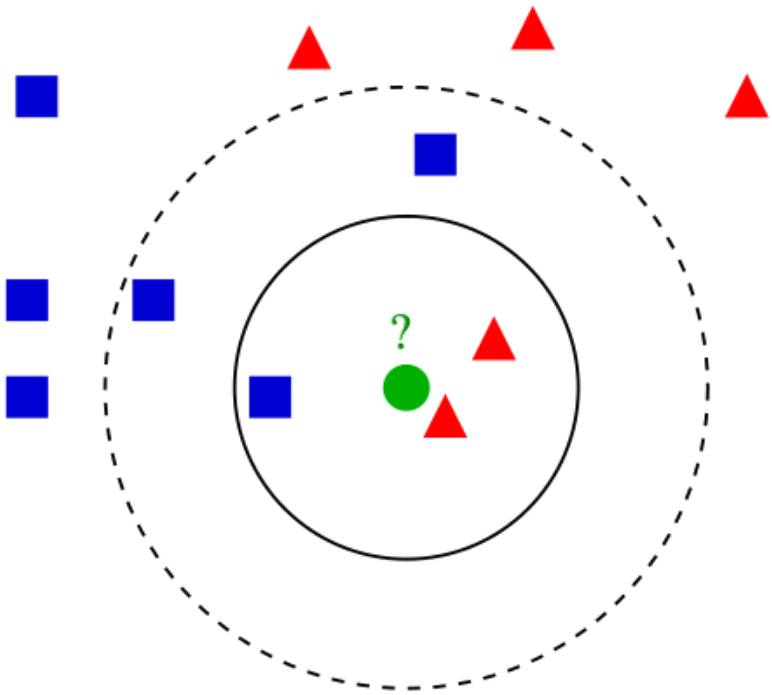
KNN 알고리즘

KNN 알고리즘 (k 최근접 이웃 알고리즘)

KNN알고리즘 정의

KNN알고리즘은 분류나 회귀에 사용되는 비모수 방식으로, **새로운 데이터를 입력으로 받았을 때 가장 가까이 있는 k개의 데이터 중 가장 많이 속하는 class로 classification하는 알고리즘.** 아래와 같은 경우에는 빨간색 세모가 2개로 파란색 네모 class보다 더 많으므로, 검증점인 초록색 동그라미는 빨간색 세모 class로 분류됨.



(비모수 통계란 통계학에서 모수에 대한 가정을 전제로 하지 않고, 모집단의 형태에 관계없이 주어진 데이터에서 직접 확률을 계산하여 통계학적 검정을 하는 분석법.)

거리 척도

새로운 데이터를 입력으로 받은 후, 가장 가까이 있는 k개의 데이터를 구하기 위해선 거리 척도를 설정해야 함 → 연속 변수에서 가장 많이 사용되는 거리척도는 **유클리디안 거리**.

유클리디안 거리

n차원 공간에서 두 점간의 거리를 알아내는 공식으로, L2 Distance라고도 한다.

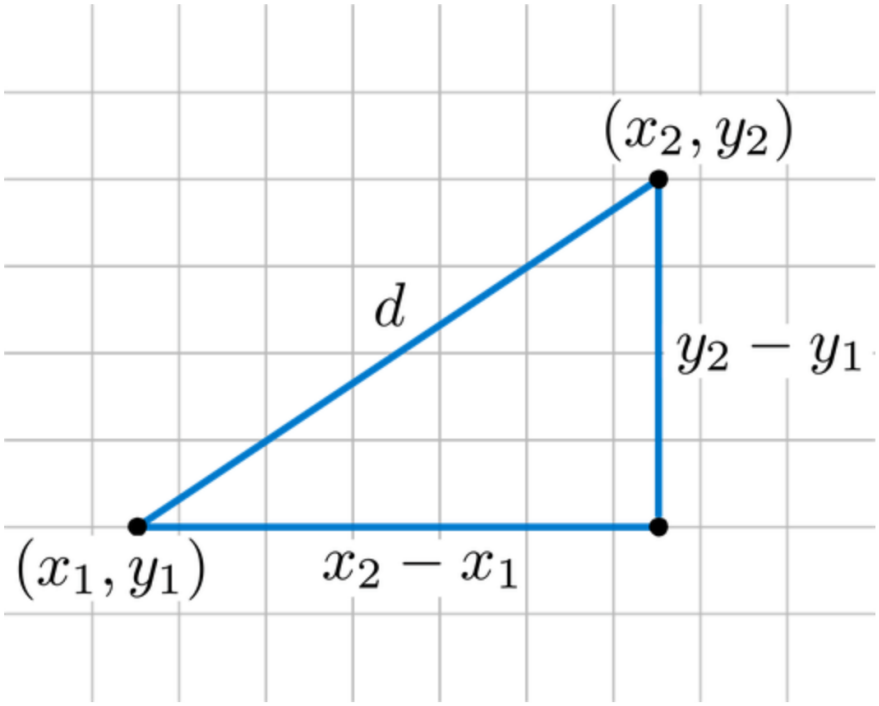
피타고라스의 정리가 이미 만들어진 삼각형을 이용한 공식이라 한다면, 유클리디안 거리는 삼각형을 만들어서 계산을 한다는 개념.

즉, 피타고라스의 정리는 2차원 공간 내 두 점간의 거리를 계산하지만 유클리디안 거리는 다차원 공간 내 두 점간의 거리 계산가능.

피타고라스 정리 공식

$$a^2 + b^2 = c^2$$

피타고라스 정리를 이용하여 거리계산 공식



$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

유클리드 거리 공식

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2 + \dots}$$

단점

"과반수 의결"에 따른 분류의 단점은 항목 분포가 편향되는 경우, 더 빈번한 항목의 데이터가 새로운 데이터의 예측을 지배하는 경향이 있음 → 이 문제는 검증점과 k개의 최근접 이웃 각각의 거리에 반비례하는 가중치를 주므로써 해결할 수 있음.

출처

- 위키백과 knn 알고리즘
- 유클리디안 거리