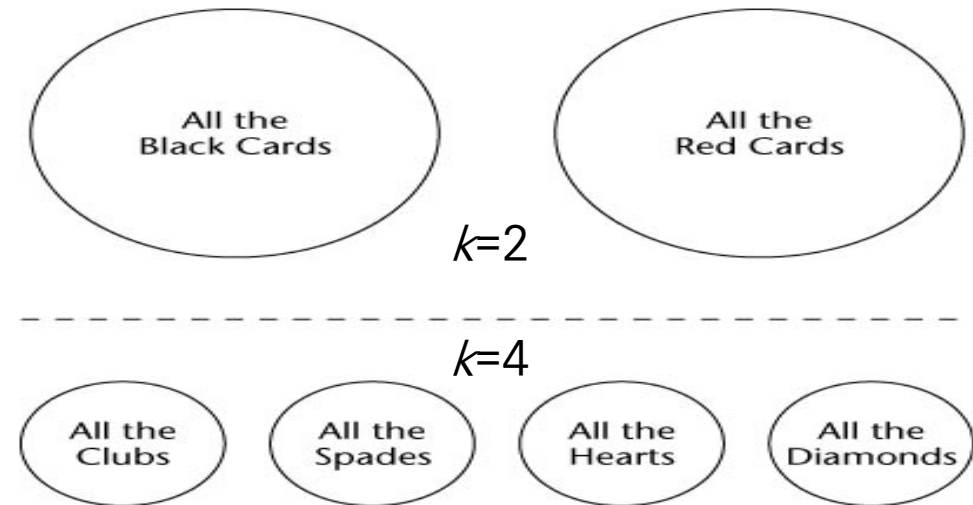


각 군집의 새로운 중심 계산

## k선택 기준

- 특정 값을 선택해야 하는 사전적(priori) 이유가 존재하지 않음
- 다양한  $k$ 값을 이용하여 얻어진 군집의 성능을 평가
  - 군집 내에서의 데이터간 평균 거리
  - 군집 중심점들간의 평균거리 비교
- 군집 내의 유사성에 대한 표준 척도: 분산
  - 낮은 분산일수록 좋은 집합
  - 군집의 크기 고려: 평균분산(=분산/군집크기)
- 주어진 상황에 맞는 주관적인 평가 기준 필요

\*elbow 방법 실습에서 추가 설명



군집화는 종종 하나 또는 몇 개의 강력한 군집을 생성하며 이 큰 군집들의 레코드들은 상당히 유사

- 강력한 군집을 보다 자세히 분석하는 것은 가치가 있음
- 이 강력한 군집을 제거하고 나머지 레코드들만으로 다시 새로운 군집들을 만들어보는 것도 유용

■ 수치형 변수:  $X = (X_1, X_2, \dots, X_n)$  ,  $Y = (Y_1, Y_2, \dots, Y_n)$

- 두 점 사이의 기하학적 거리

- 맨하탄 (Manhattan) 거리:  $d(X, Y) = \sum d(X_i - Y_i)$

- 유클리디안 (Euclidean) 거리:  $d(X, Y) = [\sum (X_i - Y_i)^2]^{1/2}$

- 두 벡터 사이의 각

- 비교의 대상이 되는 두 사물들의 크기 차이에 영향을 받지 않는 연관성에 대한 척도

- 예. 사자와 집고양이 비교

수염 길이, 꼬리 길이, 전체 신체 길이, 이빨 길이, 발톱 길이의 비가 유사 하다면 두 벡터는 거의 수평

### 단위 변환: 변수들의 범위를 동일하게 만들기

- (1) 각 변수로부터 최소값을 빼고 각 변수의 범위(최대값과 최소값의 차)로 나눈다.
- (2) 각 변수를 평균값으로 나눈다.
- (3) 각 변수로부터 평균값을 빼고 표준편차로 나눈다. (표준화)
- (4) 벡터 정규화

– 데이터 간의 차이보다는 데이터 안에서의 차이를 강조

예. 부채 \$200,000, 자산 \$100,000 vs. 부채 \$10,000, 자산 \$5,000

부채와 자산의 비율이 같으므로 동일한 것으로 봄

### 가중치(Weight)

- 특정 변수가 다른 변수보다 더 중요함을 나타냄
- 단위로 인한 편향을 없애기 위해 단위를 조정한 후, 가중치를 이용하여 비즈니스 상황의 지식을 기반으로 한 편향을 도입

### 오차 제곱합

- 군집 내의 모든 데이터로부터 중심까지의 오차의 제곱합(SSE)

$$SSE = \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - \mu_i\|^2$$

### 데이비스-볼딘 지수 (Davies-Bouldin index)

- 군집 내의 분리도와 군집 간의 분리도의 비율
- 지수값이 작을수록 좋은 군집 결과임: 군집 내의 분리도는 낮고 군집간의 분리도는 높음

$$R_{i,j} = \frac{S_i + S_j}{M_{i,j}} \quad D_i \equiv \max_{j \neq i} R_{i,j} \quad DB \equiv \frac{1}{N} \sum_{i=1}^N D_i$$

$S_i$ :  $i$  번째 군집에 속한 데이터들과 중심 간의 평균 거리  
 $M_{i,j}$ :  $i$  번째와  $j$  번째 군집의 중심 간의 거리

### 실루엣 계수 (Silhouette coefficient) \*실습에서 추가 설명

- 군집의 밀집 정도를 계산, 높을수록 좋으며 최대 점수는 1임 (-1은 잘못된 군집, 0은 중첩 군집)

$$s = \frac{b - a}{\max(a, b)}$$

$a$ : 해당 데이터 포인트와 같은 군집 내에 있는 다른 데이터 포인트와의 거리를 평균한 값

$b$ : 해당 데이터 포인트가 속하지 않은 군집 중 가장 가까운 군집과의 평균 거리

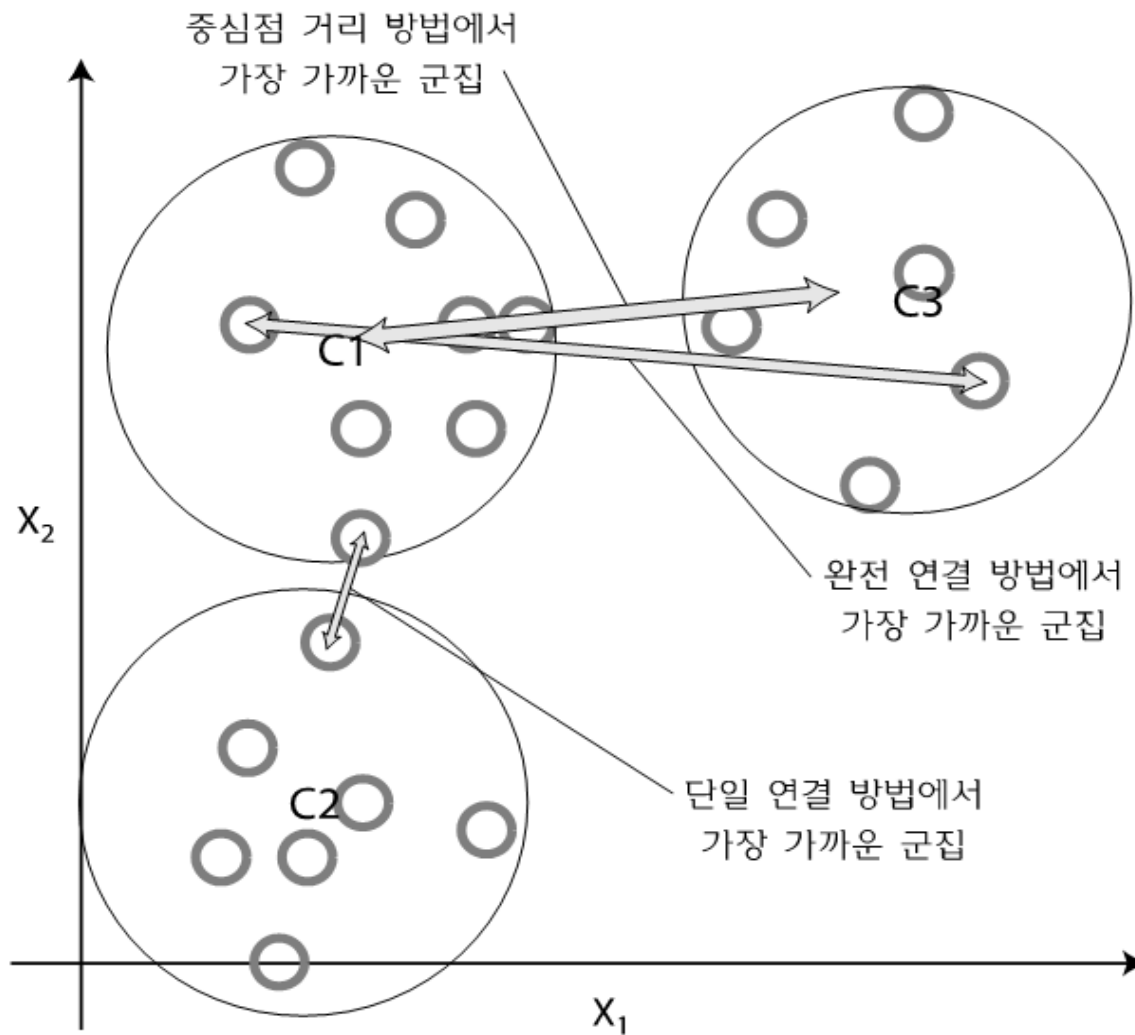
### 가까운 개체들끼리 묶거나 분할하여 군집을 만드는 방법

- 응집형 (Agglomerative)
  - 각 데이터 점이 각각 군집을 형성
  - 최종적으로 하나의 큰 군집으로 만들어질 때까지 큰 군집으로 통합해 감
- 분리형 (Divisive)
  - 데이터 전체를 하나의 군집에서 시작해서 개별 데이터로 분리해 나감

\* 군집의 개수를 미리 정할 필요가 없음

## 군집 사이의 거리 측정 방법

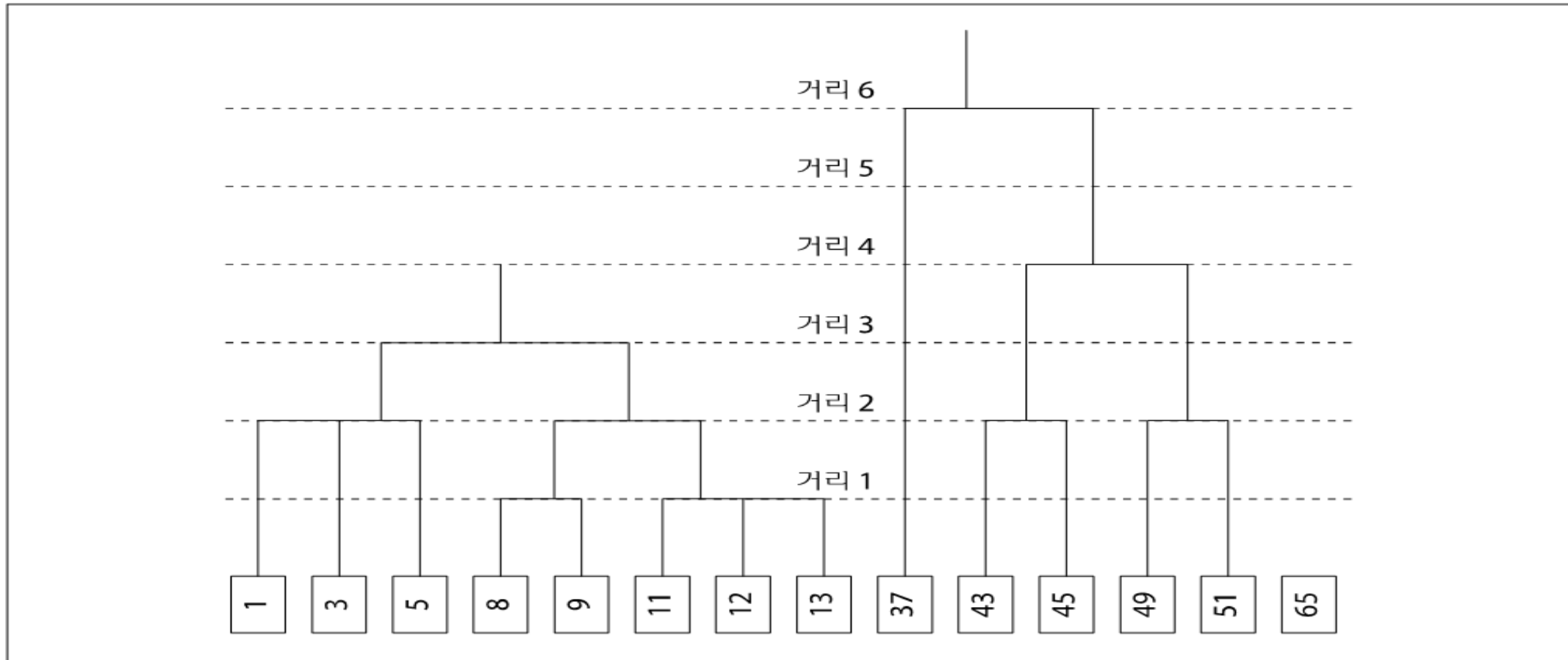
- 단일 연결(Single linkage):  
가장 가까운 개체 사이의 거리로  
군집간 거리를 측정
- 완전 연결(Complete linkage):  
거리가 가장 먼 개체들 사이의 거리로  
측정
- 중심점 거리(Centroid distance):  
두 군집의 중심점 사이의 거리로 측정

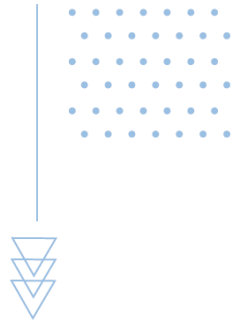




## ■ 나이에 따라 군집하기: 응집형 계층적 군집화

- 가장 일반적인 것에서 시작하여 점차 상세한 것으로 이동





---

### III. 군집 분석 활용 사례

## ■ 소비패턴 분석을 통해 고객 남녀를 9개의 코드로 정의하고 카드 상품 개발 (2014)

- 18종의 카드 출시 2년 만에 누적 발급 500만매 돌파
- 정교한 빅데이터 분석을 통해 타겟 고객군을 명확히 하고 코드별 니즈에 맞는 서비스로 상품을 구성
- 예. 신용카드 '23.5'
  - 새롭고 다양한 분야에 관심이 많은 사회초년생(Rookie)
  - 감각적 소비가 많은 호기심 많은 여성(Trend Setter)



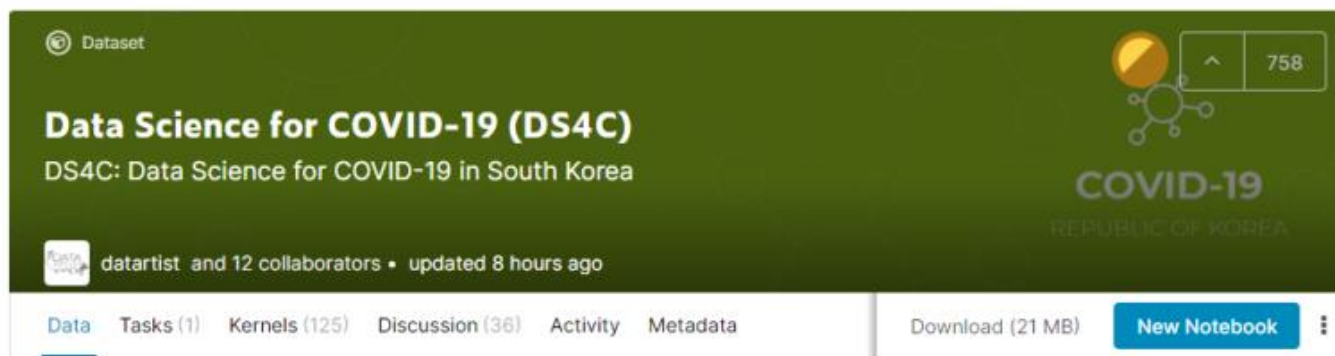
\* Source: <https://www.shinancard.com/>

## 입원 환자 특성별로 요양 병원 유형 도출

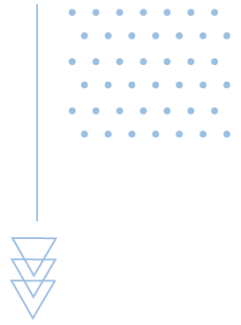
- 요양 병원관 관련한 정책 수립시 세분화된 맞춤형 정책 개발 가능
- 비교적 경증의 젊은 환자가 짧은 기간 이용하는 병원, 의학적 필요도가 높은 환자들이 주로 이용하는 병원, 고령의 치매환자 대상으로 하는 장기요양병원, 정신병원의 4개 유형으로 나눔
- 군집 분석의 변수
  - 정신질환자 비율, 치매환자 비율, 전문 재활서비스 이용환자 비율, 연령비율, 환자군 비율(의료최고도 및 의료고도, 의료중도 및 문제 행동군, 인지장애군 및 의료경도, 신체기능저하군) 등

## 코로나 확진자 군집화

- 캐글(데이터 분석 대회 플랫폼)에 한국의 코로나 확진자 정보를 배포 (2020.3)
- 확진자의 특성에 따른 군집 분석 가능
- 확진자 수 예측에 도움이 될 것으로 예상



Source: <http://www.moneews.co.kr/news/articleView.html?idxno=206798>  
<https://www.kaggle.com/kimjihoo/coronavirusdataset?select=PatientInfo.csv>



---

## IV. 정리

## ■ 자동 군집 탐지

- 자율학습 알고리즘
- 군집화 알고리즘은 특정 유사도 척도에 의존

## ■ 군집 분석의 장단점

- 사전에 그룹 분류에 대한 정보가 없는 데이터를 가지고 사용자가 추구하는 바에 맞게 그룹화
- $k$ -means 군집 분석 기법의 경우 사용자가 사전 지식 없이 그룹의 수를 정하게 되면  
분석 결과가 잘 나오지 않거나, 분석 결과에 대한 해석이 어려워짐
- 이상치 (outlier) 는 중심점을 변경하는 과정에서 군집 내의 전체 평균 값을 크게 왜곡시킬 수 있으므로 이상치 관리가 중요함
- 다른 데이터 마이닝 기법이나 통계적 분석 기법과 병행하여 사용되는 경우가 많음