STEP #1: Initialize $\{w, w_0\}$ and decide η.

initialize them to very small values

for example Uniform$(-0.001, +0.001)$

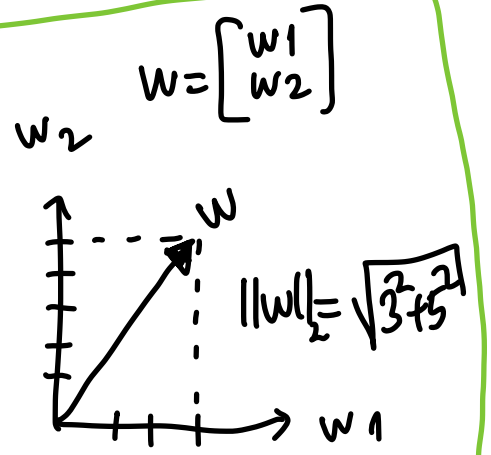
STEP #2: Calculate Δw and Δw_0

STEP #3: Update w and w_0 using Δw and Δw_0

$$w^{(t+1)} = w^{(t)} + \Delta w^{(t)}$$
$$w_0^{(t+1)} = w_0^{(t)} + \Delta w_0^{(t)}$$

STEP #4: Go to STEP#2 if there is a change in the parameters [i.e., $\|\Delta w\|_2 \neq 0$, $|\Delta w_0| \neq 0$]

If $\|\Delta w\|_2 < \varepsilon$ & $|\Delta w_0| < \varepsilon$ where ε is a very small # such as 10^{-10}, we should stop the algorithm.



Linear Discrimination (multiple classes $K > 2$)

$\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$ $\quad x_i \in \mathbb{R}^D$ $\quad y_i \in \{1, 2, \ldots, K\}$ → reference class

$$\exp\left[\log\left[\frac{Pr(y=c|x)}{Pr(y=K|x)}\right]\right] = \exp\left[w_c^T \cdot x + w_{c0}\right]$$

$$\log\left[\frac{Pr(y=c|x)}{Pr(y=K|x)}\right] = \underbrace{\log\left[\frac{p(x|y=c)}{p(x|y=K)}\right]}_{\text{is a function of } x} + \underbrace{\log\left[\frac{Pr(y=c)}{Pr(y=K)}\right]}_{\text{constant with respect to } x}$$

i) $Pr(y=c|x) \geq 0$

ii) $\sum_{c=1}^{K} Pr(y=c|x) = 1$

$$Pr(y=1|x) + Pr(y=2|x) + \ldots + Pr(y=K-1|x) + Pr(y=K|x) = 1$$

$$\frac{Pr(y=1|x) + Pr(y=2|x) + \ldots + Pr(y=K-1|x)}{Pr(y=K|x)} = \frac{1 - Pr(y=K|x)}{Pr(y=K|x)}$$

$$\sum_{c=1}^{K-1} \frac{Pr(y=c|x)}{Pr(y=K|x)} = \frac{1 - Pr(y=K|x)}{Pr(y=K|x)}$$

$$\sum_{c=1}^{K-1} \exp\left[w_c^T \cdot x + w_{c0}\right] = \frac{1}{Pr(y=K|x)} - 1$$

$$Pr(y=K|x) = \frac{1}{1+\sum_{c=1}^{K-1}\exp\left[w_c^T.x + w_{c0}\right]} \quad (1)$$

reference

$$Pr(y=c|x) = \frac{\exp\left[w_c^T.x + w_{c0}\right]}{1+\sum_{d=1}^{K-1}\exp\left[w_d^T.x + w_{d0}\right]} \quad (2)$$

nonreference

$$\theta = \{\underbrace{w_1}_{D\times 1}, \underbrace{w_{10}}_{1\times 1}, \underbrace{w_2}_{D\times 1}, \underbrace{w_{20}}_{1\times 1}, \ldots\ldots, \underbrace{w_{(K-1)}}_{D\times 1}, \underbrace{w_{(K-1)0}}_{1\times 1}\}$$

# of parameters $= (K-1)(D+1)$

$$Pr(y=c|x) = \frac{\exp\left[w_c^T.x + w_{c0}\right]}{\sum_{d=1}^{K}\exp\left[w_d^T.x + w_{d0}\right]}$$

softmax function

x_{N+1} is a new data point

$g_1(x_{N+1})$

$g_2(x_{N+1})$

$\vdots$

$g_K(x_{N+1})$

pick the maximum one

ideal / 1

$$Pr(y=1 \mid x) = \frac{\exp(2)}{\exp(2)+\exp(-2)+\exp(1)}$$ → 0.7214 / 0.9999

$g_1(x) = +2$ → 20

$g_2(x) = -2$ → -20

$g_3(x) = +1$ → 10

- $$Pr(y=2 \mid x) = \frac{\exp(-2)}{\exp(2)+\exp(-2)+\exp(1)}$$ → 0.0132 / 0.0000
- $$Pr(y=3 \mid x) = \frac{\exp(1)}{\exp(2)+\exp(-2)+\exp(1)}$$ → 0.2654 / 0.0000

$\exp(20m)$

$$\frac{\exp(20)}{\exp(20)+\exp(-20)+\exp(10)}$$

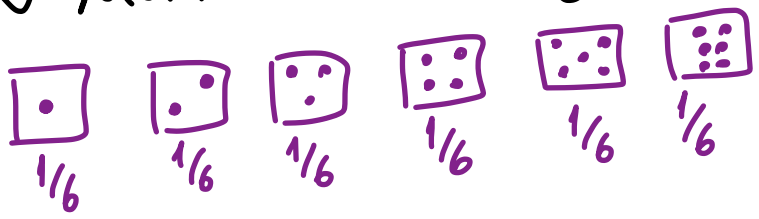
$$\frac{\exp(-20)}{\exp(20)+\exp(-20)+\exp(10)}$$

$$\frac{\exp(10)}{\exp(20)+\exp(-20)+\exp(10)}$$

$$\frac{\exp(2000)}{\exp(2000)+\exp(-2000)+\exp(1000)} \qquad \frac{\exp(-2000)}{\exp(-2000)}$$

$$= \frac{\overbrace{\exp(0)}^{1}}{\underbrace{\exp(0)}_{1}+\underbrace{\exp(-4000)}_{\simeq 0}+\underbrace{\exp(-1000)}_{\simeq 0}} \simeq 1$$

$$y_i \mid x_i \sim \text{Multinomiel}\left(y_i;\ 1,\ \{\Pr(y_i=c \mid x_i)\}_{c=1}^{K}\right)$$



$$\text{Multinomiel}\left(y_i;\ 1,\ \{1/6,\dots 1/6\}\right)$$

$$\Pr[\boxdot] = (1/6)^{0}(1/6)^{1}(1/6)^{0}(1/6)^{0}(1/6)^{0}(1/6)^{0}$$

$$\Pr[\boxdot,\boxdot] = (1/6)^{2}(1/6)^{1}(1/6)^{0}(1/6)^{0}(1/6)^{0}(1/6)^{0}$$

$$\mathbf{1}(y_i=c) \leftarrow$$

$$\text{likelihood}\left(\{w_c, w_{c0}\}_{c=1}^{K} \mid \chi\right) = \prod_{i=1}^{N}\prod_{c=1}^{K} \Pr(y_i=c \mid x_i)^{y_{ic}}$$

$$= \prod_{i=1}^{N}\prod_{c=1}^{K} \Pr(y_i=c \mid x_i)$$

$y_1 = 1 \rightarrow$ TRUE (1)
$y_1 = 2 \rightarrow$ FALSE (0)
$y_1 = 3 \rightarrow$ FALSE (0)

$$y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 3 \\ 2 \\ 1 \end{bmatrix} \xrightarrow{\text{one-hot encoding}} Y = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

$$\text{log-likelihood} = \sum_{i=1}^{N} \sum_{c=1}^{K} y_{ic} \log\left[Pr(y_i = c \mid x_i)\right]$$

y_{ic}

$$\text{Error}\left(\{w_c, w_{c0}\}_{c=1}^{K} \mid \mathcal{X}\right) = -\sum_{i=1}^{N} \sum_{c=1}^{K} y_{ic} \log(\hat{y}_{ic})$$

$0 \cdot \log(0) + 1 \cdot \log(1) + 0 \cdot \log(0)$

$0 \cdot \log(0) \to 0$; $\log(1) = 0$; $0 \cdot \log(0) \to 0$

$$\frac{s_1}{s_1+s_2+s_3}$$

$$= -\sum_{i=1}^{N} \sum_{c=1}^{K} y_{ic} \cdot \log\left[\frac{\exp\left[w_c^T x_i + w_{c0}\right]}{\sum_{d=1}^{K} \exp\left[w_d^T \cdot x_i + w_{d0}\right]}\right]$$

$\frac{\partial \text{Error}}{\partial w_c} = ?$ $\quad \frac{\partial \text{Error}}{\partial w_{c0}} = ?$

$$\frac{\partial f(s_1, s_2, s_3)}{\partial s_1} = \frac{1 \cdot (s_1+s_2+s_3) - s_1 \cdot (1)}{(s_1+s_2+s_3)^2}$$

$$\frac{\partial f(s_1, s_2, s_3)}{\partial s_2} = \frac{0 \cdot (s_1+s_2+s_3) - s_1 \cdot 1}{(s_1+s_2+s_3)^2}$$

$$\frac{\partial f(s_1, s_2, s_3)}{\partial s_i} = \frac{1(i=1) \cdot (s_1+s_2+s_3) - s_1}{(s_1+s_2+s_3)^2}$$

$$w_c^{(t+1)} = w_c^{(t)} + \Delta w_c^{(t)} \rightarrow -\eta \cdot \frac{\partial Error}{\partial w_c}$$

$$w_{co}^{(t+1)} = w_{co}^{(t)} + \Delta w_{co}^{(t)} \rightarrow -\eta \cdot \frac{\partial Error}{\partial w_{co}}$$

$$\Delta w_d = \eta \cdot \sum_{i=1}^{N} \sum_{c=1}^{K} \frac{y_{ic}}{\hat{y}_{ic}} \cdot \hat{y}_{ic} \cdot [\delta_{cd} - \hat{y}_{id}] \cdot x_i$$

$\delta_{cd} = 1(c=d)$

$$= \eta \sum_{i=1}^{N} (y_{id} - \hat{y}_{id}) \cdot x_i$$

$$\Delta w_{do} = \eta \sum_{i=1}^{N} (y_{id} - \hat{y}_{id})$$

ALGORITHM; STEP#1: Initialize $\{w_1, w_{10}, \ldots, w_K, w_{K0}\}$ randomly using Uniform$(-0.001, +0.001)$

STEP#2: Calculate gradients

STEP#3: Update $\{w_1, w_{10}, \ldots, w_K, w_{K0}\}$ using gradients

STEP#4: Go to STEP#2 if there is enough change in the parameters.