

데이터마이닝팀

4팀

황유나
문서영
김지현
위재성
이진모

CONTENTS

1. K-Nearest Neighbors Algorithm

2. Cluster Analysis

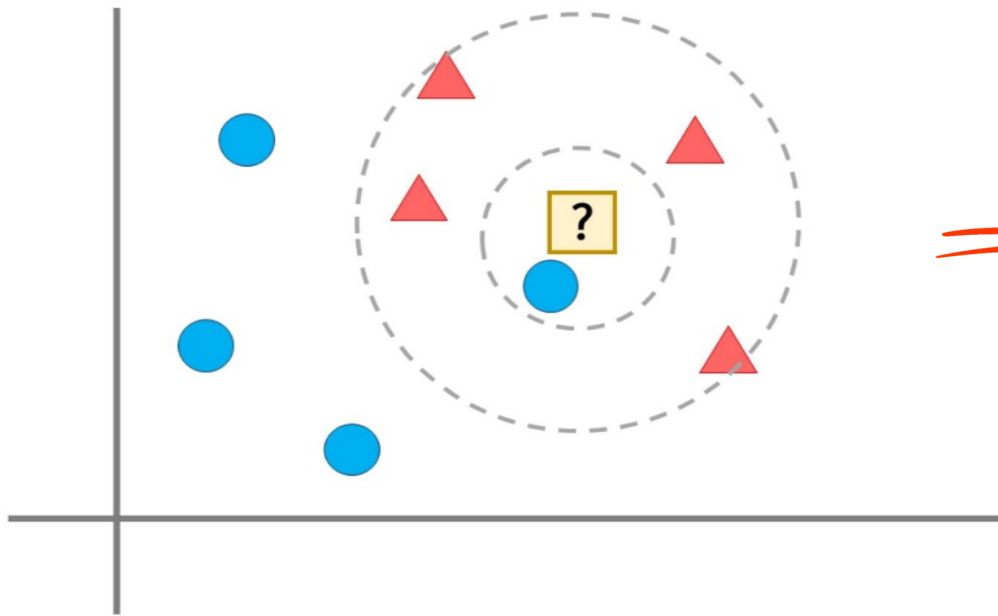
3. Data Mining Applications

1

K-NN Algorithm

Algorithm Fundamentals

K-Nearest Neighbors Algorithm



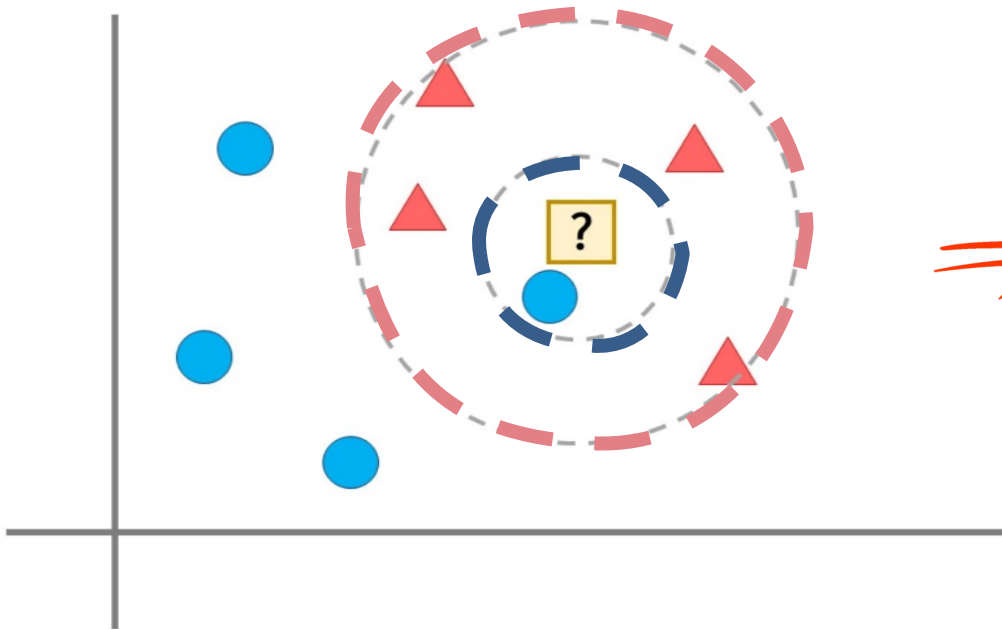
지도학습(Supervised learning)



k개의 이웃하는 기존 관측치의
최빈값을 따른다!

Algorithm Fundamentals

K-Nearest Neighbors Algorithm



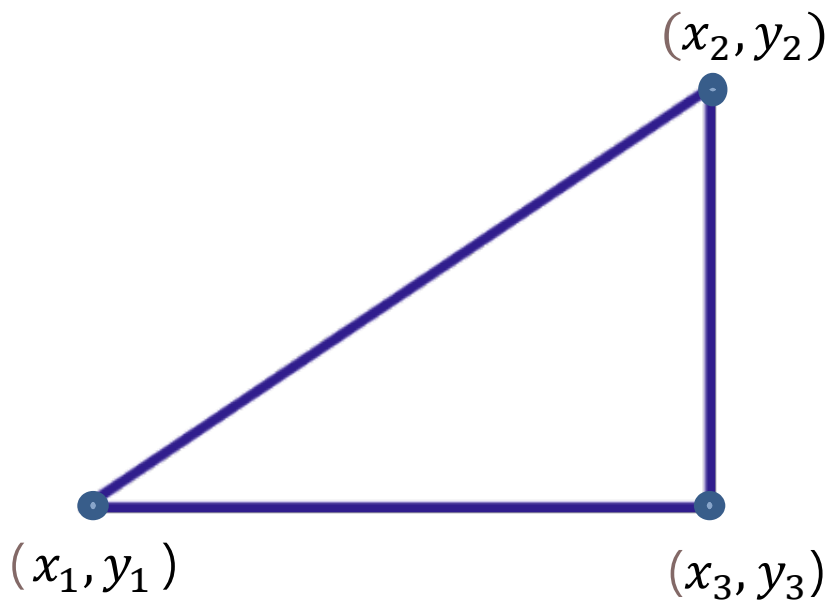
관측치 간 **거리정보**를 통해 새 관측치의
범주를 결정!



유클리드 거리
(Euclidean Distance)

Algorithm Fundamentals

유클리드 거리(Euclidean Distance)

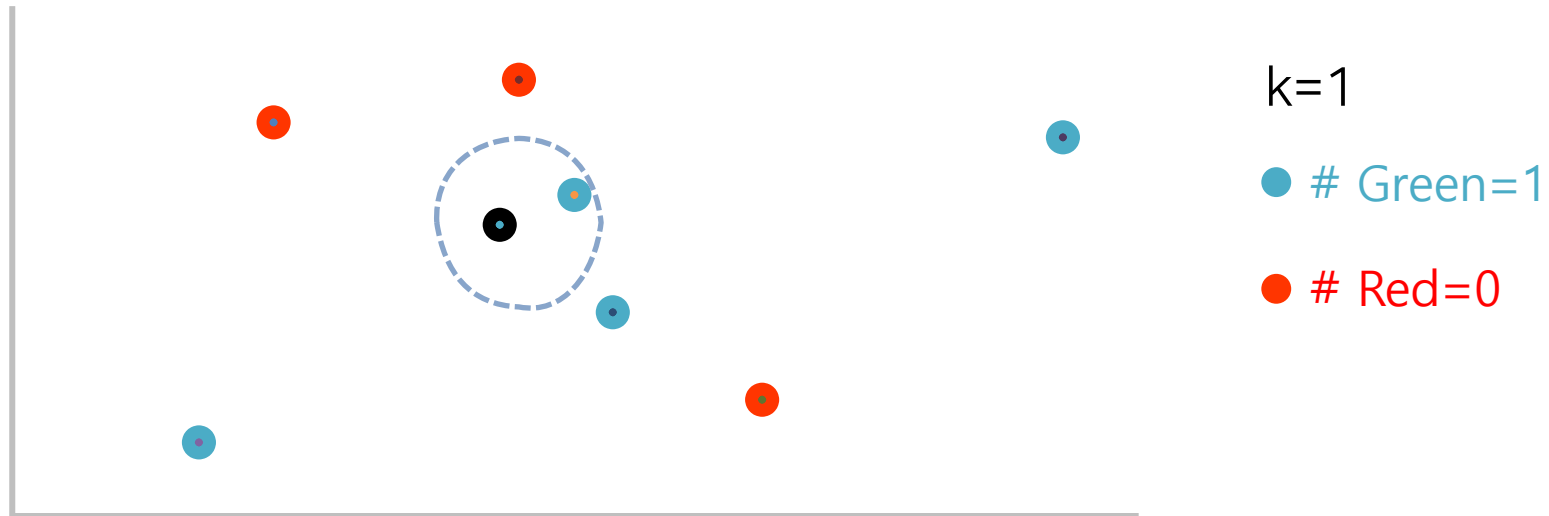


$$d(x, y) = \sqrt{\sum_{j=1}^m |x_j - y_j|^2}$$

피타고라스 공식을 떠올리면 쉽게 이해 가능!

Algorithm Fundamentals

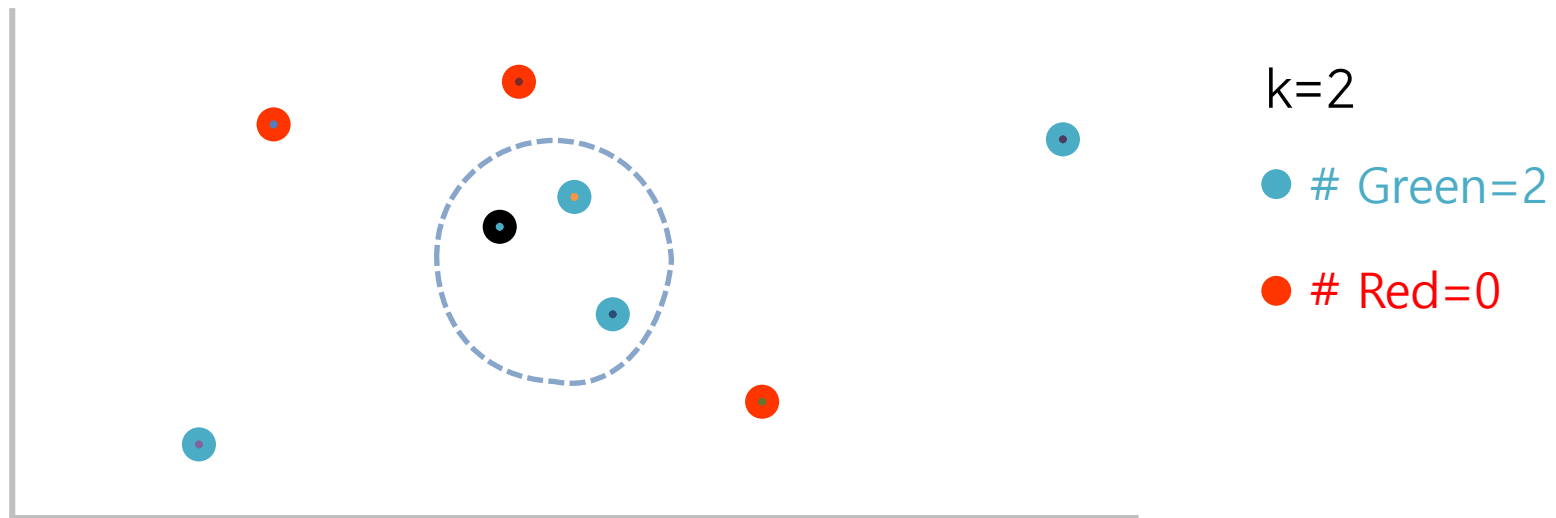
K-Nearest Neighbors Algorithm



검은 점은 가장 가까이 이웃하는 점의
범주인 초록색으로 분류!

Algorithm Fundamentals

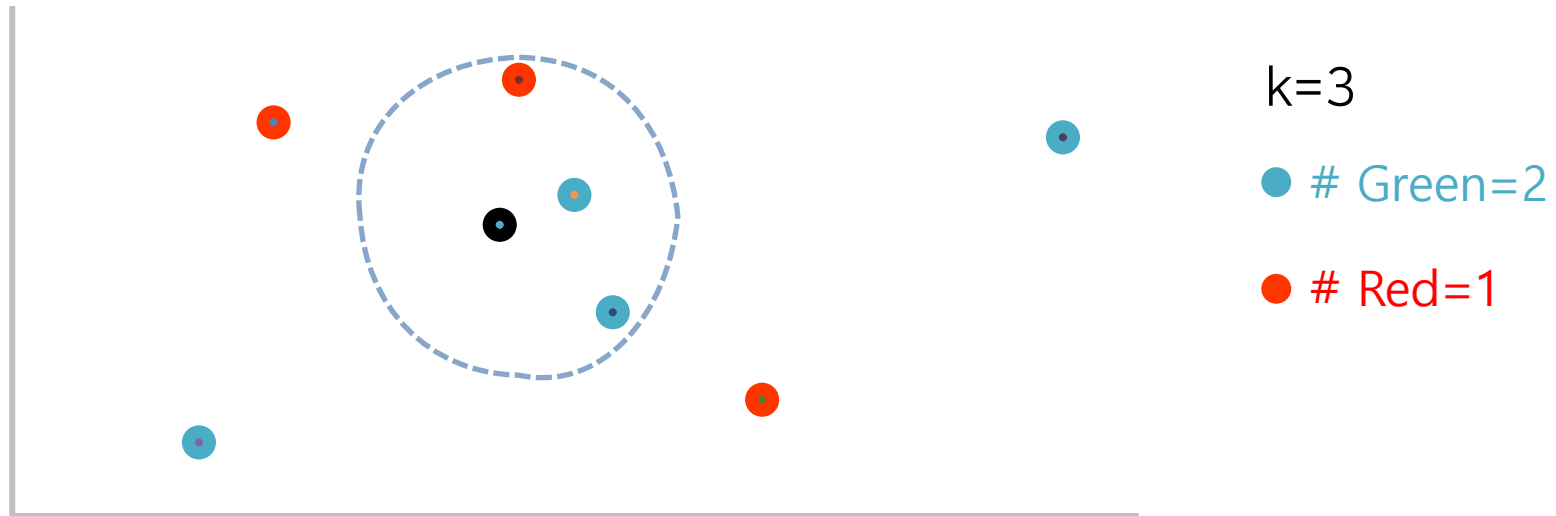
K-Nearest Neighbors Algorithm



검은 점은 가장 가까이 이웃하는 두 점의
범주 중 최빈값인 초록색으로 분류!

Algorithm Fundamentals

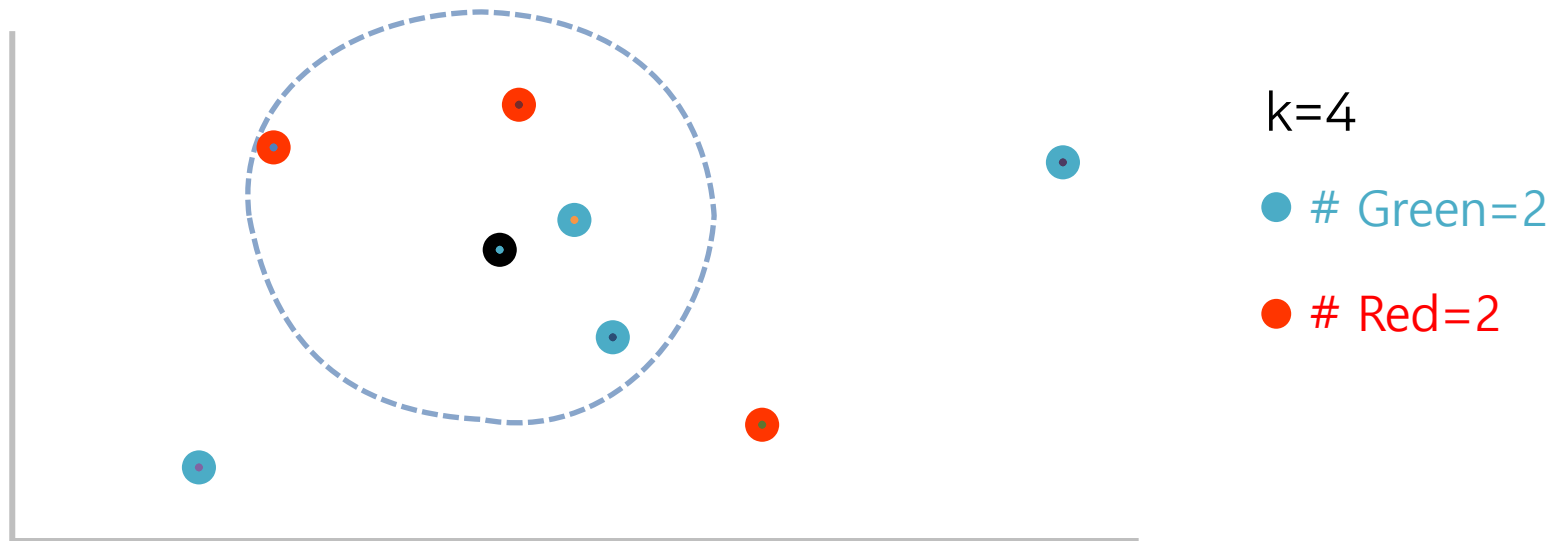
K-Nearest Neighbors Algorithm



검은 점은 가장 가까이 이웃하는 세 점의
범주 중 최빈값인 초록색으로 분류!

Algorithm Fundamentals

K-Nearest Neighbors Algorithm



이웃하는 점들의 최빈값을 결정 못함!



K값은 홀수로 권장!

Algorithm Fundamentals

K-Nearest Neighbors Algorithm

K-NN 모델에서는 **K값**에 따라 새로운 데이터의 예측값이 변한다!

$K=4$
● # Green=2

● # Red=2

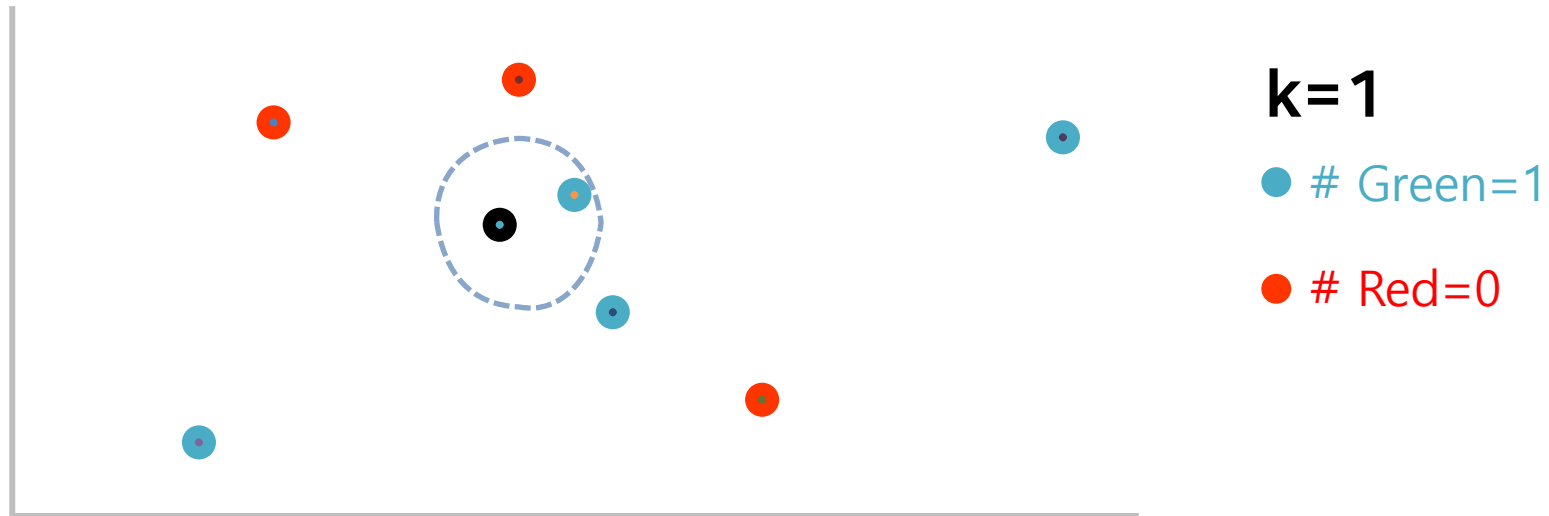
- K를 어떤 값으로 지정하느냐에 따라 **overfitting**과 **underfitting**의 문제가 있을 수 있다.
이웃하는 점들의 최근값을 결정 못함!



K값은 홀수로 권장!

Algorithm Fundamentals

K-Nearest Neighbors Algorithm



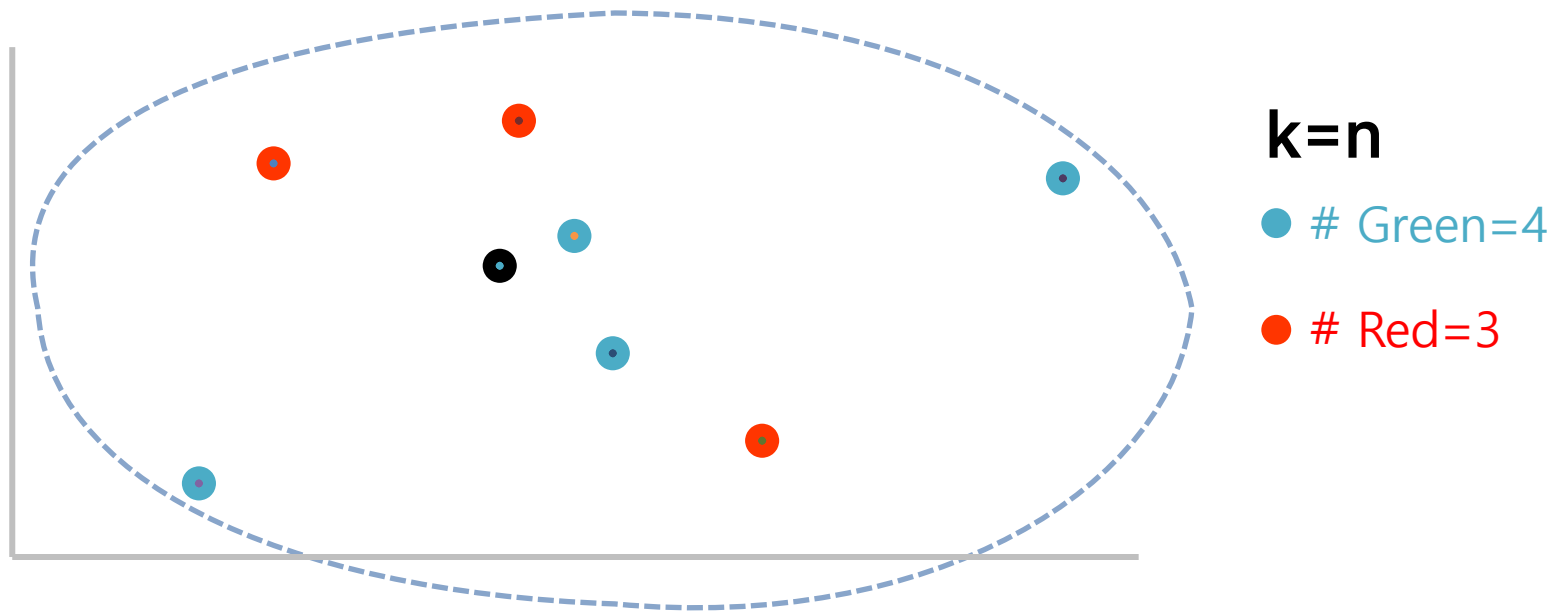
데이터 하나하나를
동일한 중요도로 받아들임!



Overfitting의 가능성

Algorithm Fundamentals

K-Nearest Neighbors Algorithm



새로운 데이터가 기존
데이터의 majority class를 따름

➡ **Underfitting**의 가능성

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)

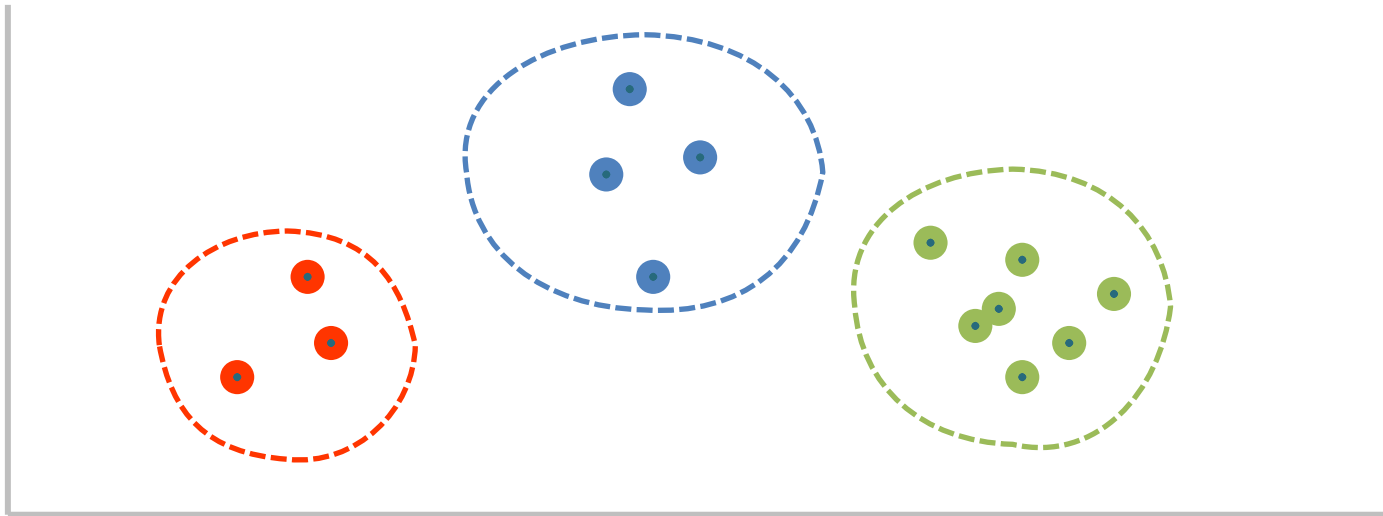
$r=2$  Euclidean Distance(유클리드거리)

2

Cluster Analysis

What is Clustering?

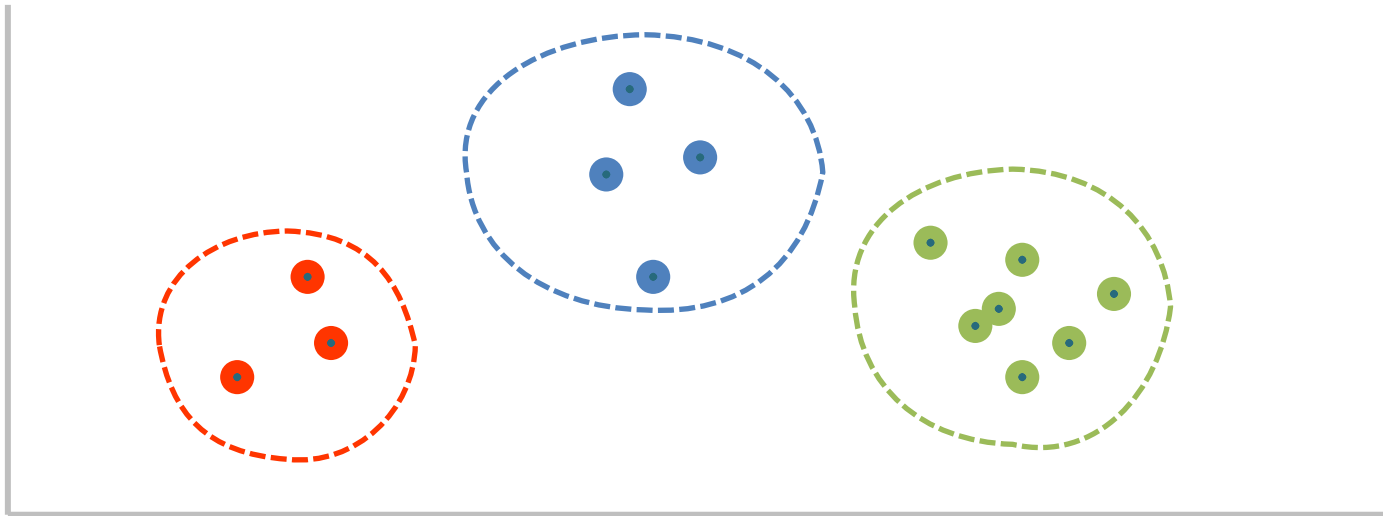
군집화(Clustering)



비지도학습
(Unsupervised Learning)

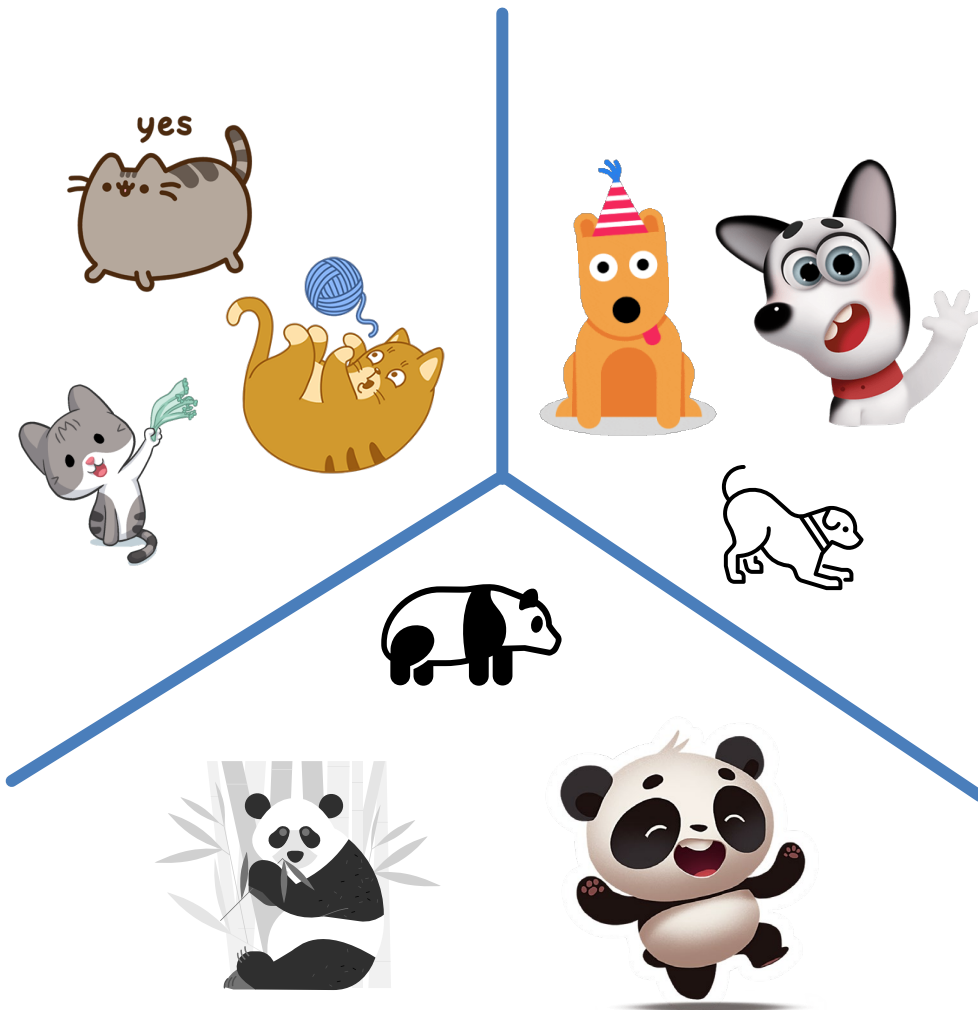
What is Clustering?

군집화(Clustering)



예측, 추론과 같은 명확한
Task가 아닌 '탐색적'인
분석상황에 적합.

What is Clustering?



속성이 유사한 관측치들끼리

지정한 Cluster의 개수로

묶어주는 방법!

Ex) 강아지, 고양이, 판다

What is Clustering?



속성이 유사한 관측치들끼리

지정한 Cluster의 개수로

좋은 클러스터링 모델을 설정하려면
어떠한 접근이 필요할까?

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

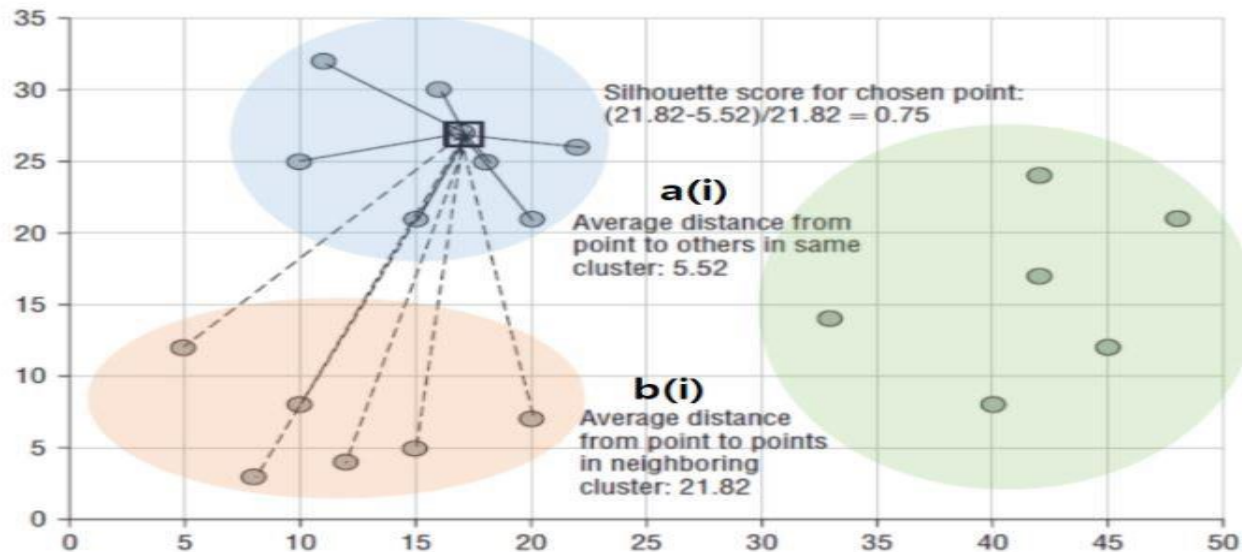
최적의 군집개수를 정하는 방법

1 Silhouette Method

2 Elbow Method

Deciding the Number of Clusters

Silhouette Method



군집내 높은 응집도
군집간 높은 분리도

Deciding the Number of Clusters

Silhouette Method

실루엣 계수

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

Deciding the Number of Clusters

Silhouette Method

실루엣 계수

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

$b(i)$

클러스터간 분리도(separation)

i번째 개체와 가장

가까운 클러스터내의 모든

데이터들과의 평균거리

Deciding the Number of Clusters

Silhouette Method

실루엣 계수

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

$b(i)$

i번째 개체가 속한 군집과 가장
가까운 군집 간 거리가 작을 수록,
즉 $b(i)$ 가 0에 가까울수록 최악!

➡ 전체 실루엣 계수 = -1

Deciding the Number of Clusters

Silhouette Method

실루엣 계수

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

$a(i)$

i번째 개체와 같은 군집에 속한

요소들 간의 평균거리,

클러스터 내의 응집도를 의미함.

Deciding the Number of Clusters

Silhouette Method

실루엣 계수

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

$a(i)$

i번째 개체가 속한 군집 내 개체들

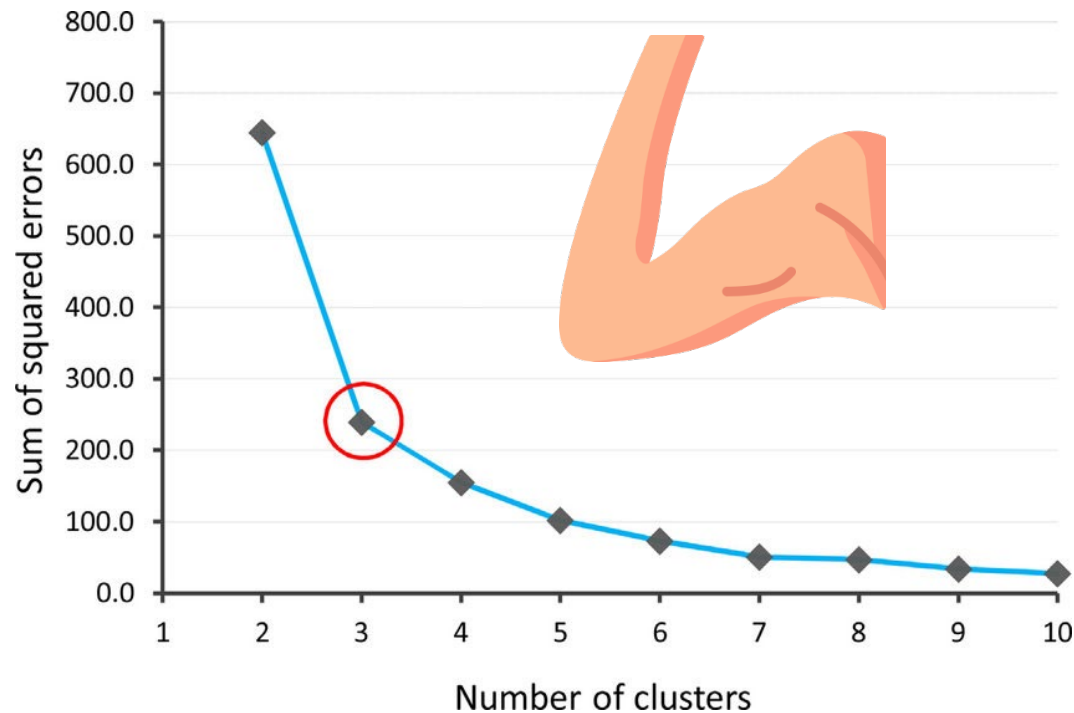
간의 거리가 가까울수록

즉, $a(i)$ 가 0에 가까울수록 최고!

➡ 전체 실루엣 계수 = 1

Deciding the Number of Clusters

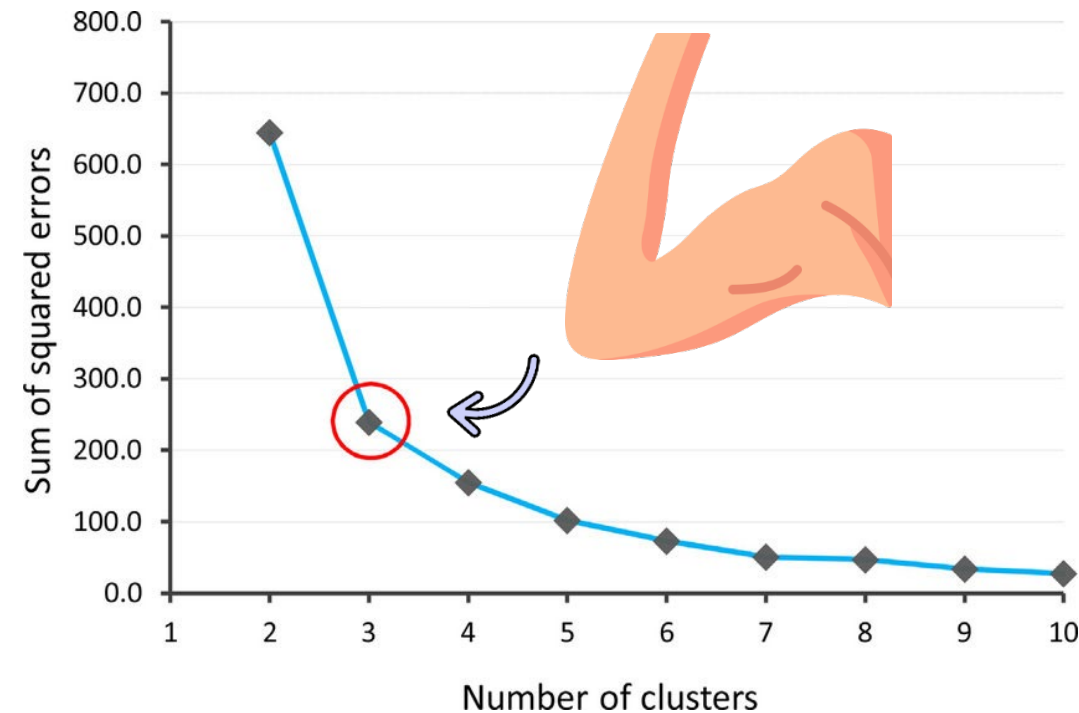
Elbow Method



클러스터 내 **오차 제곱합(RSS) 최소화!**

Deciding the Number of Clusters

Elbow Method

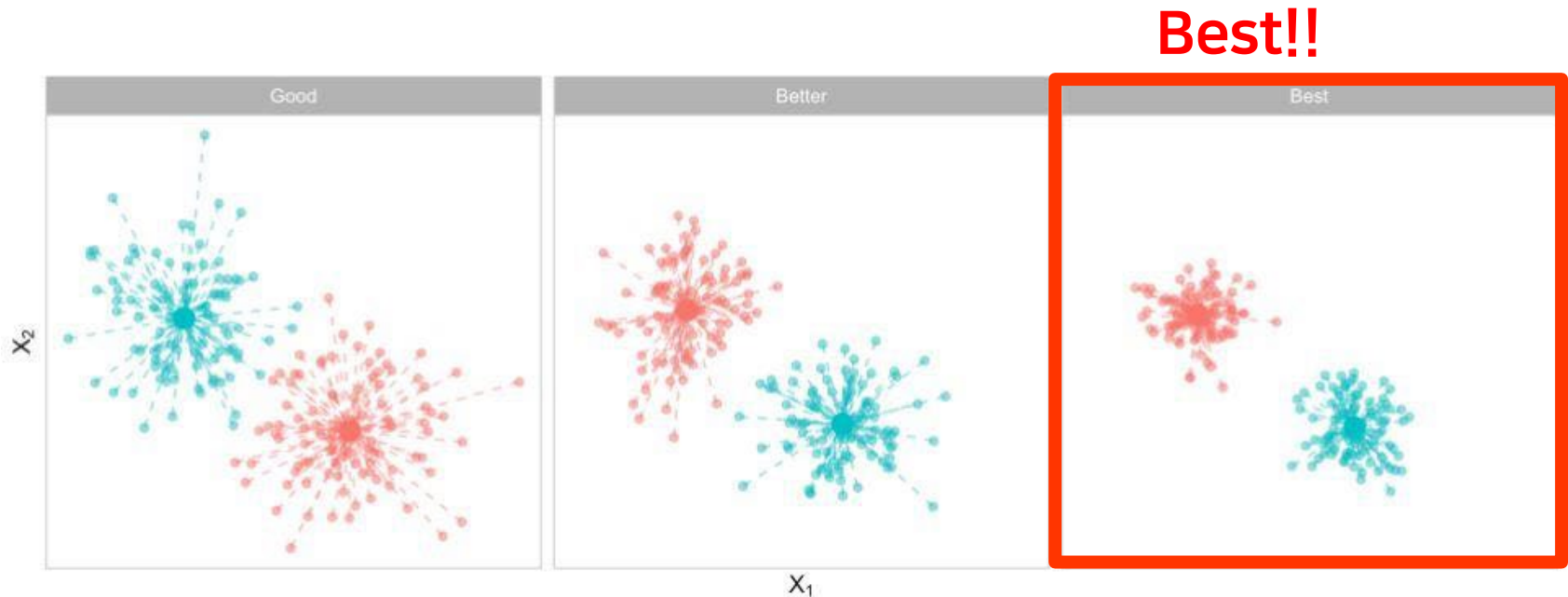


Elbow Point

그래프에서 오차의 총합이 급격히 감소하는 지점.

이 지점에서 적절한 클러스터 개수를 결정!

Deciding the Number of Clusters



compactness를 계산하고,
가장 좋은 클러스터링 결과를 찾을 수 있다.

K-means Clustering

평균값을 이용한 군집화



클러스터 내의 분산은 작게, 클러스터 간의 분산은 최대

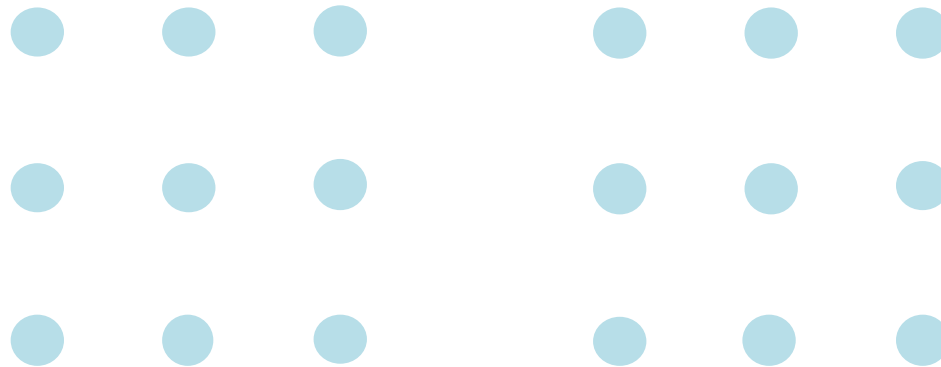
$$X = C_1 \cup C_2 \cdots C_k, \quad C_i \cap C_j = \emptyset$$

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

K-means Clustering

평균값을 이용한 군집화

* 군집 개수 $k=2$



K-means Clustering

평균값을 이용한 군집화

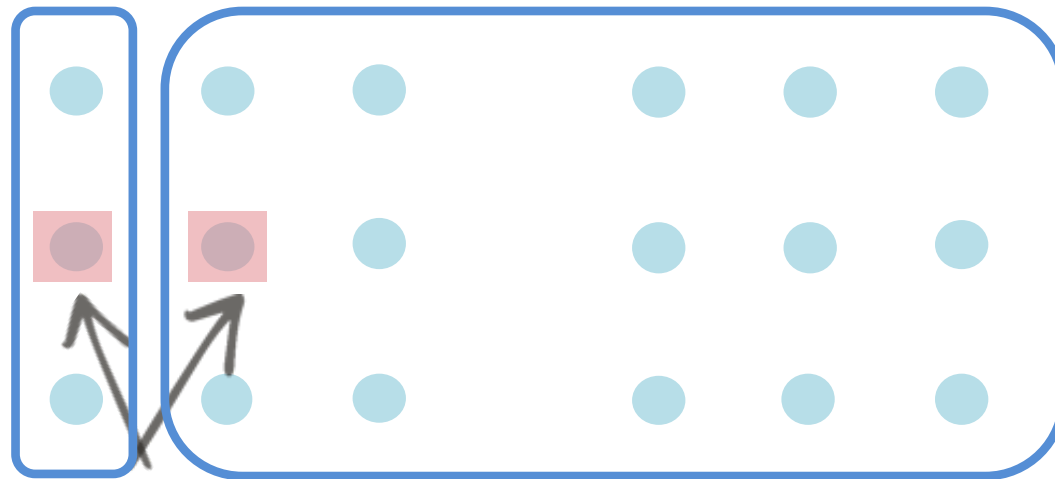
* 군집 개수 $k=2$



K-means Clustering

평균값을 이용한 군집화

* 군집 개수 $k=2$

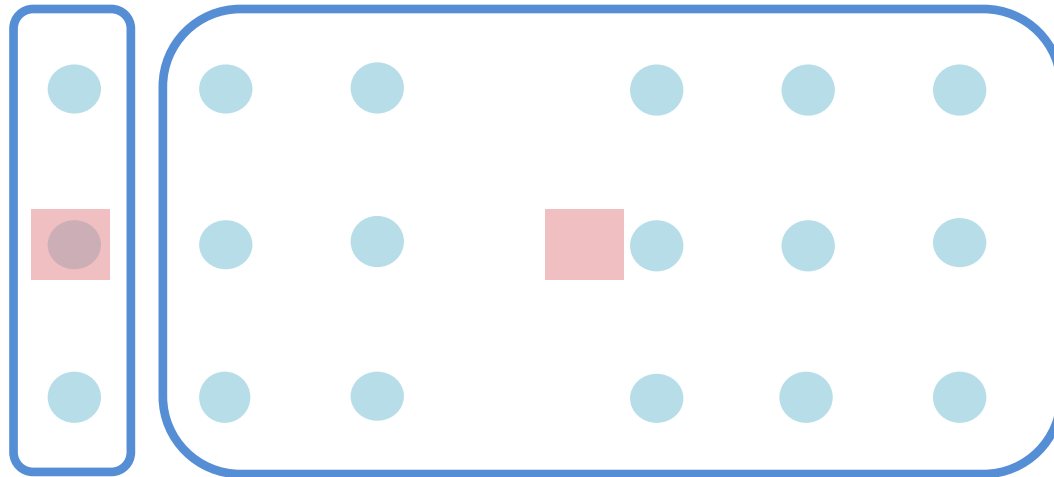


랜덤 설정된
군집의 중심

K-means Clustering

평균값을 이용한 군집화

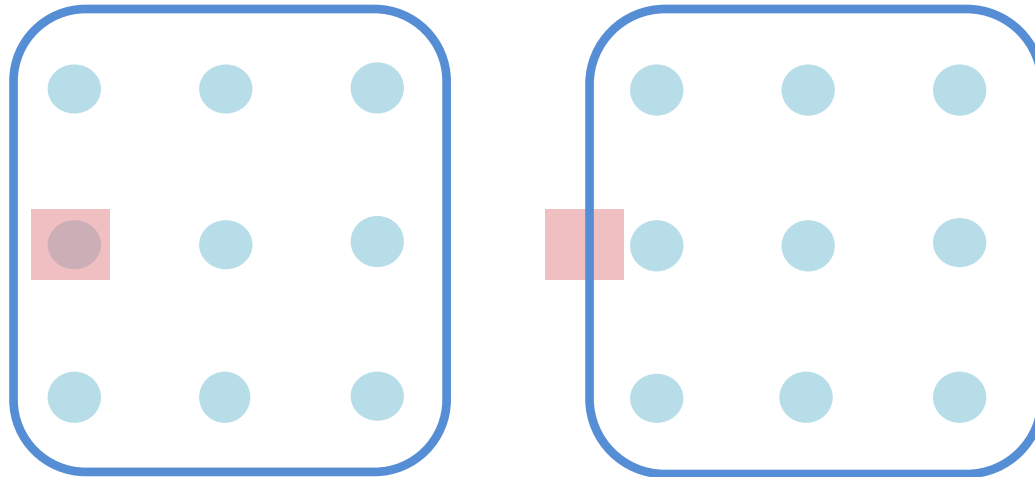
* 군집 개수 $k=2$



K-means Clustering

평균값을 이용한 군집화

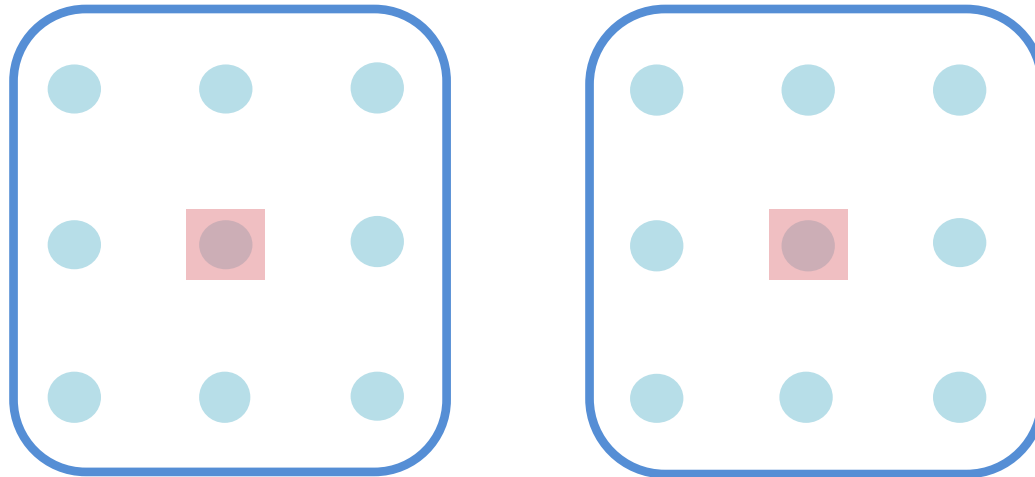
* 군집 개수 $k=2$



K-means Clustering

평균값을 이용한 군집화

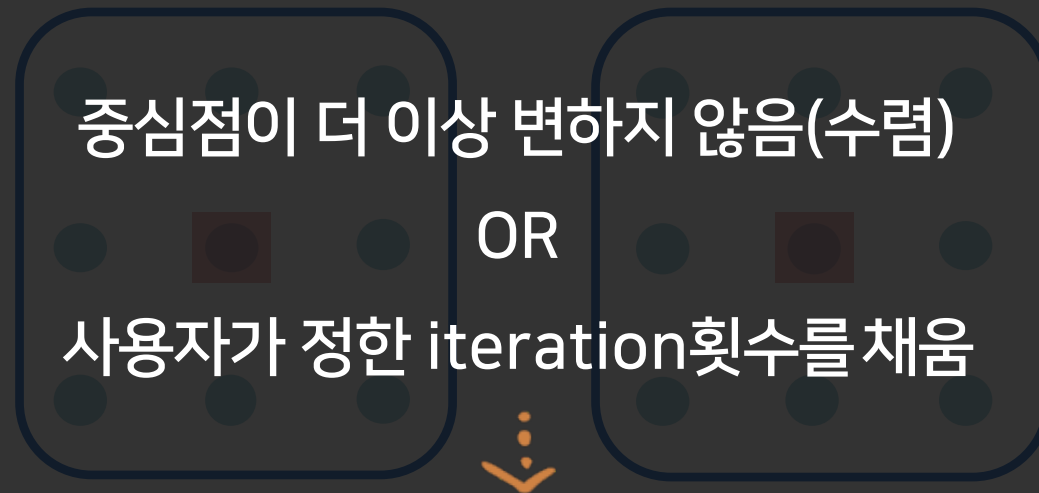
* 군집 개수 $k=2$



K-means Clustering

평균값을 이용한 군집화

* 군집 개수 $k=2$



학습 종료

K-means Clustering

평균값을 이용한 군집화

* 군집 개수 $k=2$

K-means

- Iteration 값에 따라 결과가 다르게 양산
- 처음에 중심점을 설정할 때 **랜덤성의 한계** 존재

K-means Clustering

평균값을 이용한 군집화

* 군집 개수 $k=2$

K-means

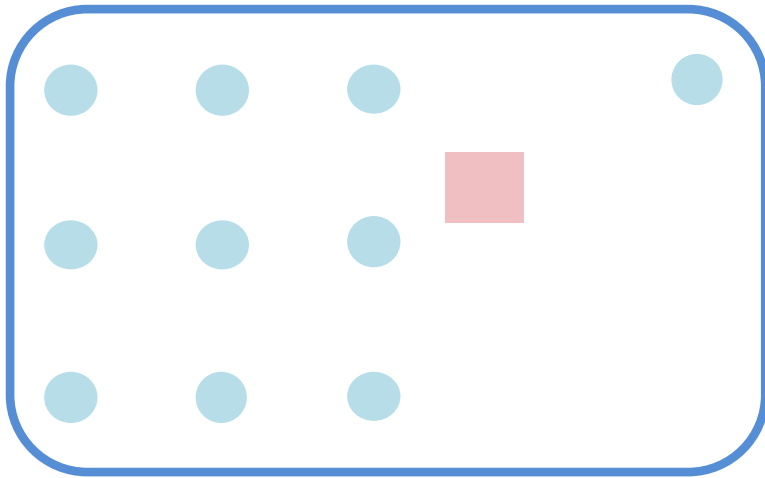
- Iteration 값에 따라 결과가 다르게 양산
- 처음에 중심점을 설정할 때 **랜덤성의 한계** 존재

Multiple random initialization을 통해
최적의 초기화 조건 찾기

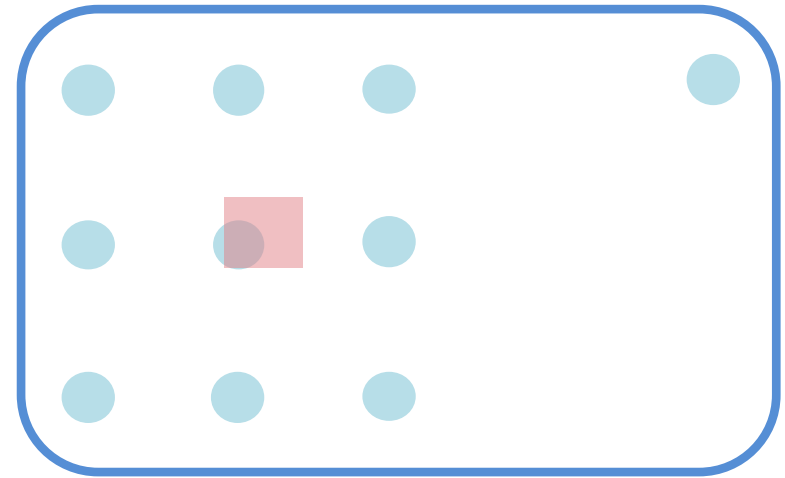
* Multiple random initialization:
For문을 통해 기존의 random initialization을
반복 수행해 최적의 초기화 조건을 찾는 방법

K-Medoids Clustering

중앙값을 이용한 군집화



K-Means Clustering



K-Medoids Clustering

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

3

Data Mining Application

NCT로 가득찼던 유나의 유튜브

	[NCT 마크] 피드백 찰떡 : 같이 알아듣고 바로 보... 지성아나라언제세워 0:32		[NCT/도영] 김도영 특유 : 말투야잇 김애용 0:11
	[NCT 해찬/유타] 엔시티 : 가 점점 정우화 되어가... 어떤말로도설명할수없어수만... 0:07		음방에서 웃음 못 참은 나 : 재민 나눈.야시즈니야 0:12
	[NCT] 엔뽕 차는 엔시티 : 녹음실 모음.zip 동구리 5:27		2020 MMA 마이클잭슨 : & 방탄 정디나 0:40
	[NCT 재현] 왕자님 : 지성아나라언제세워 0:07		설레서 뒤집 : 정재현이어렸을때영어학원에... 0:17
	[NCT] 자랑스러운 을시 : 티의 라이브 모음.zip... 엔시티에내집마련 4:20		[NCT 태용] 까딱하면 잡 : 허갈 것 같아서 말은 아... GREEN Heart 0:07
			[NCT 마크] 단짠의 정석 : 레귤러 이마크 마크느스즈췌

속보) 유나 마음에 GOT7 들어와

일편단심 NCT였던 그녀의 눈에
최근 GOT7이 들어오기 시작해 화제인데요
바뀐 그녀의 유튜브 알고리즘을 통해 알아보겠습니다.





GOT7 "ENCORE" OFFICIAL M/V
GOT7
조회수 1169만회 • 1개월 전



GOT7(갓세븐) "RUN AWAY" @ GOT7 ♥
I GOT7 6TH FAN MEETING
GOT7 ♣
조회수 108만회 • 3개월 전



GOT7(갓세븐) "Breath (넌 날 숨 쉬게 해)"
@ GOT7 ♥ I GOT7 6TH FAN MEETING
GOT7 ♣
조회수 69만회 • 3개월 전



[#빈센조] 혼자영 집에 망치를 든 괴한이?
결국 송중기 집에서 하룻밤 자게 된...
tvN D ENT ♣
조회수 21만회 • 1일 전



Adulting Diaries (in French!)
decluttering, shopping, cooking...
Leah's Fieldnotes ♣
조회수 11만회 • 4일 전



이하늬+뮤지컬 < 시카고 > 팀 - Roxie [열린 음악회/Open Concert] 20200426
KBS 골든케이팝
조회수 10만회 • 10개월 전

GOT7으로 가득 찬 유나의 유튜브



[BE ORIGINAL] GOT7 'POISON' (4K)
STUDIO CHOOM [스튜디오 춤] ♣
조회수 922만회 • 10개월 전



GOT7(갓세븐) "니가 부르는 나의 이름"
M/V
JYP Entertainment
조회수 1억회 • 1년 전



GOT7(갓세븐) "니가 하면" M/V
JYP Entertainment
조회수 1.9억회 • 5년 전

Association Rule Discovery

A priori 알고리즘

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

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, Juice가 있고 이를 각각 m, c, p, b, j]

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

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?

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, Juice가 있고 이를 각각 m, c, p, b, j]

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

< Frequent Itemset Generation >

Frequent?

: 사용자가 지정한 최소 지지도 조건
 (support threshold) 이상의 값을 취할 때,
 frequent하게 등장한다

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, Juice가 있고 이를 각각 m, c, p, b, j]

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

< Frequent Itemset Generation >

Frequent?

: 사용자가 지정한 최소 지지도 조건
 (support threshold) 이상의 값을 취할 때,
 frequent하게 등장한다

* threshold=3

{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

Association Rule Discovery

A priori 알고리즘

- Frequent itemset 찾기

Ex) if '아이템1이 구매되었다면 아이템2도 구매될 것이다.'

[다섯 개의 물품 Milk, Coke, Pepsi, Beer, Juice가 있고 이를 각각 m, c, p, b, j]

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

< Frequent Itemset Generation >

집합 구성 위해 **완전 탐색** (exhaustive search,
하나씩 다 들여다보며 특정 조건을 만족하는 요소만 취하는 방법)
실시하기엔 **연산의 복잡도가 높음**

* threshold=3

{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

Association Rule Discovery

A priori 알고리즘

- Frequent itemset 찾기

Ex) if '아이템1이 구매되었다면 아이템2도 구매될 것이다.'

[다섯 개의 물품 Milk, Coke, Pepsi, Beer, Juice가 있고 이를 각각 m, c, p, b, j]

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



< Frequent Itemset Generation >

$P(\text{특정한 조건 두가지 만족}) \leq P(\text{단일한 조건 만족})$

: 사용자가 지정한 최소 지지도 조건
(support threshold) 이상의 값을 취할 때,
frequent하게 등장한다

* threshold=3

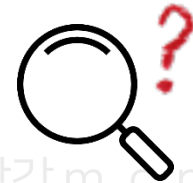
{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

Association Rule Discovery

A priori 알고리즘

- Frequent itemset 찾기

Ex) if '아이템1이 구매되었다면 아이템2도 구매될 것이다.'



[다섯 개의 물품 Milk, Coke, Pepsi, Beer, Juice가 있고 이를 각각 m, c, p, b, j]

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

$$P(\text{특정한 조건 두가지 만족}) \leq P(\text{단일한 조건 만족})$$

* 예시 {b,m}의 경우

$$P(m|b) = 0.6667 \leq P(b) = 0.75$$

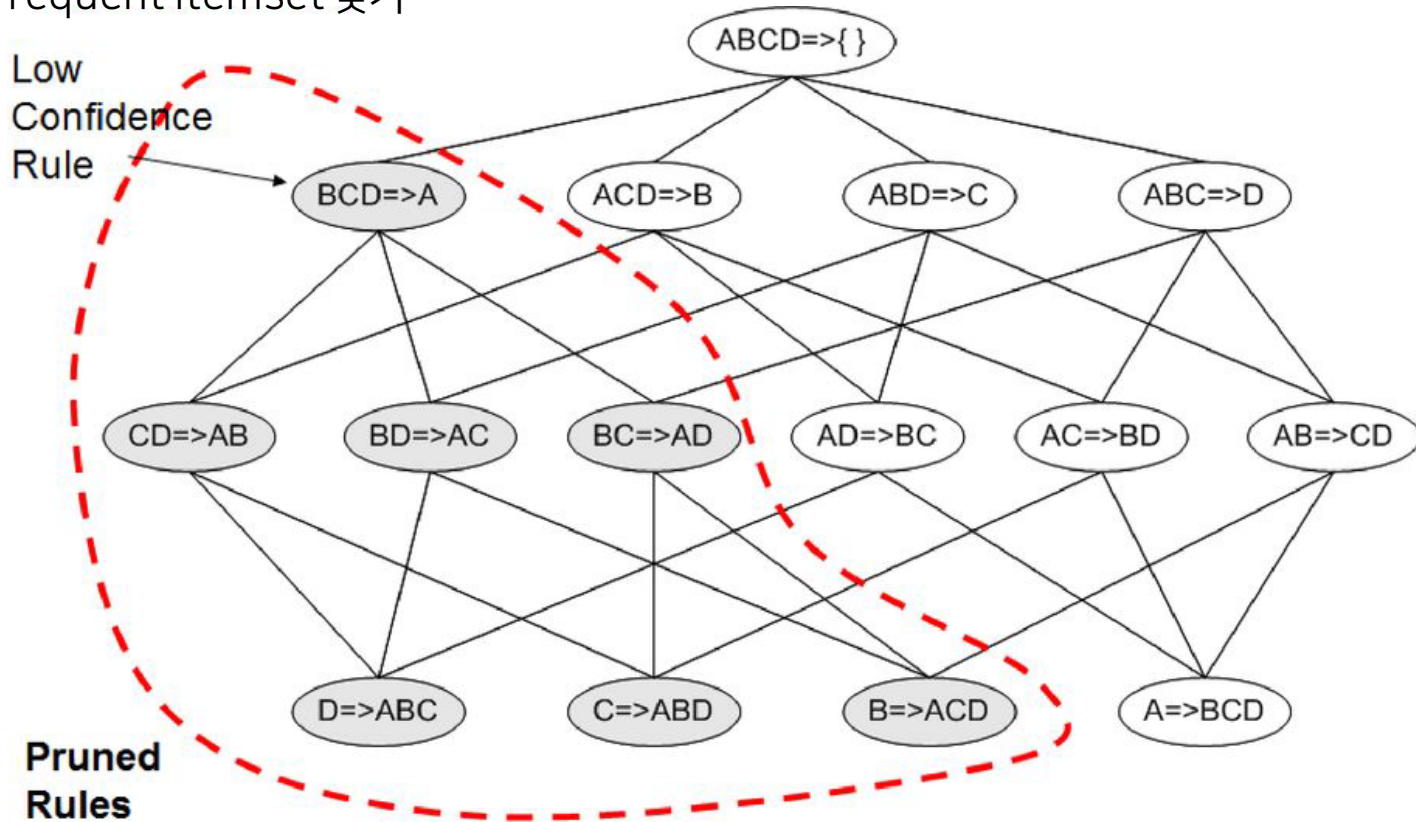
* threshold=3

{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

Association Rule Discovery

A priori 알고리즘

- Frequent itemset 찾기



A Priori 알고리즘의 계산 효율성

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?

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, Juice가 있고 이를 각각 m, c, p, b, j]

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

< Rule Generation >

Confidence

: rule을 유도하는 데 사용되는 지표로
 조건부 확률의 개념을 차용한 것

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, Juice가 있고 이를 각각 m, c, p, b, j]

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

< Rule Generation >

Confidence

: rule을 유도하는 데 사용되는 지표로
조건부 확률의 개념을 차용한 것

{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

* Confidence threshold = 0.75 (= $\frac{3}{4}$)

$$m \rightarrow b: C = \frac{4}{5}$$

$$b \rightarrow m: C = \frac{4}{6} \rightarrow \text{폐기}$$

$$b, c \rightarrow m: C = \frac{3}{5} \rightarrow \text{폐기}$$

$$b, m \rightarrow c: C = \frac{3}{4}$$

Association Rule Discovery

A priori 알고리즘

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

* Frequent threshold=3

{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

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

* Confidence threshold=0.75 ($= \frac{3}{4}$)

{b, m}, {m, c, b}

< Rule Generation >

Confidence

: rule을 유도하는데 사용되는 지표로
 조건부 확률의 개념을 차용한 것

{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

* Confidence threshold=0.75 ($= \frac{3}{4}$)

$$m \rightarrow b: C = \frac{4}{5}$$

$$b \rightarrow m: C = \frac{4}{6} \rightarrow \text{폐기}$$

$$b, c \rightarrow m: C = \frac{3}{5} \rightarrow \text{폐기}$$

$$b, m \rightarrow c: C = \frac{3}{4}$$

Association Rule Discovery

A priori 알고리즘

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

* threshold=3

{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

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

* Confidence threshold=0.75 ($= \frac{3}{4}$)

{b, m}, {m, c, b}



Interesting?

< Rule Generation >

Confidence

: rule을 유도하는데 사용되는 지표로
 조건부 확률의 개념을 차용한 것

{b, m}, {b, c}, {c, m}, {c, j}, {m, c, b}

* Confidence threshold=0.75 ($= \frac{3}{4}$)

$$m \rightarrow b: C = \frac{4}{5}$$

$$b \rightarrow m: C = \frac{4}{6} \rightarrow \text{폐기}$$

$$b, c \rightarrow m: C = \frac{3}{5} \rightarrow \text{폐기}$$

$$b, m \rightarrow c: C = \frac{3}{4}$$

Association Rule Discovery

A priori 알고리즘

: 등장 순서와 관련이 있는 조건문을 작성해 보는 것

Ex) if '아이템1'이 구매되었다면 '아이템2'도 구매될 것이다.

* threshold=3

? Interesting rule ?

$$Interest(I \rightarrow J) = conf(I \rightarrow J) - Pr[j]$$

≥ 0.5

Interesting?

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

< Rule Generation >

* Confidence threshold=0.75 (= $\frac{3}{4}$)

이것은 지표로 조건부 확률의 개념을 차용한 것

Confidence threshold=0.75 (= $\frac{3}{4}$)

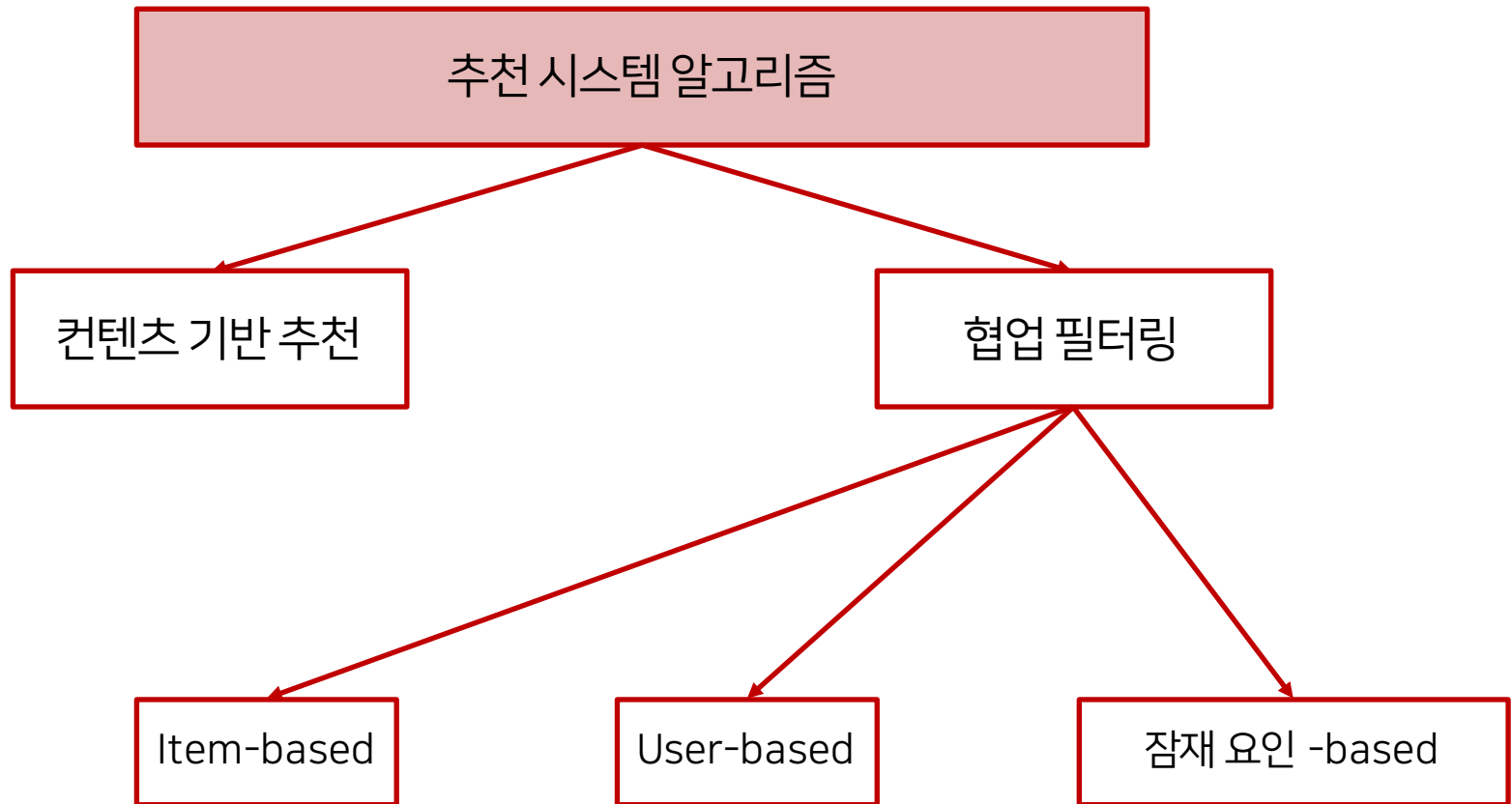
$$m \rightarrow b: C = \frac{4}{5}$$

$$b \rightarrow m: C = \frac{4}{6} \rightarrow \text{폐기}$$

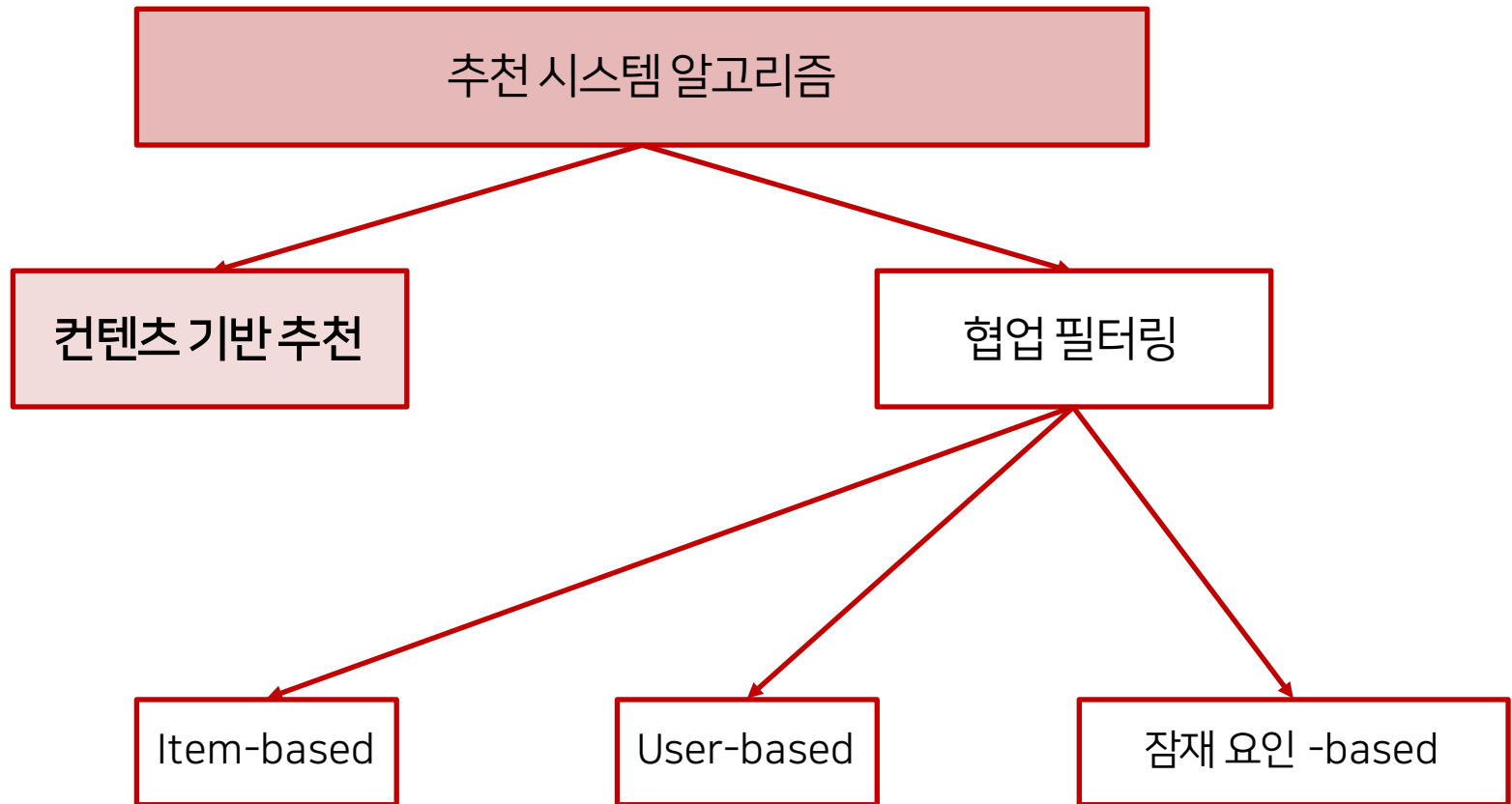
$$b, c \rightarrow m: C = \frac{3}{5} \rightarrow \text{폐기}$$

$$b, m \rightarrow c: C = \frac{3}{4}$$

Recommender System



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..!

컨텐츠 기반 추천



TMI:
유나는 슬기로운 의사생활을 세 번 봤다.

컨텐츠 기반 추천



“동일한 제작진이 연출한 드라마”

1888



"동일한 배우가 출연한 드라마"

컨텐츠 기반 추천

"동일한 제작진이 연출한 드라마 "

"동일한 배우가 출연한 영화 "



메타 데이터

컨텐츠 기반 추천

"동일한 제작진이 연출한 드라마 "



"동일한 배우가 출연한 영화 "

컨텐츠의 유사도는

어떻게 찾을까?



메타 데이터

TF-IDF

Term Frequency-Inverse Doc Frequency

-텍스트 데이터에서 feature를 뽑아내는 방법

- 1 많이 나오는 단어에 효과적으로 페널티 부과
- 2 문서를 대표할 수 있는 주요 단어들을 추출

TF-IDF

Term Frequency-Inverse Doc Frequency

-텍스트 데이터에서 feature를 뽑아내는 방법

$$TF - IDF = TF \cdot \log \frac{n_D}{1 + n_t}$$

n_D : 전체 문서 수

n_t : 단어 t가 나온 문서 수

TF-IDF

아이템의 특성과 사용자의 프로필 분석을 완료

» 이제 아이템들을 추천해줄 차례!!

TF-IDF를 통해 문서의 주요 단어와 그 빈도를 매핑



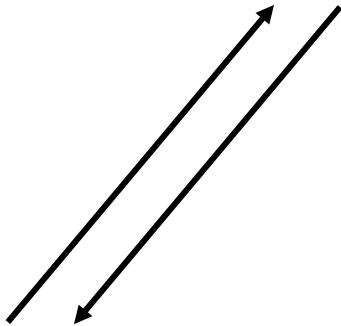
이를 컴퓨터가 이해할 수 있도록 인코딩



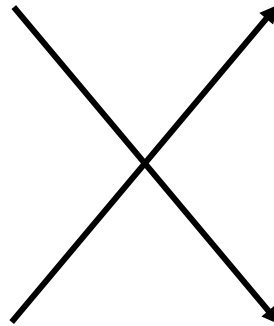
이들의 feature는 벡터의 형태를 띠

코사인 유사도

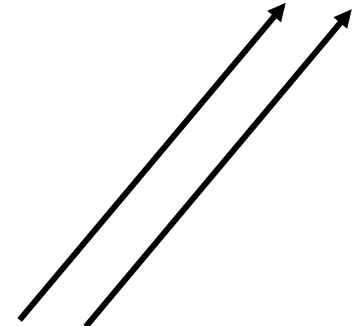
$$\text{Cosine similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$



코사인 유사도: -1



코사인 유사도: 0



코사인 유사도: 1

코사인 유사도

$$\text{Cosine similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}$$

컨텐츠 기반 필터링의 장점

Cold-Start*로 인한 데이터 부족 현상이 발생하지 않는다.

대중적이지 않은 취향을 지닌 소비자의 취향 저격도 가능하다.

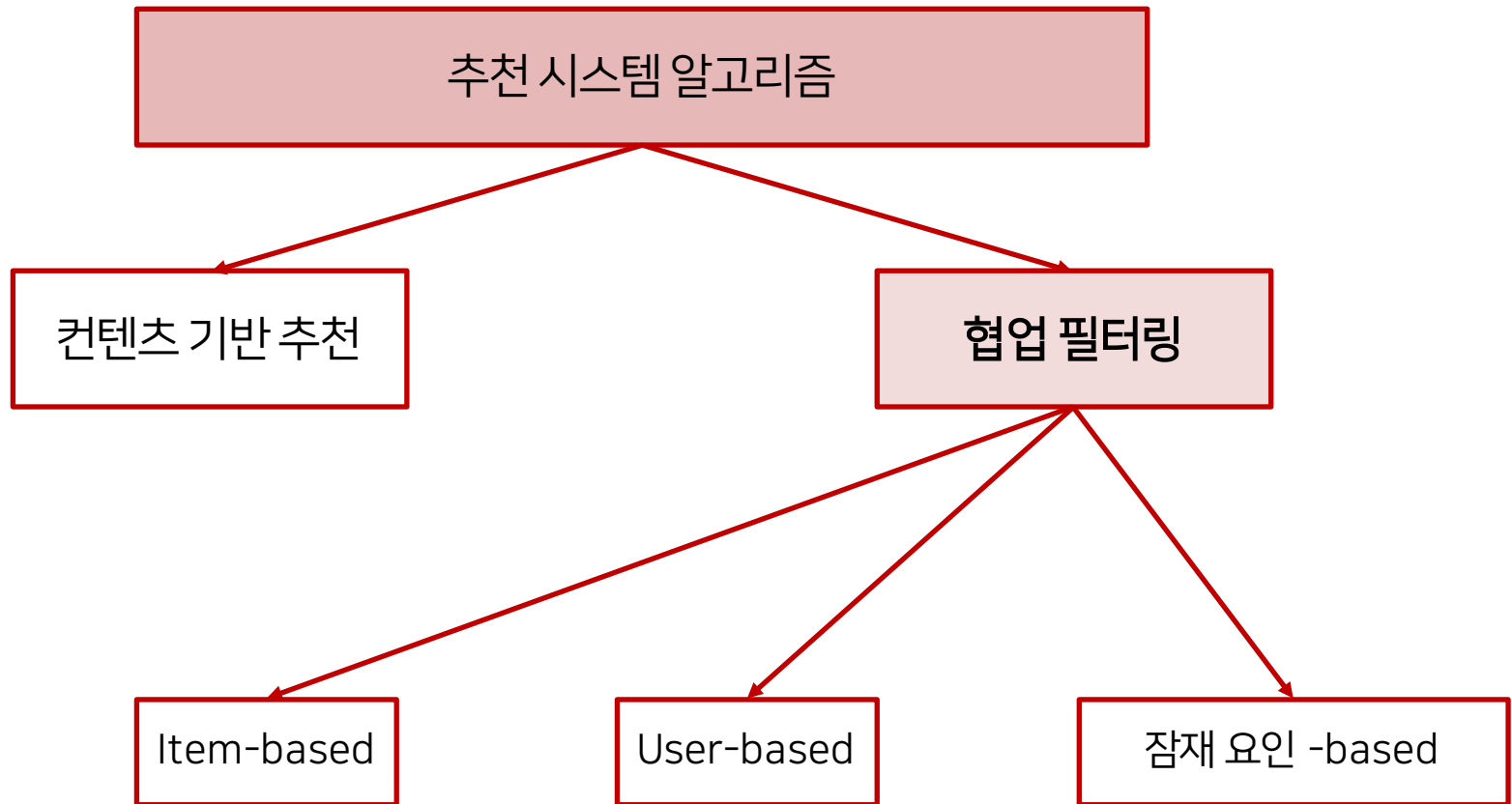
코사인 유사도: -1

코사인 유사도: 0

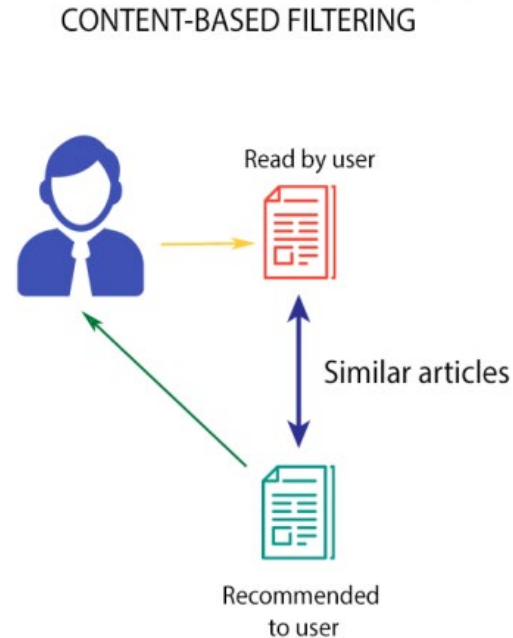
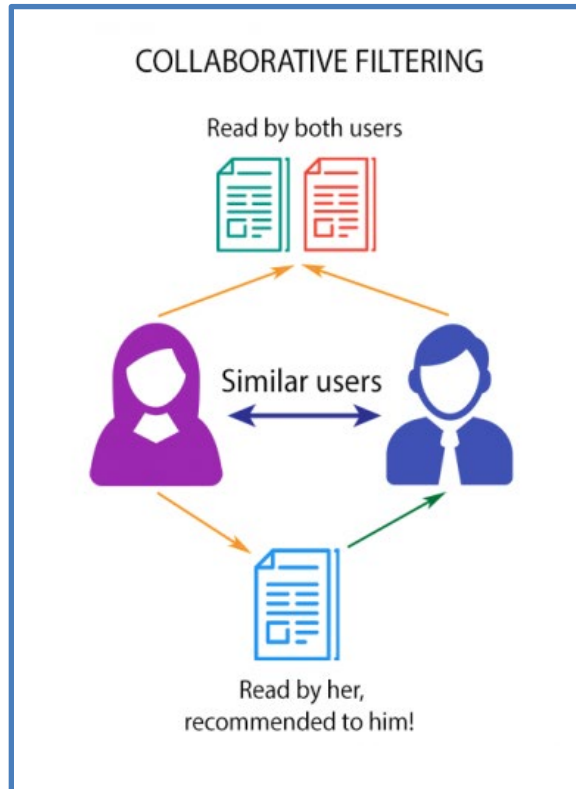
코사인 유사도: 1

*Cold-Start: 플랫폼 초기 단계의 데이터 부족

Recommender System



협업 필터링



사용자들, 아이템들 간의 협업을 통해 콘텐츠를 추천하는
협업 필터링

아이템 기반 협업 필터링

[데마 팀원들의 취미 매핑]

진모: movie, cooking
서영: movie, biking, hiking
지현: biking, cooking
재성: hiking



movie: 진모, 서영
cooking: 진모, 지현
biking: 서영, 지현
hiking: 서영, 재성

아이템 기반 협업 필터링

[데마 팀원들의 취미 매핑]

진모: movie, cooking

서영: movie, biking, hiking

지현: biking, cooking

재성: hiking



movie: 진모, 서영

cooking: 진모, 지현

biking: 서영, 지현

hiking: 서영, 재성

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

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)

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

아이템 기반 협업 필터링

헛갈려? 차이점 정리하자

컨텐츠 기반 필터링

아이템 프로필
+
사용자 프로필

아이템 기반 협업 필터링

아이템에 대한
구입내역/선호도/만족도

사용자 기반 협업 필터링

내가 평가한 아이템들에 대해
비슷한 점수를 부여한 사용자를 나와 '유사'하다고 판단



이들의 정보를 통해
새로운 아이템에 대한 나의 평가를 예측



철이 없었죠,,
예시도 안 보고 이해하려 했다는게,,,

사용자 기반 협업 필터링

	라라랜드	어벤져스	기생충	끝까지 간다	인터스텔라
유나	5	4	4	3	-
서영	1	0	1	-	4
지현	4	4	-	5	3
진모	-	2	1	4	3
재성	4		4	4	2
남택	4	2	3	-	1

사용자 기반 협업 필터링

	라라랜드	어벤져스	기생충	끝까지 간다	인터스텔라
유나	5	4	4	3	-
서영	1	0	1	-	4
지현	4	4	1	5	3
진모	5	2	1	4	3
재성	4		4	4	2
남택	4	2	3	-	1



과연 재성은
<어벤져스>에 평점을 몇 점으로 줄까?

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

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

사용자 기반 협업 필터링

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

평점	유나	서영	지현	진모	재성	남택
어벤져스	4	0	4	2		2



$$rating\ prediction = \frac{0.98 \times 4 + 0.63 \times 0 + 0.99 \times 4 + 0.85 \times 2 + 0.98 \times 2}{0.98 + 0.63 + 0.99 + 0.85 + 0.98} = 2.60$$

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"/>

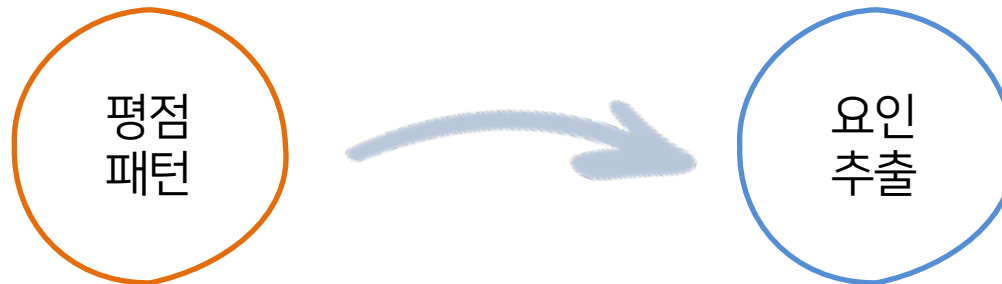
BUT, 현실적인 어려움!

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

잠재 요인 협업 필터링

잠재 요인 협업 필터링을 구현하는 대표적인 방법

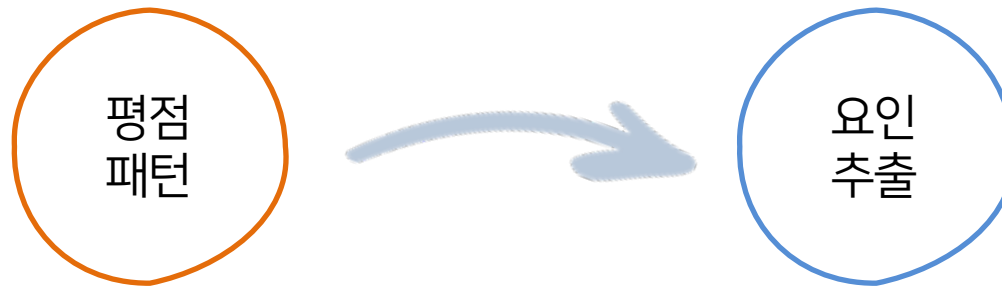
Matrix Factorization(MF)



잠재 요인 협업 필터링

잠재 요인 협업 필터링을 구현하는 대표적인 방법


Matrix Factorization(MF)



높은 정확도/높은 확장성/높은 유연성

잠재 요인 협업 필터링

MF 방법은 사용자의 선호도를 명확하게 알 수 없는
Implicit Feedback이 존재하는 경우에 최적화 되어있다.



 "sparse matrix"

$$\begin{pmatrix}
 1 & 0 & 6 & 0 & 0 & 0 \\
 0 & 2 & 0 & 0 & 0 & 8 \\
 0 & 0 & 0 & 9 & 0 & 0 \\
 0 & 0 & 6 & 0 & 7 & 0 \\
 0 & 0 & 0 & 0 & 2 & 4 \\
 4 & 0 & 0 & 0 & 5 & 0
 \end{pmatrix}$$

*sparse matrix ?
 평점 matrix에 평점이 있는 element보다
 그 값이 0으로 매겨진 element가 더 많다

잠재 요인 협업 필터링

MF 방법은
주어진 평점 데이터만을 이용해 모델링을 실시하지만,

그 안에 잠재된 사용자-아이템 사이의 상호작용을
적절히 포착하여 패턴을 찾아낸다.

잠재 요인 협업 필터링

MF 방법은

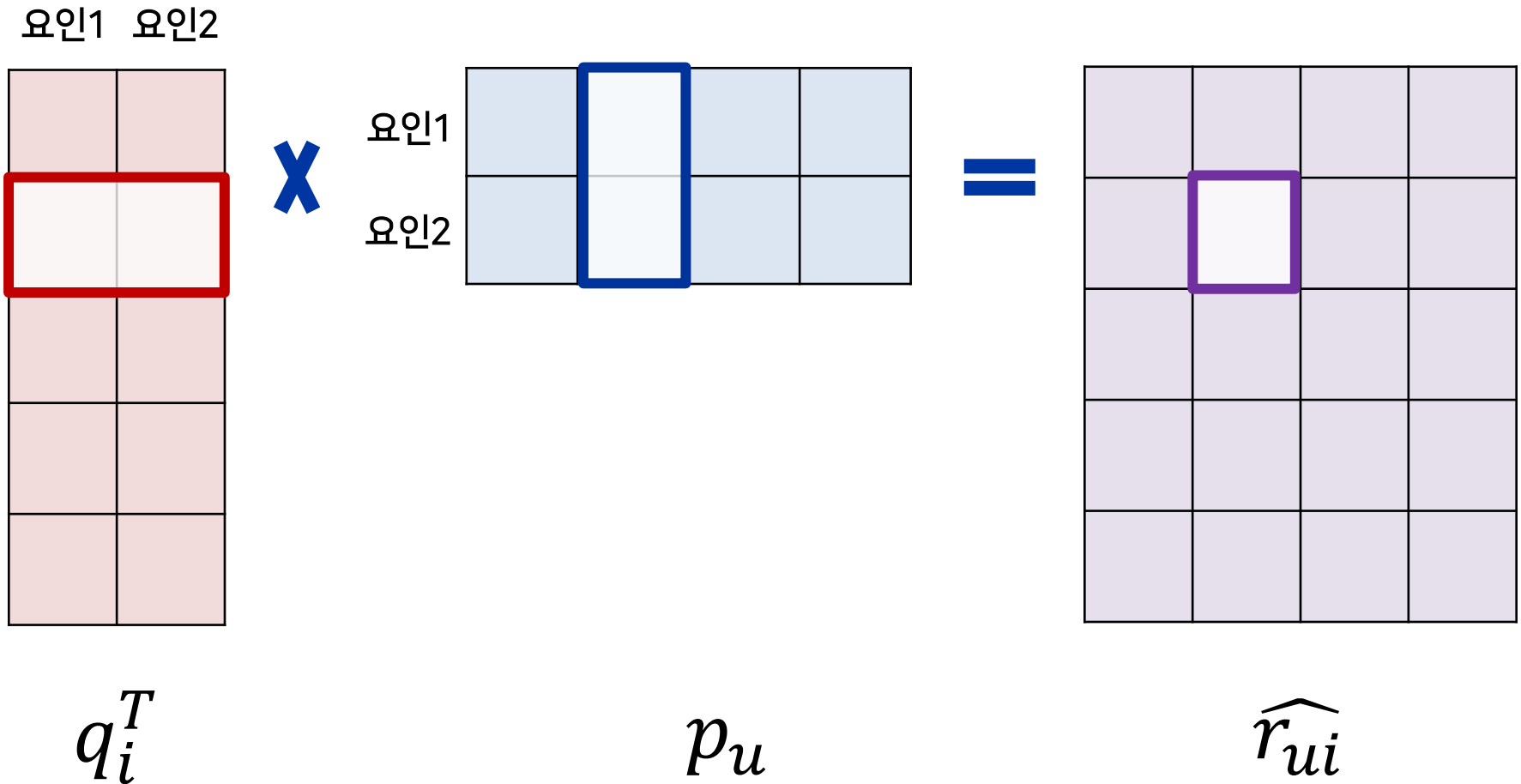
주어진 평점 데이터만을 이용해 모델링을 실시하지만,

그 안에 잠재된 사용자-아이템 사이의 상호작용을
적절히 포착하여 패턴을 찾아낸다.

$$\widehat{r}_{ui} = q_i^T p_u$$

where $q_i = \text{item}, p_u = \text{user}$

잠재 요인 협업 필터링



잠재 요인 협업 필터링

앞선 수식으로부터,

RSS를 줄이는 방식에 overfitting 방지를 위한 regularization term을 더해 손실함수를 작성할 수 있다.

$$\min_{q,p} \sum_{(u,i) \in K} (r_{ui} - q_i^T p_u)^2 + \lambda(\|q_i\|^2 + \|p_u\|^2)$$

RSS

잠재 요인 협업 필터링

앞선 수식으로부터,

RSS를 줄이는 방식에 overfitting 방지를 위한 regularization term을 더해 손실함수를 작성할 수 있다.

$$\min_{q,p} \sum_{(u,i) \in K} (r_{ui} - q_i^T p_u)^2 + \lambda (\|q_i\|^2 + \|p_u\|^2)$$



RSS

Regularization term

잠재 요인 협업 필터링

앞선 수식으로부터,

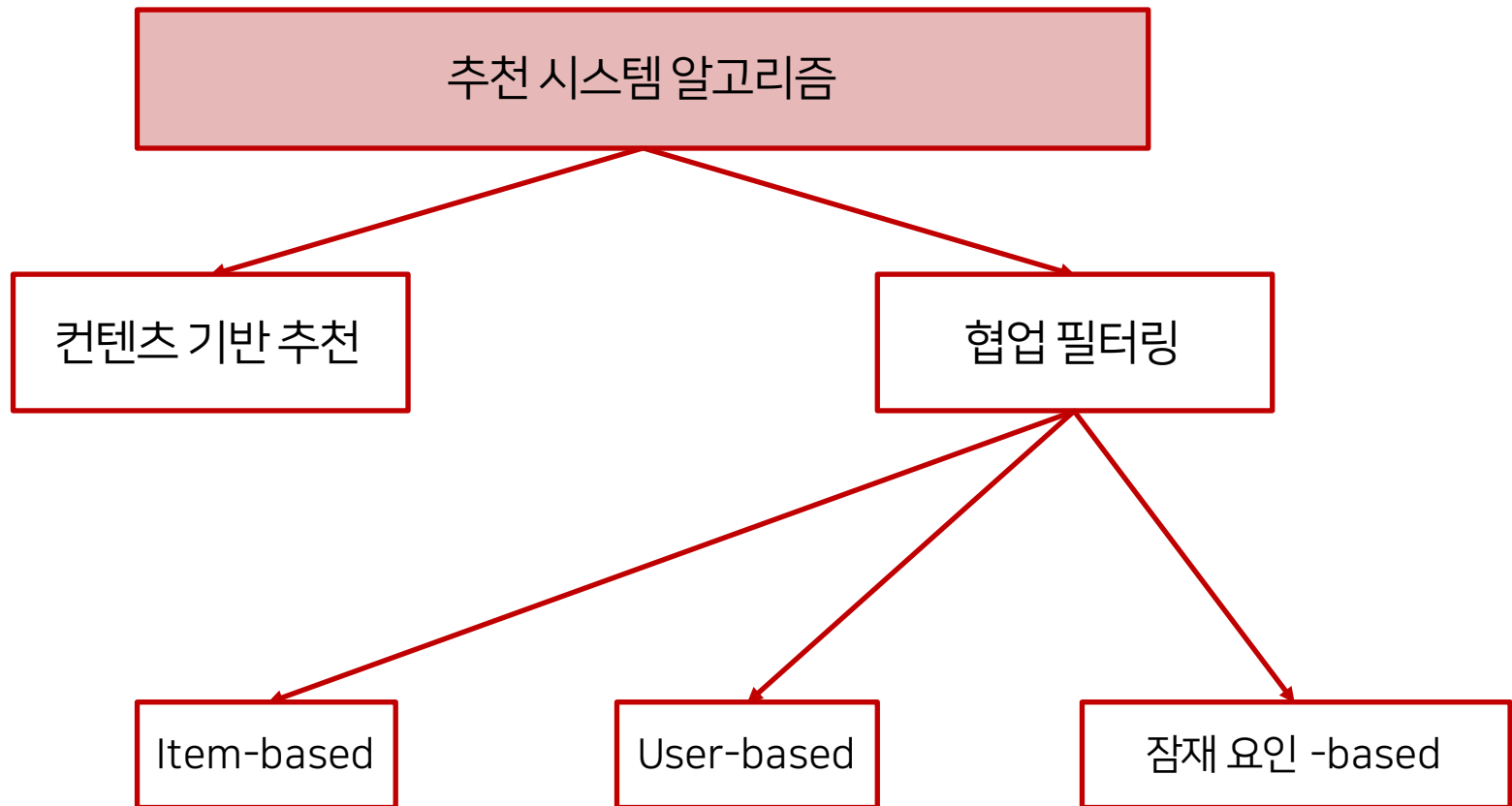
RSS를 줄이는 방식에 overfitting 방지를 위한 regularization term을 더해 손실함수를 작성할 수 있다.

$$\min_{q,p} \sum_{(u,i) \in K} (r_{ui} - q_i^T p_i)^2 + \lambda (\|q_i\|^2 + \|p_u\|^2)$$

Cross-validation을 통해 결정

이 수식의 계산 값이 minimum하게 나오는
q와 p를 찾는 게 목표

Recommender System





THANK YOU

