

- 클러스터링의 정의
 - 개인 또는 개체중에서 유사한 것들을 몇 개의 집단으로 그룹화하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 탐색적인 분석방법이며, 대표적인 비지도 학습
 - 데이터 자체에만 의존하여 자료를 탐색하고 요약하는 분석기법
 - 사전에 정의된 어떠한 특수적인 목적이 없음.
 - 대용량 데이터에 대해서는 개개의 관찰치를 요약하는 것보다 전체를 유사한 관찰치들의 군집으로 구분하여, 복잡한 전체보다는 그들을 잘 대표하는 군집들을 관찰함으로써 전체 데이터에 대한 의미 있는 정보를 얻어낼 수 있음.
 - 동일한 군집의 개체들은 유사한 성격을 갖도록, 서로 다른 군집에 속한 개체들은 서로 다른 성격을 갖도록 군집이 형성되어야함
 - 관찰단위의 성격을 표현하는 알맞은 변수를 선택한 후에 주어진 변수들을 이용해 각 관찰단위가 서로 얼마나 유사한지 또는 유사하지 않은지를 측정할 수 있는 측도가 필요
 - KEY POINT**
 - 각 개체들 간에 유사 속성을 지닌 것들끼리 모으는 방법
 - 대표적인 unsupervised learning (비지도학습)의 한 방법. Y 변수 (목표변수)가 존재하지 않음
 - 모델을 만드는것에 초점을 맞추기 보다, 개체의 속성이 유사한 것끼리 모으는 것을 목표로함
- 클러스터링의 유형
 - 상호배반적 군집
 - 각 관찰단위가 상호 배반적인 여러 군집. (한국인, 중국인, 일본인)
 - 중복군집
 - 두 개 이상의 군집에 한 관찰단위가 동시에 소속되는 것을 허용함
 - 퍼지군집
 - 관찰단위가 소속되는 특정한 군집을 표현하는 것이 아니라, 각 군집에 속할 가능성을 표현
 - 계보적 군집
 - 한 군집이 다른 군집의 내부에 포함되는 형태의 군집간의 중복은 없으며 군집들이 매단계 계층적인 구조를 형성함.
(예) 생물표본의 분류에서 '종'속'과'목'
- 클러스터링의 장/단점

장점	단점
<div>1. 탐색적인 기법</div> <div>1. 대용량데이터에 대한 탐색적인 기법이며 주어진 데이터의 내부 구조에 대한 사전정보없이, 의미 있는 자료구조를 찾아낼수 있음</div> <div>2. 다양한 형태의 데이터에 적용가능</div> <div>1. 기본적으로 관찰단위 간의 거리를 데이터 형태에 맞게 정의한다면 거의 모든 형태의 데이터에 대하여 적용가능</div> <div>3. 분석방법의 적용 용이성</div> <div>1. 대부분의 군집 방법이 분석 대상 데이터에 대해 사전 정보를 거의 요구하지 않으므로 적용하는데 큰 어려움이 없으며, 단지 관찰 단위 사이의 거리만이 분석에 필요한 입력자료가 됨</div>	<div>1. 가중치와 거리 정의</div> <div>1. 군집분석의 결과는 관찰단위 사이의 유사성(또는 비유사성)을 나타내는 거리를 어떻게 정의하는가에 크게 좌우되지만 거리 정의 및 각 변수의 가중치를 결정하는것이 매우 어려운 문제임</div> <div>2. 초기군집수의 결정</div> <div>1. K-means 군집분석에서는 사전에 정의된 군집수를 기준으로 동일한 수의 군집을 찾게 되므로, 군집 수 K가 원 데이터 구조에 적합하지 않으면, 좋은 결과를 얻을 수 없기 때문에 여러 번의 탐색적인 군집 분석 절차가 요구됨.</div> <div>3. 결과 해석의 어려움</div> <div>1. 사전에 주어진 목적이 없으므로 결과를 해석하는데 어려움이 있고, 주어진 변수에 따라 잘 구분된 군집이라 하더라도 그 결과를 충분히 이해하고 현실적으로 활용하기는 쉽지 않음</div>

- 대표적인 거리계산
 - 민코우스키 거리
$$= \left[\sum_{k=1}^p |X_{ik} - X_{jk}|^m \right]^{1/m}$$
 - 유클리드 거리
민코우스키 거리에서 m 대신 2를 대입
 - 맨하탄 거리
 - 최대좌표 거리

❖ 계층적 병합 군집화 예제 – 최단 연결법(Single Linkage Method)

➢ 5명 학생의 평점이 다음과 같이 주어져 있을 때, Hierarchical Clustering 기법 중 최단연결법을 사용하여 Dendrogram을 그린 후, 기준값 0.25를 적용했을 때 군집이 어떻게 구분되는지 보여라. (학생명은 알파벳으로, 평점은 숫자로 표현하였음)

A=4.1, B=3.2, C=2.9, D=3.9, E=2.8

➢ 평균 연결법의 두 군집 C_1, C_2 간의 거리

$$d_{C_1C_2} = d\{(C_1)(C_2)\} = \min\{d(X_i, X_j) | X_i \in C_1, X_j \in C_2\}$$

➢ 고립되어 있는 군집을 찾는 데 중점

	A	B	C	D	E
A	0.0				
B	0.9	0.0			
C	1.2	0.3	0.0		
D	0.2	0.7	1.0	0.0	
E	1.3	0.4	0.1	1.1	0.0

<초기 거리행렬 D_0 >

Step 1 : 초기 거리 행렬 D_0 에서 $d_{CE} = 0.1$ 이 최소이므로 관찰단위 C와 E를 묶어 군집 (C,E)를 생성함

Step 2 : 거리 행렬 D_1 을 생성하기 위해, 만들어진 군집 (C,E)와 다른 관찰단위와의 거리를 계산함

$$\begin{aligned} d\{(A)(C,E)\} &= \min(d_{AC}, d_{AE}) = \min(1.2, 1.3) = 1.2 \\ d\{(B)(C,E)\} &= \min(d_{BC}, d_{BE}) = \min(0.3, 0.4) = 0.3 \\ d\{(D)(C,E)\} &= \min(d_{DC}, d_{DE}) = \min(1.0, 1.1) = 1.0 \end{aligned}$$

❖ 계층적 병합 군집화 예제 – 최단 연결법(Single Linkage Method)

➢ 5명 학생의 평점이 다음과 같이 주어져 있을 때, Hierarchical Clustering 기법 중 최단연결법을 사용하여 Dendrogram을 그린 후, 기준값 0.25를 적용했을 때 군집이 어떻게 구분되는지 보여라. (학생명은 알파벳으로, 평점은 숫자로 표현하였음)

	(C,E,B)	(A,D)
(C,E,B)	0.0	
(A,D)	0.7	0.0

<Step 6 이후 거리행렬 D_3 >

Step 7 : 최종적으로 군집 (A,D)와 (C,E,B)를 묶어 전체가 한 군집을 이루게 하고 Dendrogram을 작성함

Dendrogram

0.975

0.3

0.2

0.1

0.25

기준값 0.25
군집 (A,D)
군집 (C,E)
군집 B

✓ 위의 결과를 바탕으로 분석자가 설정한 기준에 따라 군집의 수를 정할 수 있음

✓ Dendrogram을 통해 어떤 기준에서 군집들이 결합되는 지에 대한 확인이 가능함

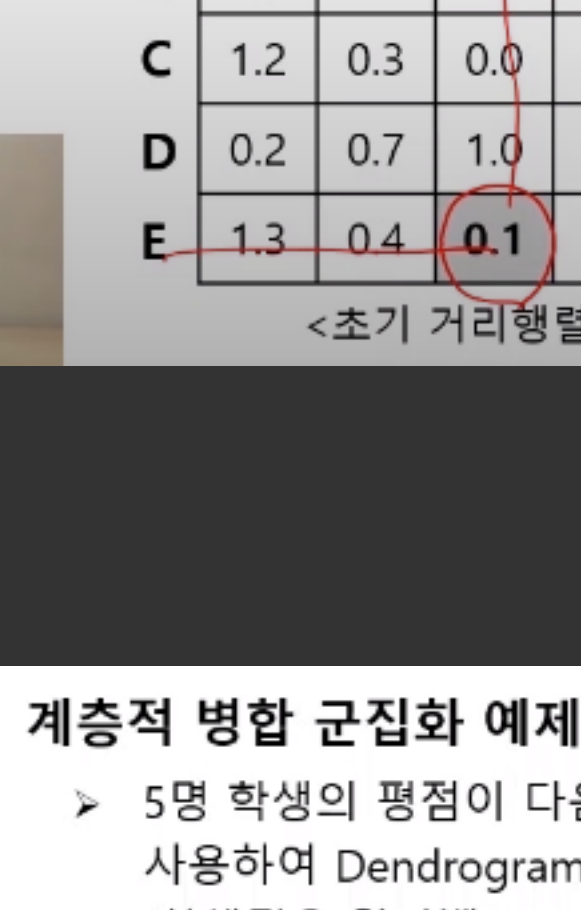
❖ K-means Clustering의 개념

➢ K-means Clustering이란 사전에 결정된 군집 수 K에 기초하여 전체 데이터를 상대적으로 유사한 K개의 군집으로 구분하는 방법임

➢ K-means Clustering 절차는 다음과 같음

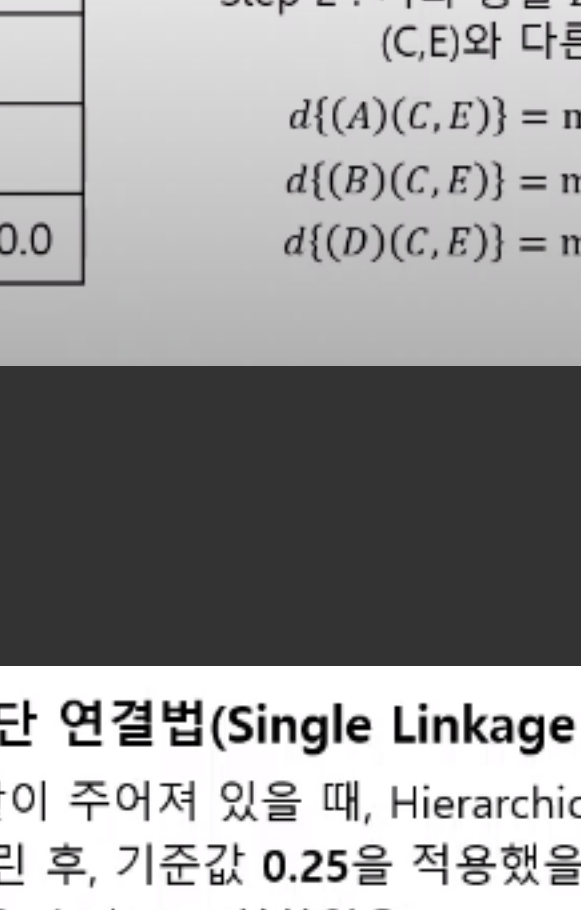
[Step 1] 군집 초기값의 선택

- 주어진 군집의 개수(K)만큼 군집 초기값(cluster seed) 선택하며 아래 예는 K=5인 경우임



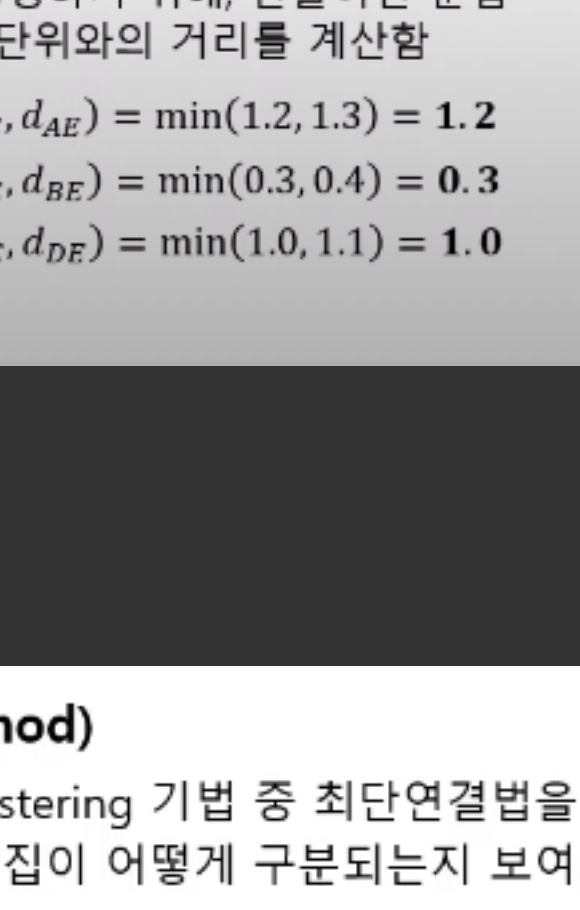
[Step 2] 초기 군집의 형성

- 각 개체들에 대하여 군집 초기값(중심)들과의 거리를 계산하고, 거리가 가장 가까운 초기값에 개체들을 할당



[Step 3] 개체들의 재할당

- 각 개체들을 가장 가까운 군집 중심(cluster center)에 재할당하고 군집의 중심(평균벡터)을 다시 계산 후 최종 군집을 형성함

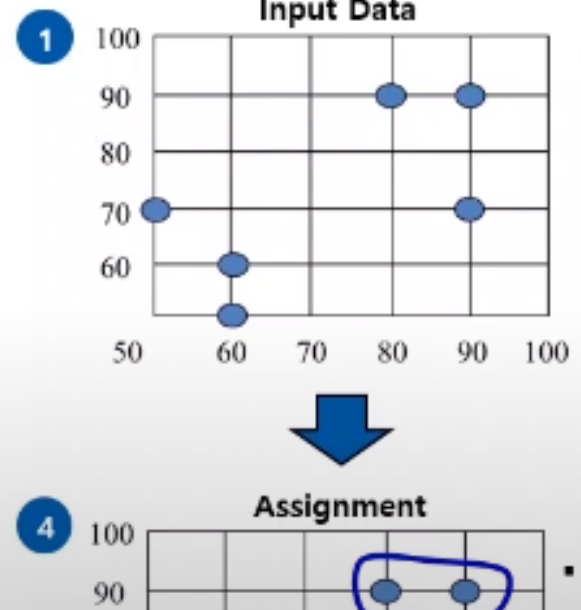


❖ K-means Clustering 예시

➢ 데이터마이닝의 중간고사 점수와 신뢰성공학의 중간고사 점수가 다음과 같이 주어졌을 때, K-means 군집분석을 시행하여라.(K=2)

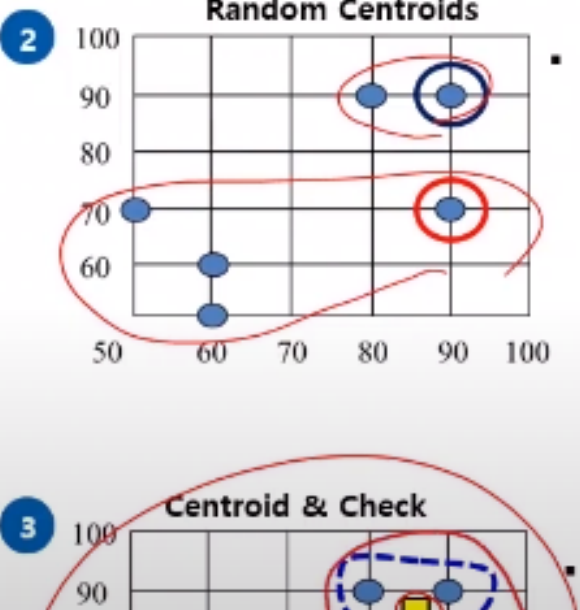
학생 A(90점, 70점), B(90점, 90점), C(80점, 90점), D(60점, 50점), E(60점, 60점), F(50점, 70점)

1. Input Data



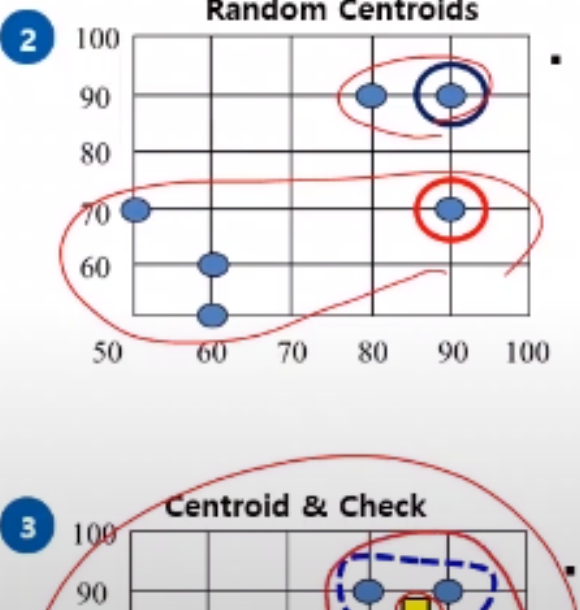
Input Data를 그래프 상에 도식화함

2. Random Centroids



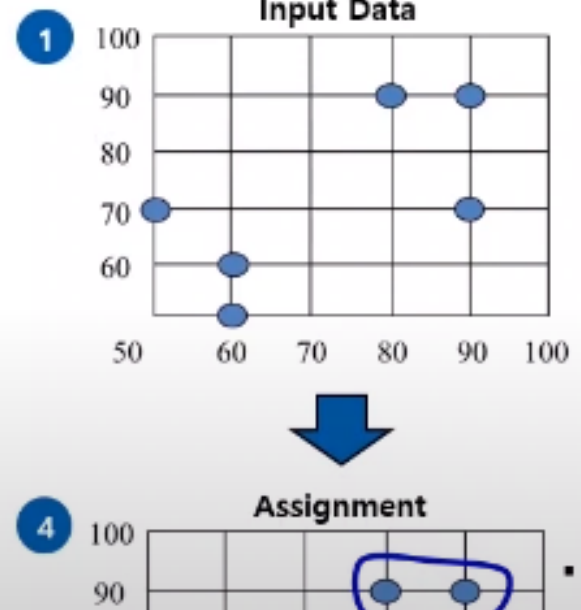
인위적으로 중심을 설정함

3. Centroid & Check



평균벡터를 다시 계산하여 새로운 군집에서 새로운 중심을 설정함

4. Assignment



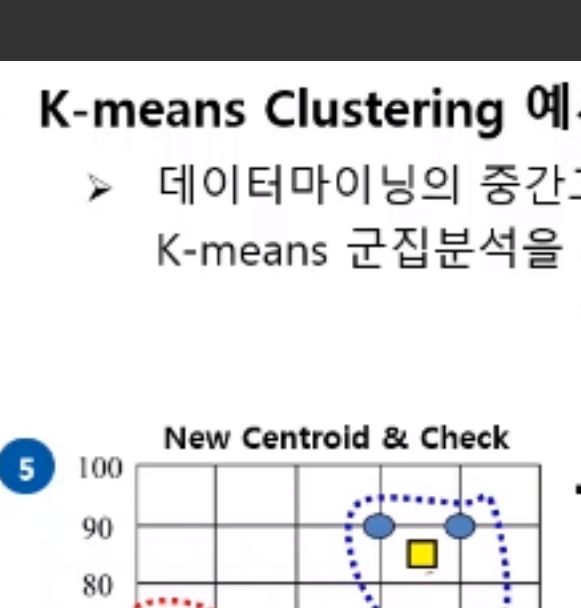
군집을 할당함

❖ K-means Clustering 예시

➢ 데이터마이닝의 중간고사 점수와 신뢰성공학의 중간고사 점수가 다음과 같이 주어졌을 때, K-means 군집분석을 시행하여라.(K=2)

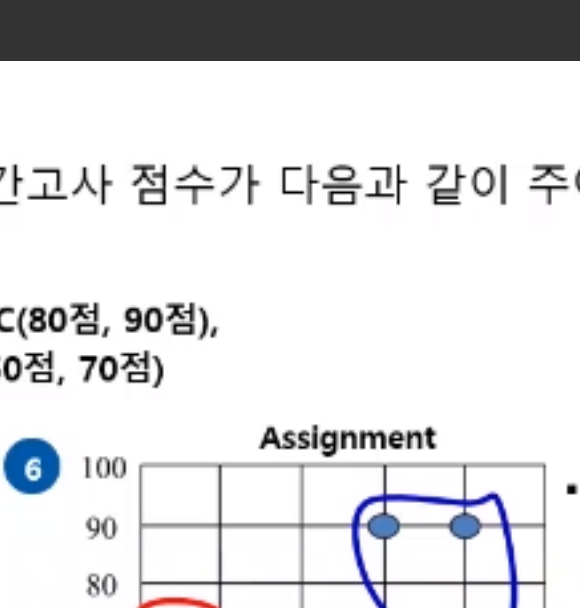
학생 A(90점, 70점), B(90점, 90점), C(80점, 90점), D(60점, 50점), E(60점, 60점), F(50점, 70점)

5. New Centroid & Check



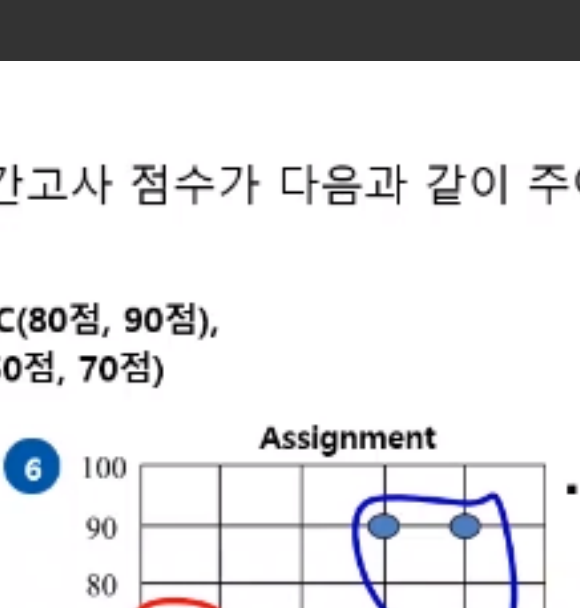
새로운 군집을 설정

6. Assignment



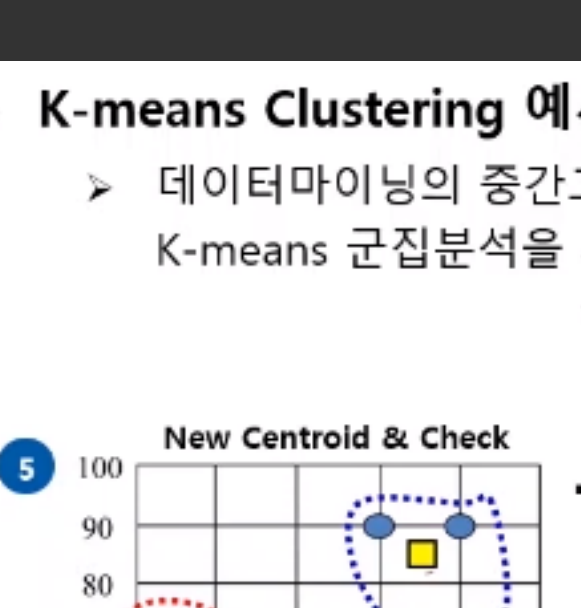
새로운 군집을 정의

7. New Centroid & Check



평균벡터를 다시 계산하여 새로운 군집에서 새로운 중심을 설정

8. Assignment



새로운 군집을 정의