

머신러닝과 빅데이터분석(R)

10주차 K-means와 KNN



박길식 교수



고려사이버대학교
THE CYBER UNIVERSITY OF KOREA



학습 목표

-  K-means Clustering을 통해 군집화를 수행할 수 있다.
-  KNN을 이용해 분류 작업을 수행할 수 있다.



학습 목차

- 1 K-means와 KNN의 이해
- 2 K-means와 KNN 실습

CHAPTER

K-means와 KNN의 이해

K-means Clustering

데이터 간의 유사도를 정의하고 그 유사도에 가까운 것부터
순서대로 병합하는 방법으로 군집(클러스터) 형성

- 유사도는 거리(Distance)를 주로 이용
- 유사도를 이용하여 분석 대상을 몇 개의 그룹으로 분류

㉔ 데이터 셋 전체를 대상으로 서로 유사한 특성들을 몇 개의
군집으로 세분화하여 대상 집단을 정확하게 이해하고
효율적으로 활용하기 위함

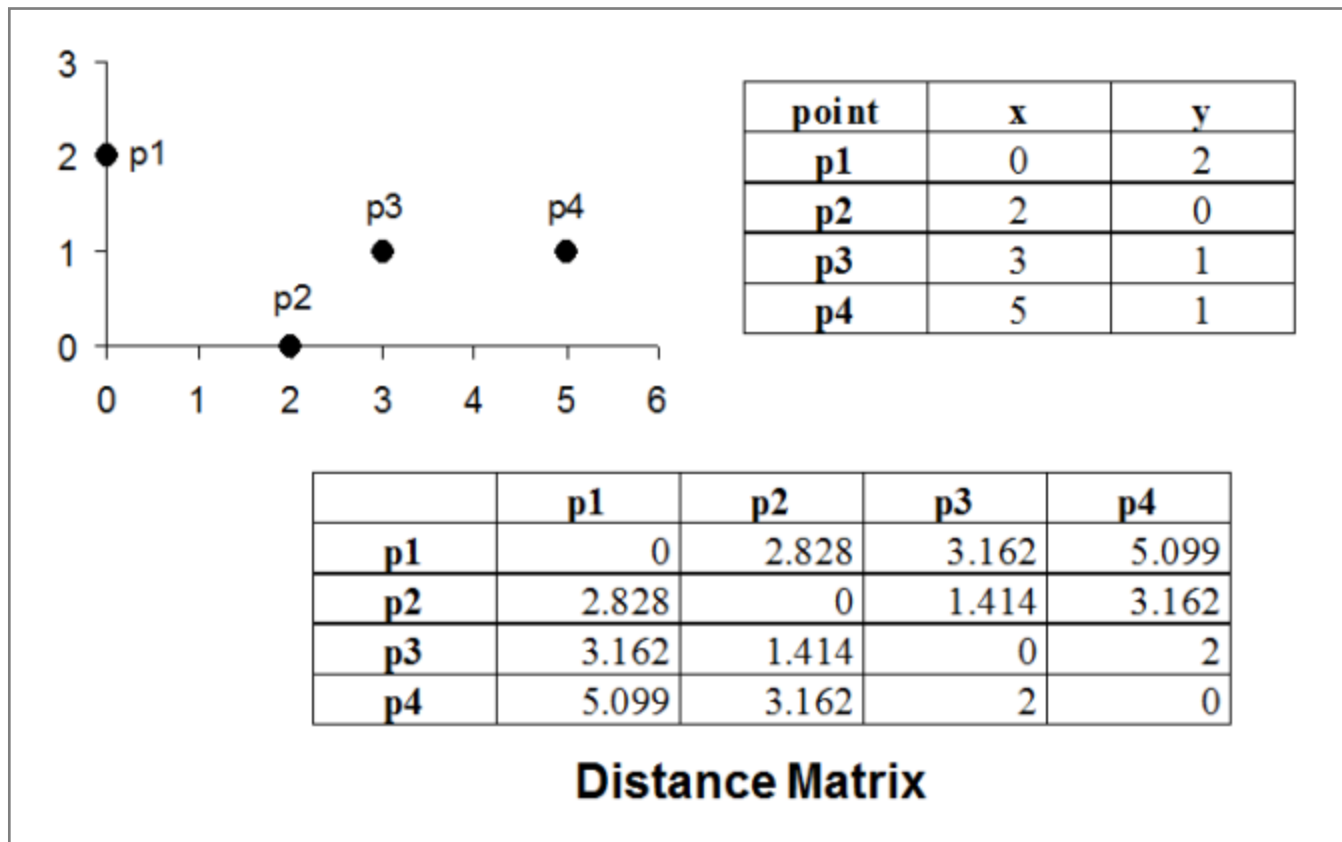
[01] K-means Clustering

 군집분석을 위한 거리 알고리즘(유클리디안)

» 유클리디안 거리의 정의

$$\text{dist} = \sqrt{\sum_{k=1}^n (p_k - q_k)^2}$$

- N = number of dimensions
- p_k, q_k = value of the k-th dimensions



[01] K-means Clustering



군집분석을 위한 거리 알고리즘(코사인 유사도 거리)

If d_1 and d_2 are two document vectors, then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / \|d_1\| \|d_2\| ,$$

where \bullet indicates vector dot product and $\|d\|$ is the length of vector d .

Example:

$$d_1 = 3 \ 2 \ 0 \ 5 \ 0 \ 0 \ 0 \ 2 \ 0 \ 0$$

$$d_2 = 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 2$$

$$d_1 \bullet d_2 = 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5$$

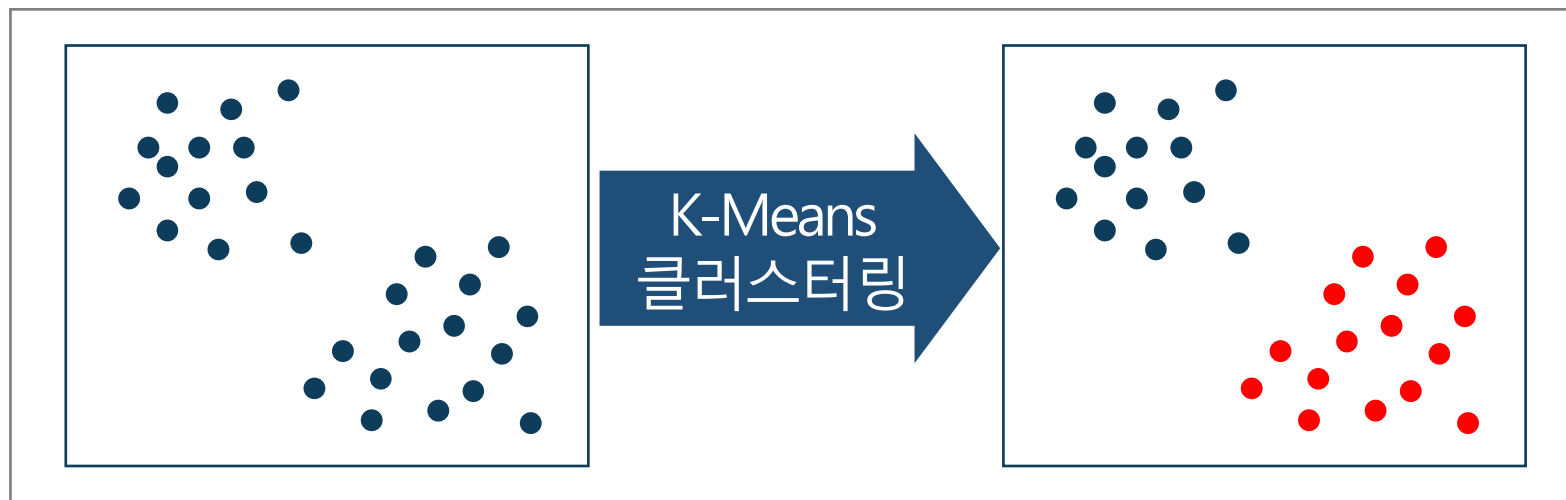
$$\|d_1\| = (3^2 + 2^2 + 0^2 + 5^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

$$\|d_2\| = (1^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 2^2)^{0.5} = (6)^{0.5} = 2.245$$

$$\cos(d_1, d_2) = .3150$$

[01] K-means Clustering

㉔ 데이터를 입력 받아 K개의 그룹으로 묶는 알고리즘

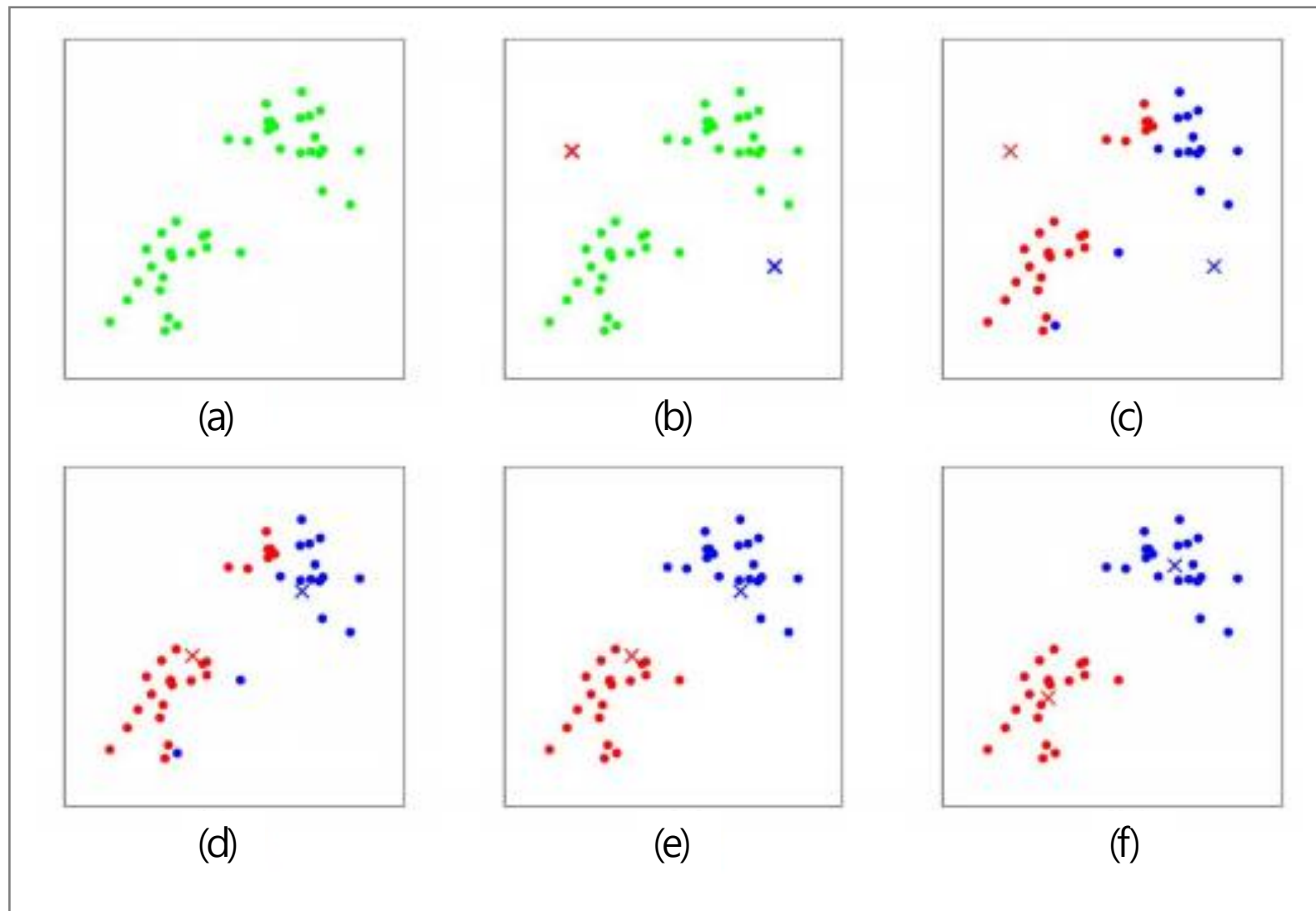


벡터 형태의 N개의 데이터에 관해
데이터가 속한 그룹의 중심과 데이터 간의 거리의
차이가 최소가 되도록 데이터들을 K개의 그룹으로
할당

— [01] K-means Clustering

- 1 클러스터별 하나의 점(Centroid)를 선택해서 클러스터를 초기화
 - 랜덤으로 점을 선택 → 단, K-1개의 다른 점들은 가능한 멀리 떨어져 있어야 함
- 2 각 점을 그 점과 가장 가까운 Centroid를 갖는 클러스터에 포함시킴
- 3 모든 점들이 할당된 후, K개의 클러스터들의 Centroid 위치를 갱신
 - 현재 클러스터에 포함된 모든 점들의 평균을 계산해서 다시 구함
- 4 2 과정으로 돌아가서 클러스터 내 데이터들이 바뀌지 않을 때까지 반복

[01] K-means Clustering



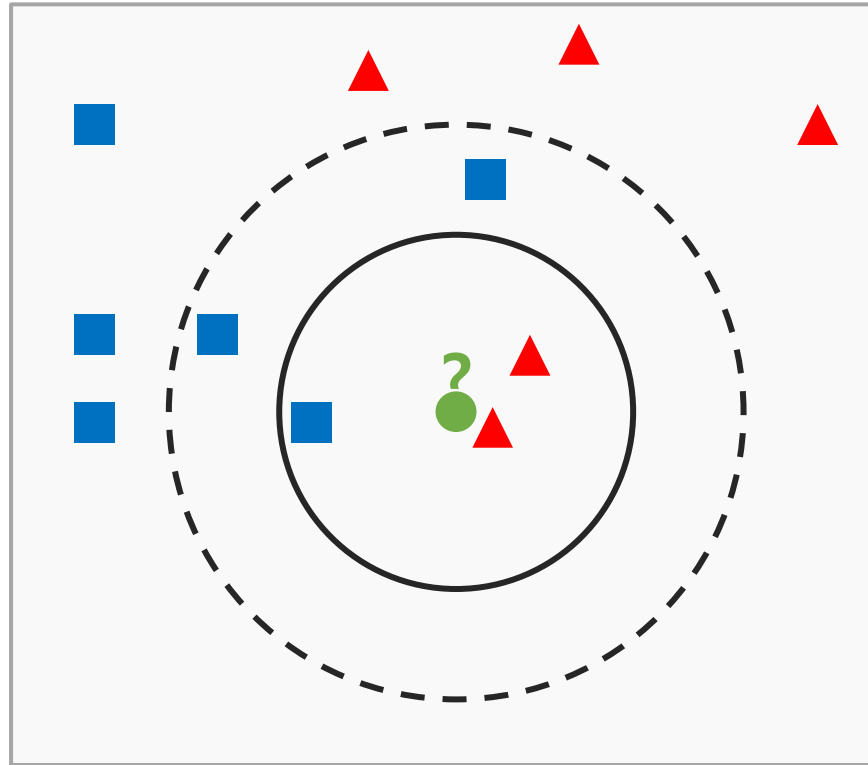
02 KNN(K-Nearest Neighbor)

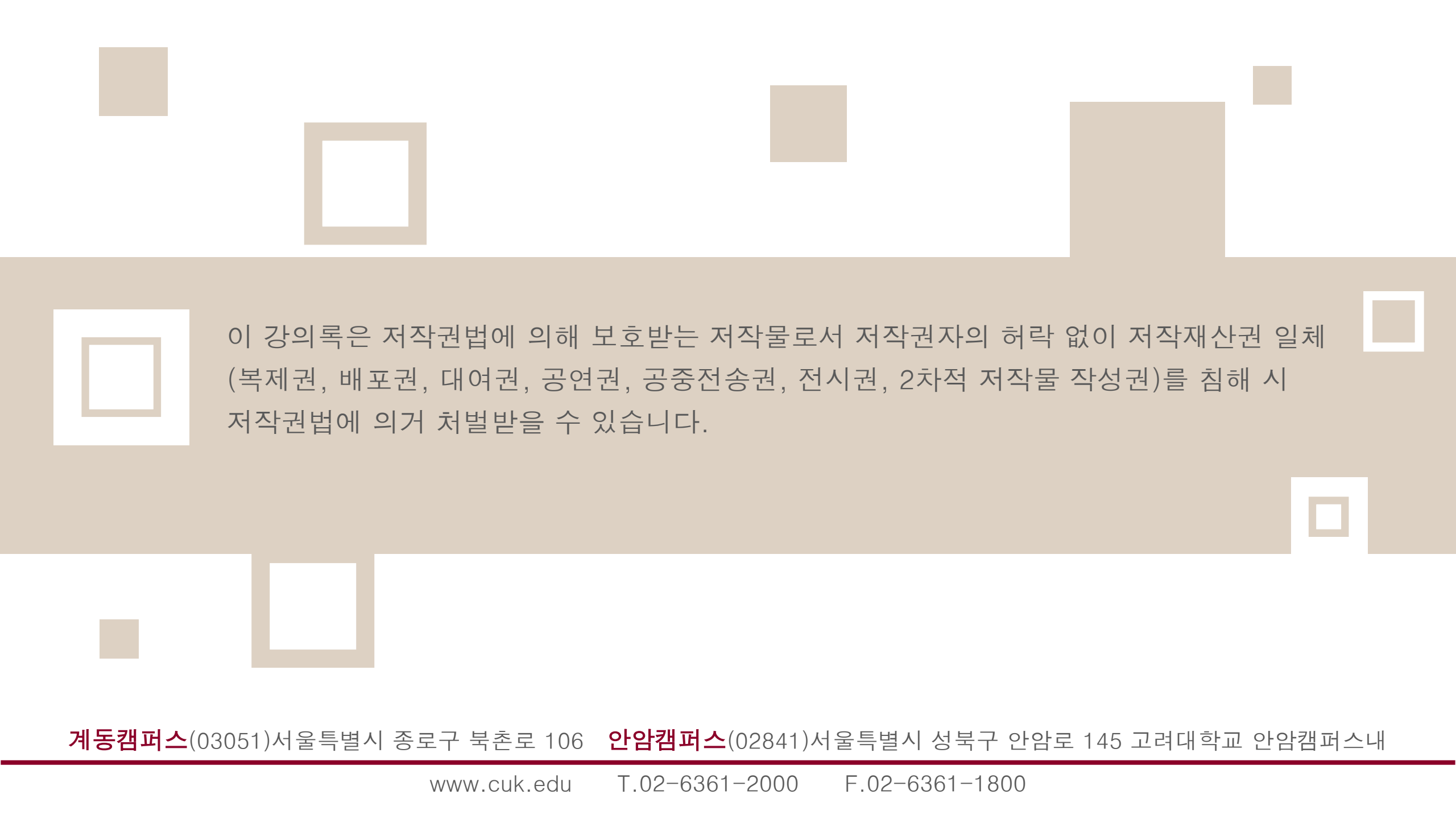
- » 새로운 데이터에 대해 가장 유사한 또는 거리가 가까운 K개의 과거 데이터를 이용하여 분류/예측 수행
- » 과거 데이터를 이용하여 미리 모델을 구성하는 것이 아니며, 과거 데이터를 저장만 하고 필요시 비교하는 방식
- » 값의 선택에 따라 새로운 데이터에 관한 예측 결과가 달라질 수 있음
- » 분류 문제와 회귀 문제에 모두 활용할 수 있음

분류	근접 데이터의 다수결로 예측
회귀	근접 데이터의 평균 값으로 예측

㉔ 평가 집합의 샘플(중앙의 초록색 점)은 1그룹(푸른색 사각형) 또는 2그룹(붉은색 삼각형)으로 분류될 수 있음

- $k=3$ (실선의 원)
→ 2그룹으로 분류
- $k=5$ (점선의 원)
→ 1그룹으로 분류
- 회귀 : 근접 데이터의 평균값으로 예측





이 강의록은 저작권법에 의해 보호받는 저작물로서 저작권자의 허락 없이 저작권재산권 일체 (복제권, 배포권, 대여권, 공연권, 공중전송권, 전시권, 2차적 저작물 작성권)를 침해 시 저작권법에 의거 처벌받을 수 있습니다.