Linear Discrimination

Classification $\Rightarrow$ $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$ $\quad y_i \in \{$

$g_1(x)$
$g_2(x)$
$\vdots$
$g_K(x)$
} score functions

$$\hat{y}_{N+1} = \arg\max_{c=1}^{K} g_c(x_{N+1})$$

$$g_c(x) = p(x \mid y=c)\, Pr(y=c)$$

$p(x \mid y=c)$ → univariate ($x_i \in \mathbb{R}$) / multivariate ($x_i \in \mathbb{R}^D$)

$Pr(y=c)$ → $\frac{N_c}{N}$: N_c = # of data points in class c, N = total # of data points

univariate: $N(x_i; \mu_c, \sigma_c^2)$ → $\hat{\mu}_c, \hat{\sigma}_c^2$

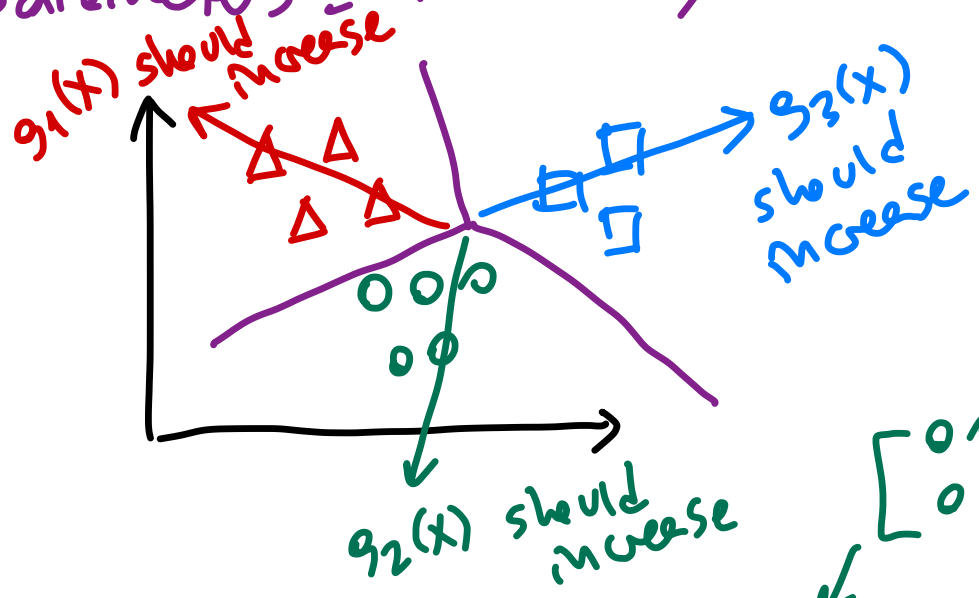
multivariate: $N(x_i; \mu_c, \Sigma_c)$ → $\hat{\mu}_c, \hat{\Sigma}_c$

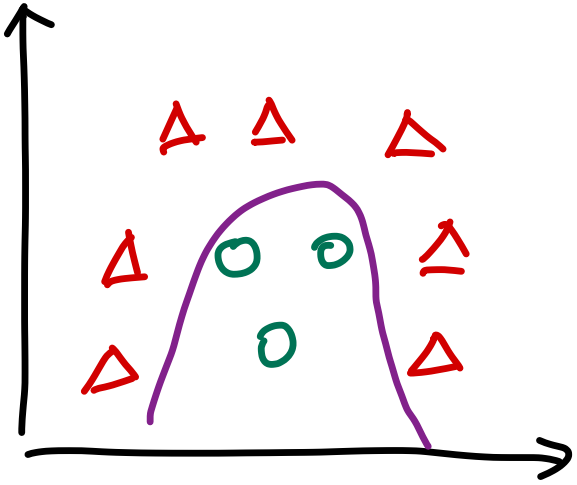
$$g_c(x \mid w_c, w_{c0}) = w_c^T \cdot x + w_{c0}$$

(w_c: $D \times 1$, w_{c0}: 1×1)

$$= \begin{bmatrix} w_{c1} & w_{c2} & \cdots & w_{cD} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} + w_{c0} = \sum_{d=1}^{D} w_{cd} x_d + w_{c0}$$

# of parameters = $K \cdot (D+1)$





$$g_c(x \mid W_c, w_c, w_{c0}) = x^T \cdot W_c \cdot x + w_c^T \cdot x + w_{c0}$$

# of parameters = $K\left(\frac{D \cdot (D+1)}{2} + D + 1\right)$

$$a^2 + 2ab + b^2$$

$$\begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

Binery Classification ($K=2$)

83 $g_1(x)$ } if $g_1(x) > g_2(x)$ $\Rightarrow \hat{y} = 1$

79 $g_2(x)$ } if $g_2(x) > g_1(x)$ $\Rightarrow \hat{y} = 2$

$g(x)$

if $g_1(x) - g_2(x) > 0$ $\Rightarrow \hat{y} = 1$

+4

if $g_1(x) - g_2(x) < 0$ $\Rightarrow \hat{y} = 2$

-3

if $g(x) > 0$ $\Rightarrow \hat{y} = 1$

if $g(x) < 0$ $\Rightarrow \hat{y} = 2$

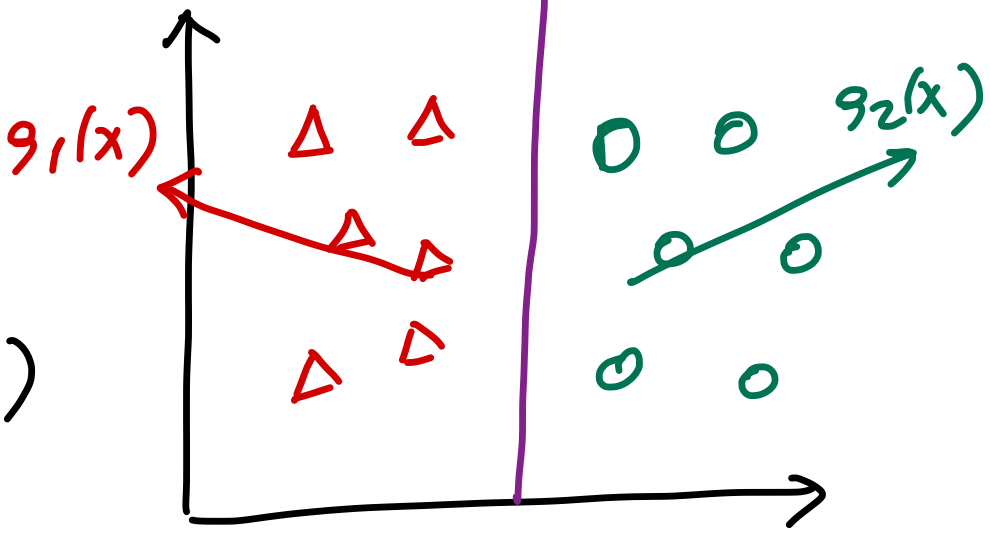
$g(x) \Rightarrow w^T x + w_0 = 0$

$g_1(x) = w_1^T \cdot x + w_{10}$

$g_2(x) = w_2^T \cdot x + w_{20}$

$g_1(x) - g_2(x) = (w_1 - w_2)^T \cdot x + (w_{10} - w_{20})$

w $\quad$ w_0



Multiclass Classification ($K > 2$)

73 tennis players

$g_1(x)$
$g_2(x)$
$\vdots$
$g_K(x)$

assume $K=3$

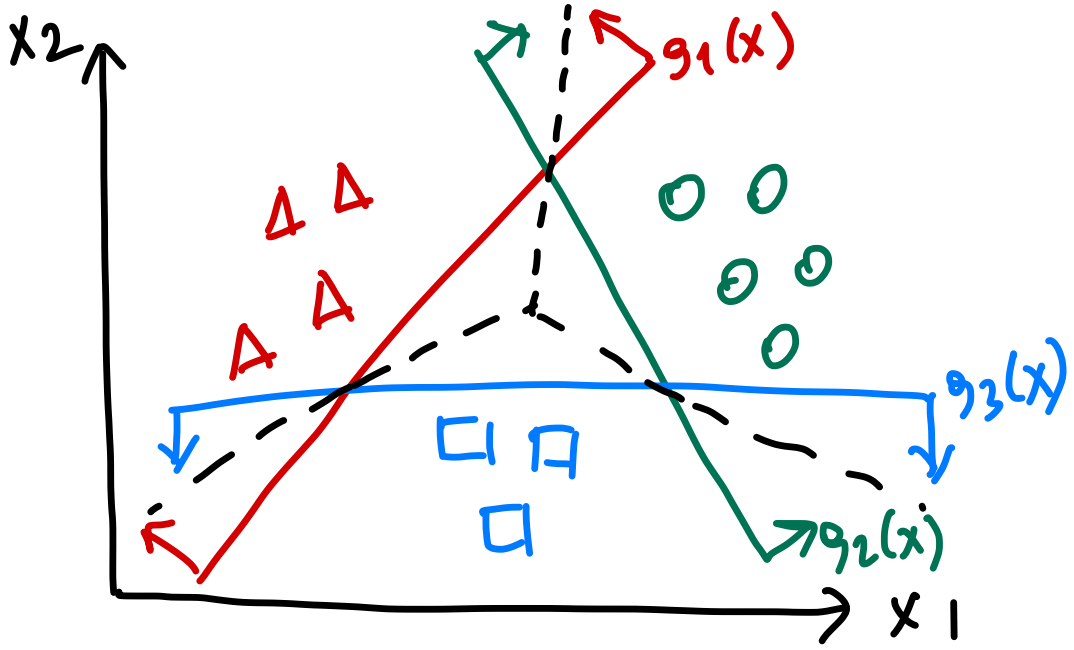
if $g_1(x) > g_2(x)$ & $g_1(x) > g_3(x) \Rightarrow \hat{y} = 1$

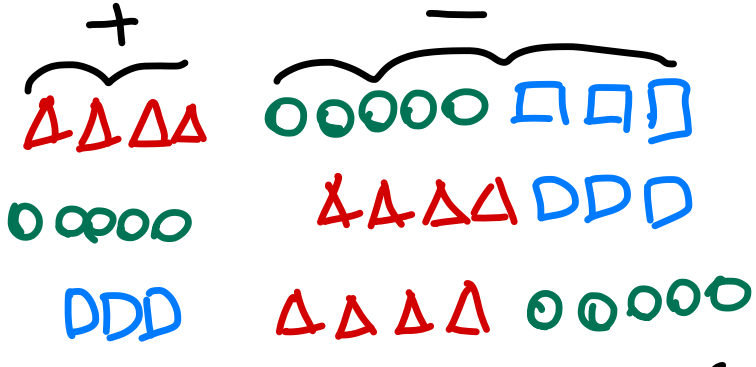
if $g_2(x) > g_1(x)$ & $g_2(x) > g_3(x) \Rightarrow \hat{y} = 2$

if $g_3(x) > g_1(x)$ & $g_3(x) > g_2(x) \Rightarrow \hat{y} = 3$

$$\hat{y} = \arg \max_{c=1}^{K} g_c(x)$$

ONE-VERSUS-ALL (OVA) APPROACH



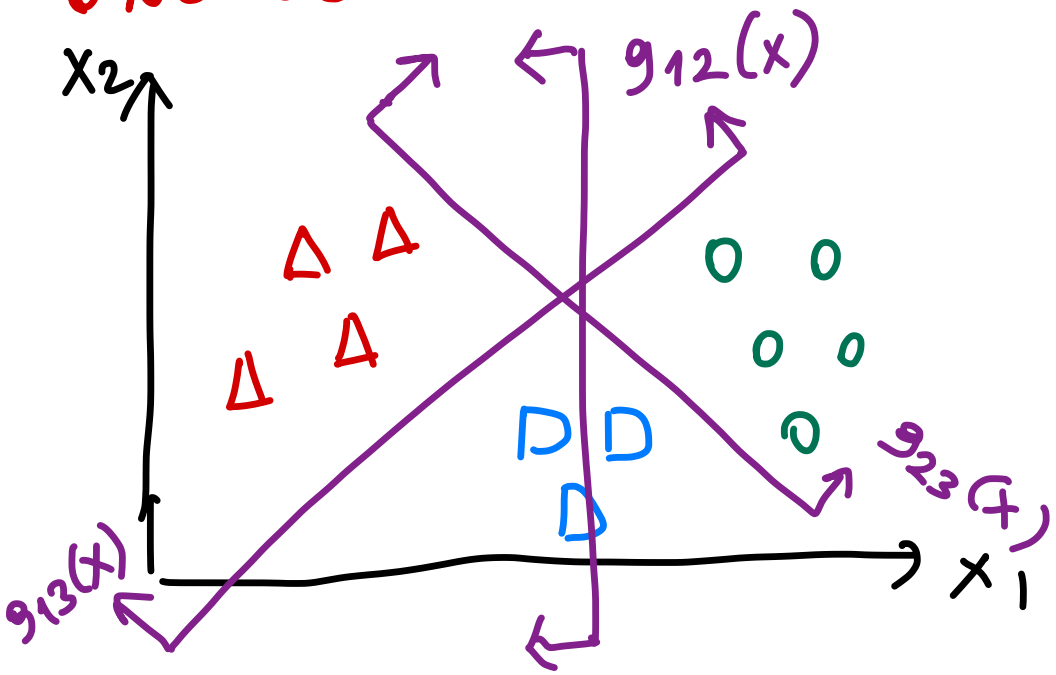


# of parameters $= K(D+1)$

data set size $= N$

# of score functions $= K$

ONE-VERSUS-OTHER (OVO) APPROACH



+	−
ΔΔΔΔ	OOOOO
ΔΔΔΔ	DDD
OOOOO	DDD

x	1	2	3
$g_{12}(x)$	1	0	∅
$g_{13}(x)$	1	∅	0
$g_{23}(x)$	∅	0	1
# of wins	2	0	1

$$\#\text{ of parameters} = \frac{K(K-1)}{2}\cdot(D+1)$$

$$\text{data set size} \simeq \frac{2N}{K}.$$

$$\#\text{ of score functions} = \frac{K(K-1)}{2}$$

K=2

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

odds ratio

log odds ratio

$$\begin{cases} \text{if } \delta > 0.5 & \Rightarrow \hat{y}=1 \\ \text{if } \dfrac{\delta}{1-\delta} > 1 & \Rightarrow \hat{y}=1 \\ \text{if } \log\left[\dfrac{\delta}{1-\delta}\right] > 0 & \Rightarrow \hat{y}=1 \end{cases}$$

$$\log\left[\frac{Pr(y=1|x)}{Pr(y=2|x)}\right] = \log\left[\frac{\frac{p(x|y=1)\,Pr(y=1)}{p(x)}}{\frac{p(x|y=2)\,Pr(y=2)}{p(x)}}\right]$$

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

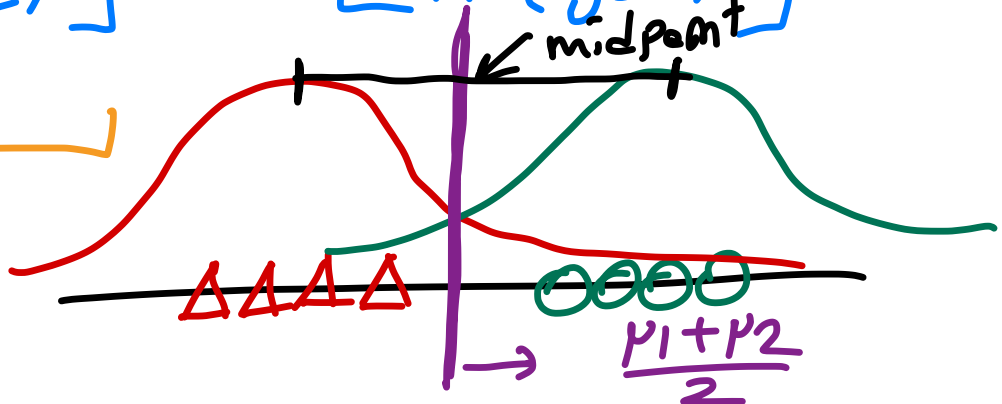
$N(x;\mu_1,\Sigma_1)$ $N(x;\mu_2,\Sigma_2)$

A B

$$\log\left(\frac{a \cdot c}{b d}\right) = \log\left(\frac{a}{b}\right) + \log\left(\frac{c}{d}\right)$$

assuming

$$\Sigma = \Sigma_1 = \Sigma_2 \;,\; Pr(y=1) = Pr(y=2)$$



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

Shared covariance assumption: $\Sigma_1 = \Sigma_2 = \Sigma$

$$\log\left[\frac{\frac{1}{\sqrt{(2\pi)^D|\Sigma_1|}} \cdot \exp\left[-\frac{1}{2}(x-\mu_1)\Sigma_1^{-1}\cdot(x-\mu_1)\right]}{\frac{1}{\sqrt{(2\pi)^D|\Sigma_2|}} \cdot \exp\left[-\frac{1}{2}(x-\mu_2)^T\Sigma_2^{-1}(x-\mu_2)\right]}\right] + \log\left[\frac{Pr(y=1)}{Pr(y=2)}\right]$$

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

(A)

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

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

linear · constant · linear · constant

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

constant

$$= \underbrace{w^T \cdot x}_{} + \underbrace{w_0}_{}$$

$\hat{\Sigma}$ = shared covariance estimate (calculated on the whole dataset)

$$\underbrace{w}_{D\times 1} = \underbrace{\hat{\Sigma}^{-1}}_{D\times D} \cdot \underbrace{(\hat{\mu}_1 - \hat{\mu}_2)}_{D\times 1}$$

$$\underbrace{w_0}_{1\times 1} = -\frac{1}{2}\underbrace{(\hat{\mu}_1 + \hat{\mu}_2)^T}_{1\times D}\underbrace{\hat{\Sigma}^{-1}}_{D\times D} \cdot \underbrace{(\hat{\mu}_1 - \hat{\mu}_2)}_{D\times 1} + \underbrace{\log\left[\frac{\hat{Pr}(y=1)}{\hat{Pr}(y=2)}\right]}_{1\times 1}$$