# Linear classification

선형회귀 방법론을 이용해서 어떻게 이진 분류기를 설계할 수 있는지 공부할 것이다.

설명할 순서: Linear regression 🡪 Logistic regression 🡪 binary classification 🡪 multi class classify

선형회귀는 binary classification에 적용하기에는 단점이 있기 때문에 Logistic을 사용하여

binary classification을 하며, 이를 확장해서 multi class classify를 할 수 있다.

**Linear regression에 대한 복습**

학생 1이 10시간 공부하고 수업을 5번 들었더니 score는 90.

그렇다면 10시간 공부하고 수업을 3번 들었다면 score는?

- 데이터 기반의 예측을 하고 싶을 때 linear regression을 사용한다.

1. 가설 설정

위의 예에서는 wx1 + wx2 = y

1. Cost : 에러 값의 평균
2. Gradient decent : cost가 가장 작은 w값을 찾기.

Binary classification

Grouping 할 수 있는 것들을 class라고 할 수 있다.

Ex) 사기거래 detection, 합법, 불법에 해당하는 상황 등

( 지금까지는 숫자로 구성 되어있었다. x1=10, x2=5, score=90 등의 )

<-> 범주형 클래스 (범주형 라벨을 가진 데이터)

= 숫자가 아닌 비정형 타입 Ex) 스팸이냐 아니냐

수학적으로 가능하도록 반드시 인코딩하기. ex) 1 or 0으로

6p

Pass를 1, Fall을 0으로 인코딩

2,3,4 시간 공부하면 불합, 6,7,8시간 공부하면 합격인 데이터가 존재할 때,

10시간 공부하는 사람의 합격여부는??

7p **linear regression을 이용해서 분류하기**

가설 설정 WX를 이용해서

WX < 0.5이면 불합, WX>0.5이면 합격으로 binary classification으로 할 수 있다.

8p 하지만 linear regression을 바로 classification으로 활용하면 단점이 존재한다.

Noise(outlier)인 50시간도 합격으로 분류된다.

이 noise 데이터에 선형회귀함수를 이용해 비용함수를 설계하면 noise에 bias되있는 함수가 나온다. 이 함수를 이용해 WX>0.5를 기준으로 binary classification 하면, 6시간 공부한 사람은 합격이어야 하지만 불합격이라는 잘못된 결과가 나온다.

Y를 0 or 1로 두었다.

하지만 예측 값이 1보다 크고, 0보다 작을 수도 있다.

가설에 문제가 있는 것이 아닐까?

🡪 가설을 재설계. 0 <= H(x) <= 1

sigmoid함수를 가설설정으로 변경 = **Logistic regression**

**11p sigmoid함수**

Sigmoid 함수는 0과 1을 주로 갖는 형태.

H(x)=WX 를 = Z로 두고, sigmoid function α(z) 에 넣으면, z값은 **0< z <1** 이 된다.

모든 값을 0 과 1 사이로 bound시키는 것이다.

🡺 **H(x) = 1/1+e(-wx)** 🡨 z=wx

**🡺 0 <= H(x) <= 1**

13p

# 가설설정에 맞는 cost function을 적용해보기.

Cost function은 어떤 곳에서 시작하든지 최소값에 도달하는 형태여야 한다.

Linear regression의 cost 함수는 Global minimum(극소점)을 찾을 수 있었다.

하지만 sigmoid의 cost함수는 그림과 같은 형태가 그려져서, global minimum을 찾을 수 없다.

Local minimum(지역 극소점)에 빠져서 빠져나올 수 없기 때문에.

🡪 cost 함수를 재설계--

예측 값이 얼마나 정답에 가까운가를 측정하기.

정답에 가까워질수록 cost function값은 작고, 멀어질수록 cost function값은 크게 설계하기!

- y=1 일 때, 예측 값 H(x)=1 이라면, cost 값은 작다.

y=1: 파란색선을 보면 H(x)=1 일 때 cost는 0이다.

- y=1 일 때, 예측 값 H(x)=0 이라면, cost 값은 크다.

y=1: 파란색선을 보면 H(x)=0 일 때 cost는 무한대이다.

- y=0 일 때, 예측 값 H(x)=0 이라면, cost 값은 작다.

y=0: 빨간색선을 보면 H(x)=0 일 때 cost는 0이다.

- y=0 일 때, 예측 값 H(x)=1 이라면, cost 값은 크다.

y=0: 파란색선을 보면 H(x)=1 일 때 cost는 무한대이다.

🡪 두 cost 함수를 하나로 만들면, **𝑪(𝑯(𝒙),𝒚)=−𝒚𝒍𝒐𝒈(𝑯(𝒙))−(𝟏−𝒚)𝐥𝐨𝐠(𝟏−𝑯(𝒙))**

이 cost function은 Global minimum(극소점)을 찾을 수 있다.

🡪 따라서 **Gradient descent algorithm**을 적용하여 W를 구할 수 있고,

🡪 이렇게 구해진 W를 통해서 **binary classification**을 할 수 있다.

20p 0.5를 기준으로 분류하면 그림과 같이 분류된다.

이진분류기를 이용해서.

# Multi-class classification

Grade를 숫자가 아닌 A, B, C로 바꾸기.

총 클래스는 3개. Grade를 예측하기.

- 아래 세번의 질문을 통해 클래스를 예측하게 된다. = binary classification을 여러 번 질문하는 것.

A or not(B,C)

B or not(C,A) 파란색 선을 기준으로

C or not(C)

24p 같은 데이터 X를 분류기 3개 모두에 넣어서 나오는 Y값을 통해 예측할 수 있다.

25p

3개의 변수가 존재할 때, 매트릭스 형태로 분류할 수 있고,

이 3개의 분류기는 각각 A, B, C분류기에 해당한다.

28p

분류기 3개를 합쳐서 3,3 매트릭스로 표현할 수 있다.

W\*X 한 값을 sigmoid를 통해서 1과 0으로 표현하게 된다.

주의할 점

2, 0, 0 을 sigmoid 🡪 1, 1, 0

정답은 1, 0, 0 (1이 나온 것만 해당 클래스.) 이지만 1이 하나가 아닌 두개가 나와벼렸다.

Sigmoid를 바로 통과시켜서 나온 결과는 multi classification에 사용하기에는 부적합하다.

하나의 클래스에만 1을 띄울 수 있는 방법

🡪 1) Softmax function을 이용한 가설설정

모든 값을 0~1로 만들어주고, 전체 합이 1이되는 확률정규화를 만들어 준다.

2, 1, 0.1를 softmax function에 넣으면 0.7, 0.2, 0.1

🡪 argument 중 가장 큰 값만 1로 바꾸고 나머지는 0으로. 1, 0, 0

= One-hot encoding: 가장 큰 한가지의 값만 1로 변화시키는 방법

특정 클래스에 대해서만 레이블을 만들기 위해 사용

32p

Sigmoid를 하지 않은 H(x)=2, 1, 0.1

🡪 softmax function에 넣어서 0.7, 0.2, 0.1

🡪 One-hot encoding 1, 0, 0 = (a클래스, x, x)

33p Softmax 상황에 대해

**2) Cross-entropy를 이용해 cost 함수 설정하기.**

예측 값이 정답에 가까워질수록 cost function값은 작고, 멀어질수록 cost function값은 크게 설계하기!

34p

Softmax에 의해서 Y= 0,1 🡪 B클래스가 정답

- 가설이 참일 때, 내가 예측한 값도 B가 나왔을 때.

식을 풀어보면 log[0,1]을 파란색log함수에 넣어보면, [∞,0]

이를 element wise 곱(0\*∞, 1\*0)을 해보면 [0,0]. **cost가 0.**

- 가설이 거짓일 때, 내가 예측한 값은 A

식을 풀어보면 log[1,0]을 파란색log함수에 넣어보면 [0, ∞]

(0\*0, 1\*∞)로 값은 [0, ∞]. Cost가 ∞

35p

Cost function은 gradient descent 알고리즘을 이용해서 풀 수 있다.

D : Cross-entropy

𝑆(𝑊𝑋+𝑏) : softmax function H(x) : sigmoid

Li : Linier regression

🡺 Linier regression 값을 softmax를 통과시키고

이 값을 cross-entropy로 loss를 구하기.

: Linear regression 🡪 Logistic regression 🡪 binary classification 🡪 multi class classify

선형회귀는 binary classification에 적용하기에는 단점이 있기 때문에 Logistic을 사용하여 binary classification을 하며, 이를 확장해서 multi class classify를 할 수 있다.

- 숫자로 구성 되어있는 데이터를 linear regression으로 해결해왔다.

- 하지만 이제 0과 1에 해당하는 회귀문제 일 때, 가설 설정을 sigmoid 함수를 이용하는 방법으로 변경 = Logistic regression

- Logistic을 사용하여 binary classification을 할 수 있지만 문제가 있기 때문에

Cost function을 정답에 가까워질수록 cost function값은 작고, 멀어질수록 cost function값은 크게 재설계한다. 이로써 gradient descent를 사용할 수 있게 된다.

- binary classification을 확장해서 multi class classify를 할 수 있게 된다.

위의 내용을 손으로 다 증명할 수 있어야 함.

# 6-2 logistic regression classification 실습

Logistic regression 은 범주형 데이터를 다룬다. (선형회귀와 유사하지만,)

따라서 로지스틱 회귀문제는 🡪 이진분류 혹은 여러 클래스의 분류 문제가 된다.

Logistic regression을 이용해 클래스 분류를 하게 되면 Noise에 더 bias된 함수가 나옴으로써 로지스틱 0.5를 기준으로 분류하면 잘못된 분류결과가 도출된다.

따라서 예측 값을 0~1 사이로 bound 해야 한다. == **sigmoid**함수로 가설설정.

🡪 이렇게 가설설정과 cost를 정하게 되면, gradient descent로 문제해결이 가능하다.

# 6-2 Multiclass classification 실습

각 클래스를 0, 1, 2 등으로 표현.

**\* softmax function**: argmax 중 가장 큰 값만 1로 바꾸고 나머지는 0으로. 1, 0, 0

**\* One hot encoding:** [0,1,0] -> B클래스, [1,0,0] -> A클래스