Linear Discrimination

$$\Pr(y=1 \mid x) = \delta$$
$$\Pr(y=2 \mid x) = 1-\delta$$

choose the first class if $\begin{cases} \delta > 0.5 \\ \frac{\delta}{1-\delta} > 1 \\ \log\left[\frac{\delta}{1-\delta}\right] > 0 \end{cases}$

$$\underbrace{\log\left[\frac{\Pr(y=1|x)}{\Pr(y=2|x)}\right]}_{\log\left[\frac{\delta}{1-\delta}\right]} = \log\left[\frac{p(x|y=1)}{p(x|y=2)}\right] + \log\left[\frac{\Pr(y=1)}{\Pr(y=2)}\right]$$

$p(x|y=1) \to N(x;\mu_1,\Sigma_1)$

$p(x|y=2) \to N(x;\mu_2,\Sigma_2)$

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$\frac{\exp(a)}{\exp(b)} = \exp(a-b)$$

$$N(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^D |\Sigma|}} \exp\left[-\frac{1}{2}(x-\mu)^T \cdot \Sigma^{-1} \cdot (x-\mu)\right]$$

$$\log\left[\frac{\cancel{(2\pi)^{-D/2}} \cdot \cancel{|\Sigma|^{-1/2}} \exp\left[-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right]}{\cancel{(2\pi)^{-D/2}} \cdot \cancel{|\Sigma|^{-1/2}} \exp\left[-\frac{1}{2}(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)\right]}\right] + \log\left[\frac{\Pr(y=1)}{\Pr(y=2)}\right]$$

$$= \underbrace{\left[\Sigma^{-1}(\mu_1-\mu_2)\right]^T}_{w} \cdot x + \underbrace{\left[-\frac{1}{2}(\mu_1+\mu_2)^T\Sigma^{-1}(\mu_1-\mu_2) + \log\left[\frac{\Pr(y=1)}{\Pr(y=2)}\right]\right]}_{w_0}$$

$$= w^T \cdot x + w_0 \quad \text{where} \quad \hat{w} = \hat{\Sigma}^{-1} \cdot (\hat{\mu}_1 - \hat{\mu}_2)$$

$\hat{\mu}_2$ → sample mean of the second class

$\hat{\Sigma}$ → sample covariance matrix of all data points

$\hat{\mu}_1$ → sample mean of the first class

$$\hat{w}_0 = -\frac{1}{2} (\hat{\mu}_1 + \hat{\mu}_2)^T \hat{\Sigma}^{-1} (\hat{\mu}_1 - \hat{\mu}_2) + \log\left[\frac{\hat{Pr}(y=1)}{\hat{Pr}(y=2)}\right]$$

$\hat{Pr}(y=1)$ → frequency of the first class

$\hat{Pr}(y=2)$ → frequency of the second class.

$$\exp\left[\log\left[\frac{\delta}{1-\delta}\right]\right] = \exp\left[w^T \cdot x + w_0\right]$$

variable of interest: δ

$$\frac{\delta}{1-\delta} = \exp[w^T x + w_0] \Rightarrow \delta = \exp[w^T x + w_0] - \delta . \exp[w^T . x + w_0]$$

$$\delta\left[1 + \exp[w^T x + w_0]\right] = \exp[w^T . x + w_0]$$

$$\delta = \frac{\exp[w^T . x + w_0] \exp[-(w^T . x + w_0)]}{1 + \exp[w^T . x + w_0] \exp[-(w^T . x + w_0)]}$$

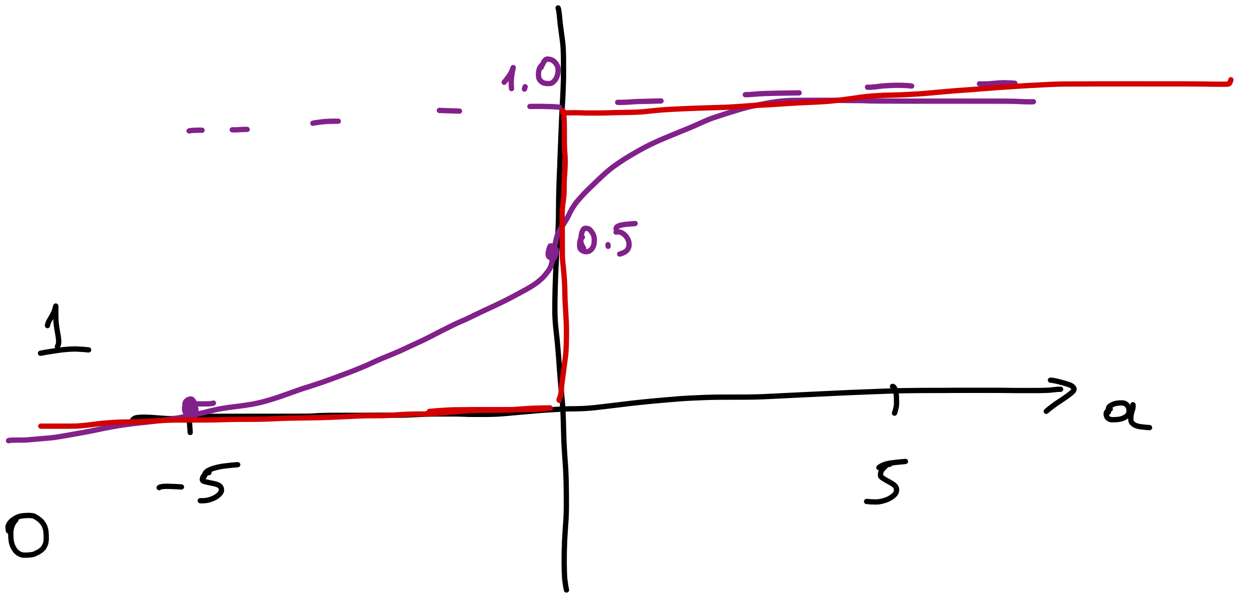
$$\text{Sigmoid} \Rightarrow \delta = \frac{1}{1 + \exp[-(w^T . x + w_0)]}$$

a) if $w^T . x + w_0 > 0 \Rightarrow \delta > 0.5 \Rightarrow 1 - \delta < 0.5$

b) if $w^T . x + w_0 = 0 \Rightarrow \delta = 0.5 \Rightarrow 1 - \delta = 0.5$

c) if $w^T . x + w_0 < 0 \Rightarrow \delta < 0.5 \Rightarrow 1 - \delta > 0.5$

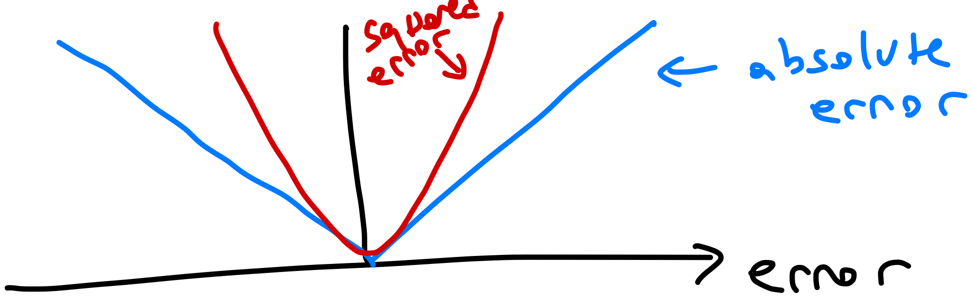
$$\delta(a) = \frac{1}{1 + \exp(-a)}$$

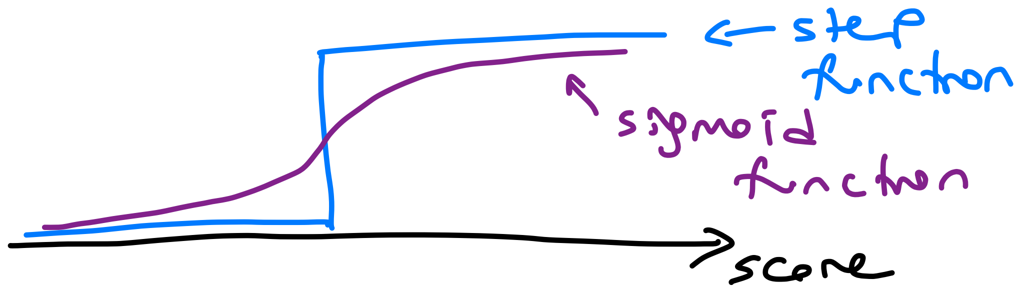


if $a = 5 \Rightarrow \frac{1}{1 + \exp(-5)} \cong 1$

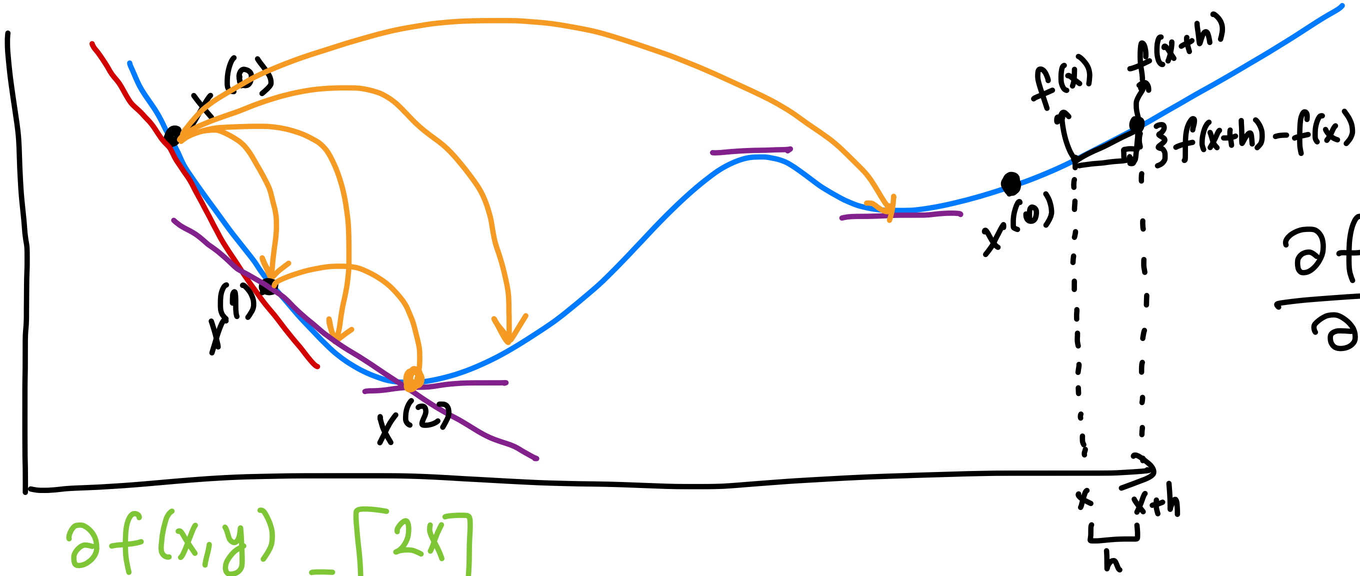
if $a = -5 \Rightarrow \frac{1}{1 + \exp(+5)} \cong 0$

if $a = 0 \Rightarrow \frac{1}{1 + \exp(0)} = \frac{1}{2} = 0.5$





Gradient Descent / Gradient Ascent



$f(x)$

$f(w)$

$$\frac{\partial f(x)}{\partial x} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{(x+h) - x}$$

$$= \lim_{h \to 0} \frac{f(x+h/2) - f(x-h/2)}{(x+h/2) - (x-h/2)}$$

$f\begin{pmatrix} x \\ y \end{pmatrix} = x^2 + y^2$

$$\frac{\partial f(x,y)}{\partial \begin{bmatrix} x \\ y \end{bmatrix}} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\underbrace{\Delta x}_{\substack{\text{step} \\ \text{(update)}}} = - \underbrace{\eta}_{\substack{\text{step} \\ \text{size}}} \cdot \underbrace{\frac{\partial f(x)}{\partial x}}_{\text{derivative / slope}}$$

$$x^{(t+1)} = x^{(t)} + \Delta x$$

$$= x^{(t)} - \eta \cdot \frac{\partial f(x)}{\partial x}$$

$$(w^*, w_0^*) = \arg\min_{(w,w_0)} E[w, w_0 \mid \underbrace{\mathcal{X}}_{\text{training set}}]$$

T H

Bernoulli(0.5)

$$\boxed{p(x) = \pi^x \cdot (1-\pi)^{1-x}}$$

$$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^N \qquad y_i \in \{0, 1\} \qquad x_i \in \mathbb{R}^D$$

1 → success

0 → failure

$$y_i \mid x_i \sim \text{Bernoulli}\left(y_i \,;\, \underbrace{\hat{P}r(y=1 \mid x_i)}_{\text{success probability} = \hat{y}_i}\right)$$

$$\text{Likelihood}(w, w_0 \mid \mathcal{X}) = \prod_{i=1}^N \left[\hat{y}_i^{(y_i)} \cdot (1-\hat{y}_i)^{(1-y_i)}\right]$$

$$\text{loglikelihood}(w, w_0 \mid \mathcal{X}) = \sum_{i=1}^N \left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$$

$$E[w, w_0 \mid \mathcal{X}] = \underbrace{-\sum_{i=1}^N \left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]}_{\text{negative loglikelihood / error / loss}}$$

minimize $\quad -\sum_{i=1}^{N}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right]$

with respect to w, w_0

$$\hat{y}_i = \frac{1}{1+\exp\left[-\left[w^T.x_i + w_0\right]\right]}$$

$$\frac{\partial \text{Error}}{\partial w} = ? \qquad \frac{\partial \text{Error}}{\partial w_0} = ?$$

$$\text{sigmoid}(a) = \frac{1}{1+\exp(-a)} \quad \begin{matrix}\rightarrow f(a)\\ \rightarrow g(a)\end{matrix}$$

$$\frac{\partial \log(a)}{\partial a} = \frac{1}{a}$$

$$\frac{\partial \frac{f(a)}{g(a)}}{\partial a} = \frac{\frac{\partial f(a)}{\partial a}.g(a) - \frac{\partial g(a)}{\partial a}.f(a)}{[g(a)]^2}$$

Exercise

$$\frac{\partial\, \text{sigmoid}(a)}{\partial(a)} = \text{sigmoid}(a)\left[1-\text{sigmoid}(a)\right]$$

Hint:

$$\frac{\partial\, \text{sigmoid}(a)}{\partial(a)} = \frac{0.[1+\exp(-a)] - 1.\frac{\partial(1+\exp(-a))}{\partial(a)}}{[1+\exp(-a)]^2}$$

$-1.\exp(-a)$

$$= \frac{[\exp(-a)]}{[1+\exp(-a)]}.\frac{1}{[1+\exp(-a)]}$$

$$\log[\hat{y}_i] = \log[\underbrace{\text{sigmoid}(\underbrace{w^T . x_i + w_0}_{c})}_{d}]$$

$$\frac{\partial\left(w_1 . x_1 + w_2 . x_2\right)}{\partial \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$w_1^T . x$

$$\frac{\partial \log(\hat{y}_i)}{\partial w} = \underbrace{\frac{\partial \log(d)}{\partial d}}_{1/\cancel{d}} \cdot \underbrace{\frac{\partial d}{\partial c}}_{\cancel{d}.(1-d)} \cdot \underbrace{\frac{\partial c}{\partial w}}_{x_i}$$

[previous page]

$$= (1-\hat{y}_i) . x_i$$

$$\frac{\partial \log(1-\hat{y}_i)}{\partial w} = ?$$

Exercise

$$\frac{\partial \text{Error}}{\partial w} = -\sum_{i=1}^{N} (y_i - \hat{y}_i) . x_i \qquad \frac{\partial \text{Error}}{\partial w_0} = -\sum_{i=1}^{N} (y_i - \hat{y}_i)$$

$$\boxed{\begin{aligned} \Delta w &= -\eta . \frac{\partial \text{Error}}{\partial w} = -\eta\left[-\sum_{i=1}^{N} (y_i - \hat{y}_i) . x_i\right] = \eta \sum_{i=1}^{N} (y_i - \hat{y}_i) . x_i \\ \Delta w_0 &= -\eta \frac{\partial \text{Error}}{\partial w_0} = -\eta\left[-\sum (y_i - \hat{y}_i)\right] = \eta \sum_{i=1}^{N} (y_i - \hat{y}_i) \end{aligned}}$$