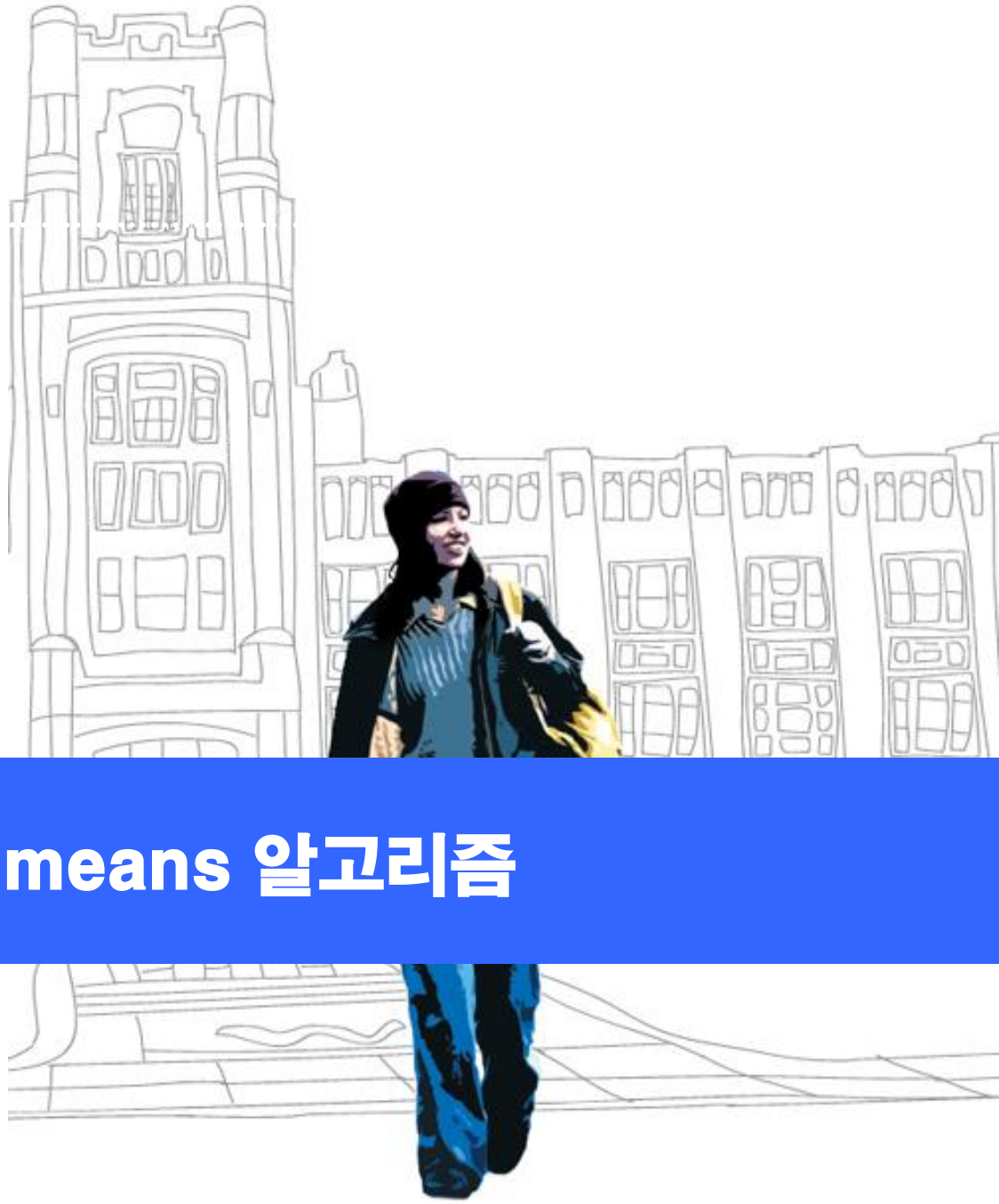
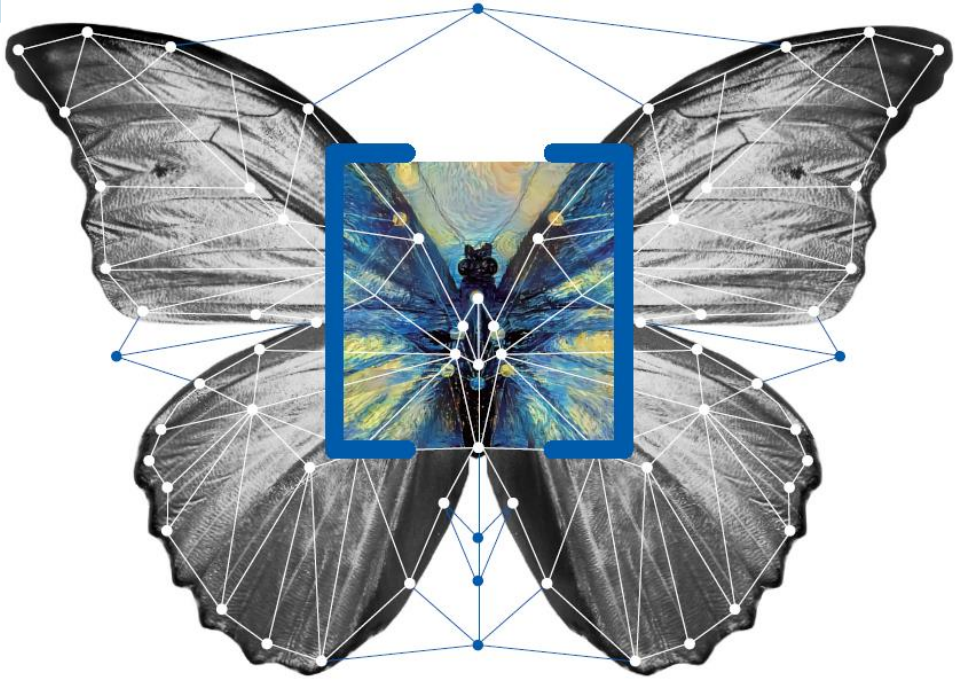


06_ *k*-means 알고리즘



MACHINE LEARNING 기계 학습

오일석 지음

6장. 비지도 학습

6.1 지도 학습과 비지도 학습, 준지도 학습

세 가지 유형의 학습

- 지도 학습: 모든 훈련 샘플이 레이블 정보를 가짐
- 비지도 학습: 모든 훈련 샘플이 레이블 정보를 가지지 않음
- 준지도 학습: 레이블을 가진 샘플과 가지지 않은 샘플이 섞여 있음

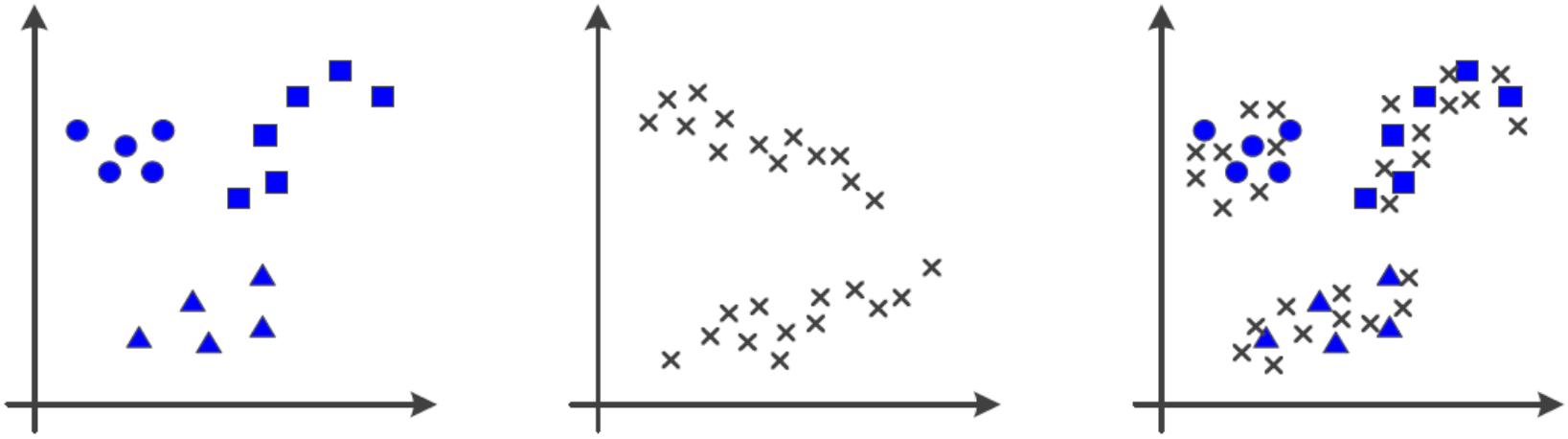


그림 6-1 기계 학습의 유형(속이 찬 샘플은 레이블이 있고, x 표시된 샘플은 레이블이 없음)

6.2.1 비지도 학습의 일반 과업

- 세 가지 일반 과업
  - 군집화: 유사한 샘플을 모아 같은 그룹으로 묶는 일
  - 밀도 추정: 데이터로부터 확률분포를 추정하는 일
  - 공간 변환: 원래 특징 공간을 저차원 또는 고차원 공간으로 변환하는 일
- 데이터에 내재한 구조를 잘 파악하여 새로운 정보를 발견해야 함

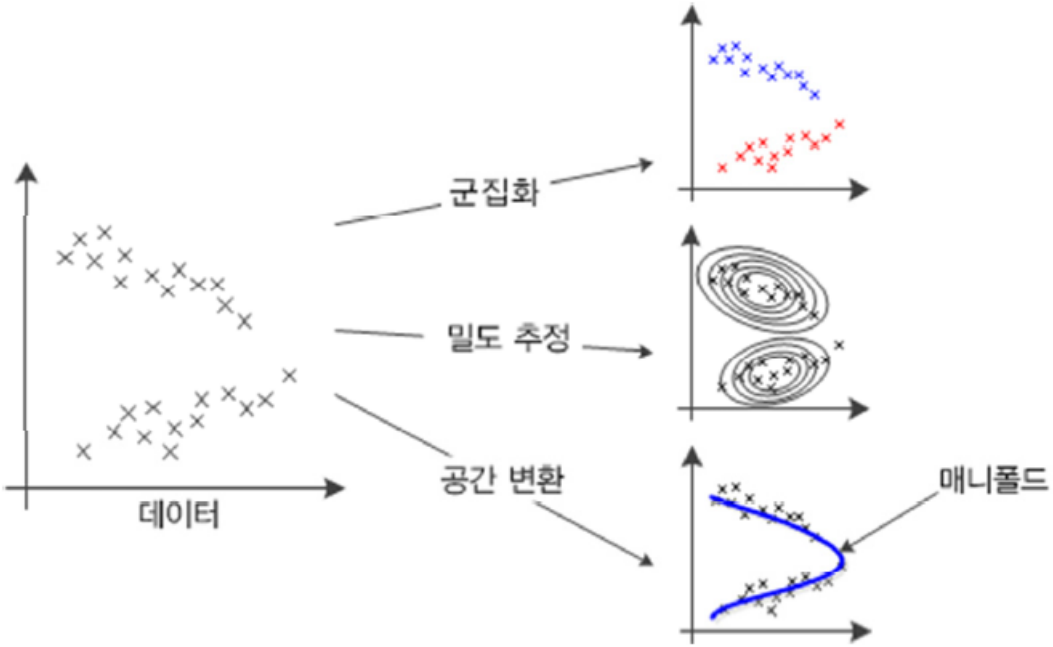


그림 6-2 비지도 학습의 군집화, 밀도 추정, 공간 변환 과업이 발견하는 정보

6.2.2 비지도 학습의 응용 과업

- **아주 많은 응용(서로 밀접하게 연관)**
  - 군집화의 응용
    - 맞춤 광고, 영상 분할, 유전자 데이터 분석, SNS 실시간 검색어 분석하여 사람들의 관심 파악 등
  - 밀도 추정의 응용
    - 분류, 생성 모델 구축 등
  - 공간 변환의 응용
    - 데이터 가시화, 데이터 압축, 특징 추출(표현 학습) 등

6.3 군집화

군집화 문제

- $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$에서 식 (6.1)을 만족하는 군집집합 $C = \{c_1, c_2, \cdots, c_k\}$를 찾아내는 작업

$$\left.\begin{array}{l} c_i \neq \emptyset, i = 1,2,\cdots,k \\ \bigcup_{i=1}^{k} c_i = \mathbb{X} \\ c_i \bigcap c_j = \emptyset, i \neq j \end{array}\right\} \quad (6.1)$$

K가 주어지지 않을 경우도 있음.

- 군집의 개수 k는 주어지는 경우와 자동으로 찾아야 하는 경우가 있음
- 군집화를 부류 발견 작업이라 부르기도 함

군집화의 주관성

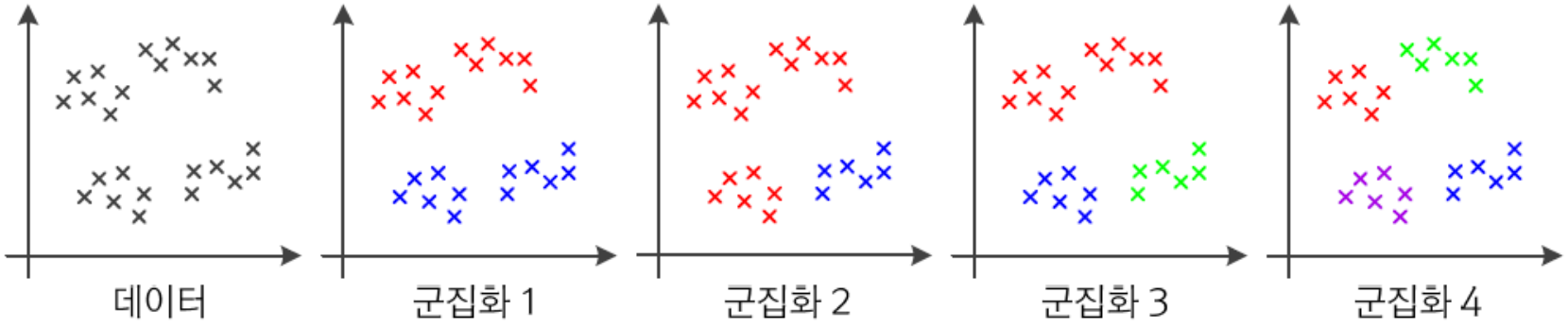


그림 6-3 군집화의 주관성

6.3.1 k-평균 알고리즘

k-평균 알고리즘의 특성

- 원리 단순하지만 성능이 좋아 인기 좋음
- 직관적으로 이해하기 쉽고 구현 쉬움
- 군집 개수 k를 알려줘야 함

알고리즘 6-1 k-평균

입력: 훈련집합 $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, 군집의 개수 k
출력: 군집집합 $C = \{c_1, c_2, \cdots, c_k\}$

```
1  k개의 군집 중심 Z = {z_1, z_2, ..., z_k}를 초기화한다.
2  while (true)
3     for (i=1 to n)
4           x_i를 가장 가까운 군집 중심에 배정한다.
5     if (라인 3~4에서 이루어진 배정이 이전 루프에서의 배정과 같으면) break
6     for (j=1 to k)
7           z_j에 배정된 샘플의 평균으로 z_j를 대치한다.
8  for (j=1 to k)
9     z_j에 배정된 샘플을 c_j에 대입한다.
```

반복을 멈출지 체크

6.3.1 k-평균 알고리즘

k-평균과 k-medoids

- k-평균은 [알고리즘 6-1]의 라인 7에서 샘플의 평균으로 군집 중심을 갱신
- k-medoids는 대표를 뽑아 뽑힌 대표로 군집 중심을 갱신(k-평균에 비해 잡음에 둔감)

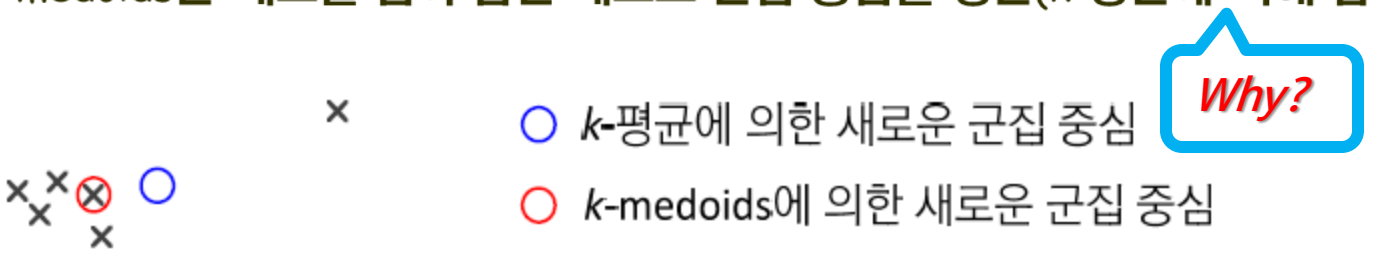


그림 6-4 k-평균과 k-medoids가 군집 중심을 갱신하는 과정

최적화 문제로 해석

- k-평균은 식 (6.2)의 목적함수를 최소화하는 알고리즘
- 행렬 A는 군집 배정 정보를 나타내는 $k*n$ 행렬(i번째 샘플이 j번째 군집에 배정되었다면 a_{ji}는 1, 그렇지 않으면 0)

$$J(Z,\mathbf{A}) = \sum_{i=1}^{n}\sum_{j=1}^{k} a_{ji}\,dist(\mathbf{x}_i,\mathbf{z}_j) \qquad (6.2)$$

6.3.1 k-평균 알고리즘

예제 6-1 ***k*-평균의 동작**

[그림 6-5]는 훈련집합이 7개의 샘플을 가진 n=7인 예를 보여 준다. 좌표는 다음과 같다.

$$\mathbf{x}_1 = \begin{pmatrix} 18 \\ 5 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 20 \\ 9 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 20 \\ 14 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 20 \\ 17 \end{pmatrix}, \mathbf{x}_5 = \begin{pmatrix} 5 \\ 15 \end{pmatrix}, \mathbf{x}_6 = \begin{pmatrix} 9 \\ 15 \end{pmatrix}, \mathbf{x}_7 = \begin{pmatrix} 6 \\ 20 \end{pmatrix}$$

군집의 개수 k=3이라 하자. 맨 왼쪽 그림은 초기 군집 중심을 보여 준다. [알고리즘 6-1]의 라인 3~4는 7개 샘플을 아래와 같이 배정할 것이다.

$$\{\mathbf{x}_1\}\text{은 } \mathbf{z}_1,\ \{\mathbf{x}_2\}\text{은 } \mathbf{z}_2,\ \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6, \mathbf{x}_7\}\text{은 } \mathbf{z}_3$$

이 배정을 행렬 **A**로 표현하면 다음과 같다.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

6.3.1 k-평균 알고리즘

[그림 6-5]의 가운데 그림은 새로 계산한 군집 중심이다. $\mathbf{z}_1 = (18, 5)^{\mathrm{T}}$, $\mathbf{z}_2 = (20, 9)^{\mathrm{T}}$, $\mathbf{z}_3 = (12, 16.2)^{\mathrm{T}}$ 이고, 식 (6.2)에 대입하면 $J = 244.80$이 된다. 이때 거리함수 dist로 식 (1.7)의 유클리디언 거리를 사용한다.

두 번째 루프를 실행하면 행렬 **A**는 아래와 같이 바뀐다. 군집 중심은 $\mathbf{z}_1 = (18, 5)^{\mathrm{T}}$, $\mathbf{z}_2 = (20, 13.333)^{\mathrm{T}}$, $\mathbf{z}_3 = (6.667, 16.667)^{\mathrm{T}}$이다. 이것을 식 (6.2)에 대입하면 $J = 58.00$이 된다. [그림 6-5]의 맨 오른쪽 그림은 두 번째 루프 수행 후의 상황이다.

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

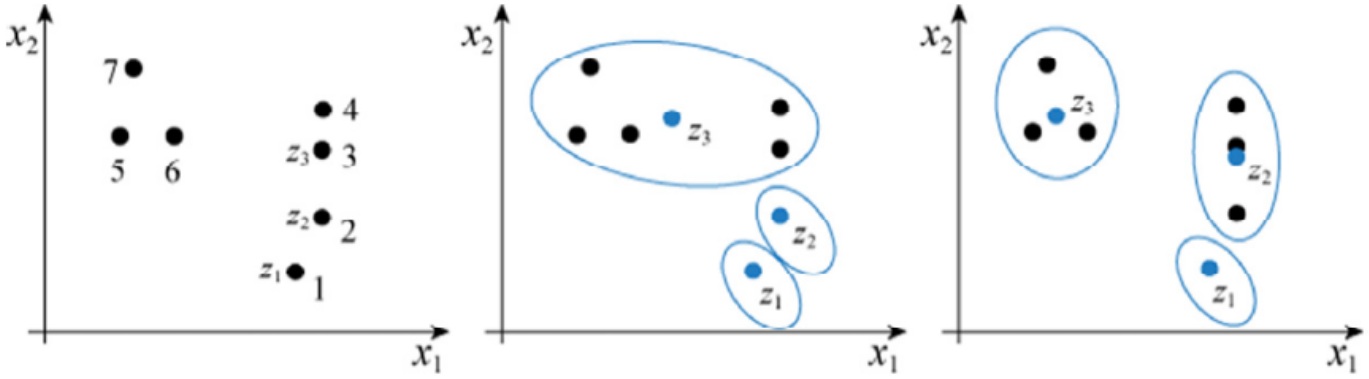


그림 6-5 k-평균의 동작 예제

6.3.1 k-평균 알고리즘

다중 시작 k-평균

- k-평균은 [알고리즘 6-1]의 라인 1에서 초기 군집 중심이 달라지면 최종 결과가 달라짐
- 다중 시작은 서로 다른 초기 군집 중심을 가지고 여러 번 수행한 다음, 가장 좋은 품질의 해를 취함

알고리즘 6-2 다중 시작 k-평균

입력: 훈련집합 $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$, 군집의 개수 k, 다중 시작 횟수 t
출력: 군집집합 $C = \{c_1, c_2, \cdots, c_k\}$

```
1  for (i=1 to t)
2      𝕏에서 임의로 k개 샘플을 뽑는다.
3      라인 2에서 뽑은 샘플을 초기 군집 중심으로 삼고, [알고리즘 6-1]의 k-평균을 수행한다.
4      k-평균이 출력한 해를 가지고 식 (6.2)의 목적함숫값을 계산한다.
5  t개의 해 중 목적함숫값이 가장 작은 해를 최종해로 취한다.
```