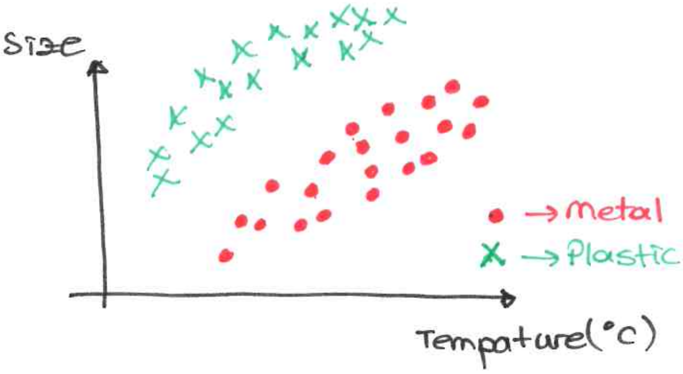
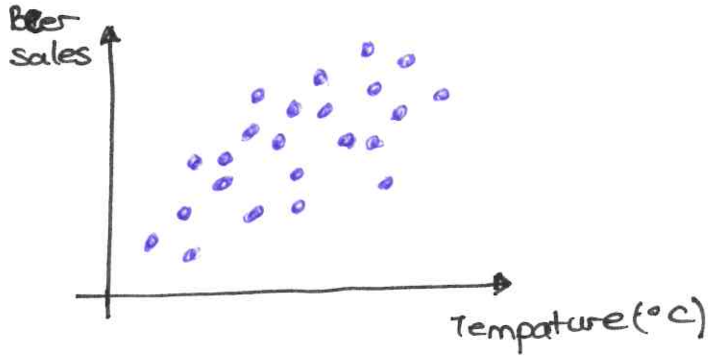
# Exercise - 2 / Question - 1

①




## ② Classification

⇒ We have training set $\mathbf{X} = \{x_1, \dots, x_N\}$, i.e we have N data points.

⇒ we know target values $\mathbf{T} = \{t_1, \dots, t_N\}$ for each data point in training set $\mathbf{X}$.

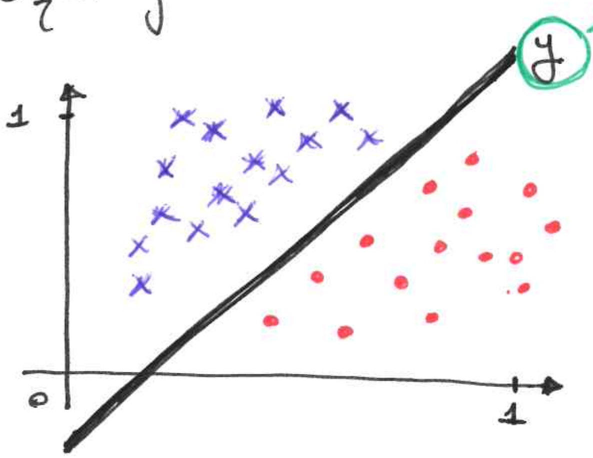
**Goal:** for a given new data $x$, classify these new data to one of K classes $C_k$.

**2-class**

$t \in \{0,1\}$

**K-class**

$t_n = (0,1,0,0)^T$




⇒ In this question, we have 2-class problem

⇒ Define a hyperplane $y$ such that:

$$\boxed{y(x) = w^T x + w_0}$$

**Goal:** Finding the hyperplane that seperates the data points into two classes (blue or red).

⇒ Let's consider a single data point:

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} \qquad w = \begin{pmatrix} w_1 \\ \vdots \\ w_D \end{pmatrix}$$

$x$: * $D \times 1$ "D dimensional vector"

$w$: * $D \times 1$ "D dimensional vector"

$$\Rightarrow \quad y(x) = w^T x + w_0 \qquad 1 \times D \times D \times 1$$

$$= \sum_{d=1}^{D} w_i x_i + w_0$$

$$\underbrace{(w_1 \; \cdots \; w_D) \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix}}_{\substack{1 \times D \times D \times 1 \\ 1 \times 1}} + w_0 \; !$$

$$\tilde{x} = \begin{pmatrix} x_0 = 1 \\ x_1 \\ \vdots \\ x_D \end{pmatrix} \; ! \qquad \tilde{w} = \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix} \; !$$

$$\boxed{\begin{aligned} y(x) &= w^T x + w_0 \\ &= \sum_{d=1}^{D} w_i x_i + w_0 \\ &= \sum_{d=0}^{D} w_i \tilde{x}_i \\ &= \tilde{w}^T \tilde{x} \end{aligned}}$$

! In that way, this equation becomes the multiplication of two vectors of size $(D+1) \times 1$.

⇒ Now, let's generalize it to K-class classification problem:

$$\boxed{y_k(x) = w_k^T x + w_{k0} \qquad k = 1, \ldots, K}$$

⇒ Turn this equation to vector notation:

(1)
$$y_k(x) = w_k^T x + w_{k0} \qquad k = 1, \dots, K$$

(2)
$$y(x) = \tilde{W}^T \tilde{x} = [\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_K]^T \tilde{x}$$

We know that, $\tilde{w}_k$ is a vector of size $(D+1) \times 1$

(3)
$$\tilde{W}^T = [\tilde{w}_1, \tilde{w}_2, \dots, \tilde{w}_K]^T$$

$$= \begin{array}{c} \left( \begin{array}{cccc} w_{10} & w_{20} & \cdots & w_{K0} \\ w_{11} & w_{21} & & w_{K1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1D} & w_{2D} & \cdots & w_{KD} \end{array} \right)^T \\ \begin{array}{cccc} (D+1)\times 1 & (D\times 1)\times 1 & & (D\times 1)\times 1 \end{array} \end{array}$$

when, we take transpose of this.

$$= \underbrace{\begin{pmatrix} w_{10} & w_{11} & \cdots & w_{1D} \\ w_{20} & w_{21} & \cdots & w_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K0} & w_{K1} & \cdots & w_{KD} \end{pmatrix}}_{K \times (D+1)}$$

(4)
$$y(x) = \tilde{W}^T \cdot \tilde{x}$$

$$K \times \cancel{(D+1)} \quad \cancel{(D+1)} \times 1$$

$$= K \times 1$$

As the output's a vector of $K \times 1$; we can directly compare it with target values.

⇒ Until now, we have talked about the equations for a single data point $\tilde{x}$.

⇒ Let's write them for the entire dataset with N data points:

(1)

$$\boxed{Y(\tilde{X}) = \tilde{X}\tilde{W}}$$

The entire dataset with N data points

A little explanation: why it is written like $\tilde{X}\tilde{W}$ and ~~why not $\tilde{W}^T\tilde{X}$~~

Actually, when we take transpose of $(\tilde{W}^T\tilde{X})$

$$(\tilde{W}^T\tilde{X})^T = \tilde{X}^T\tilde{W}$$

In the notation, we assume that $\tilde{X}^T$ is $\tilde{X}$.

(2)

$$\tilde{W} = [\tilde{w}_1, \text{-----}, \tilde{w}_K]$$

$\tilde{w}_1$: $(D+1) \times 1$ ; $\tilde{w}_K$: $(D+1) \times 1$

$\Rightarrow (D+1) \times K$ *

(3)

$$\tilde{X} = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} = \begin{bmatrix} x_{10} & x_{20} & \cdots & x_{N0} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1D} & x_{2D} & \cdots & x_{ND} \end{bmatrix}^T$$

$$= \begin{bmatrix} x_{10} & \cdots & x_{1D} \\ x_{20} & \cdots & x_{2D} \\ \vdots & \ddots & \vdots \\ x_{N0} & \cdots & x_{ND} \end{bmatrix}$$

$x_1 \rightarrow (D+1)$

$x_1^T \rightarrow 1 \times (D+1)$

$\Rightarrow N \times (D+1)$ *

4

$$T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix}$$

where $\tilde{t}$ is the target value for a single data point $\tilde{x}$ which is expressed as $\tilde{t} = [t_1, \ldots, t_K]^T$

$(K \times 1)$

for a single data point

$$T = \begin{bmatrix} t_1^T \\ \vdots \\ t_N^T \end{bmatrix} = \begin{bmatrix} t_{11} & \cdots & t_{N1} \\ t_{12} & \cdots & t_{N2} \\ \vdots & \ddots & \\ t_{1K} & \cdots & t_{NK} \end{bmatrix}^T = \begin{bmatrix} t_{11} & \cancel{t_{12}} & \cdots & t_{1K} \\ t_{21} & t_{22} & \cdots & t_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ t_{N1} & t_{N2} & \cdots & t_{NK} \end{bmatrix}$$

$t_1 \rightarrow K \times 1$

$t_1^T \rightarrow 1 \times K$

when we ~~take~~ transpose

$\Rightarrow N \times K$ *

5

$$Y(\tilde{X}) = \tilde{X} \cdot \tilde{W}$$

$$\boxed{N \times \cancel{(D+1)} \quad \cancel{(D+1)} \times K}$$

$$\boxed{N \times K}$$

$$Y(\tilde{X}) = \tilde{X} \cdot \tilde{W} \quad \Rightarrow N \times K \text{*}$$

⇒ As the output vector T has the size of N x K we can directly compare $\tilde{X}\tilde{W}$ with T:

$$\boxed{y(\tilde{X}) - T = \underbrace{\tilde{X}\tilde{W}}_{(N \times K)} - \underbrace{T}_{(N \times K)} \quad \Rightarrow (N \times K)}$$

**Goal**: Choose $\tilde{W}$ such that this is minimal!

⇒ We directly try to minimize the sum of squares error:

(1)
$$E_D(\tilde{w}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} \left( y_k(x_n; w) - t_{kn} \right)^2$$

$$= \underbrace{\frac{1}{2} \sum_{k=1}^{K} \sum_{n=1}^{N} \left( w_k^T x_n - t_{kn} \right)^2}$$

↳ $\frac{1}{2} (\tilde{X}\tilde{W} - T)^2$

(2)
$$\boxed{E_D(\tilde{w}) = \frac{1}{2} \operatorname{Tr}\left( (\tilde{X}\tilde{W} - T)^T (\tilde{X}\tilde{W} - T) \right)}$$

Use:

$$\boxed{\sum_{i,j} a_{ij}^2 = \operatorname{Tr}\{A^T A\}}$$

where $a_{ij} \rightarrow w_k x_n - t_{kn}$

and A is $(\tilde{X}\tilde{W} - T)$

⇒ Take the derivative of $E_D(\tilde{w})$ for minimizing

(1)
$$\boxed{\frac{\partial E_D(\tilde{w})}{\partial \tilde{w}} = \frac{1}{2}\frac{\partial}{\partial \tilde{w}} Tr\left((\tilde{X}\tilde{w}-T)^T(\tilde{X}\tilde{w}-T)\right)}$$

use chain rule

$$\boxed{\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}}$$

where $z \to Tr\left((\tilde{X}\tilde{w}-T)^T(\tilde{X}\tilde{w}-T)\right)$

$x \to \tilde{w}$

$$\frac{\partial z}{\partial x} = \frac{\partial Tr\left((\tilde{X}\tilde{w}-T)^T(\tilde{X}\tilde{w}-T)\right)}{\partial \tilde{w}}$$

So, y ??

y will be $(\tilde{X}\tilde{w}-T)^T(\tilde{X}w-T)$ so

we get that:

(2)
$$\boxed{\frac{\partial E_D(\tilde{w})}{\partial \tilde{w}} = \frac{1}{2}\frac{\partial Tr\left((\tilde{X}\tilde{w}-T)^T(\tilde{X}\tilde{w}-T)\right)}{\partial\left((\tilde{X}\tilde{w}-T)^T(\tilde{X}\tilde{w}-T)\right)} \cdot \frac{\partial (X\tilde{w}-T)^T(\tilde{X}\tilde{w}-T)}{\partial \tilde{w}}}$$

use:

$$\frac{\partial}{\partial A} Tr\{A\} = I$$

for us, $A \to (\tilde{X}\tilde{w}-T)^T(\tilde{X}\tilde{w}-T)$

(3)
$$\boxed{\frac{\partial E_D(\tilde{w})}{\partial \hat{w}} = \frac{1}{2} \cdot I \cdot \frac{\partial}{\partial \tilde{w}}(\tilde{X}\tilde{w}-T)^T(\tilde{X}\tilde{w}-T)}$$

4

$$\frac{\partial E_D(\tilde{w})}{\partial \tilde{w}} = \frac{1}{\cancel{2}} \cdot \cancel{2} \cdot \tilde{X}^T(\tilde{X}\tilde{w} - T)$$

$$= \tilde{X}^T(\tilde{X}\tilde{w} - T)$$

⇒ In order to minimize the sum-of-squares error, set the derivative $\frac{\partial E_D(\tilde{w})}{\partial \tilde{w}}$ to 0.

5

$$\frac{\partial E_D(\tilde{w})}{\partial \tilde{w}} = \cancel{\tilde{X}^T}(\tilde{X}\tilde{w} - T) \overset{!}{=} 0$$

Since $\tilde{x}^T$ cannot be 0 as it is our data points, only $(\tilde{X}\tilde{w} - T)$ can be 0.

6

$$\tilde{X}\tilde{w} - T = 0$$

$$\tilde{X}\tilde{w} = T$$

$$\tilde{X}^T \cdot \tilde{X} \cdot \tilde{w} = \tilde{X}^T T$$

In order to leave $\tilde{w}$ alone first take transpose of $\tilde{X}$ on both side of the equation

7

$$\tilde{X}^T\tilde{X}\tilde{w} = \tilde{X}^T T$$

$$\underbrace{(\tilde{X}^T\tilde{X})^{-1}(\tilde{X}^T\tilde{X})}_{I} \cdot \tilde{w} = \underbrace{(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T}_{\tilde{X}^+} \cdot T$$

multiply both side of equation w/ $(\tilde{X}^T\tilde{X})^{-1}$.

$$I\,\tilde{w} = \tilde{X}^+ T$$

$$\boxed{\tilde{w} = \tilde{X}^+ \cdot T}$$

where $\tilde{X}^+$ is "pseudo-inverse"