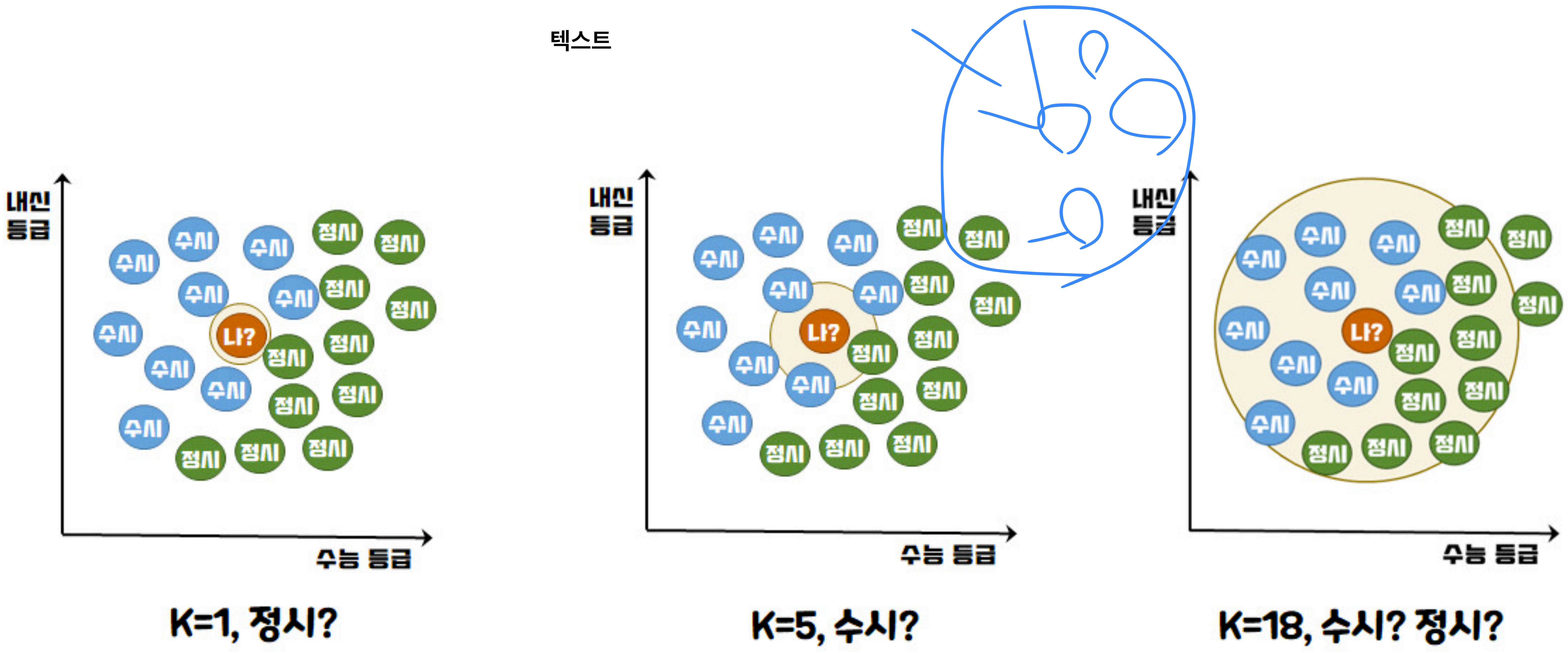
KNN 알고리즘

240324_8기_모델링반 (ML1)

KNN , Kmeans의 차이
지도학습과 비지도학습의 차이

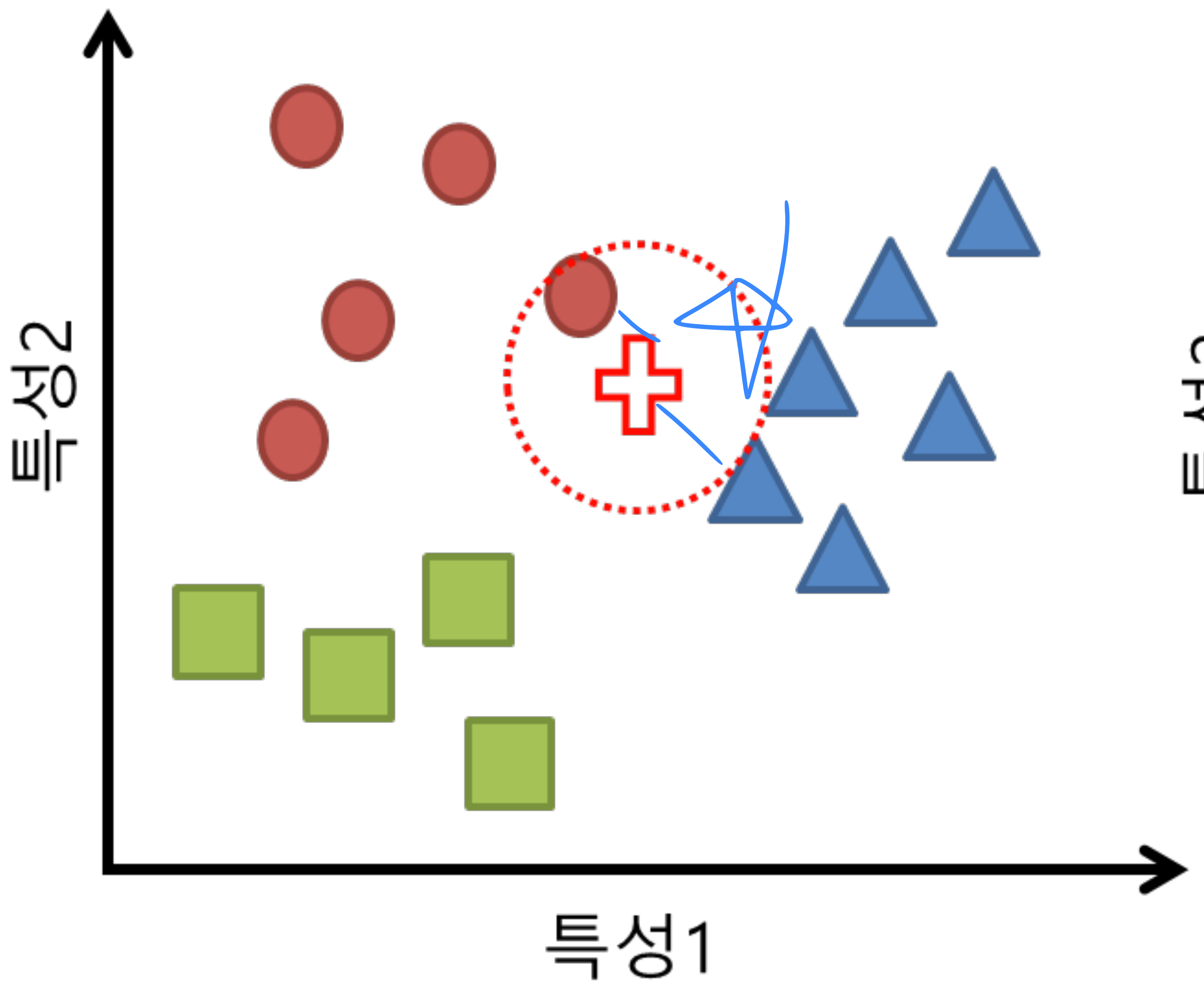
K-Nearest Neighbors : KNN

텍스트

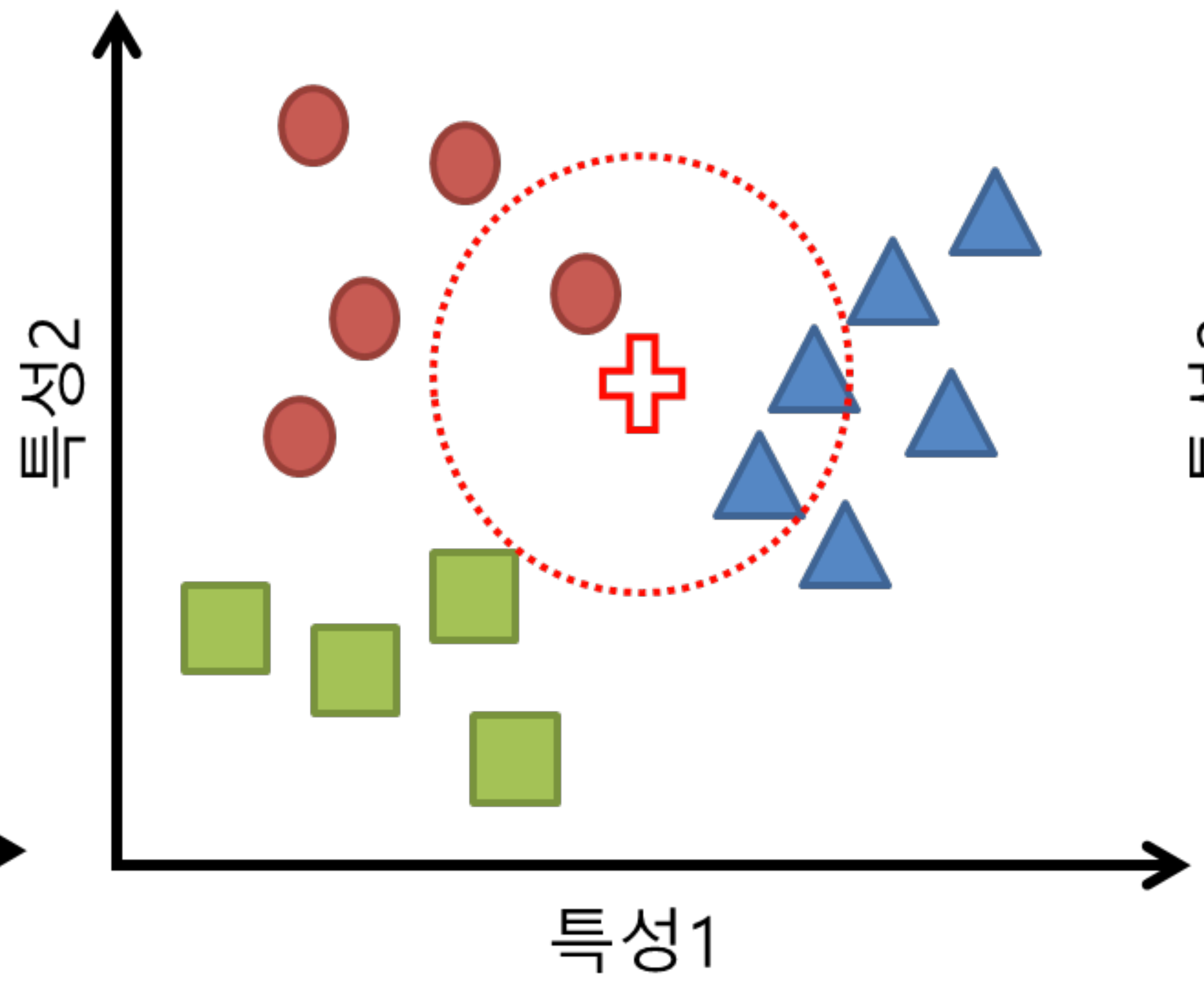


회귀, 분류 모두 가능한 Memory-based learning

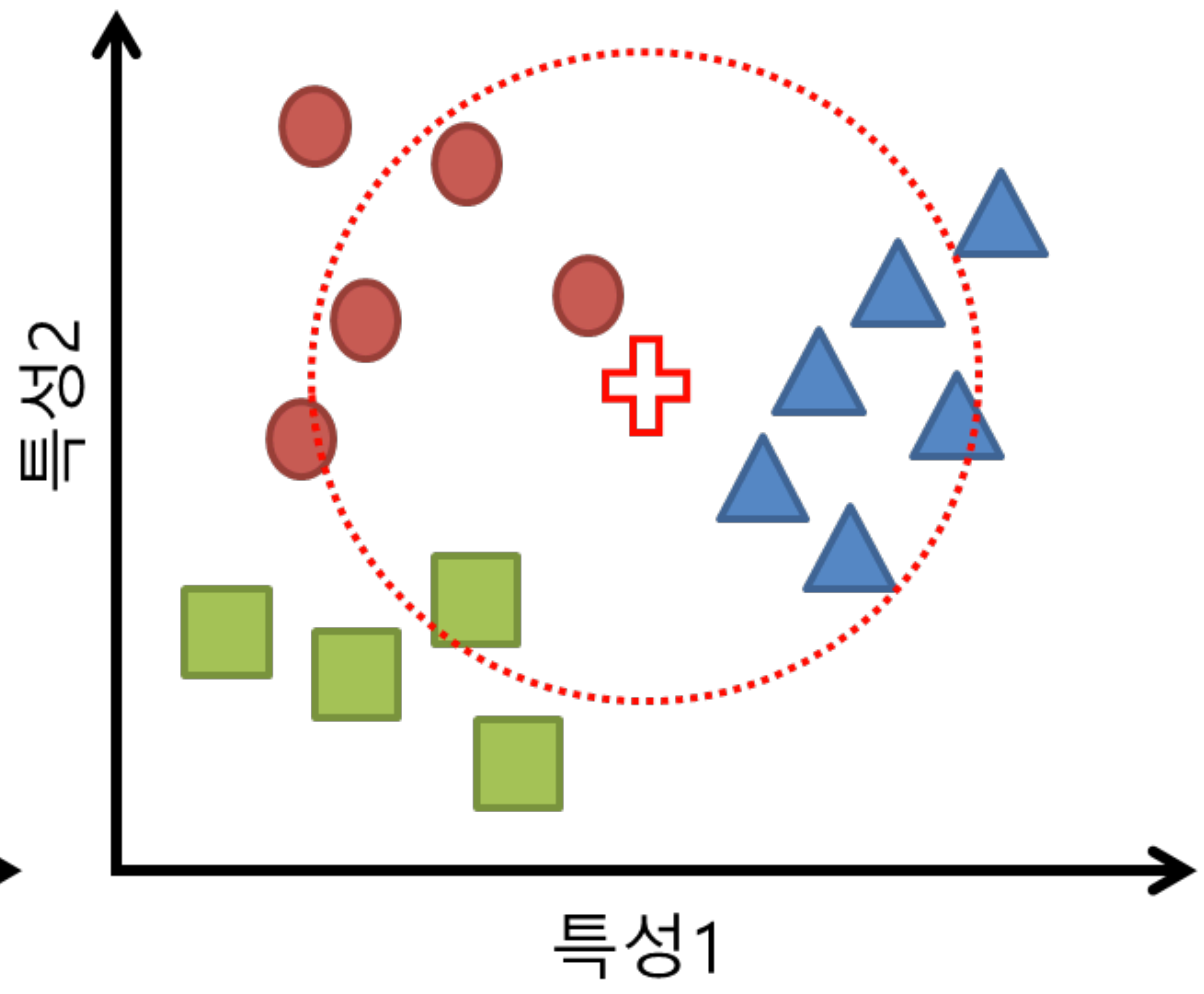
이웃의 수



k=1일 때

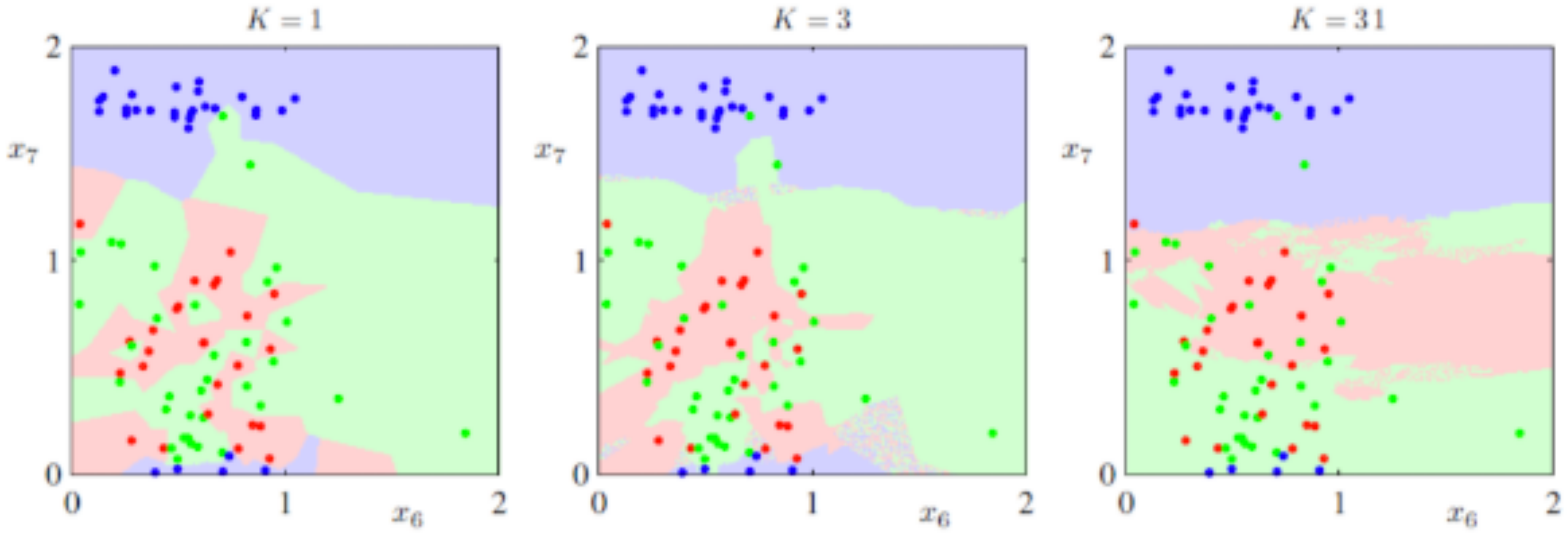


k=3일 때

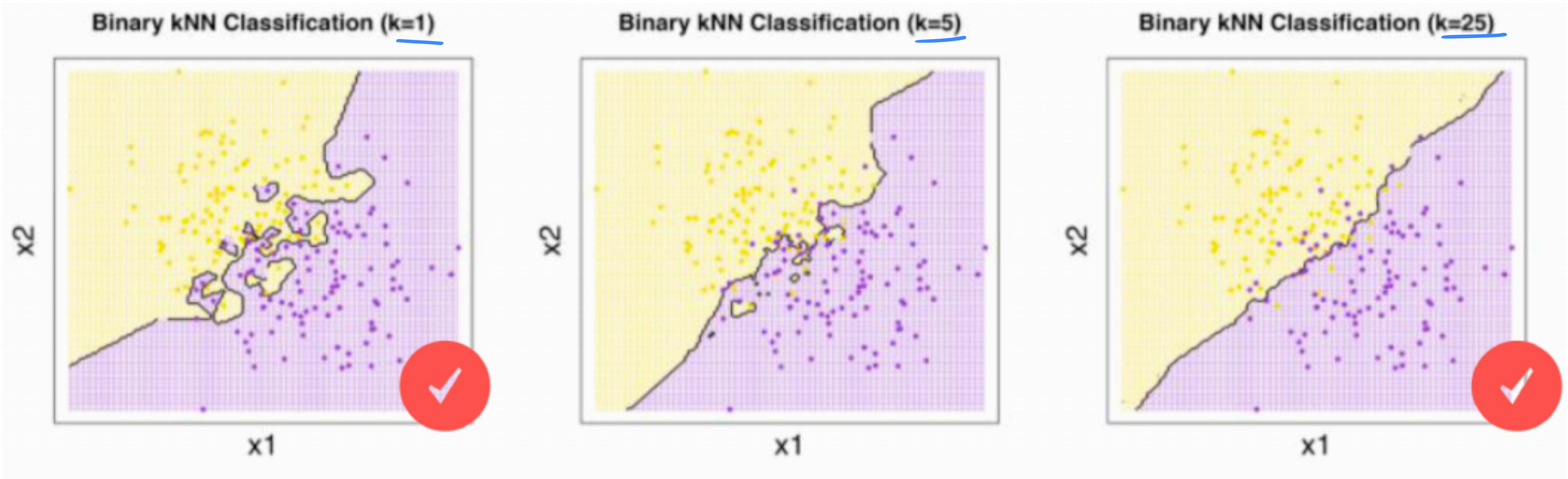


k=9일 때

적절한 K의 수?



$K=1$
$K=3$
$K=31$
x_7
x_6

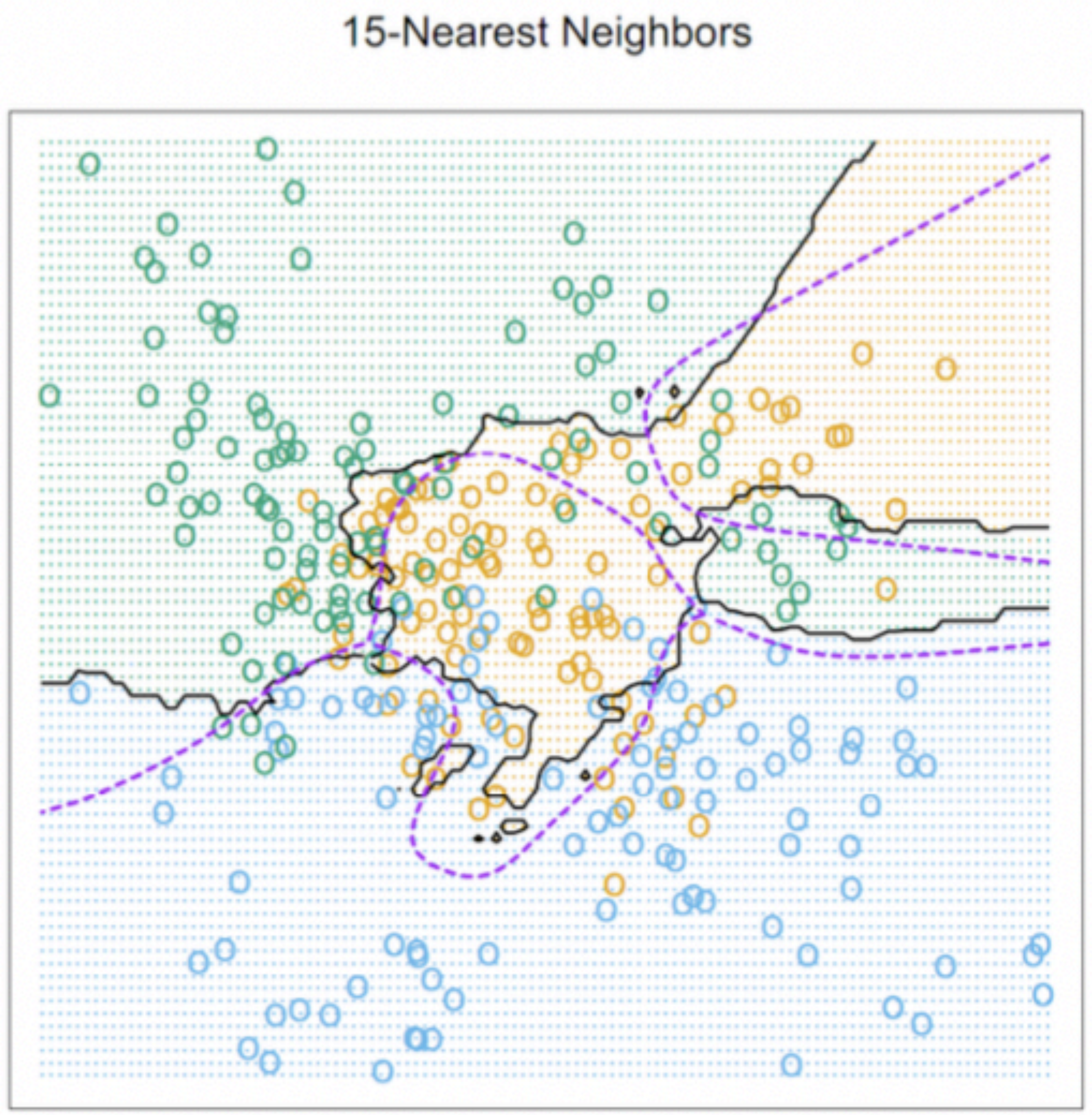


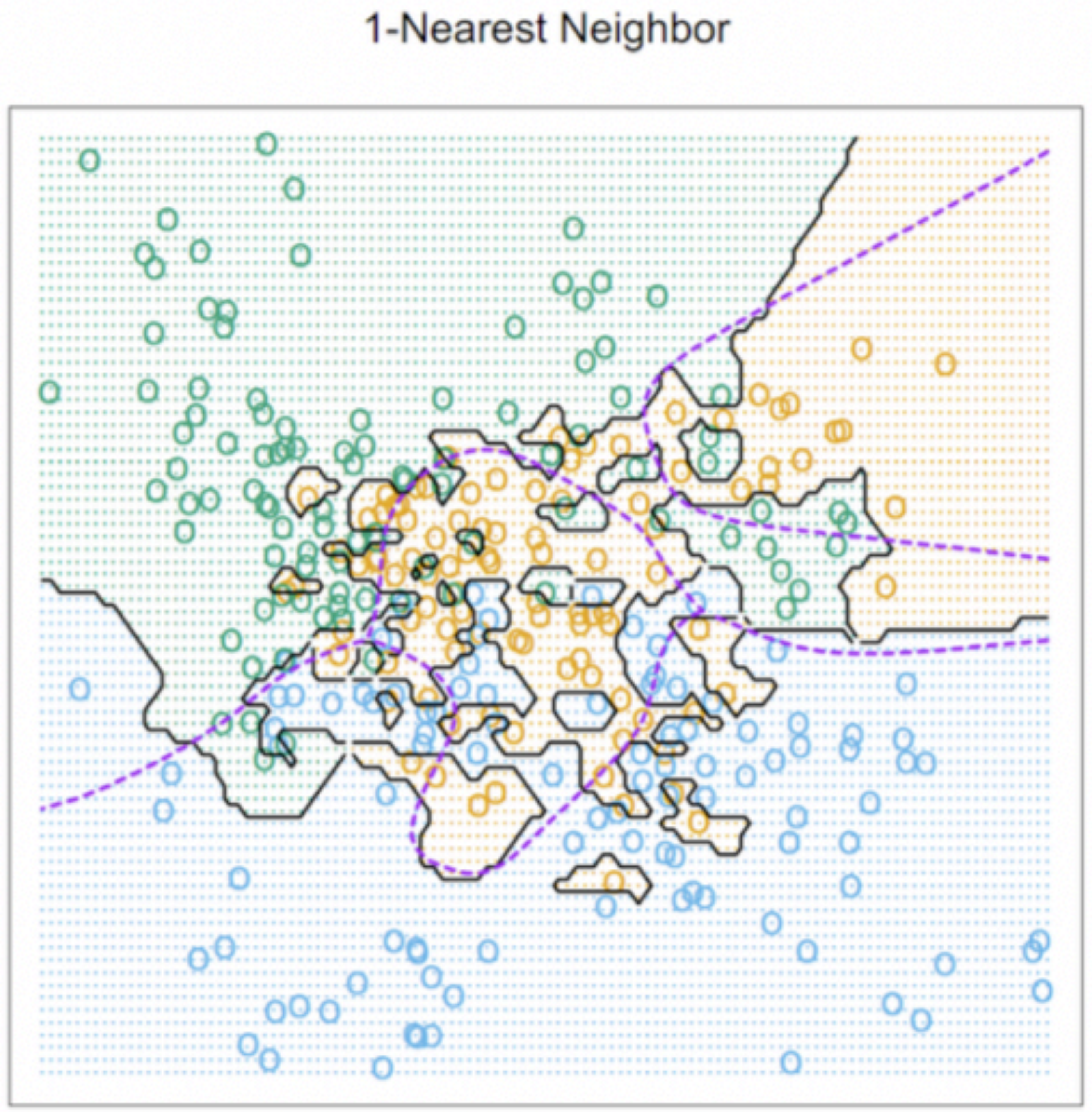
Binary kNN Classification (k=1)
Binary kNN Classification (k=5)
Binary kNN Classification (k=25)
x2
x1

I k의 결정

- 너무 큰 k
  - 미세한 경계부분 분류가 아쉬울 것.

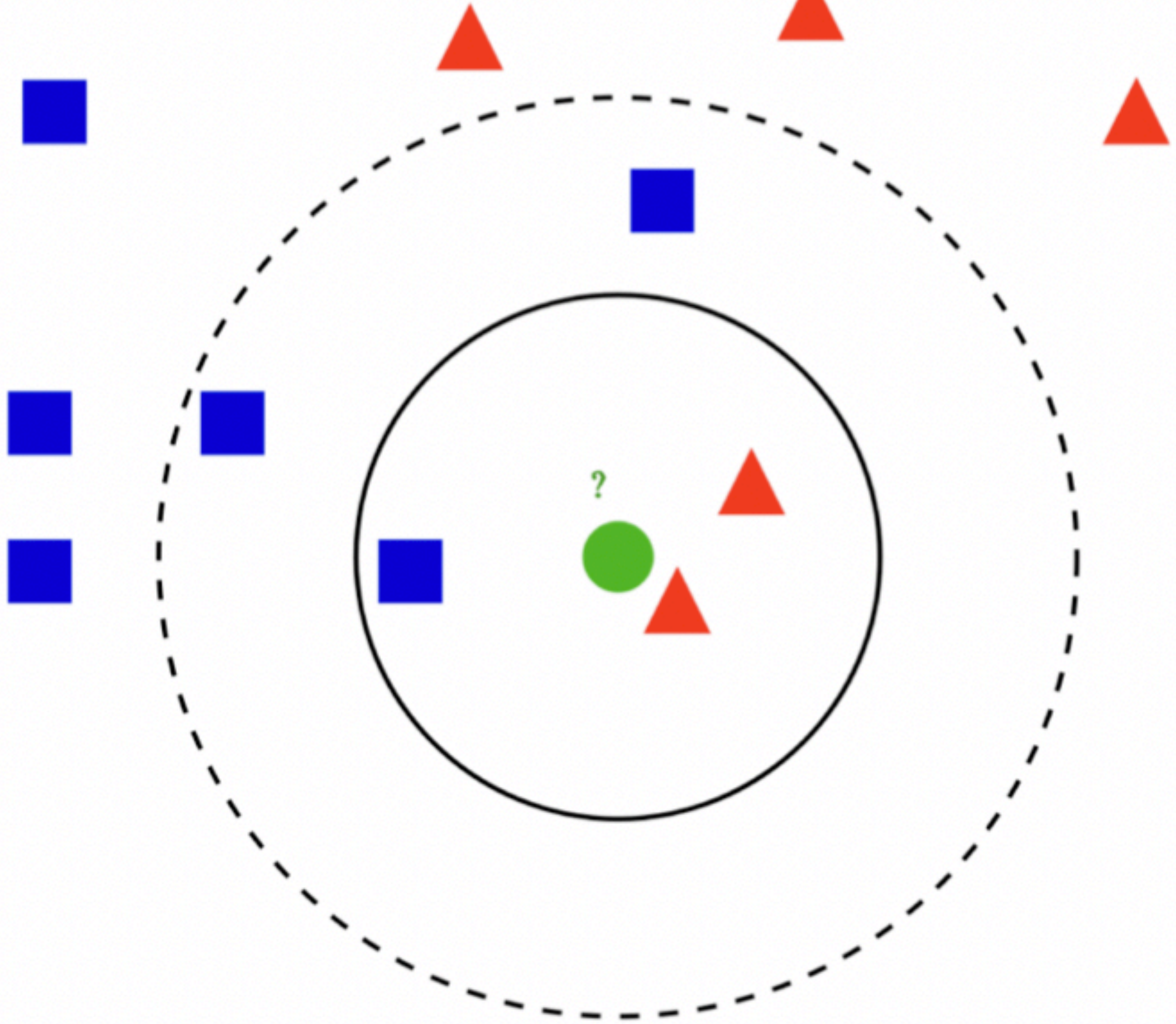
- 너무 작은 k
  - 과적합 우려
  - 이상치의 영향을 크게 받을 것.
  - 패턴이 직관적이지 않을 것.





I k-Nearest neighborhood

- k는 어떻게 정하는가?
  - 너무 큰 k
    - 미세한 경계부분을 잘못 분류할 것
  - 너무 작은 k
    - 이상치의 영향을 크게 받을 것.
    - 패턴이 직관적이지 않을 것.



- Majority voting
  - Blue가 red에 비해 훨씬 많다면?
  - 거리에 반비례하는 Weight를 줄 필요가 있음

- 중요한 변수와 불필요한 변수가 섞여 있다면?
  - 중요한 변수를 선별할 필요가 있음.

I k-Nearest neighborhood

- 종속 변수
  - 범주형 변수
    - k-nearest neighbors 중 가장 많이 나타나는 범주로 y를 추정.
    - Tie 문제를 막기 위해 k는 홀수로 정하는 것이 좋다.

  - 연속형 변수
    - k-nearest neighbors의 대표값 (평균)으로 y를 추정.
    - Inverse distance weighted average 고려 가능.

k-Nearest neighborhood

- 거리는 어떻게 구하나?
  - 설명 변수
    - 범주형 변수
      - Hamming distance $= D_H = \sum_{j=1}^{J} I(x_j \neq y_j)$
      - 예시) '1011101'과 '1001001'사이의 해밍 거리는 2
    - J개의 연속형 변수, j=1, … , J
      - Euclidian distance = $\sqrt{\sum_{j=1}^{J} (x_j - y_j)^2}$
      - Manhattan distance = $\sum_{j=1}^{J} |x_j - y_j|^2$

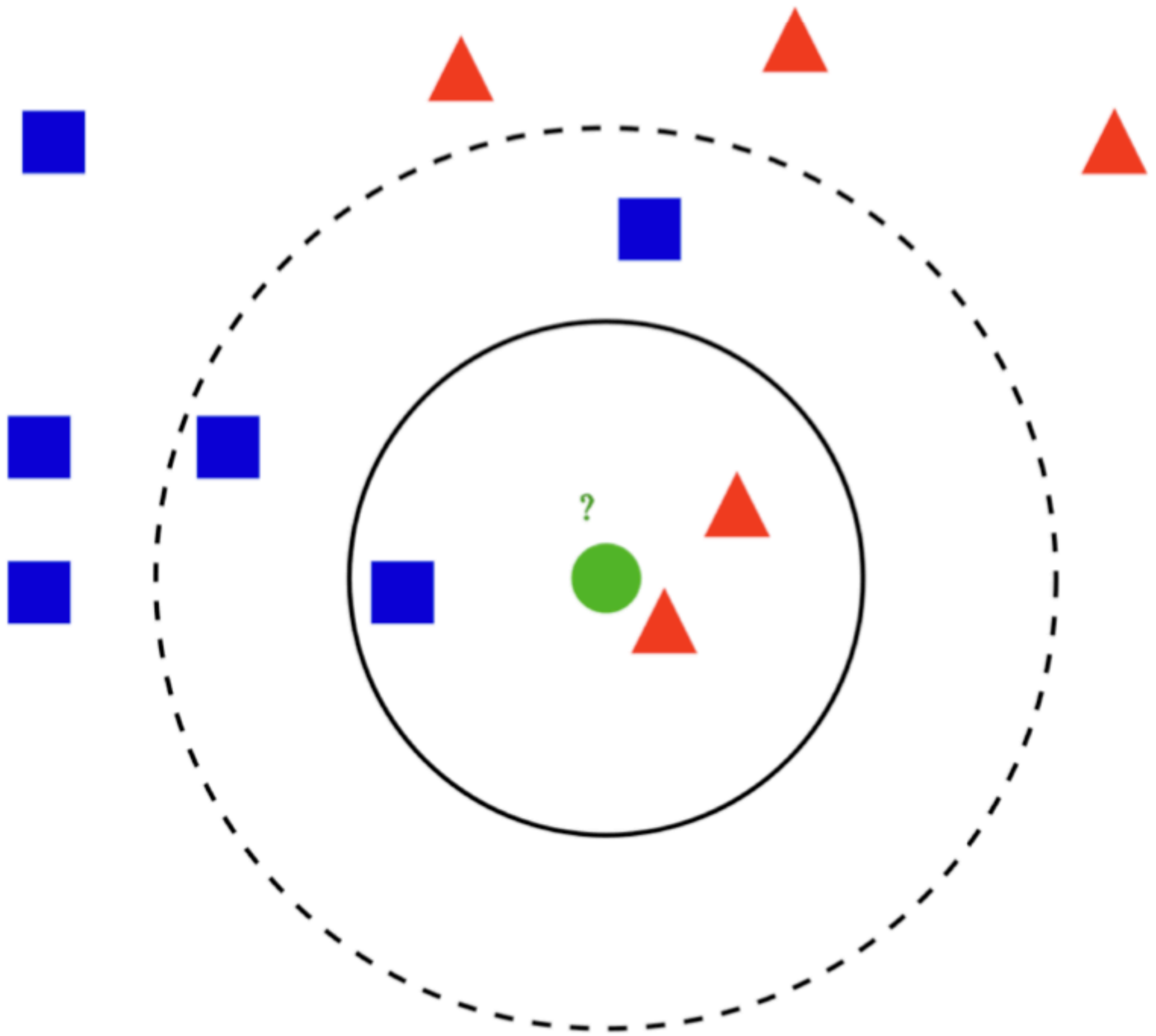
k-Nearest neighborhood

- 점 (x, y), N개의 Training 관측치 (X_i, Y_i), i= 1,…,N에 대하여,
  - $\left(X_{(1)}, Y_{(1)}\right), \ldots, \left(X_{(n)}, Y_{(n)}\right)$
    - 다음 조건에 따라 정렬되어 있음.
      - $d\left(X_{(1)}, x\right) \leq \ldots \leq d\left(X_{(n)}, x\right)$

- Distance $d(a, b)$의 선택
  - 범주형 변수
    - Hamming distance
  - 연속형 변수
    - Euclidian distance, Manhattan distance

I k-Nearest neighborhood

- 종속 변수
  - 범주형 변수 m=1,…,M
    - 근처 k개 중에 가장 많은 범주를 선택.
    - $\hat{p}_m = \frac{\sum_{i=1}^{k}(Y_{(i)}=m)}{k}$
    - $\hat{y} = \underset{m=1,\ldots,M}{\operatorname{argmax}}\, \hat{p}_m$



I k-Nearest neighborhood

- 종속 변수
  - 연속형 변수
    - 근처 k개의 평균을 선택
    - $\hat{y} = \frac{\sum_{i=1}^{k} Y_{(i)}}{k}$
    - Inverse distance weighted average 고려
      - $\hat{y} = \frac{\sum_{i=1}^{k} \frac{1}{d\left(X_{(i)}, x\right)} \cdot Y_{(i)}}{k}$