5장
로지스틱 회귀

3팀 강호연 김범주 정유찬 송지원 김수환

TABLE OF CONTENTS

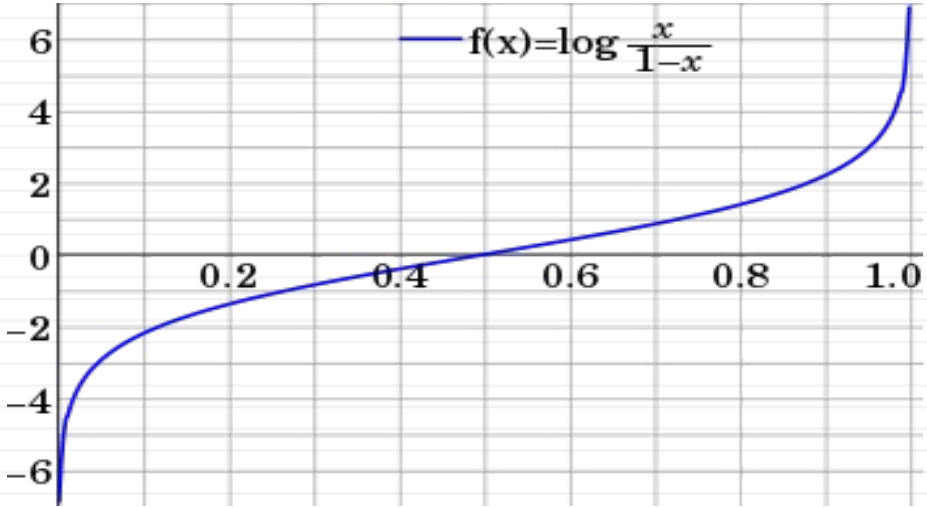
로지스틱 회귀

어원

"Logistic" => Logit function 에서 파생

Q : Logit function(함수) 란?

1. log-odd function과 동일.
   (x : 0<=x<=1, y : All R.)
2. 시그모이드 함수의 역함수



로지스틱 회귀의 원리 (1)

〈 시그모이드 함수가 만들어지는 과정 〉

오즈 비 -> 로짓변환 -> 시그모이드 함수

$$OR(odds\,ratio) = \frac{p}{1-p} (p = 성공\ 확률)$$

- 실패 비율 대비 성공 비율
- 시그모이드 함수는 오즈 비 통계를 기반으로 만들어짐

오즈 비 사용 이유

1. 일반적으로 모집단 크기를 알 수 없다.
   ->전체 크기는 미지수, 선택한 표본들을 통해 정보 얻음

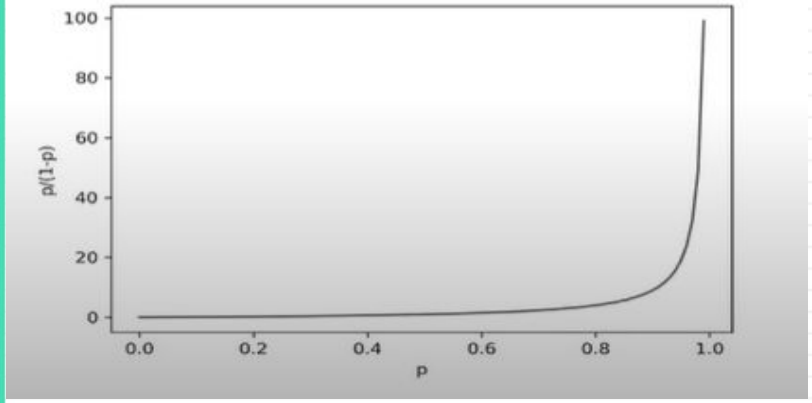
2. 표본의 크기를 선택하여 성공과 실패 횟수를 얻음.
   ->통계 처리시 자주 사용

로지스틱 회귀의 원리 (2)

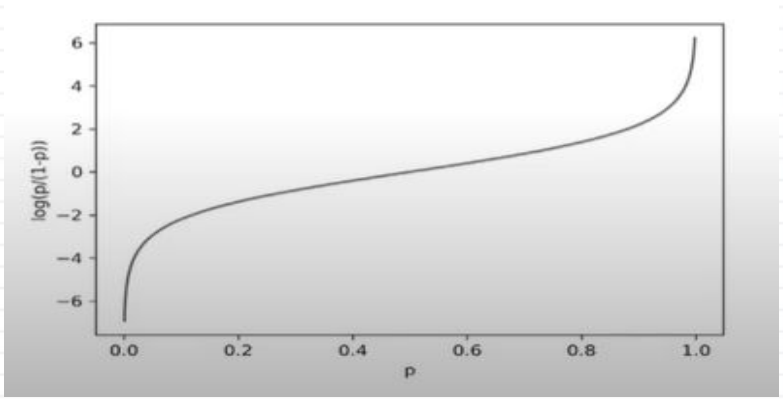
로짓 함수 : 오즈 비에 로그 함수를 취하여 만든 함수

$$logit(p) = \log(\frac{p}{1-p})$$

- p가 0.5일 때 0이 되고 p가 0, 1일 때 각각 무한대로 발산한다. (반대 방향)



오즈 비 함수



로짓 함수

잠깐, 왜 로짓 변환을? ---- 비선형모형의 선형화

$$E(Y|X) = p(X) = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}$$

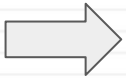
$$= \frac{e^{b_0 + b_1 X}}{1 + e^{b_0 + b_1 X}}$$

선형화 시키는 이유!

- 선형모형에서만 사용할 수 있는 모형의 해석, 확장, 수정 등의 방법을 사용하기 위함
- 비선형모형의 경우, 다루는 방법의 제한이 심하고 새로운 데이터에 민감

잠깐, 왜 로짓 변환을? ---- 범위 제한 X

선형 모델의 특징
y값(종속 변수)의 범위가 실수 전체!



선형 모델의 종속 변수 범위[-infinity, infinity]와 같이
확률 p의 종속변수 범위를 같게 만드는 것이 목적!

확률 p의 범위는 **[0, 1]**

Odds(p)의 범위는 **[0, infinity]**

log(Odds(p))의 범위는 **[-infinity, infinity]**

잠깐, 왜 로짓 변환을? ---- 선형 회귀의 식 만들기

$$\log(\text{Odds}(p))=wx+b$$

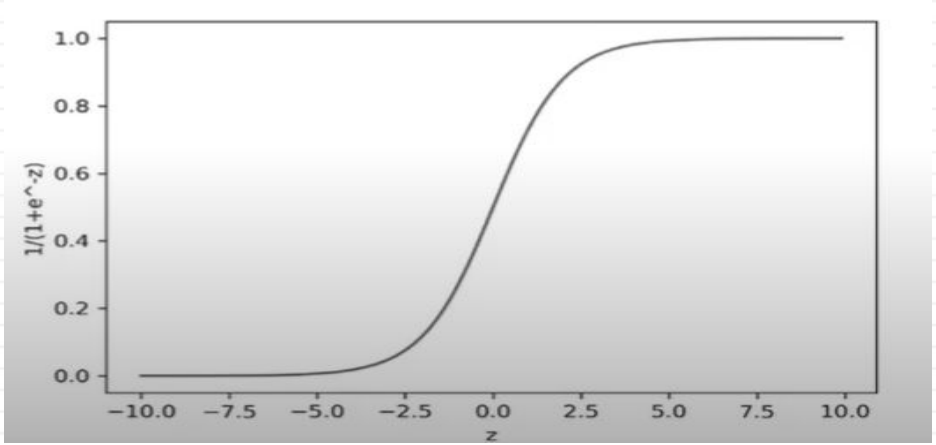
선형 회귀의 식(y=wx+b)의 형태와 매우 흡사!

선형 분석이 가능해졌고, 선형 회귀와 마찬가지로
함수의 w(기울기)와 b(절편)를 찾는 문제로 변경

로지스틱 회귀의 원리 (3)

〈정리〉

1. [0,1]의 범위인 확률을 로짓을 통해 [-infinity, infinity] 범위로 넓혀줌.
2. 범위의 변경으로 선형 분석이 가능해짐.
3. 다시 확률을 나타내 주기 위해 역함수를 씌워서 [0,1]로 범위를 제한



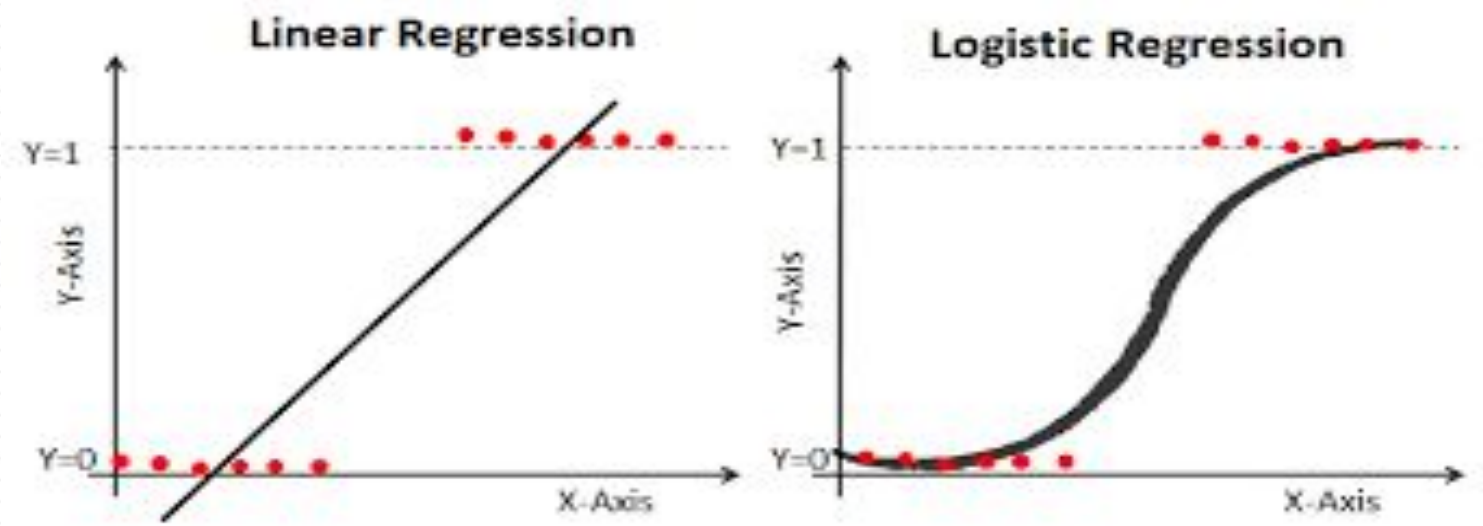
시그모이드 함수

When?

합격? or 불합격?

스팸으로 분류?

이진 분류 + 선형 회귀... 괜찮을까?

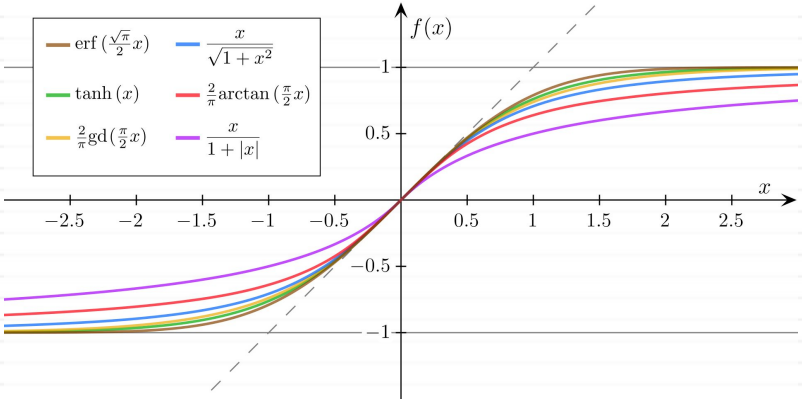


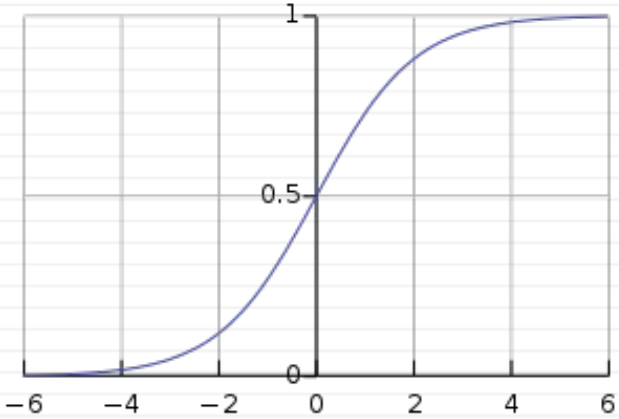
1. 원칙상 확률은 0과 1 사이의 범위를 벗어날 수 없다, 하지만 선형 회귀는...?

2. 훈련 데이터의 표준 편차가 높을 경우, 분류 기준점의 위치가 바뀐다.
   (타 데이터에도 영향을 주어 분류 성능 down) -> 그림판 설명

시그모이드 함수(Sigmoid function)

- S자형 곡선을 갖는 함수
- 반환값은 일반적으로 0에서 1까지의 범위
- 일반적으로 단조증가 함수
- 로지스틱 회귀에서 사용되는 시그모이드 함수는 분모에 자연상수 e가 있는 형태

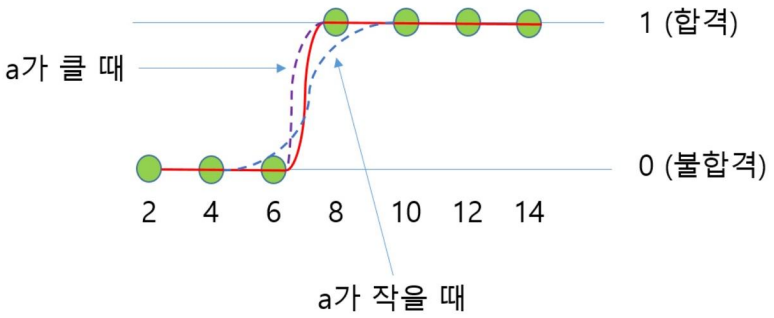


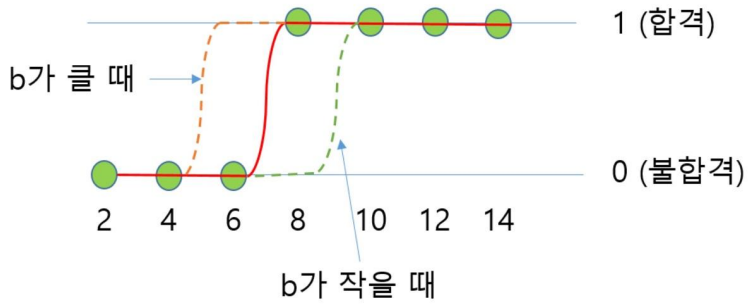


시그모이드 함수(Sigmoid function)

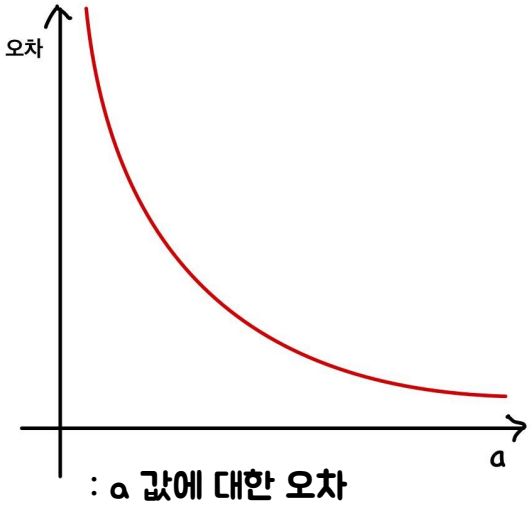
$$y = \frac{1}{1 + e^{-(ax+b)}}$$

- a의 값은 경사도 결정, b는 그래프의 좌우 이동 의미
- a와 b의 값에 따라 오차가 변함

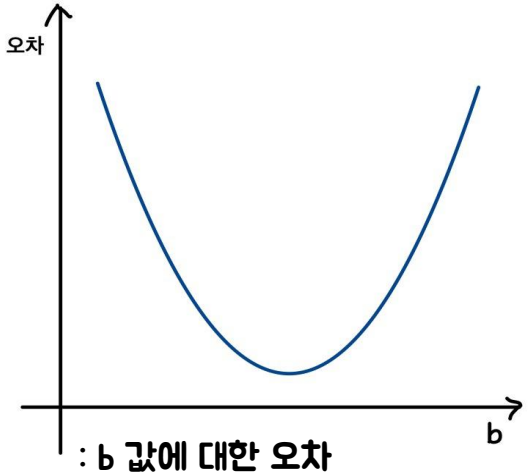




시그모이드 함수(Sigmoid function)



: a 값에 대한 오차



: b 값에 대한 오차

오차 공식

선형 회귀와 동일하게 로지스틱 회귀에서도 a, b를 구해야한다

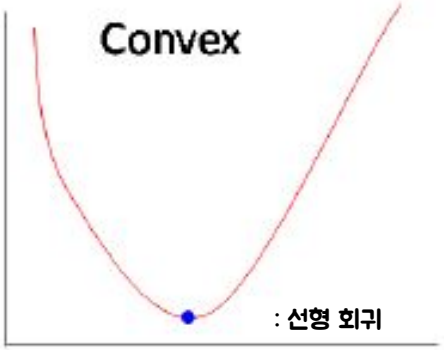
- 경사 하강법 이용

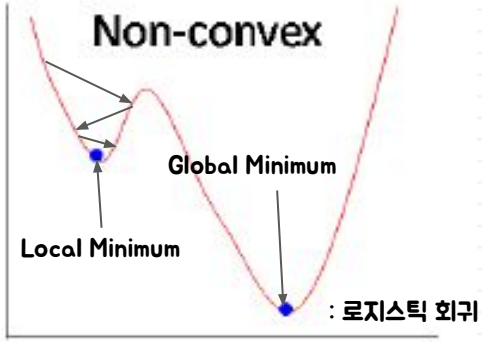
- 경사 하강법은 먼저 오차를 구한 후에 오차를 줄이는 쪽으로 이동시키는 방법

- 오차를 구하기 위해 어떤 공식을 사용해야할까?

평균 제곱 오차(MSE)?

선형 회귀에서 사용한 MSE를 이용하여 오차를 구한다면?





로지스틱 회귀에서 MSE를 이용한 오차와의 관계를 그래프로 나타내면 Non-convex(비볼록) 형태

➡ 경사 하강법으로 오차가 최소가 되는 지점을 Local Minumum 지점으로 착각 가능

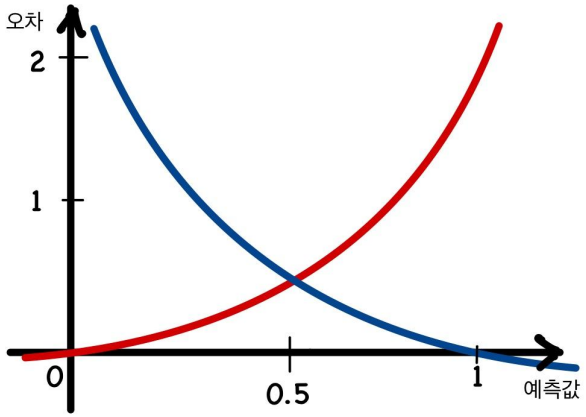
➡ 오차를 구하는 다른 공식이 필요

로그 함수

예측값이 실제값과 가까워질수록 오차가 감소하고, 멀어질수록 오차 증가하도록 설계

$$c(h, y\_data) = \begin{cases} -\log h & : \ y\_data = 1 \\ -\log(1 - h) & : \ y\_data = 0 \end{cases}$$

$$-\{\underbrace{y\_data \log h}_{A} + \underbrace{(1 - y\_data) \log(1 - h)}_{B}\}$$



실제값이 0일때: 빨간색 그래프

실제값이 1일때: 파란색 그래프

실습 문제

pre-class quiz 해설

1번. 로지스틱 회귀 정의 관련 문제

1. 로지스틱 회귀를 사용하기 적절하지 않은 것을 고르시오. *

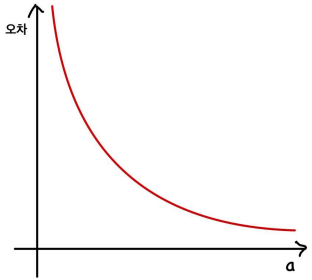
- (1) 제품의 상태가 불량인지 양품인지 분류하는 문제
- (2) 나이에 따른 혈압수치를 분석하는 문제
- (3) 나이에 따른 고혈압 유무를 판별하는 문제
- (4) 이메일이 스팸인지 정상메일인지 분류하는 문제

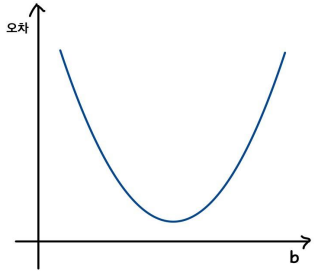
로지스틱 회귀는 종속 변수가 이분형일 때 사용하기 적합하다.
! 혈압수치는 이분형으로 나타내기 힘든 연속적인 숫자이다.

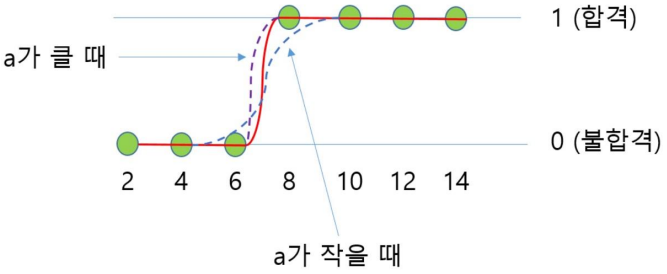
2번. 시그모이드 함수 관련 문제

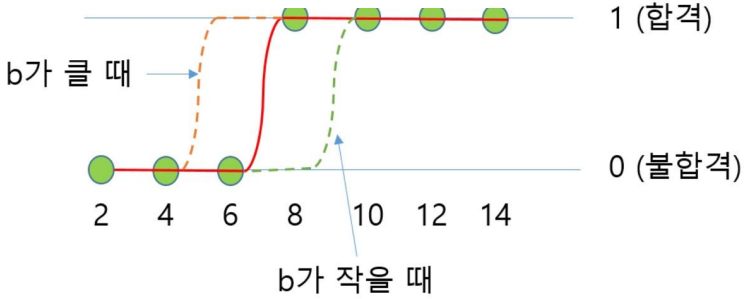
2. 시그모이드 함수 1/1+e**-(ax+b)에서 a는 (가)를(을) 나타내고 b는 (나)를(을) 의미한다 *
(가)의 값이 작으면 오차는 (다)에 수렴하고, b에 대한 오차의 그래프는 (라)의 형태로 나타

- (1) (가) : 경사도, (나) : 좌우 이동, (다) : 무한대, (라) : 2차 함수
- (2) (가) : 경사도, (나) : y절편, (다) : 0, (라) : 시그모이드 함수
- (3) (가) : 경사도, (나) : 좌우 이동, (다) : 무한대, (라) : 시그모이드 함수
- (4) (가) : 경사도, (나) : y절편, (다) : 0 ,(라) : 2차함수



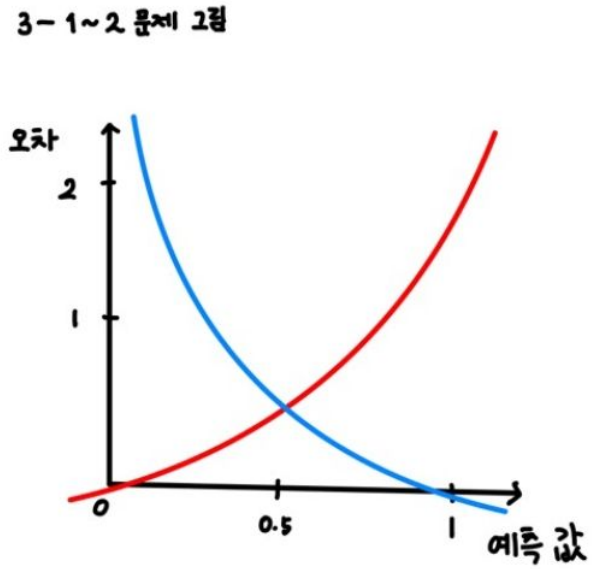






3번. 손실 함수 유도 관련 문제

3-1. 이 그래프에서 파란선은 실제값이 1일 때, 빨간선은 실제값이 0일 때 사용할 수 있는 그래프이다. 이는 각각 -logh와 -log(1-h)의 식으로 표현할 수 있다. y라는 실제값이 주어졌을 때, 이 그래프를 실제값에 따라 적절히 사용할 수 있는 함수 식을 찾아라. *



$$c(h, y\_data) = \begin{cases} -\log h & : \ y\_data = 1 \\ -\log(1 - h) & : \ y\_data = 0 \end{cases}$$

(2) -{y*logh + (1-y)log(1-h)}

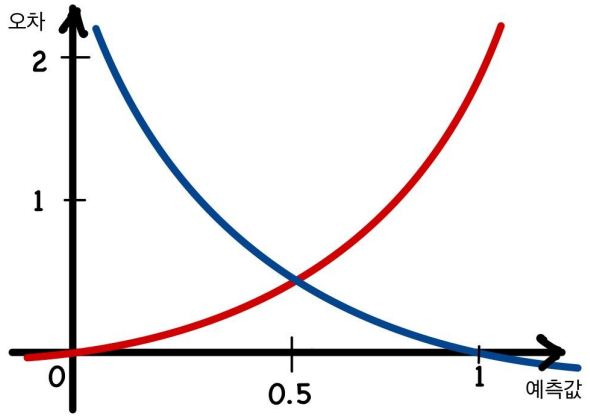
4번. 오차 공식 유도 관련 문제

3-2. 실제값이 0이고 예측값이 (a) 일 때, 오차는 0 이다. 예측값이 (b)에 가까울 수록 오차는 커진다. *
이때 사용하는 그래프의 함수 식은 (c)이다.

- (1) (a) = 1 (b) = 1 (c) = -log(1 - 예측값)
- (2) (a) = 1 (b) = 0 (c) = -log(1 - 예측값)2
- ✓ (3) (a) = 0 (b) = 1 (c) = -log(1 - 예측값)
- (4) (a) = 0 (b) = 1 (c) = -log(예측값)
- (5) (a) = 0 (b) = 0 (c) = -log(예측값

실제값이 0일때: 빨간색 그래프

실제값이 1일때: 파란색 그래프



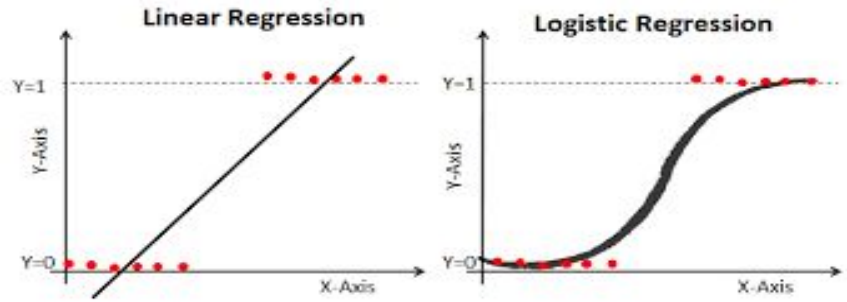
5번. 시그모이드 함수 관련 문제

5. 시그모이드 함수와 관련된 설명 중 가장 거리가 먼 것은? *

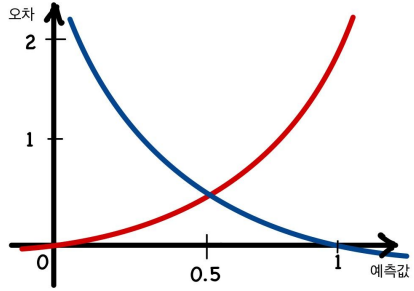
- (1) 활성화 함수(Activation Function)의 일종으로, 비선형 함수이다.
- (2) 기존의 선형회귀가 아닌 이진 분류에 더 특화된 함수이다.
- (3) 입력값에 상관 없이 결과값은 항상 제한된 범위 내에서 출력된다.
- (4) 3개 이상의 입력 값을 다룰 시, 시그모이드 함수가 아닌 Softmax 함수를 사용해야 한다.
- (5) 시그모이드 함수의 오차를 구하기 위해 사용되는 함수는 삼각함수 계열이다.

***softmax 함수**

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \; for\; j = 1, \ldots, k$$



로그함수가 사용된다



추가 예제

타이타닉에서 살아남기

sklearn의 로지스틱 회귀 라이브러리를 이용하여 타이타닉 실제 승객의 데이터를 학습시켜보고, 직접 다른 데이터를 넣어 생존자를 예측해보자

->실습 ㄱㄱ!