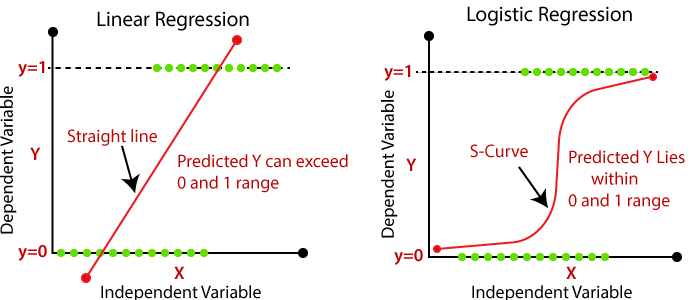
**LogisticRegression**

회귀를 사용하여 데이터가 어떤 범주에 속할 확률을 0에서 1사이 값으로 예측하고 그 확률에 따라 가능성이 더 높은 범주에 속하는 것으로 분류해주는 지도학습 알고리즘.

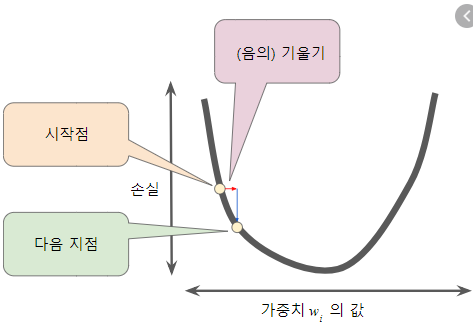


장점 : 선형회귀를 사용하면 확률이 음과 양의 방향으로 무한대까지 뻗어 나간다. X축을 학습시간이라고 하고 y를 시험 합격 확률이라고 가정하면 선형회귀에서는 특정값 이하의 학습시간은 확률이 0이 안된다, 이런 문제를 해결한 것이 로지스틱 회귀이다.

실제 많은 자연, 사회현상에서 특정 변수에 대한 확률값이 선형이 아닌 s-커브 형태를 따르는 경우가 많다. 이러한 S 커브를 함수로 표현해낸 것이 로지스틱 함수

**SgdClassifier (Stochastic Gradient Descent) 확률적 경사 하강법**

예측모델은 손실을 줄여 예측하게 될 확률을 높이는데 SGD는 손실을 줄이기 위한 최적값을 찾기 위해 무작위로 아무 값이나 넣어보면서 최적값을 구하는 분류보델이다,



확률적 경사 하강법은 훈련 세트에서 랜덤하게 하나의 샘플을 선택하여 가파른 경사를 조금 내려간다. 다음 훈련 세트에서 다시 샘플을 하나 추출하여 경사를 조금 내려가고, 전체 샘플이 소진될 때까지 이를 반복한다. 샘플이 전부 소진되는 한 사이클을 에포크라고 부른다. 그리고 이 에포크 이후에 전체 샘플이 최적점에 도달하지 못했을 경우 계속해서 전부 도달할 때까지 에포크를 반복한다.

KNeighborsClassifier

**Knn ( k-최근접 이웃 알고리즘)** :

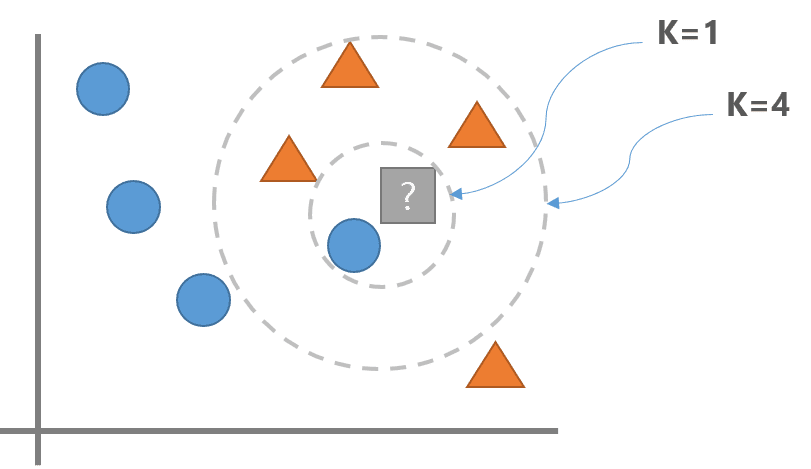
새로운 데이터를 입력받았을 때 가장 가까이 있는 것이 무엇이냐를 중심으로 새로운 데이터를 분류하는 알고리즘이다.

장점 : 단순하고 효율적이며 훈련 단계가 빠르다, 기저 데이터 분포에 대한 가정을 하지 않는다

단점 : 주변 데이터의 개수 즉 k를 무엇으로 지정하냐에 따라 분류예측 성능이 크게 좌우됨.

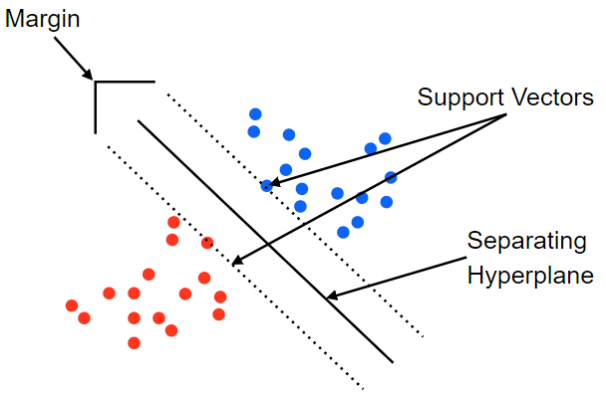
예를 들면, 새로운 데이터 회색 네모를 새로 분류하고자 하는데,주변이 다 주황색이므로 주황색으로 분류되는 것이 더 알맞을 것이다. 하지만 k =1로 할 경우 물음표 박스에 가장 가까운 한 개의 요소만 바라보게 될 테니 파란색으로 분류하게 된다.

knn알고리즘은 지도학습의 분류 알고리즘의 하나로 로직이 간단하여 구현하기 쉽다. 하지만 학습 모델이 따로 없고, 전체 데이터를 스캔하여 데이터를 분류하기 때문에 데이터의 양이 많아지면 분류 속도가 현저하게 느려진다. 그래서 게으른 알고리즘이라고 한다.



Svm (support vector machine)

Support vector와 hyperplane(초평면)을 이용해서 분류를 수행하게 되는 알고리즘



장점 : 신경망보다 사용이 간결함. 범주나 수치 예측 문제에 사용이 가능, 오류 데이터에 대한 영향이 없다. 과적합 되는 경우가 적다.

SVM은 다양한 데이터 분포에서도 잘 작동하는 분류기법 중 최상의 기법으로 널리 이용되고 있음.

최상의 기법으로 평가받는 이유는 정확도 측면에서 다른 분류 기법들 보다 우수한 결과를 보여주기 때문,

단점 :의사결정나무처럼 직관적인 해석이 불가능하다 (어떤 이유로 데이터들이 분류됐는지 알 수가 없다), 학습속도가 느림, 최적의 모델을 찾기 위해서 커널과 모델에서 다양한 테스트가 필요하다. 따라서 여러 연산이 필요하고 입력 데이터 셋이 많을 경우에 학습 속도가 느립니다.