기계학습 (2-1): Logistic Classification

[redacted] 임 경 태

CONTENTS

CONTENTS

❑ 점수(정수)가 아닌 Binary(Pass/fail) 분류 문제

- 우리반 학생들의 H 대학 지원 결과가 아래와 같다. 나는 7,3 했는데 합격할 수 있을까?

이름	수학 공부시간	영어 공부시간	합/불
야쓰오	2	3	불합격
오공	4	8	불합격
블리츠	6	2	합격
페이커	8	4	합격

04. Logistic Regression = Logistic Classification (Classification의 예제)

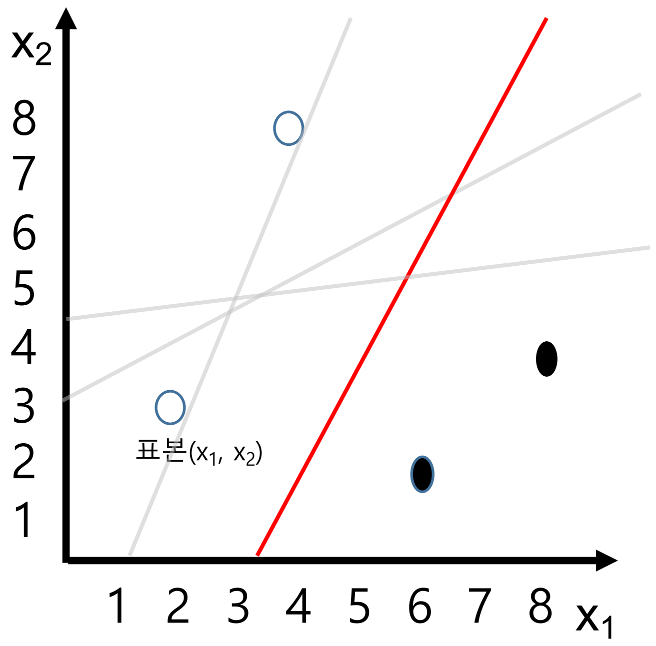
- Binary Classification (분류) 문제는 무엇인가?
  - 스팸메일 검출: Yes, No
  - 영화 추천: 추천, 비추천
  - 수능 합/불: 합격, 불합격
  - Where y={0, 1}

- Multinominal Classification (분류) 문제?
  - 수능 등급 분류: 1등급, 2등급, 3등급, 4등급
  - 객체 인식: 고양이, 강아지, 사자
  - 감정 인식: 기쁨, 슬픔, 냉소, 분노
  - Where y={0, 1, 2, .. n}

❑ 합격 불합격을 어떻게 나눌 수 있을까?

- **로지스틱 회귀**의 목적은 일반적인 회귀 분석의 목표와 동일하게 종속 변수와 독립 변수간의 관계를 구체적인 함수로 나타내어 향후 예측 모델에 사용하는 것이다 (위키)

이름	수학 (x_1) 공부시간	영어 (x_2) 공부시간	합/불 (y)
야쓰오	2	3	불합격
오공	4	8	불합격
블리츠	6	2	합격
페이커	8	4	합격

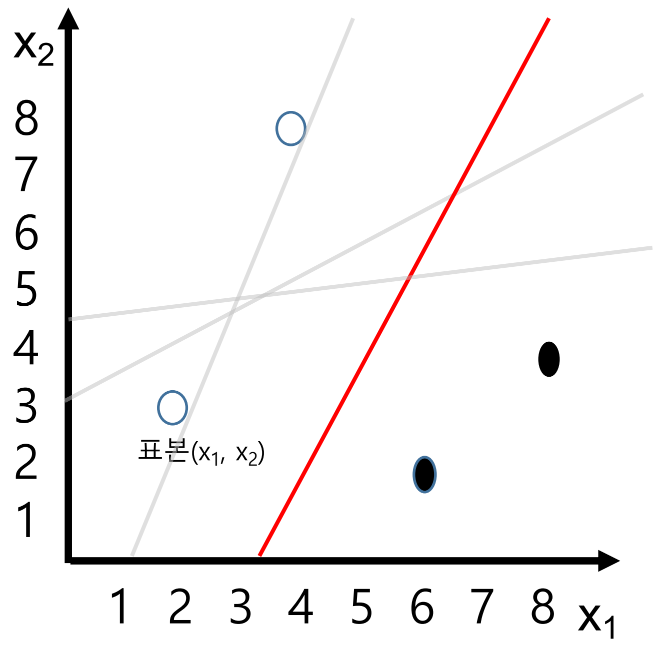


04. Logistic Classification (개요)

❑ 합격 불합격은 숫자가 아닌데?

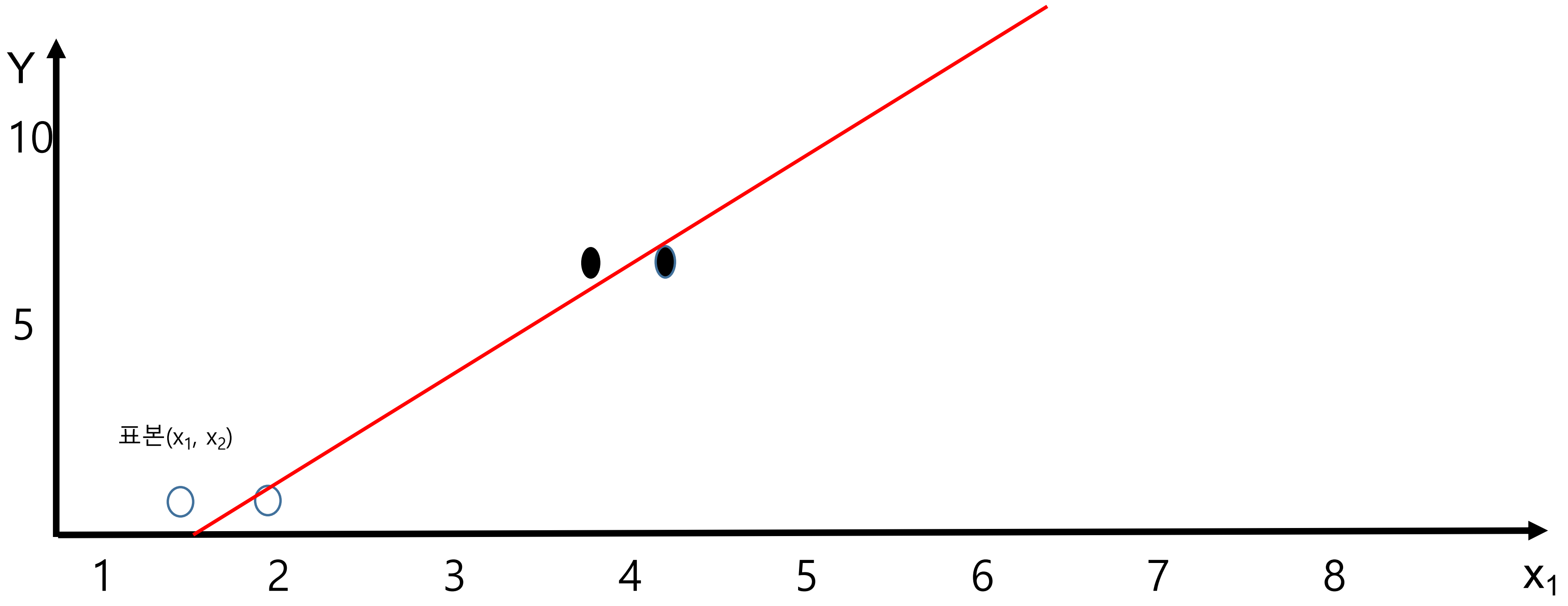
- ❑ 숫자로 만들지 뭐! 종속 변수 y 결과의 범위를[0,1]로 제한해서 0이면 불합격 1이면 합격으로 만들자!
- ❑ 종속 변수 y가 이진적이기 때문에 조건부 확률(P(y|x))의 분포가 이항 분포를 따른다
- ❑ 종속 변수 y의 결과는 0과 1, 두 개의 경우만 존재하는데 반해, 단순 선형 회귀를 적용하면 범위[0,1]를 벗어나는 결과가 나오기 때문에 오히려 예측의 정확도만 떨어뜨리게 된다. (위키)

이름	수학 (x_1) 공부시간	영어 (x_2) 공부시간	합/불 (y)
야쓰오	2	3	0
오공	4	8	0
블리츠	6	2	1
페이커	8	4	1



04. Logistic Classification (개요)

- ❑ Linear Regression으로 Classification 을 해볼 수 있지 않을까?



이미지출처: 위키피디아

04. Logistic Classification (개요)

- ❑ y값이 0 or 1이 나와야 하는데..
  - ❑ 가능한 모든 X에 f(X)가 0 or 1값을 반환하는 함수를 어떻게 찾을 수 있을까?

$$H(X)=f(X) = w_1{*}x_1 + w_2{*}x_2 + b$$
$$= 0 \text{ or } 1$$

$$Pr(Y_i = y_i | x_{1,i}, \ldots, x_{m,i}) = \begin{cases} p_i & \text{if } y_i = 1 \\ 1 - p_i & \text{if } y_i = 0 \end{cases}$$

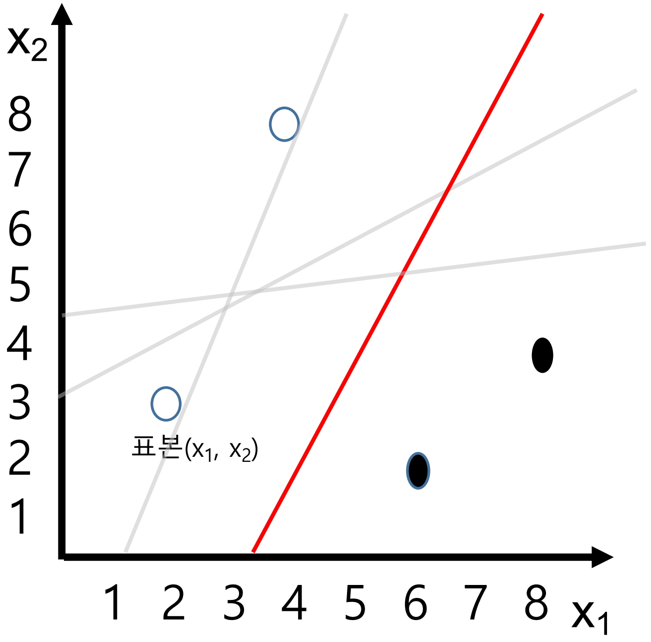
이름	수학 (x_1) 공부시간	영어 (x_2) 공부시간	합/불 (y)
야쓰오	2	3	0
오공	4	8	0
블리츠	6	2	1
페이커	8	4	1

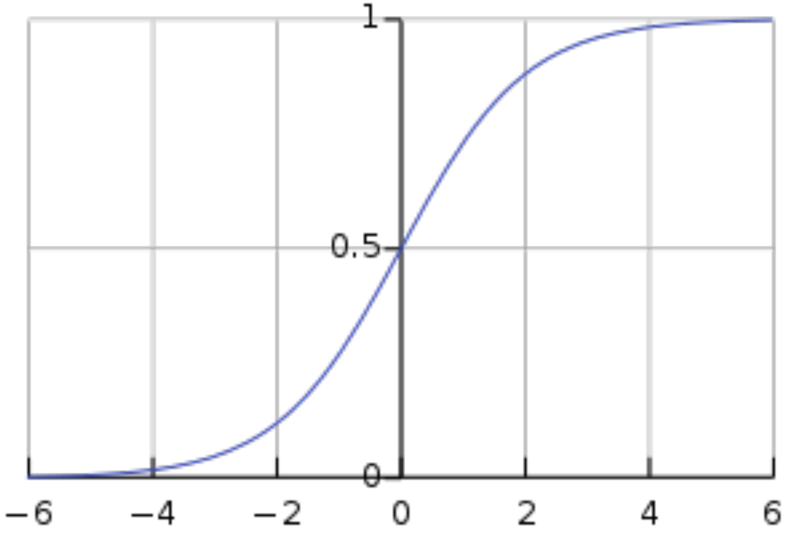
- 어떤 값이든 0~1사이로 변경해주는 Sigmoid 함수를 활용하자

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

$H(X)=f(X) = w_1{*}x_1 + w_2{*}x_2 + b$ → $S(H(X)) = S(w_1{*}x_1 + w_2{*}x_2 + b)$

$= 1/1+e^{-(w1{*}x1 + w2{*}x2 + b)}$





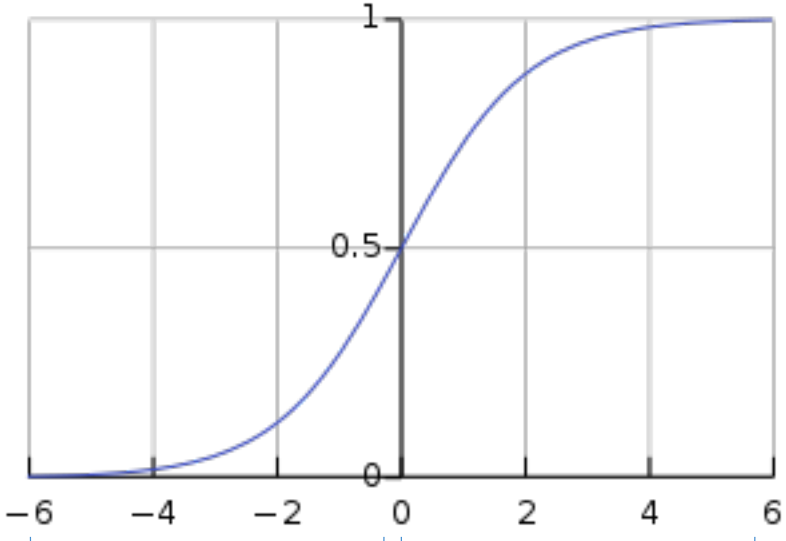
이미지출처: 위키피디아

04. Logistic Classification (개요)

- 어떤 값이든 0~1사이로 변경해주는 Sigmoid 함수를 활용하자
  - S(X) 값이 0.5보다 크면 1, 작으면 0으로 가정해보자

$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$$

$S(H(X)) = S(w_1 * x_1 + w_2 * x_2 + b)$
$= 1/1+e^{-(w1*x1 + w2*x2 + b)}$



$\underbrace{\qquad\qquad}$ H(X) 값이 음수면 S(H(X))< 0.5

$\underbrace{\qquad\qquad}$ H(X) 값이 양수면 S(H(X))=>0.5

즉
H(X) =>0 이면 y=1로 예측
H(X) < 0 이면 y=0으로 예측

즉
$w_1 * x_1 + w_2 * x_2 + b$ =>0 이면 y=1로 예측
$w_1 * x_1 + w_2 * x_2 + b$ < 0 이면 y=0으로 예측

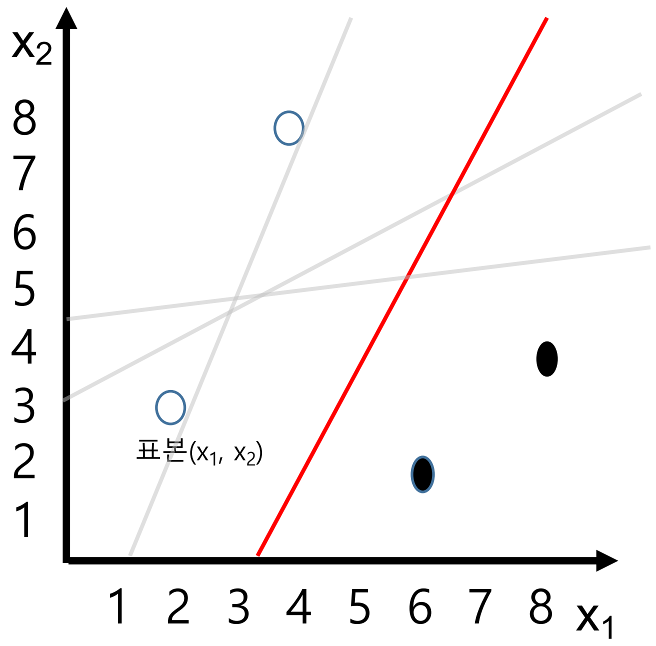
Sigmoid 데모: https://www.desmos.com/calculator/bgontvxotm?lang=ko

이미지출처: 위키피디아

04. Logistic Classification (개요)

❑ 어떤 값이든 0~1사이로 변경해주는 Sigmoid 함수를 활용하자

$H(X)=f(X) = w_1*x_1 + w_2*x_2 + b$



즉

$w_1*x_1 + w_2*x_2 + b$ =>0 이면 y=1로 예측

$w_1*x_1 + w_2*x_2 + b < 0$ 이면 y=0으로 예측

이 말은

$w_1*x_1 + w_2*x_2 + b = 0$ 인 선을 기준으로

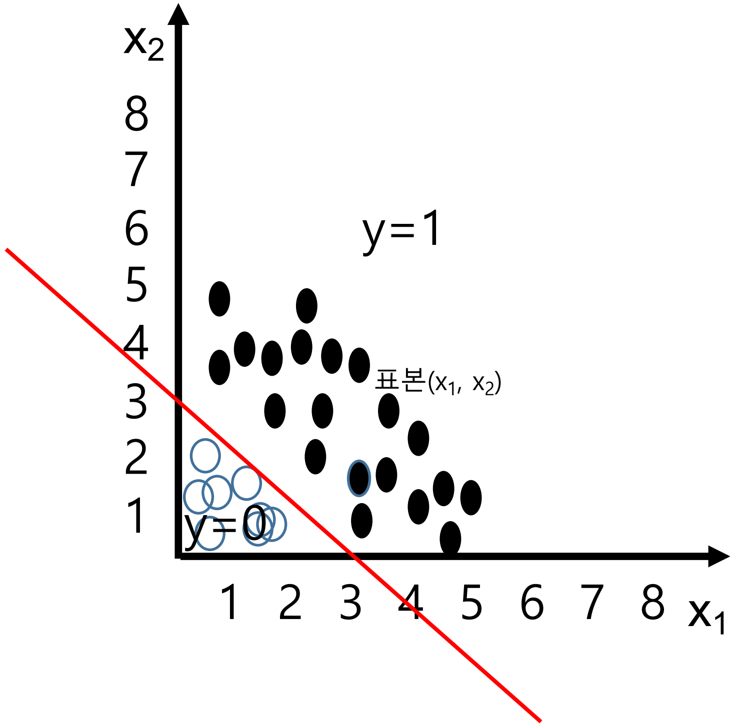
- 한쪽은 1
- 다른 한쪽은 0

우리는 이 선을 decision boundary라고 함

04. Logistic Classification (graphical 예제)

❑ 데이터 A에 대해 Linear decision boundary를 찾았다고 가정해 보자

$H(X)=f(X) = w_1*x_1 + w_2*x_2 + b$



w_1, w_2, b_1 을 1, 1, -3로 둔다면

$x_1 + x_2 - 3$ =>0 이면 y=1로 예측
$x_1 + x_2 - 3 < 0$ 이면 y=0으로 예측

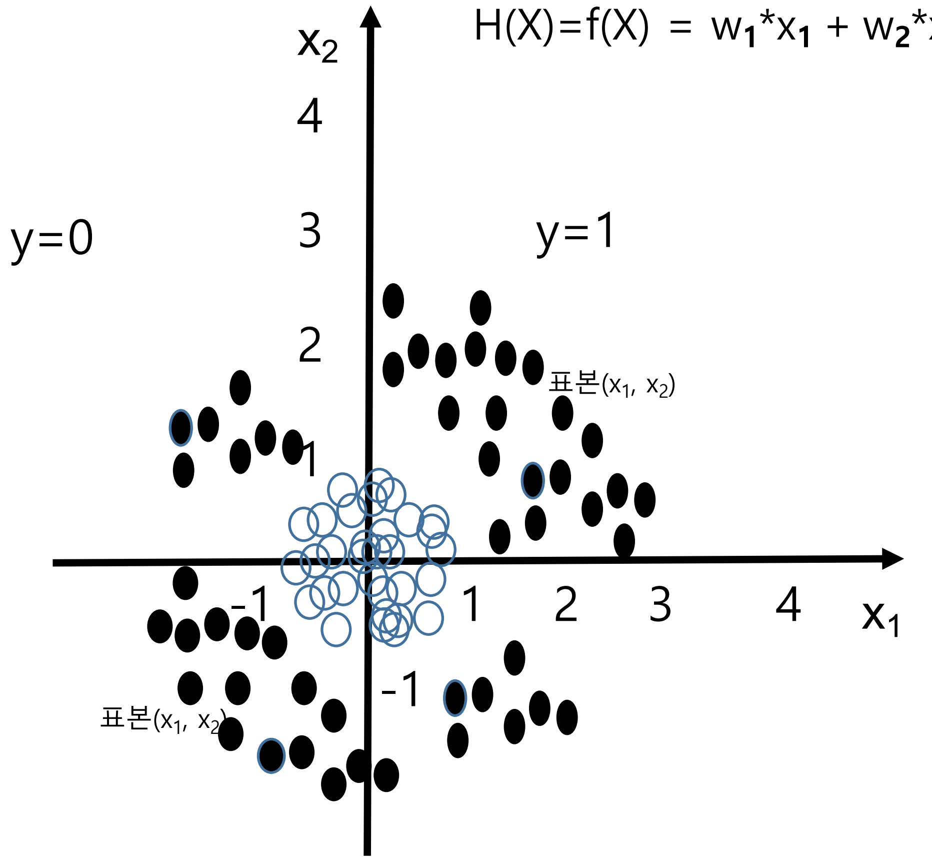
$x_1 + x_2$ =>3 이면 y=1로 예측
$x_1 + x_2 < 3$ 이면 y=0으로 예측

이 말은
$x_1 + x_2 - 3 = 0$ 인 선을 기준으로
- 한쪽은 1
- 다른 한쪽은 0

04. Logistic Classification (graphical 예제)

❑ Linear로 분류가 안될 수도 있겠지?
 - 실제 예제 Non-linear decision boundary를 보자

$H(X)=f(X) = w_1*x_1 + w_2*x_2 + w_3*x^2_1 + w_4*x^2_2 + b$



w_1, w_2, w_3, w_4, b 을 0, 0, 1, 1, -1로 둔다면

$x^2_1 + x^2_2 - 1 => 0$ 이면 y=1로 예측
$x^2_1 + x^2_2 - 1 < 0$ 이면 y=0으로 예측

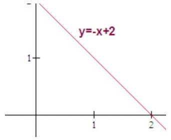
$x^2_1 + x^2_2 => 1$ 이면 y=1로 예측
$x^2_1 + x^2_2 < 1$ 이면 y=0으로 예측

이 말은
$x^2_1 + x^2_2 - 1 = 0$ 인 원을 기준으로
- 한쪽은 1
- 다른 한쪽은 0

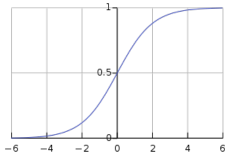
04. Logistic Classification (Cost 함수)

❑ Logistic Classification의 Cost 함수를 MSE로 해볼까?

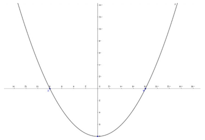
❑ H(x) $= w_1 * x_1 + w_2 * x_2 + b$



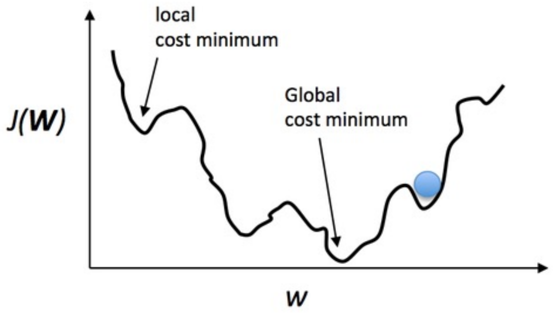
❑ $S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1}$



❑ MSE $cost(W,b) = \frac{1}{m}\sum_{i=1}^{m}(H(x^{(i)}) - y^{(i)})^2$



❑ 1차원 함수에 MSE를 하면 Cost값은 2차원 함수로 표현됨, 하지만 1차원 함수에 Sigmoid를 취하면??



Cost 함수에 log를 취해주면?
- 주름이 펴진다 (convex)

04. Logistic Classification (Cost 함수)

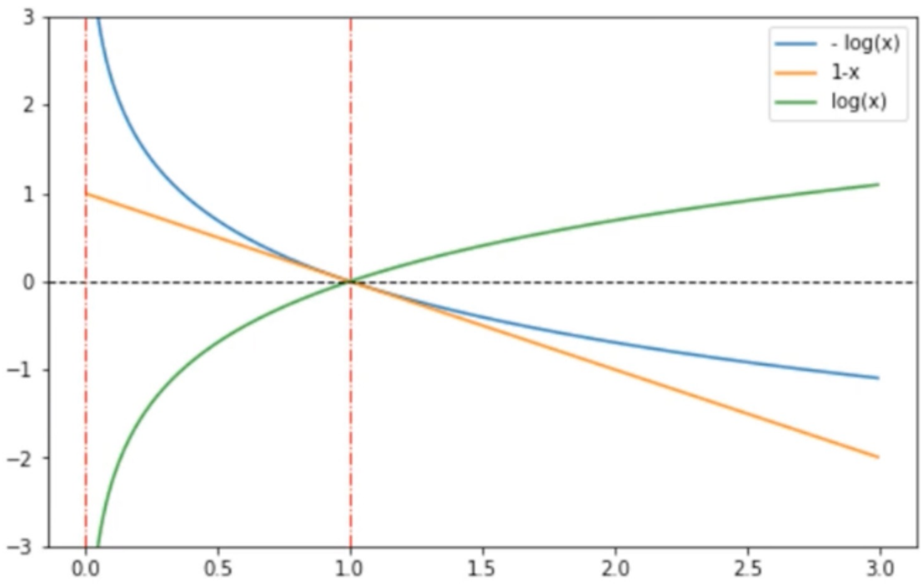
- ❑ 그렇다면 Logistic Classification의 Cost 함수는 어떻게 정의하면 될까?

  - ❑ Cost함수의 목적에 충실해지자!
    - ❑ 예측이 틀릴 경우 Cost = 매우 높음
    - ❑ 예측이 맞을 경우 Cost = 0

```
if: y==1 and S(x)==1: #정답과 예측이 일치
    return loss=0
if: y==0 and S(x)==0: #정답과 예측이 일치
    return loss=0
if: y==1 and S(x)==0: #정답과 예측이 불일치
    return loss=무한대
if: y==0 and S(x)==1: #정답과 예측이 불일치
    return loss=무한대
```

- 그렇다면 Logistic Classification의 Cost 함수는 어떻게 정의하면 될까?

  - Cost함수의 목적에 충실해지자!
    - 예측이 틀릴 경우 Cost = 매우 높음
    - 예측이 맞을 경우 Cost = 0



$h_\theta(x) = S(H(X))$,where θ = {W, b}로 간략화 하자!

$$\mathrm{cost}\big(h_\theta(x), y\big) = \begin{cases} -\log\big(h_\theta(x)\big) & \text{if } y = 1 \\ -\log\big(1 - h_\theta(x)\big) & \text{if } y = 0 \end{cases}$$

```
if: y==1 and S(x)==1: #정답과 예측이 일치
    return loss=0
if: y==0 and S(x)==0: #정답과 예측이 일치
    return loss=0
if: y==1 and S(x)==0: #정답과 예측이 불일치
    return loss=무한대
if: y==0 and S(x)==1: #정답과 예측이 불일치
    return loss=무한대
```

04. Logistic Classification (Negative Log Likely-hood Cost 함수)

- Y == 0, Y==1 두 경우의 손실 함수를 하나로 합쳐보자!

  - $Cost(W,b) = \frac{1}{m}\sum_{i=1}^{m} -ylog\left(h_\theta(x)\right) - (1-y)\log(1-h_\theta(x))$

$$J(w) = \sum_{i=1}^{m} y^{(i)} \log P(y=1) + (1-y^{(i)}) \log P(y=0)$$

$$\left.\begin{aligned} \mathbb{P}(y=1|z) &= \sigma(z) = \frac{1}{1+e^{-z}} \\ \mathbb{P}(y=0|z) &= 1-\sigma(z) = \frac{1}{1+e^{z}} \end{aligned}\right\} \mathbb{P}(y|z) = \sigma(z)^y(1-\sigma(z))^{1-y}$$

04. Logistic Classification (Gradient Decent)

- Cost값을 기반으로 Gradient Decent 알고리즘을 이용해 최적의 파리미터 W값 을 찾자!
  - H(x) = $w_1 * x_1 + w_2 * x_2 + b$
  - $Cost(W,b) = \frac{1}{m}\sum_{i=1}^{m} -ylog\left(h_\theta(x)\right) - (1-y)\log(1-h_\theta(x))$
  - $\underset{W,b}{\text{minimize}}\ cost(W,b)$ $\quad W := W - \alpha\frac{\partial}{\partial W}cost(W)$
  - $\frac{\partial}{\partial\theta}Cost(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(Sh_\theta\left(x^i\right) - y^i\right)x^i$

04. Logistic Classification 정리!

- Logistic Classification을 정리하면!
  - $H(x) = w_1 * x_1 + w_2 * x_2 + b$
  - $S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x + 1}$
  - $S(H(X)) = \frac{1}{1+e^{-(w1*x1 + w2*x2 + b)}}$
  - $Cost(W,b) = \frac{1}{m}\sum_{i=1}^{m} -y\log(h_\theta(x)) - (1-y)\log(1-h_\theta(x))$
  - $\underset{W,b}{\text{minimize}}\, cost(W,b)$ $\quad W := W - \alpha \frac{\partial}{\partial W} cost(W)$

04. Logistic Classification 정리!

Appendix: Partial derivative of $J(\theta)$

우선 logistic funtion의 미분을 구해두자.

$$\begin{aligned}\sigma(x)' &= \left(\frac{1}{1+e^{-x}}\right)' = \frac{-(1+e^{-x})'}{(1+e^{-x})^2} = \frac{-1' - (e^{-x})'}{(1+e^{-x})^2} = \frac{0-(-x)'(e^{-x})}{(1+e^{-x})^2} = \frac{-(-1)(e^{-x})}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2} \\ &= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}}{1+e^{-x}}\right) = \sigma(x)\left(\frac{+1-1+e^{-x}}{1+e^{-x}}\right) = \sigma(x)\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) = \sigma(x)(1-\sigma(x))\end{aligned}$$

이제 편미분을 구해보자.

$$\begin{aligned}\frac{\partial}{\partial\theta_j}J(\theta) &= \frac{\partial}{\partial\theta_j}\frac{-1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\log(1-h_\theta(x^{(i)}))\right] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\frac{\partial}{\partial\theta_j}\log(h_\theta(x^{(i)})) + (1-y^{(i)})\frac{\partial}{\partial\theta_j}log(1-h_\theta(x^{(i)}))\right] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}\frac{\partial}{\partial\theta_j}h_\theta(x^{(i)})}{h_\theta(x^{(i)})} + \frac{(1-y^{(i)})\frac{\partial}{\partial\theta_j}(1-h_\theta(x^{(i)}))}{1-h_\theta(x^{(i)})}\right] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}\frac{\partial}{\partial\theta_j}\sigma(\theta^T x^{(i)})}{h_\theta(x^{(i)})} + \frac{(1-y^{(i)})\frac{\partial}{\partial\theta_j}(1-\sigma(\theta^T x^{(i)}))}{1-h_\theta(x^{(i)})}\right] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}\sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)}))\frac{\partial}{\partial\theta_j}\theta^T x^{(i)}}{h_\theta(x^{(i)})} + \frac{-(1-y^{(i)})\sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)}))\frac{\partial}{\partial\theta_j}\theta^T x^{(i)}}{1-h_\theta(x^{(i)})}\right] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[\frac{y^{(i)}h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))\frac{\partial}{\partial\theta_j}\theta^T x^{(i)}}{h_\theta(x^{(i)})} - \frac{(1-y^{(i)})h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))\frac{\partial}{\partial\theta_j}\theta^T x^{(i)}}{1-h_\theta(x^{(i)})}\right] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}(1-h_\theta(x^{(i)}))x_j^{(i)} - (1-y^{(i)})h_\theta(x^{(i)})x_j^{(i)}\right] \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}(1-h_\theta(x^{(i)})) - (1-y^{(i)})h_\theta(x^{(i)})\right]x_j^{(i)} \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} - y^{(i)}h_\theta(x^{(i)}) - h_\theta(x^{(i)}) + y^{(i)}h_\theta(x^{(i)})\right]x_j^{(i)} \\
&= -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)} - h_\theta(x^{(i)})\right]x_j^{(i)} \\
&= \frac{1}{m}\sum_{i=1}^{m}\left[h_\theta(x^{(i)}) - y^{(i)}\right]x_j^{(i)}\end{aligned}$$

출처 https://wikidocs.net/4289

감사합니다.