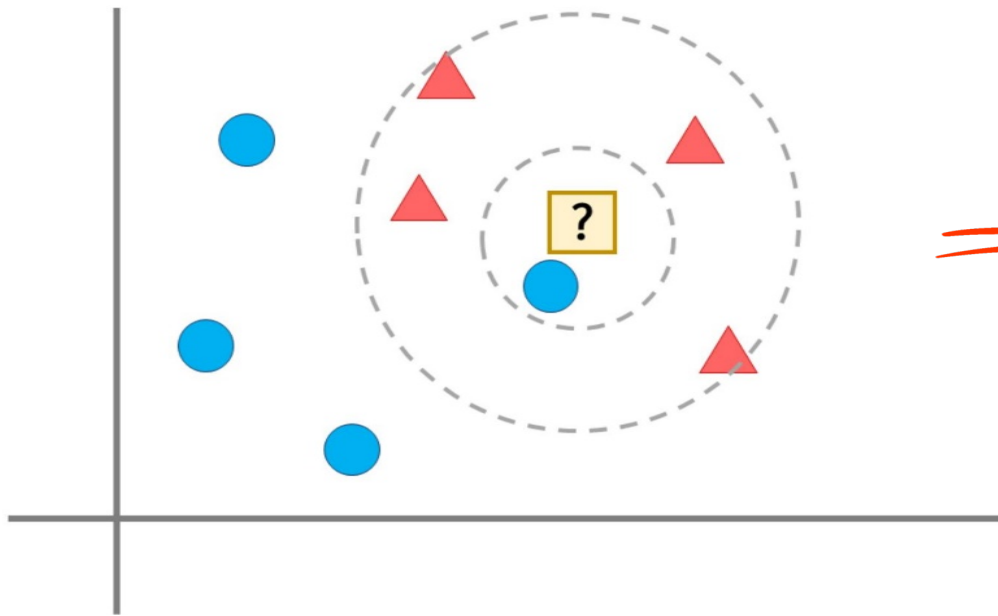


## Algorithm Fundamentals

### K-Nearest Neighbors Algorithm



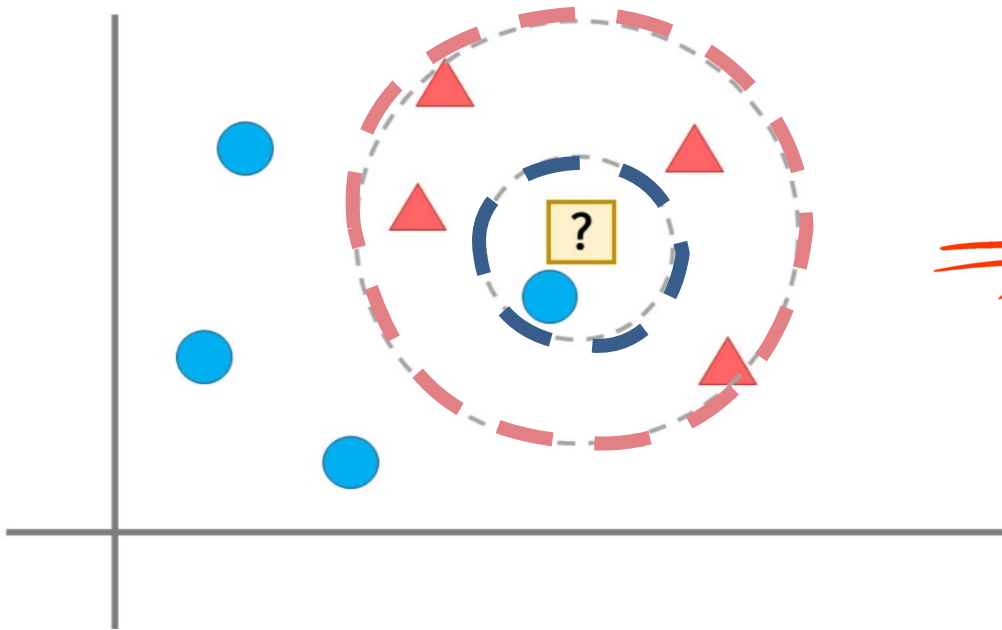
지도학습(Supervised learning)



k개의 이웃하는 기존 관측치의  
**최빈값**을 따른다!

## Algorithm Fundamentals

### K-Nearest Neighbors Algorithm



관측치 간 **거리정보**를 통해 새 관측치의  
범주를 결정!



**유클리드 거리**  
(Euclidean Distance)

## Distance Metric


관측치들이 서로 얼마나 떨어져 있는 지에 대한 거리지표

### Minkowski distance (민코우스키 거리)

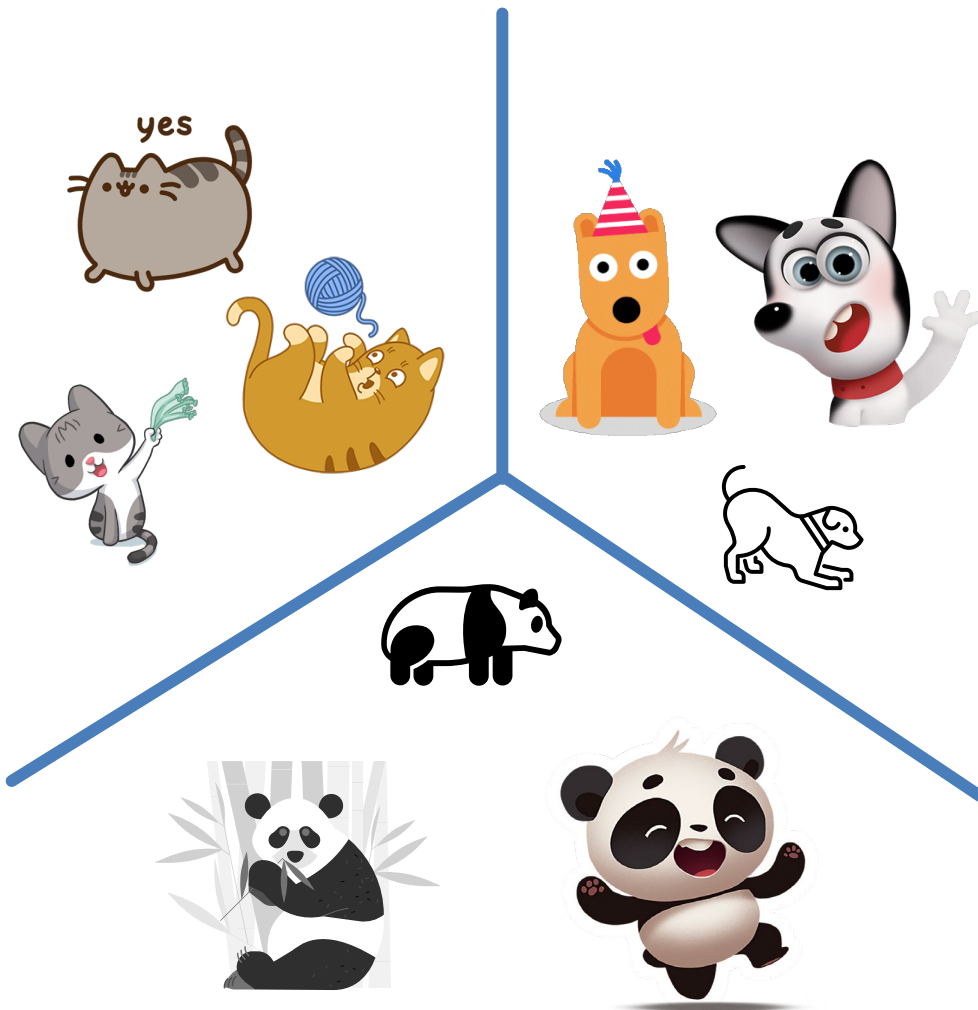
$$d_{minkowski}(x, y) \equiv \left( \sum_{j=1}^m |x_j - y_j|^r \right)^{1/r}$$

Norm들의 일반화 된 표현

$r=1$   Manhattan Distance(맨하탄거리)

$r=2$   Euclidean Distance(유클리드거리)

## What is Clustering?



**속성이 유사한** 관측치들끼리

지정한 Cluster의 개수로

묶어주는 방법!

Ex) 강아지, 고양이, 판다

## What is Clustering?

군집 간 분산  
(inter-cluster variance)

최대

속성이 유사한 관측치들끼리

군집 내 분산  
(intra-cluster variance)

최소

지정한 Cluster의 개수로

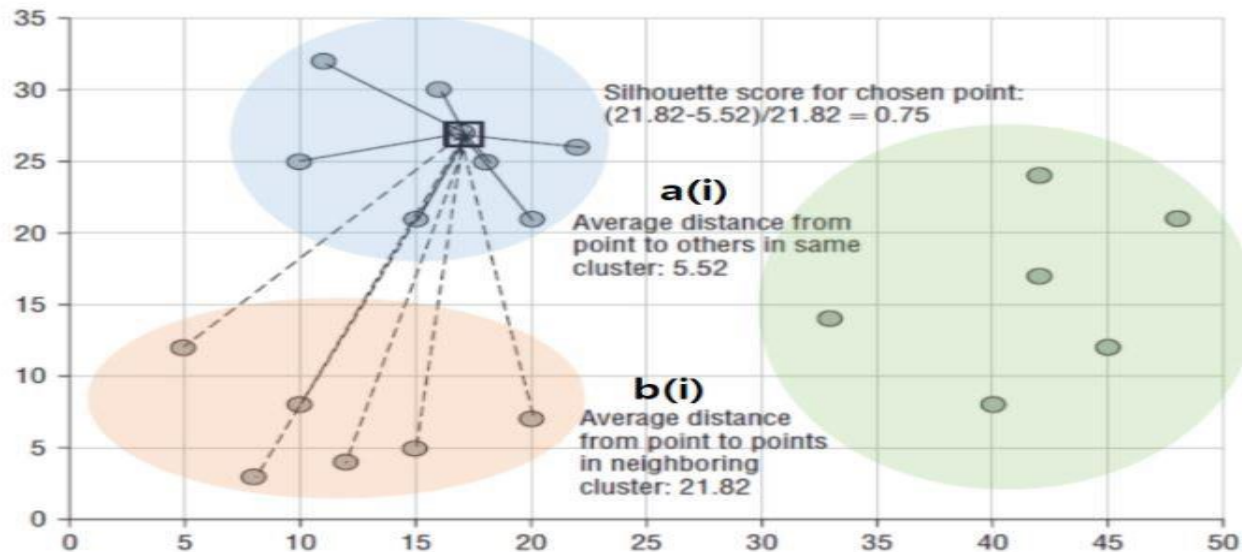
묶어주는 방법!

Ex) 강아지, 고양이, 판다

최적의 군집개수를 정해야 한다!

## Deciding the Number of Clusters

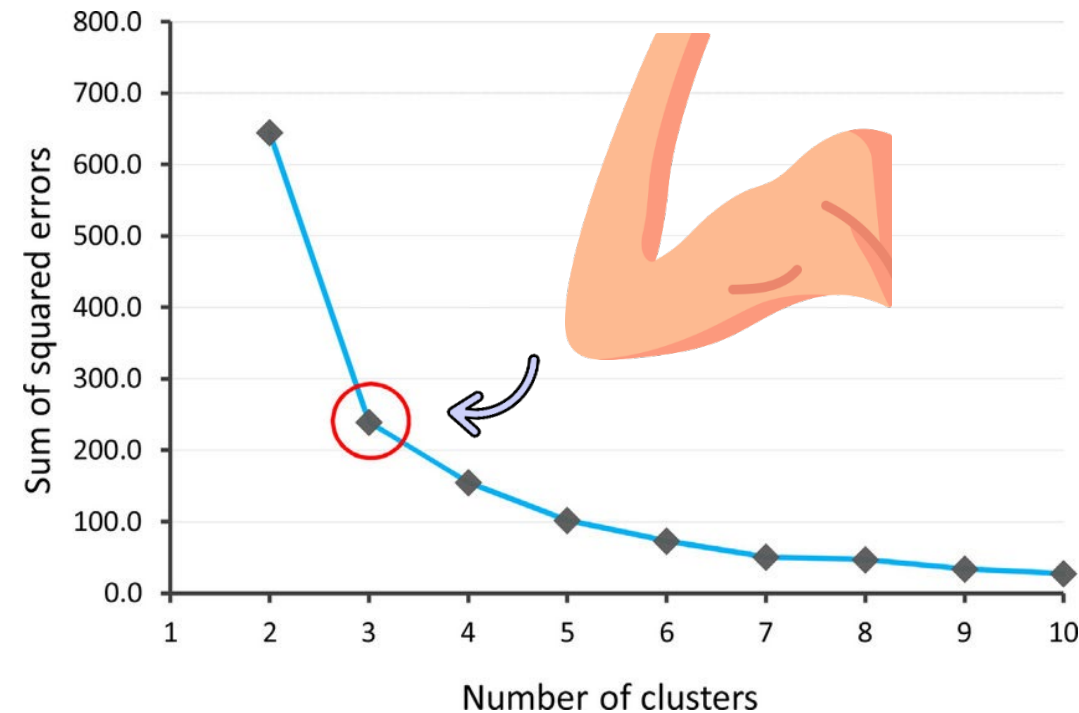
### Silhouette Method



군집내 높은 응집도  
군집간 높은 분리도

## Deciding the Number of Clusters

### Elbow Method



### Elbow Point

그래프에서 오차의 총합이 급격히 감소하는 지점.

이 지점에서 적절한 클러스터 개수를 결정!

## K-means Clustering

평균값을 이용한 군집화



클러스터 내의 분산은 작게, 클러스터 간의 분산은 최대

$$X = C_1 \cup C_2 \cdots C_k, \quad C_i \cap C_j = \emptyset$$

$$\operatorname{argmin}_c \sum_{i=1}^K \|x_j - c_i\|^2$$



## K-Medoids Clustering

중앙값을 이용한 군집화

- 이상치로부터 받는 영향이 적다
- 클러스터에서 대표적인 객체를 찾는 것이 쉬움
- 해석이 용이
- 처음에 중심점을 설정할 때 **랜덤성의 한계** 존재

Multiple random initialization을 통해  
최적의 초기화 조건 찾기

K-Means Clustering

K-Medoids Clustering

## Association Rule Discovery

### A priori 알고리즘

: 등장 순서와 관련이 있는 조건문을 작성해 보는 것  
 Ex) if '아이템1이 구매되었다면 아이템2도 구매될 것이다.'

[ 다섯 개의 물품 Milk, Coke, Pepsi, Beer, Juice의 구매 여부를 각각 m, c, p, b, j ]

거래 순번 (transaction number)	구매 물품
T1	m, c, b
T2	m, c, b, n
T3	m, c, b, n
T4	m, c, b, j
T5	m, p, b
T6	m, c, b, j
T7	c, b, j
T8	b, c

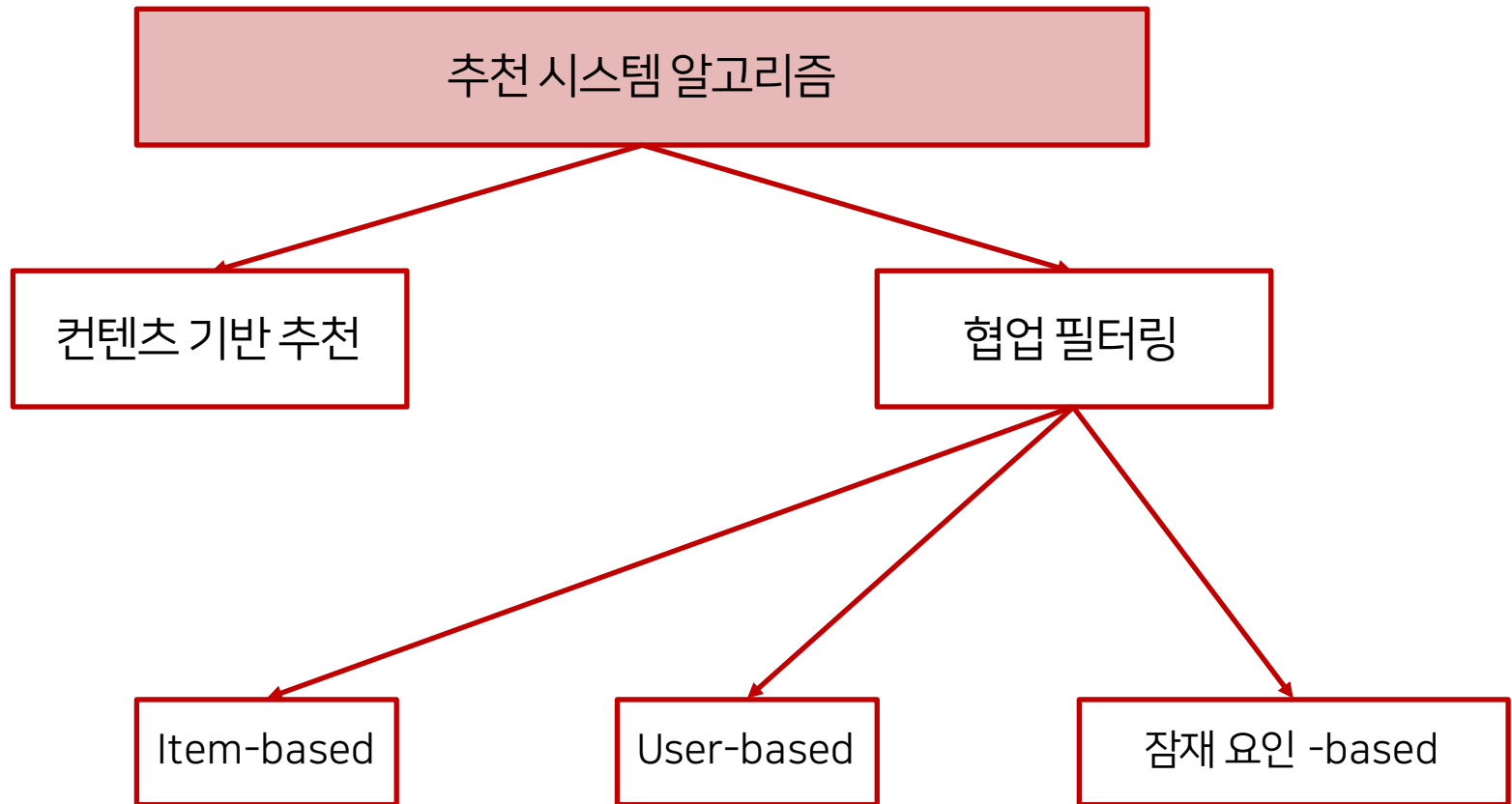
### Frequent Itemset Generation

: 어떤 아이템들끼리 같이 많이 구입되는가?

### Rule Generation

: 위의 결과를 바탕으로 어떤 규칙을 추론해낼 수 있는가?

## Recommender System



## 컨텐츠 기반 추천

사용자가 과거에 소비한 컨텐츠의 특성을 분석해  
이와 유사한 특성을 지닌 컨텐츠를 추천

데이터 획득

컨텐츠 분석

유저  
프로필 파악

유사  
아이템 선택

추천  
리스트 생성

예시를 하나 들어보자..!

## 아이템 기반 협업 필터링

테마 토픽원들의 취미 매핑



진모: movie, cooking

서영: movie, hiking, biking

지현이는 biking, cooking을 좋아해,,

지현: biking, cooking

movie: 진모, 서영

biking: 진모, 지현

cooking: 서영, 지현

Score(movie) = 유사도(movie, biking) + 유사도(movie, cooking)

Score(hiking) = 유사도(hiking, biking) + 유사도(hiking, cooking)

$$\text{유사도 (movie, cooking)} = \frac{\text{진모}}{\text{진모} + \text{서영} + \text{지현}}$$

## 사용자 기반 협업 필터링

먼저, 코사인 유사도 계산을 통해  
각 사용자가 얼마나 비슷한 지 기술해야 한다.

유사도	유나	서영	지현	진모	재성	남택
재성	0.98	0.63	0.99	0.85	1	0.98

1. 가장 유사한 몇 명의 점수만을 사용

OR

2. 전체를 대상으로 유사도 기반의 **weighted sum** 값을 예측 점수로 사용

## 잠재 요인 협업 필터링

각 요인에 대한  
사용자의 선호도를 열거할 수 있다면

어떤 요인들이 얼마나 반영되어야 할지  
직관적으로 파악할 수 있다.

In general, how much do you like watching movies from the following genres?

	Really dislike	Dislike	Neither like nor dislike	Like	Really like	Not sure of genre definition
Action	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Adventure	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Animation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>
Comedy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>
Crime/Gangster	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Documentary	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Drama	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fantasy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Film-Noir	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Foreign	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Horror	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Indie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Musical	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Mystery	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

**BUT, 현실적인 어려움!**

겉으로 보이지 않는 **잠재적 요인들(latent factor)**을 바탕으로  
추천 평점 예측 작업을 수행하는 것이 더욱 선호되곤 한다.