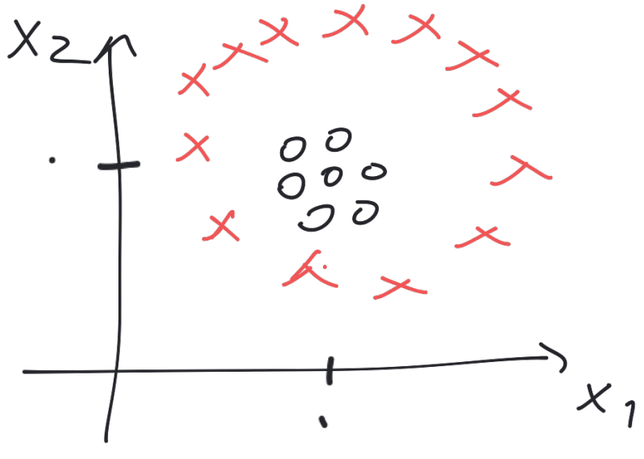
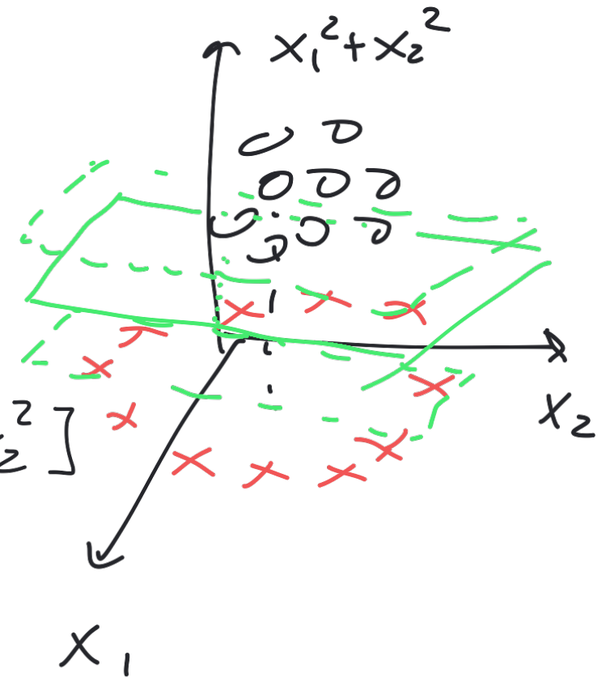
SVM



Transform flat data in higher-dimensional space.

$$\phi(x) = [x_1, x_2, x_1^2 + x_2^2]$$



SVM: ① Linear Classifier

② Works with a data transformation
— MAPPING: $\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$

③ $\phi(x)$ can be any $x \longmapsto \phi(x)$ transformation, $\phi(x)$ is also has basis functions.

$\phi(x)$ can be infinite dimensional!!

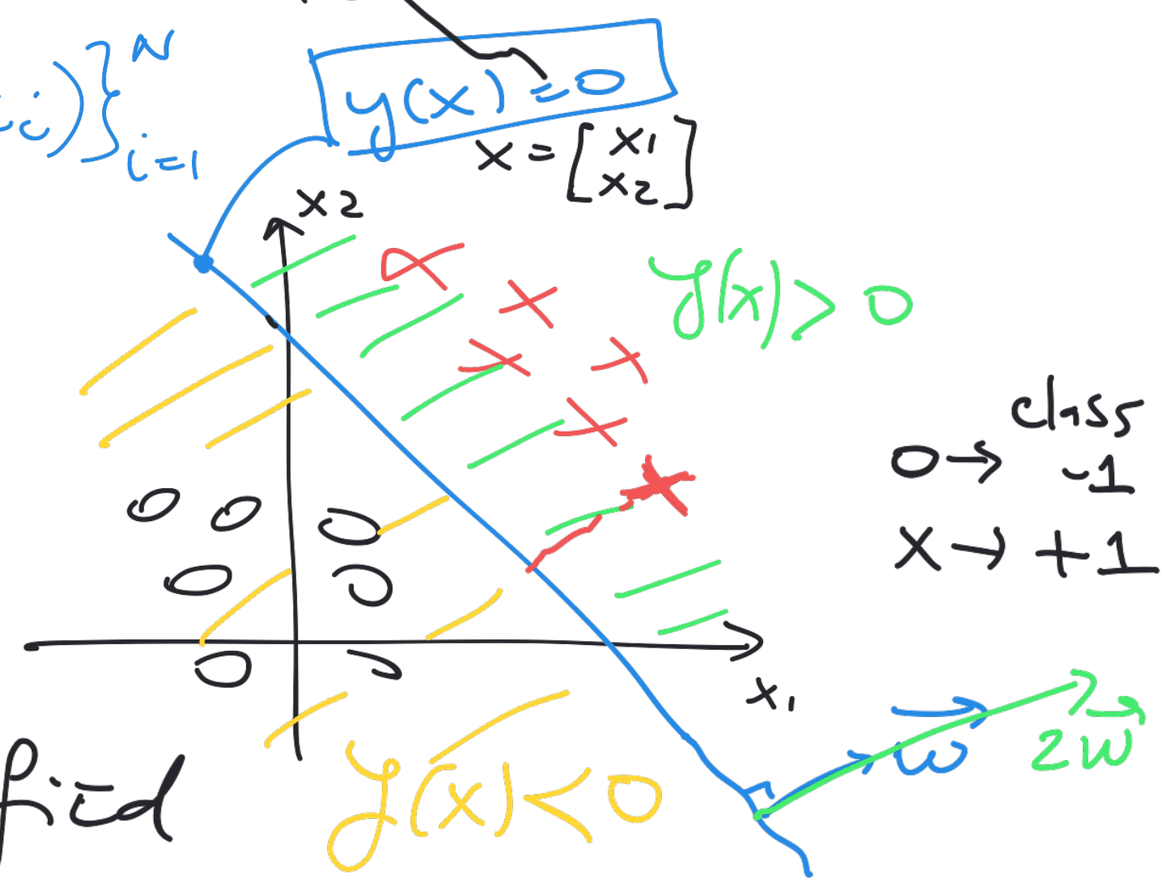
$$y(x) = \underline{w}^T \cdot \underline{\phi(x)} + b$$

if $\phi(x)$ is 1-D then: $y(x) = m\phi(x) + b$

Suppose we have data $\{(x_i, t_i)\}_{i=1}^{N}$ and $t_i \in \{-1, 1\}$

$$y(x_n) \cdot t_n > 0$$

$\forall x_n$ correctly classified



Distance of a point x_n to decision surface:

$$\frac{t_n \cdot y(x_n)}{\|w\|} = \frac{t_n(w^T\phi(x_n)+b)}{\|w\|}$$

↳ we want to maximize this distance $\forall x_n$!

↳ But we only the values x_n for support vectors!!

↳ they have smallest dist. to surface.

$$\text{dist.} = \frac{t_n y(x_n)}{\|w\|}$$

$$\arg\max_{w,b} \left\{ \frac{1}{\|w\|} \min_n \left[t_n (w^T \phi(x_n) + b) \right] \right\}$$

Scaling both $\begin{cases} w \to \kappa w \\ b \to \kappa b \end{cases}$ and

it does not affect direction nor the distance to the decision surface.

Then, we will consider support vectors to have distance:

$$t_n(w^T\phi(x_n)+b)=1$$

$\forall x_n \equiv$ support vectors.

Any other x_n will have distance:

$$t_n(w^T\phi(x_n)+b)>1$$

For all data points:

$$t_n(w^T\phi(x_n)+b) \geq 1, \forall x_n$$

Find w that defines maximum margin such that (constrained to)

$$t_n(w^T\phi(x_n)+b) \geq 1.$$

$$\arg\max_{w,b} \frac{1}{\|w\|} \quad \text{subject to} \quad t_n(w^T\phi(x_n)+b) \geq 1$$

$$\Leftrightarrow \boxed{\arg\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad t_n(w^T\phi(x_n)+b) \geq 1}$$

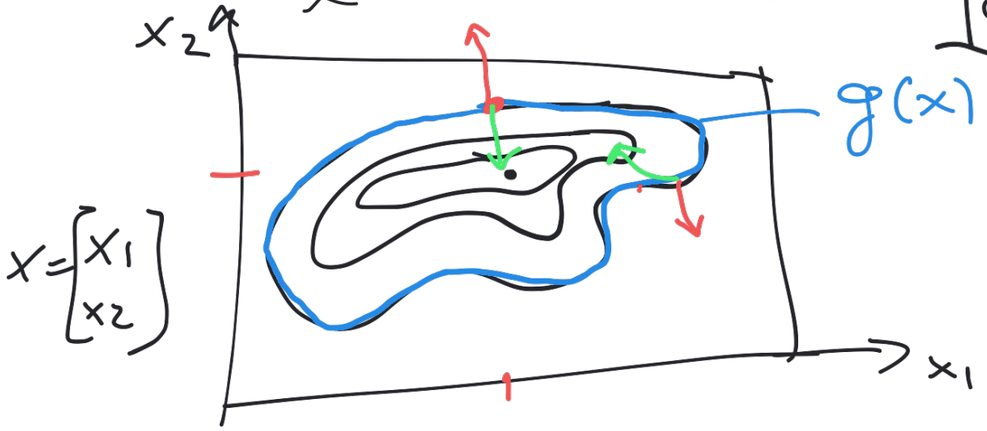
Optimization of a function with equality/inequality constraint.

① $\min_x f(x)$

NECESSARY Condition

$f'(x) = 0$

② $\min_x f(x)$ s.t. $g(x) = 0$



$$\mathcal{L}(x, \lambda) = f(x) - \lambda \cdot g(x)$$

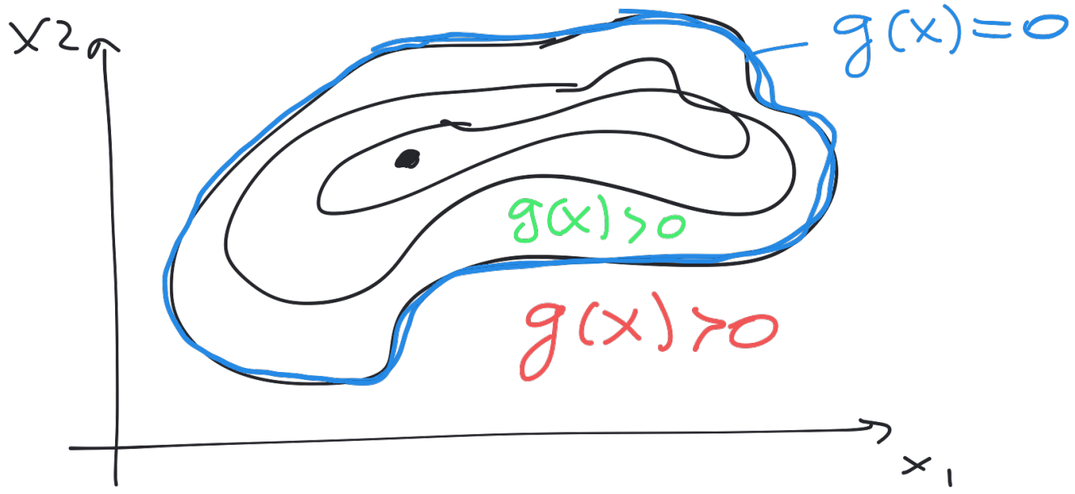
Lagrange multiplier

NECESSARY condition:

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \Rightarrow f'(x) = \lambda \cdot g'(x)$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \Rightarrow + g(x) = 0$$

(3) $\min_x f(x)$ s.t. $g(x) \geq 0$



Necessary condition:

$$\mathcal{L}(x,\lambda) = f(x) - \lambda g(x)$$

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = f'(x) - \lambda g'(x) = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} = -g(x) = 0 \end{cases}$$

Case 1: INACTIVE constraint: $g(x) > 0$, $\lambda = 0$

Case 2: Active constraint: $g(x) = 0$, $\lambda > 0$

KKT condition:

$$\begin{cases} g(x) \geq 0 \\ \lambda \geq 0 \\ \lambda \cdot g(x) = 0 \end{cases}$$

Lagrangian Multiplier for our SVM constrained optimization:

$$\mathcal{L}(w, b, a) = \frac{1}{2}\|w\|^2 - \sum_{n=1}^{N} a_n \left(t_n(w^T \phi(x_n) + b) - 1\right)$$

PRIMAL

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0 \Leftrightarrow \underline{w} = \sum_{n=1}^{N} a_n \cdot t_n \cdot \phi(x_n) \\ \frac{\partial \mathcal{L}}{\partial b} = 0 \Leftrightarrow \boxed{\sum_{n=1}^{N} a_n t_n = 0} \end{cases}$$

↳ note that if $\phi(x)$ is infinite dimensional, we CANNOT compute this.

By plugging in the solution for w, we are able rewrite PRIMAL $\mathcal{L}$ into its DUAL representation.

DUAL :

$$\mathcal{L}(a) = \sum_{n=1}^{N} a_n - \sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m K(x_n, x_m)$$

such that $\sum_{n=1}^{N} a_n t_n = 0$ and

$a_n \geq 0$, $\forall n = 1, 2, \ldots, N$

$K(x_n, x_m)$ known as GRAM MATRIX

↓

KERNEL operation between points x_n and x_m.

The most popular KERNEL function is RADIAL BASIS FUNCTIONS (RBF)

RBF

$$K(x,y) = \exp\left(-\gamma \|x-y\|^2\right)$$

γ controls the variance

$\gamma \to \infty \to$ smaller variance

$\gamma \to 0 \to$ large variance.

Proximity (or similarty) between two points x and y.