지도 학습 알고리즘

1. **최근접 이웃 (K-Nearest Neighbor)**
- 레이블(정답)이 없는 예시를 **분류**하기 위한 알고리즘.
- 가장 고전적이고 직관적이라는 특징이 있음

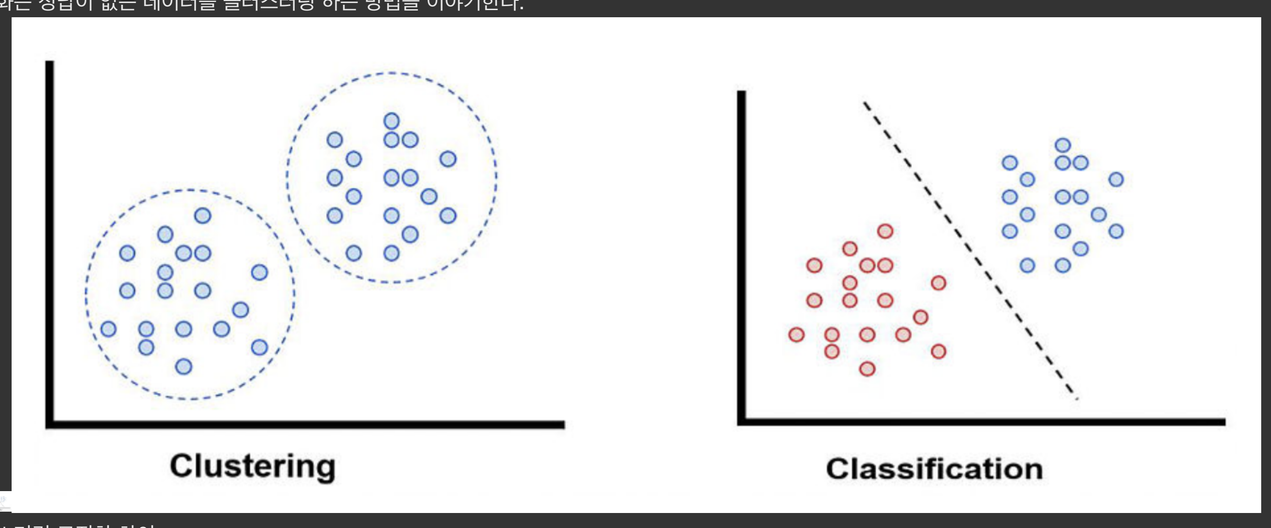
지도 학습에는 크게 9가지 정도 대표적인 알고리즘이 있는데, 그 중에서도 KNN은 가장 고전적이면서 직관적인 방법으로 널리 알려져 있다

분류와 군집화

분류 : 라벨이 있음, 지도 학습

군집화 : 라벨이 없음, 비지도 학습

먼저 KNN에 대해서 알기 전에 분류와 군집화의 차이점을 알아볼 필요가 있다.
분류는 말 그대로 정답이 있는 데이터(라벨)을 지도학습을 통해 분류하는 것을 이야기하고,
군집화는 정답이 없는 데이터를 클러스터링 하는 방법을 이야기한다.



클러스터링 군집화 차이

여기서 KNN 알고리즘은 라벨이 있는 데이터 속에서 라벨이 없는 데이터를 어떻게 **분류할 것인지**를 찾기 위한 **분류**에 속한다

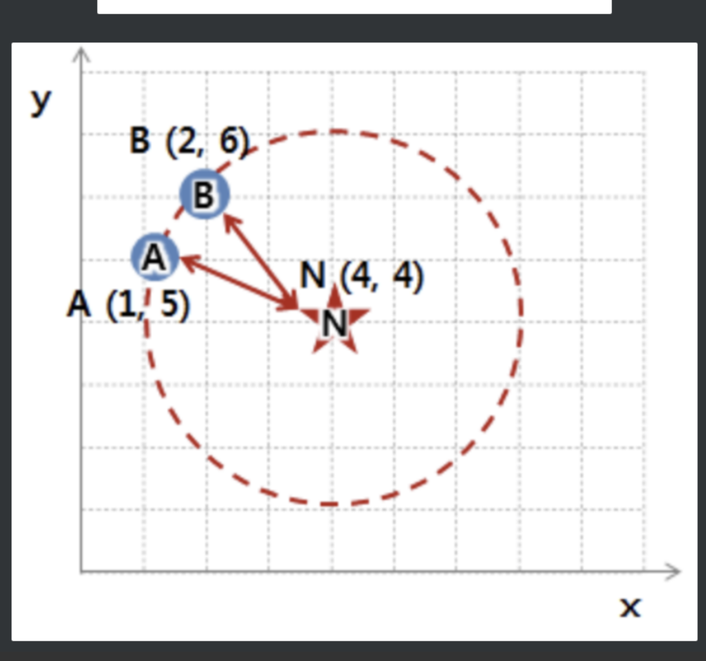
KNN 최근접 이웃법

KNN의 기본적인 로직은 다음과 같다.
- 새로운 샘플 **가까운 거리**에 있는 몇 가지 라벨을 함께 본다.
- 그리고 가장 **빈도가 높은 것**을 통해 분류한다.

그럼 여기서, **가까운 거리와 빈도가 높은 것**의 기준을 알아보자.

(1) 가까운 거리 척도의 단위 — 표준화 (유클리드 거리)

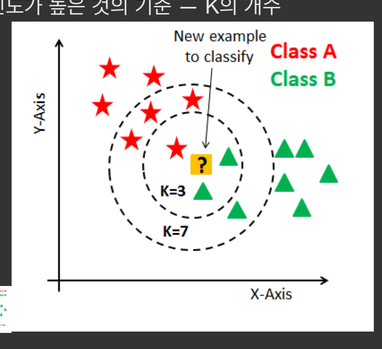
$$\sqrt{(Ax - Bx)^2 + (Ay - By)^2}$$



ex1) A-N간의 유클리드 거리는 루트10 -> 3.1xx
ex2) B-N간의 유클리드 거리는 루트8 -> 2.8xx

-> 여기서 N은 B가 더 가까운 것으로 측정이 된다.

(2) 빈도가 높은 것의 기준 — K의 개수



- 여기서 K=3일 때는 새로운 샘플이 세모로 분류되지만 K=7일 때는 별표로 분류된다.
  여기서 주의할 점은 **K의 개수가 짝수일 경우에는 동점이 발생할 수 있기에 홀수로 지정해줘야 한다는 것**이다.

(3) 장점
- 단순하다. 성능이 좋다. 모델 훈련 시간이 필요 없다

출처 :
https://medium.com/@john_analyst/knn-%EC%B5%9C%EA%B7%BC%EC%A0%91-%EC%9D%B4%EC%9B%83-%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98-b397a0b2030e

샘플 코드:
http://parrot_work:8888/notebooks/users/voiz2men/test/KNN.ipynb

표시