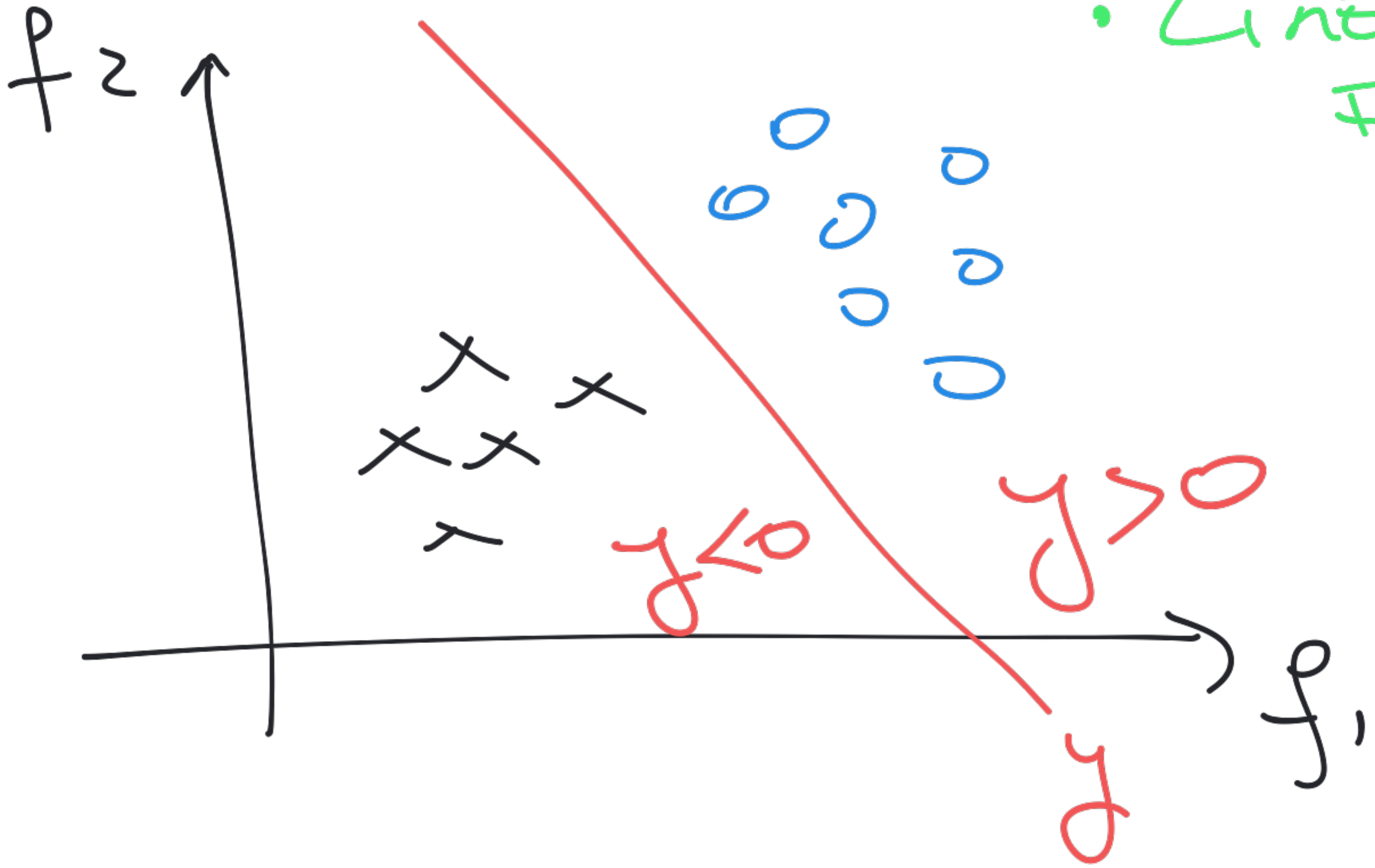
Discriminative Classifiers

↳ partition the feature into regions for each class.



- Linear Discriminant Functions
  ↳ finds a linear boundary

$$y = w^T x + b$$

One approach to find a linear discriminant function / boundary is by using

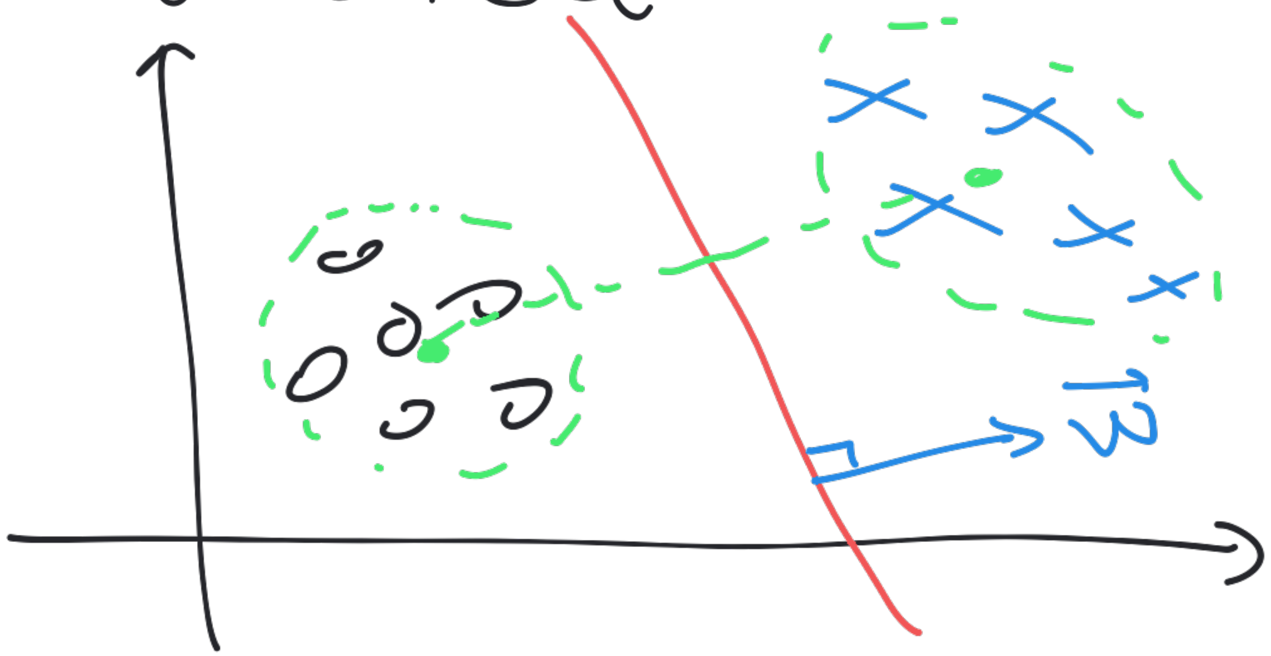
Fisher's Linear Discriminant

↳ Goal: maximize distance of means
AND
minimize within-class variance.

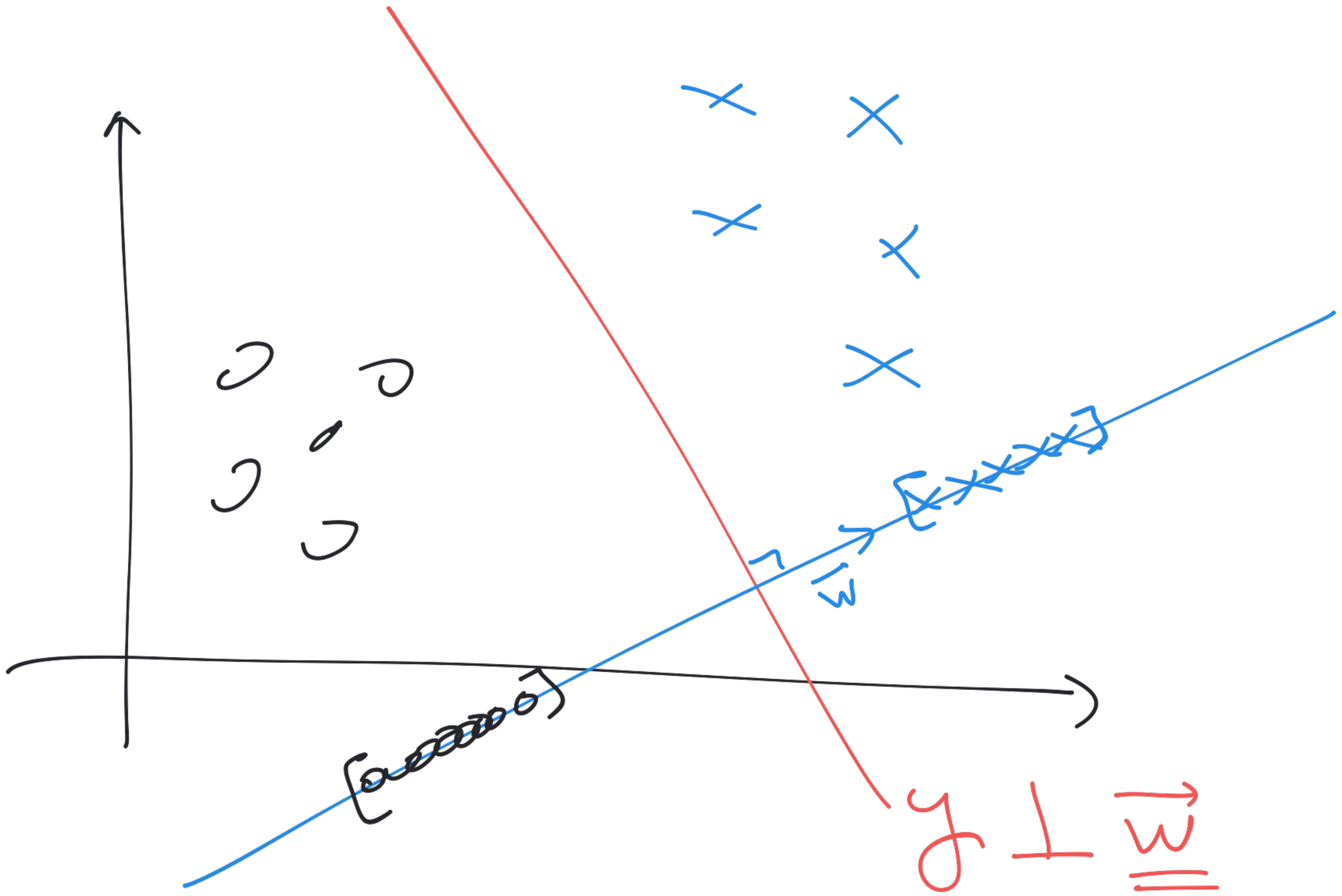
$$J(w) = \frac{w^T \cdot S_B\, w}{w^T \cdot S_{W} \cdot w}$$

$$S_B = (\vec{m_2} - \vec{m_1})(\vec{m_2} - \vec{m_1})^T \equiv \text{between class scatter matrix}$$

$$S_W = \sum_{i=1}^{2} \left[\sum_{n \in C_i} (\vec{x_n} - \vec{m_i})(\vec{x_n} - \vec{m_i})^T \right]$$



$\vec{w}$: direction of projection onto which class remain separable



$y \perp \underline{\underline{\vec{w}}}$

$\parallel$

boundary or discriminant function

$$y = w^T . x + \underline{\underline{b}}$$

In vector/matrix notation:

$$y = X.w \quad , \text{where } X = \underbrace{\begin{bmatrix} 1 & x_{1,1} & x_{1,2}, \cdots & x_{1,D} \\ 1 & x_{2,1} & x_{2,2}, \cdots, & x_{2,D} \\ \vdots & & & \\ 1 & & & \end{bmatrix}}_{N \times (D+1)}$$

$$\text{and } w = \begin{bmatrix} b \\ \hline w_1 \\ w_2 \\ \vdots \\ w_D \end{bmatrix}$$

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Optimize J :

$$\arg\max_{w} J(w)$$

$$\frac{\partial J(w)}{\partial w} = 0 \iff \frac{2(w^T S_W w) S_B w - 2(w^T S_B w) S_W \cdot w}{(w^T S_W w)^2} = 0$$

$$\iff \frac{S_B \cdot w}{w^T S_W \cdot w} - \frac{(w^T S_B \cdot w) S_W \cdot w}{(w^T S_W w)^2} = 0$$

$$\iff S_B \cdot w = \boxed{\frac{w^T S_B w}{w^T S_W \cdot w}} \cdot S_W \cdot w \iff$$ next page

multiply numerator by $w^T S_W \cdot w$

λ

$$S_B w = \lambda \cdot S_w w \quad , \quad \lambda \text{ is constant}$$

$$\lambda = \frac{w^T S_B w}{w^T S_w w}$$

$$\Longleftrightarrow \underbrace{S_w^{-1} S_B}_{\text{matrix}} w = \lambda \cdot w$$

if S_w^{-1} Exists

matrix

Linear Transformation

constant

This is the generalized Eigenvalue problem.

$\Longrightarrow$ w is Eigenvector of $S_w^{-1} \cdot S_B$.

$$S_W^{-1} = (\Sigma_1 + \Sigma_2)^{-1}$$

$$S_B = (\vec{m_2} - \vec{m_1})(\vec{m_2} - \vec{m_1})^T$$

Plugging in the values:

$$S_W^{-1}(\vec{m_2} - \vec{m_1})\underbrace{(\vec{m_2} - \vec{m_1})^T \vec{w}}_{\alpha \text{ is constant}} = \lambda \cdot \vec{w}$$

$$\Leftrightarrow \boxed{S_W^{-1}(\vec{m_2} - \vec{m_1}) = \frac{\lambda}{\alpha} \cdot \vec{w}}$$

Because we only care about direction of $\vec{w}$, not its magnitude,

$$\underline{\underline{\vec{w}}} = \underline{\underline{S_w^{-1}}} \underbrace{(\vec{m_2} - \vec{m_1})}$$

but we need to make it a unit vector, i.e.

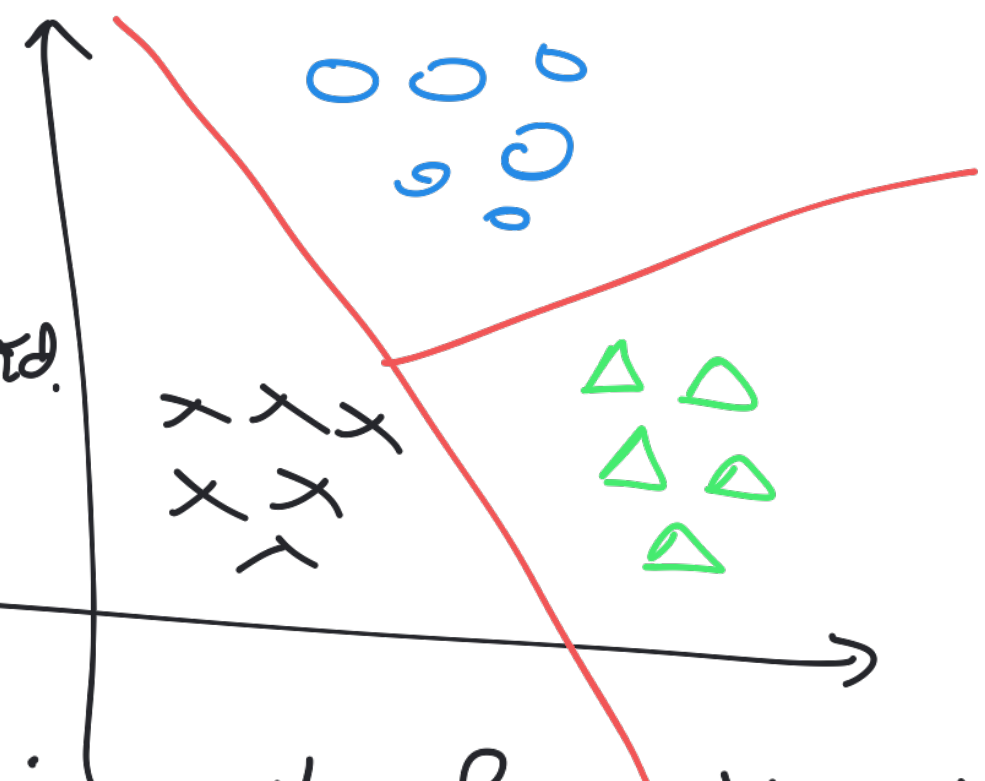
$$\boxed{\vec{w}^* = \frac{\vec{w}}{\underline{\underline{\|\vec{w}\|}}}}$$

gives us the direction of projection, which is the $\perp$ direction of the discriminant function y.

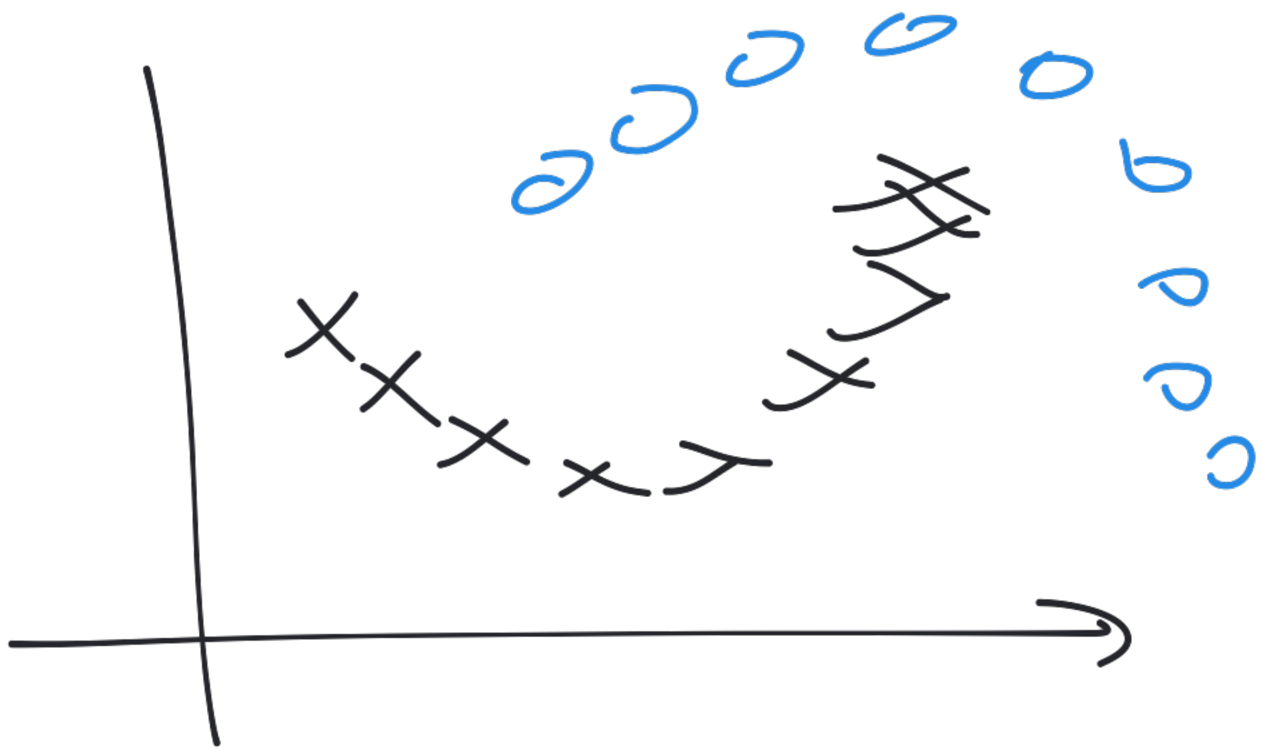
Limitations of Fisher's Discriminant Function:

① For a multi-class problem, we will find each boundary as a one-vs-all approach

② Assumes classes are Gaussian distributed.



③ It can only find linear discriminant functions.

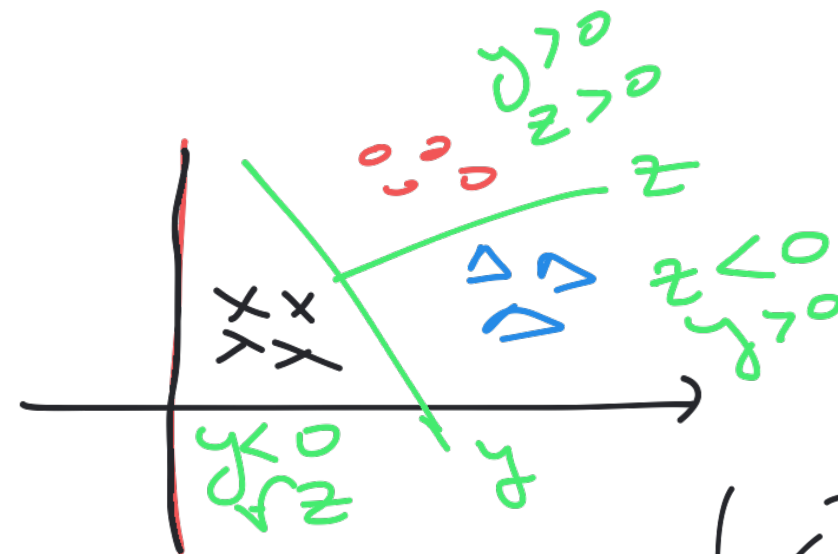


→ when data has a non-linear relationship, Fisher's cannot linearly separate the two classes.

↳ A popular variant Fisher's are Multi-Layer Perceptron (MLP)

Another approach is to find $\vec{w}$ as the Least Squares solution.

$\{x, t\}$, $\underline{y = w^T x + b}$



$$\boxed{J(w) = \sum_{i=1}^{N} (t_i - y_i)^2} \quad \underset{w}{\arg\min}\, J(w) \Rightarrow \underline{w = (X^T X)^{-1} X^T t}$$

page 185 Bishop

→ If we consider y as the linear regression, then y is a continuous-valued number ≠ class label

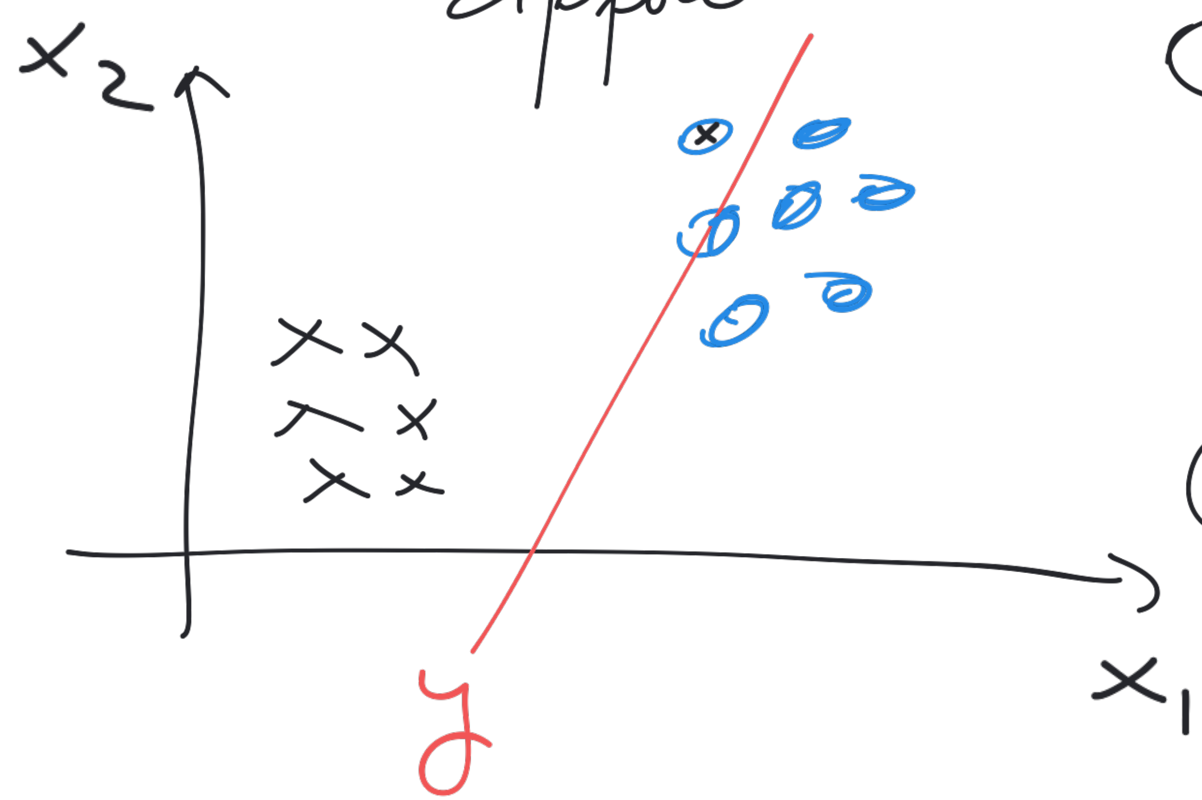
→ $y > 0 \Rightarrow$ assign to class 1
$y < 0 \Rightarrow$ assign to class 2.

→ From Bayesian interpretation, we saw that the errors of Least Squares solution are Gaussian-distribution.
↳ Does not hold when we use regression as a classification problem.

∴ Avoid using regression for a classification problem.

The Perceptron Algorithm

↳ search parameter space
by gradient-descent
approach



① Initialize the parameters of discriminant function

② Use gradient descent to find which direction to move discriminant function