Linear Discrimination

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

initializing them to very small values

for example $\text{Uniform}(-0.001, +0.001)$

STEP #2: Calculate Δw and Δw_0.

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

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

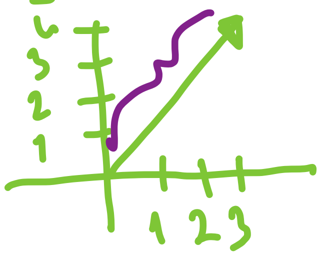
$$w_0^{(t+1)} = w_0^{(t)} + \Delta w_0^{(t)}$$

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.

$\|w\|_2 = ?$

$\|w\|_2 = 5 \quad w = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$



Linear Discrimination (Multiple classes $K > 2$)

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

$\hookrightarrow$ reference class

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

is a function of x — constant with respect to x

$$\log\left[\frac{Pr(y=c\mid x)}{Pr(y=K\mid x)}\right] = \log\left[\frac{p(x\mid y=c)}{p(x\mid y=K)}\right] + \log\left[\frac{Pr(y=c)}{Pr(y=K)}\right]$$

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

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

$$Pr(y=1\mid x) + Pr(y=2\mid x) + \ldots + Pr(y=K-1\mid x) + \boxed{Pr(y=K\mid c)} = 1$$

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

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

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

reference

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

nonreference

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

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

\# of peremeters $= (K-1)(D+1)$

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

x_{N+1} is a new data point

$$\left.\begin{aligned} g_1(x_{N+1}) &= w_1^T . x_{N+1} + w_{10} \\ g_2(x_{N+1}) &= w_2^T x_{N+1} + w_{20} \\ &\vdots \\ g_K(x_{N+1}) &= w_K^T . x_{N+1} + w_{K0} \end{aligned}\right\} \text{pick the maximum one}$$

IDEAL		SCORES		
1	$Pr(y=1 \mid x)$	$+2$	$= \dfrac{\exp(2)}{\exp(2)+\exp(-2)+\exp(1)}$	0.7214
0	$Pr(y=2 \mid x)$	-2	$= \dfrac{\exp(-2)}{\exp(2)+\exp(-2)+\exp(1)}$	0.0132
0	$Pr(y=3 \mid x)$	$+1$	$= \dfrac{\exp(1)}{\exp(2)+\exp(-2)+\exp(1)}$	0.2654
1	$Pr(y=1 \mid x)$	$+20$	$= \dfrac{\exp(20)}{\exp(20)+\exp(-20)+\exp(10)}$	0.9999
0	$Pr(y=2 \mid x)$	-20	$= \dfrac{\exp(-20)}{\exp(20)+\exp(-20)+\exp(10)}$	0.0000
0	$Pr(y=3 \mid x)$	$+10$	$= \dfrac{\exp(10)}{\exp(20)+\exp(-20)+\exp(10)}$	0.0000

2000

-2000

1000

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

$$\frac{\mathrm{Inf}}{\mathrm{Inf}} = \Pr(y=1|x)$$

$$+2000-2000=0$$

$$-2000-2000=-4000$$

$$+1000-2000=-1000$$

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

$$\Downarrow$$

$$\Pr(y=1|x) = 1$$

$$y_i | x_i \sim \text{Multinomial}(y_i ; 1, \{\Pr(y_i = c | x_i\}_{c=1}^{K})$$

⚀ ⚁ ⚂ ⚃ ⚄ ⚅ Multinomial$(y_i ; 1, \{1/6, \dots 1/6\})$

$$\Pr[⚁] = (1/6)^0 (1/6)^1 (1/6)^0 (1/6)^0 (1/6)^0 (1/6)^0$$

$$y_1 \{ 0 \ 1 \ 0 \ 0 \ 0 \ 0$$

$$\Pr[⚁, ⚀, ⚀] = (1/6)^2 (1/6)^1 (1/6)^0 (1/6)^0 (1/6)^0 (1/6)^0$$

$$\begin{array}{l} y_1 \\ y_2 \\ y_3 \end{array} \left\{ \begin{array}{cccccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{array} \right.$$

$$\text{likelihood}(\underline{\{w_c, w_{c0}\}}_{c=1}^{K} | \mathcal{X}) = \prod_{i=1}^{N} \prod_{c=1}^{K} \Pr(y_i = c | x)^{1(y_i = c)}$$

$$= \prod_{i=1}^{N} \prod_{c=1}^{K} \Pr(y_i = c | x)^{y_{ic}}$$

$y_1 = 2$

$y_2 = 1$

$y_3 = 3$

$y_4 = 2$

$$y = \begin{bmatrix} 2 \\ 1 \\ 3 \\ 2 \end{bmatrix} \text{ or } Y = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

one-hot encoding

$$\text{loglikelihood} = \sum_{i=1}^{N} \sum_{c=1}^{K} y_{ic} \cdot \log\Big[\underbrace{Pr(y_i = c \mid x)}_{\hat{y}_{ic}}\Big]$$

$$\text{Error}\left(\{w_c, w_{c0}\}_{c=1}^{K} \mid \mathcal{X}\right) = -\sum_{i=1}^{N} \sum_{c=1}^{K} \underbrace{y_{ic}}_{0} \cdot \underbrace{\log(\hat{y}_{ic})}_{\log(0)}$$

$$0 \cdot \log(0) \equiv 0$$

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

d e

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

$$\frac{\partial f_1(s_1, s_2, s_3)}{\partial s_1} = \frac{\partial \frac{\exp(s_1)}{\exp(s_1)+\exp(s_2)+\exp(s_3)}}{\partial s_1} = \frac{\exp(s_1) \cdot [\exp(s_1)+\exp(s_2)+\exp(s_3)] - \exp(s_1) \cdot \exp(s_1)}{[\exp(s_1)+\exp(s_2)+\exp(s_3)]^2}$$

$$\frac{\partial f_1(s_1, s_2, s_3)}{\partial s_2} = \frac{0 \cdot [\exp(s_1)+\exp(s_2)+\exp(s_3)] - \exp(s_2) \cdot \exp(s_1)}{[\exp(s_1)+\exp(s_2)+\exp(s_3)]^2}$$

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

$$w_{c_0}^{(t+1)} = w_{c_0}^{(t)} + \Delta w_{c_0}^{(t)} \rightarrow -\eta \frac{\partial \text{Error}}{\partial w_{c_0}}$$

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

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

$$\Delta w_{d_0} = \eta \sum_{i=1}^{N} (y_{id} - \hat{y}_{id})$$

ALGORITHM:

STEP#1: Initialize $\{w_1, w_{10}, \ldots, w_K, w_{K0}\}$ randomly to small values.

STEP#2: Calculate gradients. STEP#3: Update parameters using gradients.

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