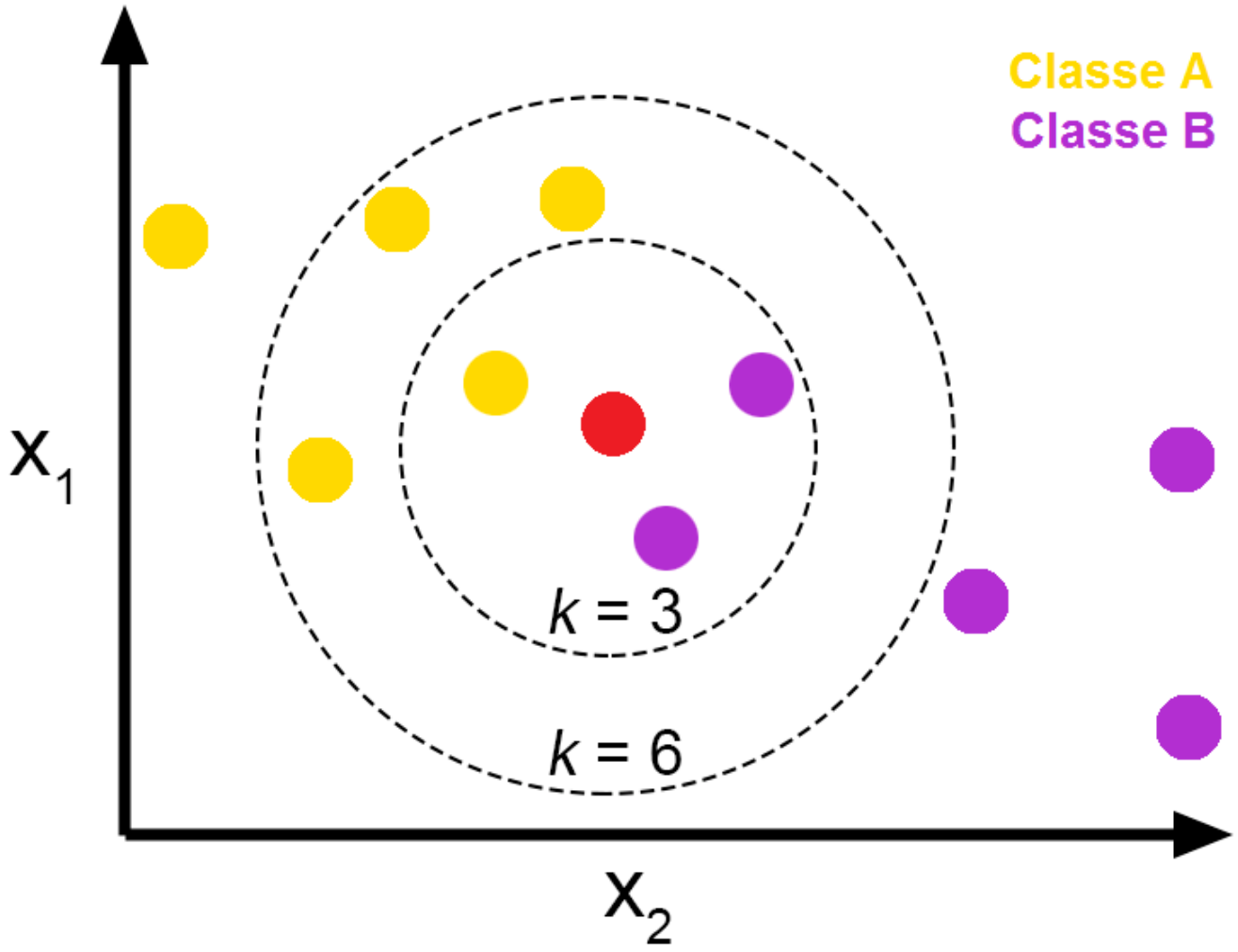
KNN(K-Nearest Neighbor, K-최근접 이웃)

1.설명

지도 학습 알고리즘 중 하나로, 특정 데이터가 주어지면 주변(이웃) 데이터를 살펴본후 더많은 데이터가 포함되어 있는 범주로 분류하는 방식.주변 가까운 데이터를 이용해 특정 데이터를 유츄하는 방식.



출처 : towardsdatascience

새로운 데이터 (붉은 점)이 주어졌을때 Class A와 B 중 어떤데이터로 분류할지 판단하는 문제에서 k=3일 경우, 주변의 3개 데이터(노랑 1, 보라 2)를 살펴 본 후, 주변에 더 많이 포함된 범주로 데이터 분류하여 class B로 유추하게 된다. k=6 일 경우, 주변 6개의 데이터(노랑 4, 보라 2)로 주변 데이터 분포가 바뀌므로 class A로 유추하게 된다.즉 K값에 따라 분류 결과가 바뀌게 된다. 그러므로 K 값을 어떻게 정하느냐에 따라 결과 값이 바뀌므로 **K 값의 적합한 지정이 중요하다. 너무 작아서도 커서도 안된다. 또한 짝수일 경우 동점으로 인한 결과 도출 실패를 방지하기 위해 주로 홀수로 지정한다.**

2.특징

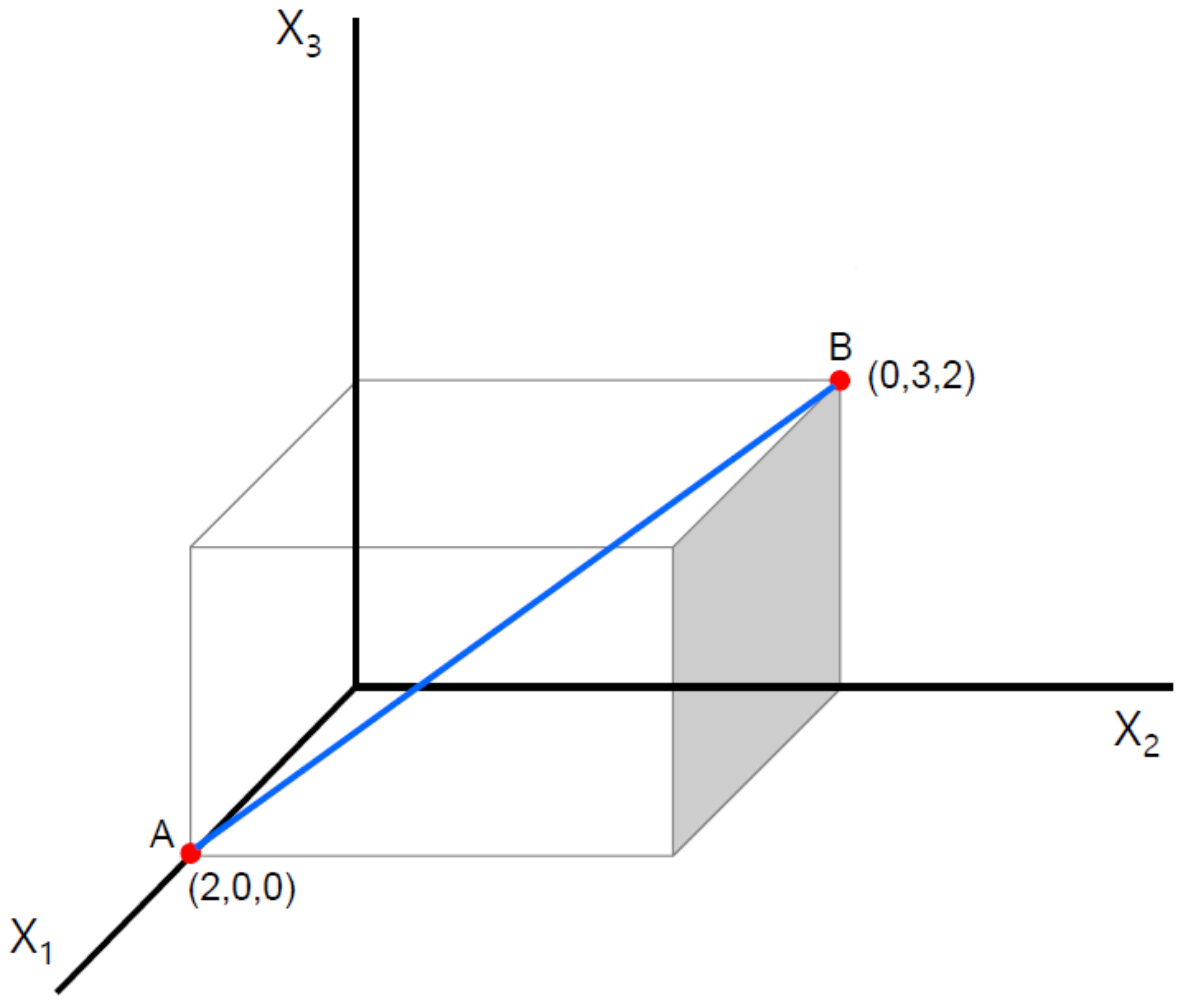
2-1.Lazy Model

KNN은 훈련이 필요 없다는 것이 특징이다. 다른 알고리즘은 훈련 데이터를 기반으로 모델을 만들고 테스트 데이터로 테스트를 한다. 하지만, KNN은 새로운 데이터가 주어지면 그때야 주변의 K개의 데이터를 보고 새로운 데이터를 분류한다. 즉, 사전 모델링이 필요없고 real-time예측이 이루어진다. 따라서 모델을 별도로 구축하지 않기 때문에 **Lazy Model**이라고 부른다. 그러므로 SVM이나 선형 회귀보다 빠르다.

2-2.거리 계산

KNN에서 데이터간의 거리를 구해야하는데, 이때 두가지 방법이 있다.

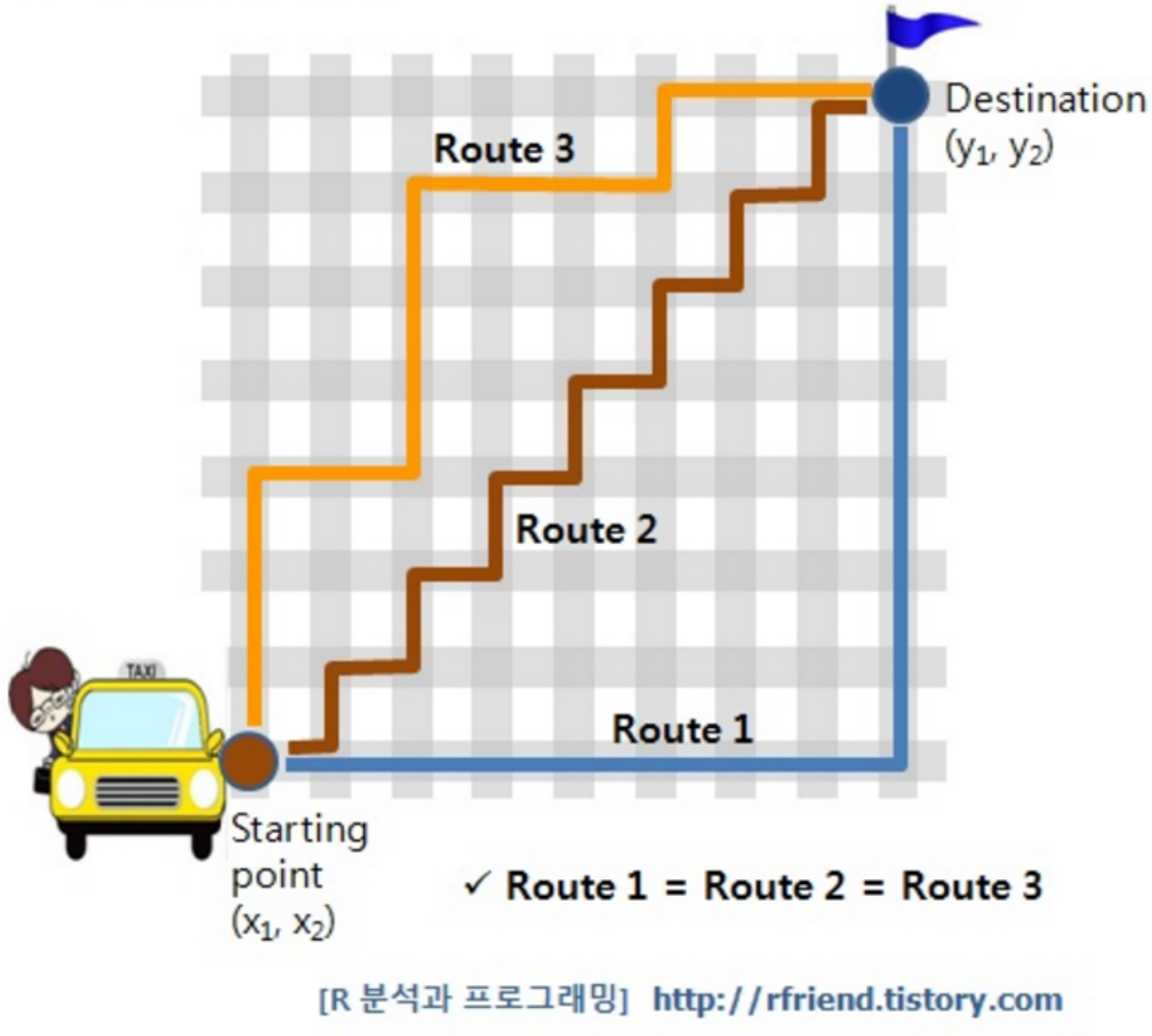
- 유클리드 거리(Euclidean Distance) : 일반적으로 점과 점 사이의 거리를 구하는 방법이다.



출처: ratsgo's blog

$$d_{(A,B)} = \sqrt{(0-2)^2 + (3-0)^2 + (2-0)^2} = \sqrt{17}$$

- 맨해튼 거리(Manhattan Distance) : X축,Y축을 따라 간거리. x축 n1 만큼, y축 n2 만큼 간거리를 의미한다.



출처: friend.tistory.com

2-3. 장단점

- 장점 : 단순하고 효율적이고 기저 데이터 분포에 대한 가정을 하지 않는다.훈련 단계가 빠르다. 수치 기반 데이터 분류 작업에서 성능이 우수하다.
- 단점 : 모델을 생성하지 않아 특징과 클래스간 관계를 이해하는데 제한적이다. 적절한 k선택이 필요하다. 데이터가 많아지면 분류 단계가 느리다. 누락데이터를 위한 추가 처리가 필요하다.

적합한 사례

- 데이터가 너무 많지 않고, 구분과 분류가 명확한 데이터에 적합한 것.
- Ex 당도와 아삭함에 따른 채소분류, 암진단.