**Logistic Regression**

선형 분석과 가장 큰 차이점은 ‘값’이 아닌 ‘확률’로서 분류한다는 점이다.

기존의 선형분석에서는 단순히 입력한 값을 그대로 사용하다보니 평균보다 차이가 큰 값이 입력되면 종속변수의 값이 크게 달라져 버리는 문제점이 있다. 이는 허용할 수 있는 한계치를 벗어나버리는데 로지스틱에서는 시그모이드라는 계산을 사용하여 입력값을 비율로 변환하여 항상 0과 1사이 값으로 한정시키기 때문에 범위를 벗어나 정확도가 떨어지는 오류를 방지한다.

장점

값이 아닌 비율로 계산하는 만큼 일반적인 범주형 데이터를 받을 수 있다.

분류모델일 뿐만 아니라 확률 또한 제공한다 이는 이산형 뿐만 아니라 다항형 또는 분화형으로 계산할 수 있어 예측과 분류 2가지 모델을 만들 수 있다.

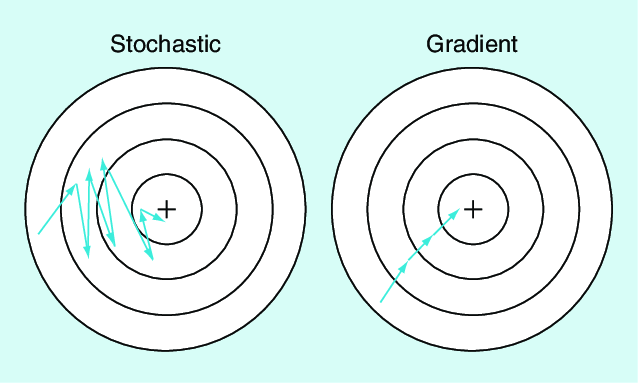
단점

로지스틱 회귀는 많은 사람들에게 사용되고 있지만 교호 작용을 따로 추가해야 한다는 것과 같이 표현적인 제한이 있고 다른 모델들이 더 예측 성능이 좋을 수 있습니다.

가중치 해석이 덧셈식이나 곱셈식이 아니기 때문에 해석이 더 어렵다는 것입니다.

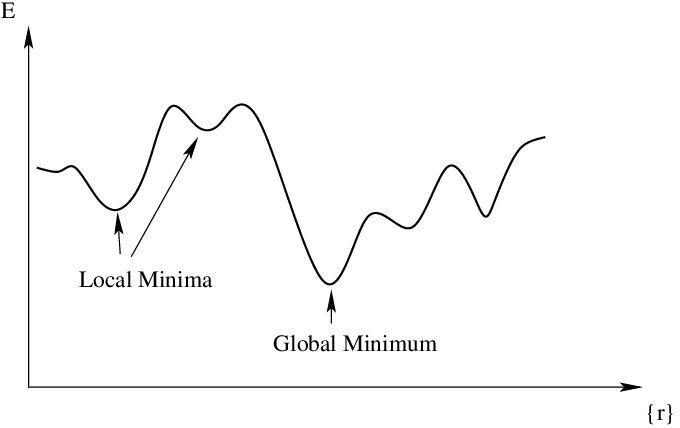
Sgd 전체 data를 가지고 한번의 loss function을 계산하는 것이 아니라 batch단위로 loss function을 계산

Sgd와 gd의 비교



장점: batch단위로 loss를 계산하기 때문에 loss function을 여러 번 빨리 계산할 수있으며

Local minima에 쉽게 빠지지 않고 global minima를 찾을 확률이 높다



단점 : 학습률이 낮으면 곧장 최적화하지 못하고 지그재그로 이동하게 되면서 지역 최솟값이 갇혀 빠져나오지 못하는 경우가 있고, 학습률이 높으면 최적화 자체를 실패할 수 있는 문제가 있다.

**Knn ( k-최근접 이웃 알고리즘)** :

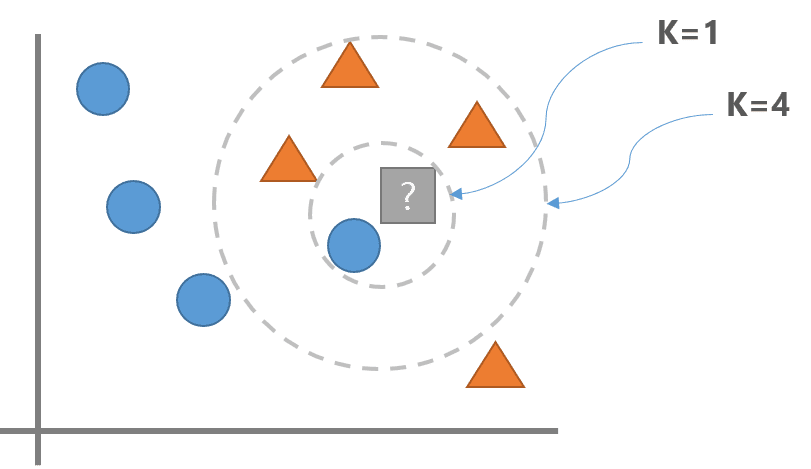
새로운 데이터를 입력받았을 때 가장 가까이 있는 것이 무엇이냐를 중심으로 새로운 데이터를 분류하는 알고리즘이다.

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.

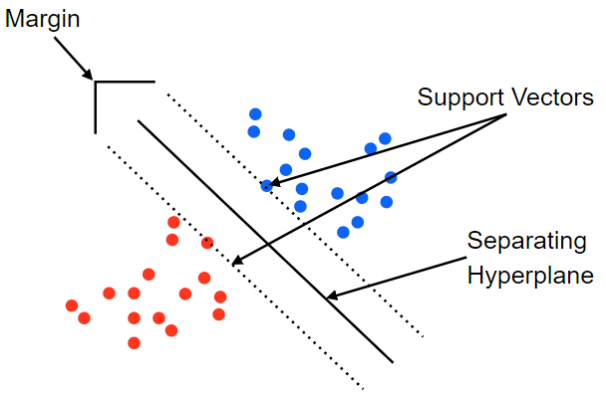
예를 들면, 새로운 데이터 회색 네모를 새로 분류하고자 하는데,주변이 다 주황색이므로 주황색으로 분류되는 것이 더 알맞을 것이다. 하지만 k =1로 할 경우 물음표 박스에 가장 가까운 한 개의 요소만 바라보게 될 테니 파란색으로 분류하게 된다.

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.