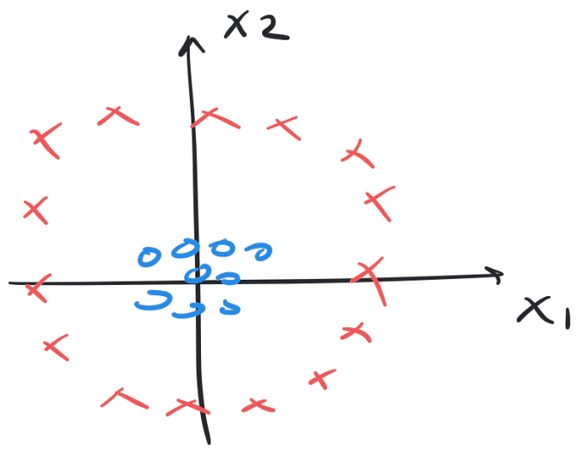
HARD-MARGIN SUPPORT VECTOR MACHINE (SVM)

MAPPER: $y(x) = \omega^T \phi(x) + b$

$\phi(x) \equiv$ transformation space

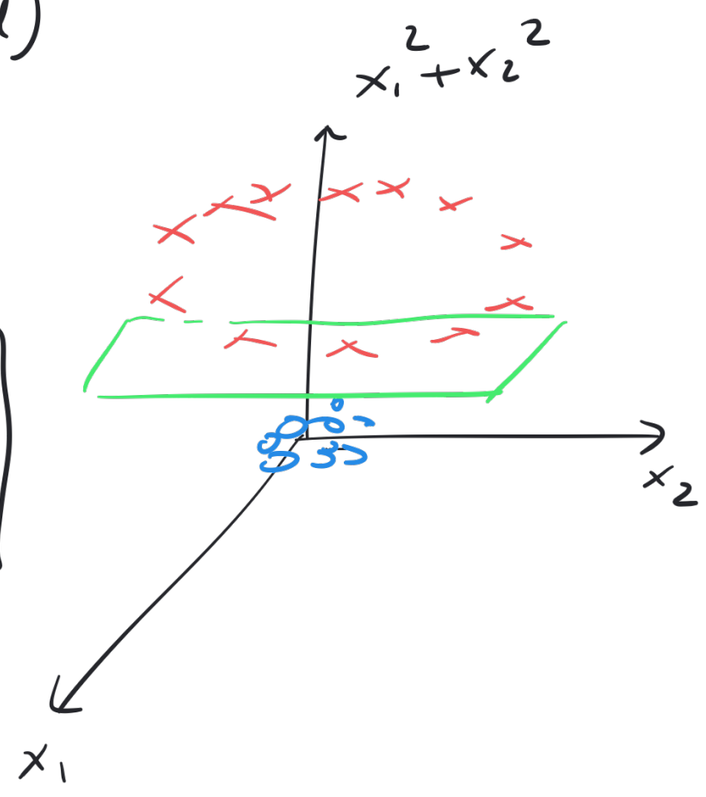
typically $\phi(x)$ maps to a higher-dimensional space

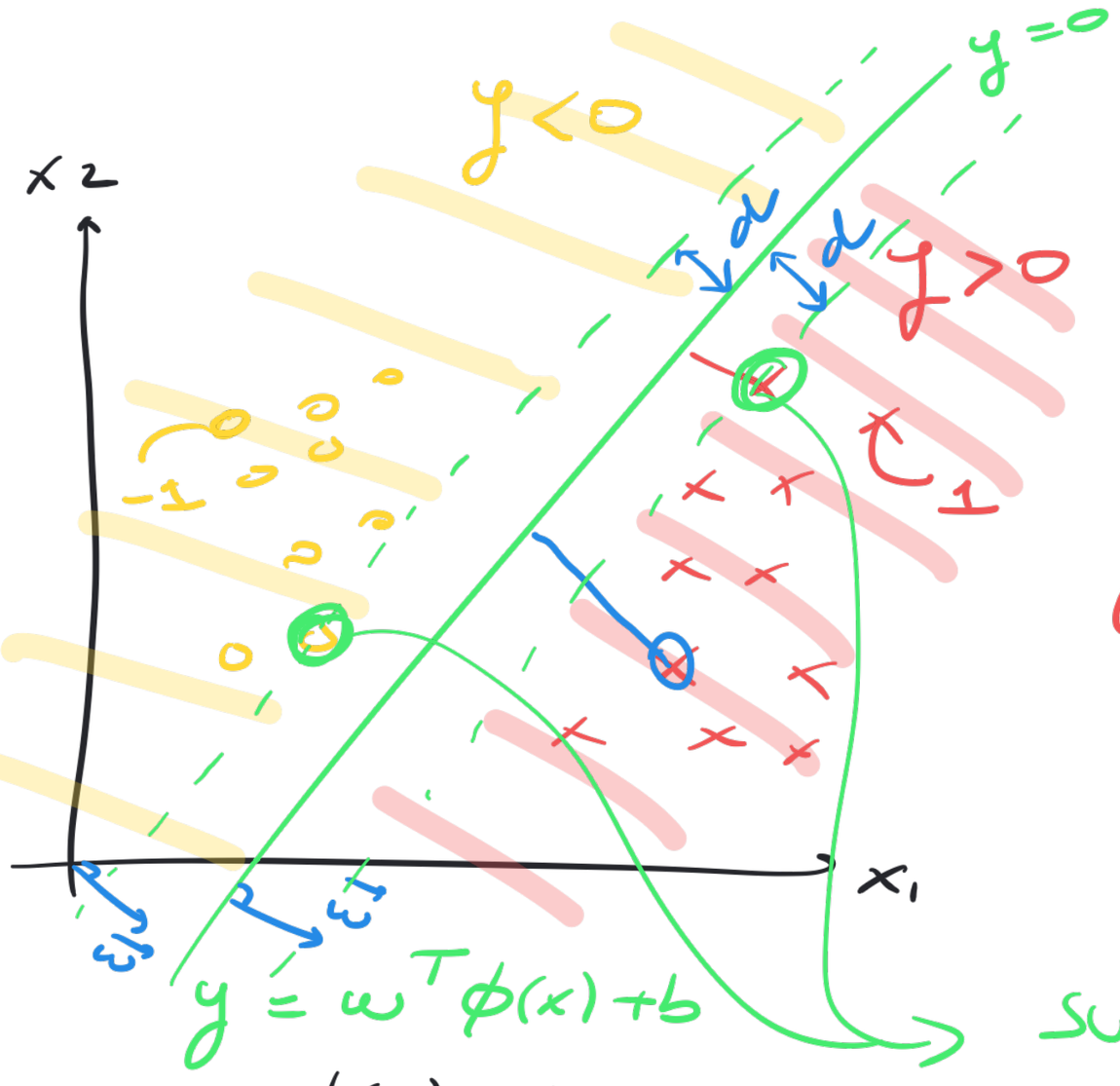
$\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^D \quad (D > d)$



$$\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \longmapsto \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 + x_2^2 \end{bmatrix}$$





x - class 1

o - class -1

$t_n \equiv$ label (class) for point $x_n \in \{1, -1\}$

① $y(x_n) \cdot t_n > 0$

$\forall x_n$ that are correctly classified

SUPPORT VECTORS (S.V.s)

Distance of point x_n to discriminant function $y(x_n)$:

$$\frac{y(x_n) \cdot t_n}{\|w\|} \geq 0$$

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

So we want to <u>MAXIMIZE</u> this

distance for all points x_n.

$$\arg_{\{w,b\}} \max \left\{ \frac{1}{\|w\|} \underbrace{\min_n \{t_n \cdot (w^T \phi(x_n) + b)\}}_{\downarrow} \right\}$$

we are looking to preserve the smallest amount of points that are closest to $y(x_n)$.

we will scale the parameters w and b

such that the distance of

Support vectors (S.V.s) to $y(x)$

is 1, i.e.,

$$\frac{t_n \cdot y(x_n)}{\|w\|} = 1 \quad \forall\, x_n \text{ that are S.V.s}$$

$$w \longleftarrow k \cdot w$$

$$b \longleftarrow k \cdot b$$

with this, we have:

$$t_n (w^T \phi(x_n) + b) = 1 \text{ , } \forall x_n \text{ that are S.V.s}$$

and

$$t_n . (w^T \phi(x_n) + b) > 1 \text{ , } \forall x_n \text{ that are not S.V.s}$$

In general,

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

OPTIMIZATION PROBLEM w/ inEquality constraints

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

$\Leftrightarrow$

$$\arg\min_{\{w,b\}} \|w\|^2 \quad \text{subject to} \quad t_n \cdot (w^T \phi(x_n) + b) \geq 1$$

// (APPENDIX E — Bishop textbook)

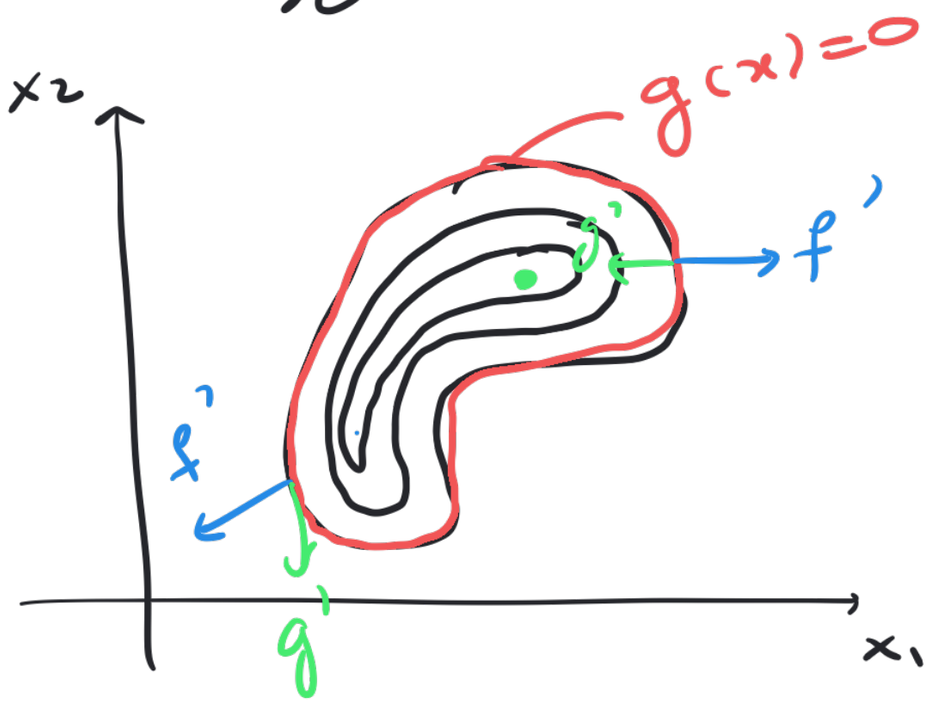
OPTIMIZATION function with EQUALITY / INEQUALITY CONSTRAINTS

① $\min_x f(x)$

NECESSARY condition:

$$f'(x) = 0$$

② $\arg\min_x f(x)$ sub. to $g(x) = 0$



Lagrangian function:

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

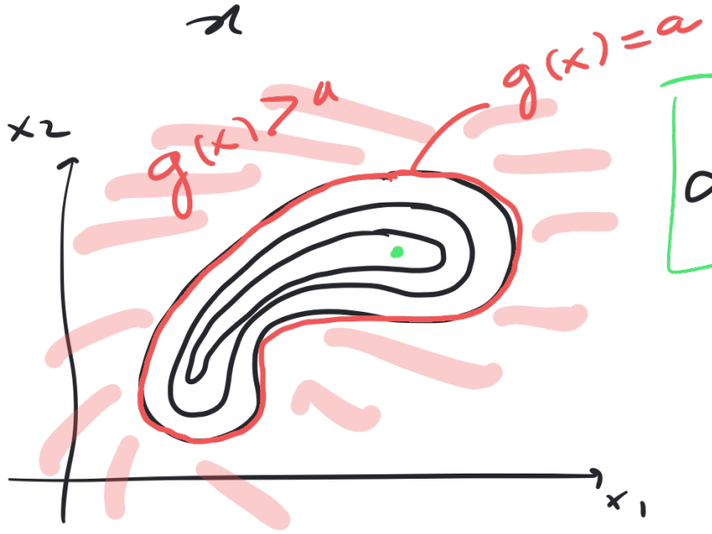
↑ Lagrange multiplier

NECESSARY condition:

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

(3) $\arg\min_x f(x)$ sub. to $g(x) \geq a$

$\Leftrightarrow g(x) - a \geqslant 0$

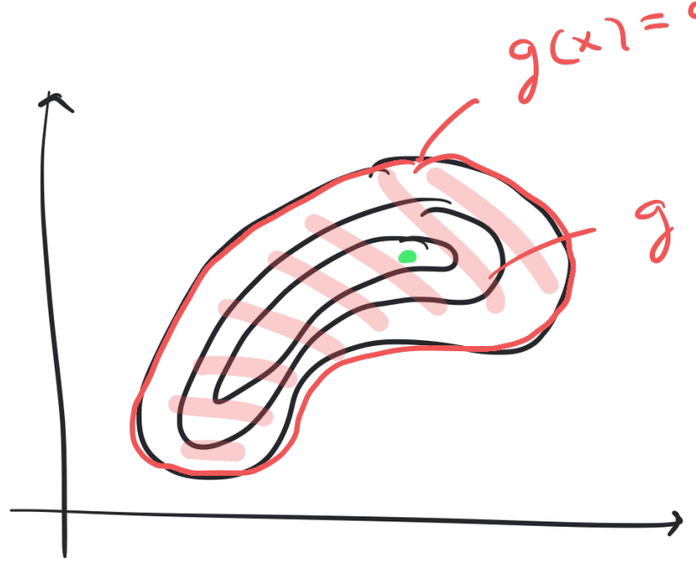


$$\mathcal{L}(x, \lambda) = f(x) - \lambda(g(x) - a)$$

3.1) ACTIVE CONSTRAINT

In this case the closest x that minimizes $f(x)$ is on $g(x) = a$

$\lambda > 0, \; g(x) = a$



3.2) INACTIVE CONSTRAINT

In this case, the minima of $f(x)$ already satisfies $g(x) \geq a$ thus:

$\lambda = 0, \; g(x) > a$

KARUSH - KUHN - TUCKER (KKT)

CONDITIONS

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

← at least one of them is zero.

Coming back to the SVM optimization:

$$\arg\min_{\{w,b\}} \|w\|^2 \quad \text{sub. to} \quad t_n \cdot y(x_n) \geq 1$$

Lagrangian function:

$$\mathcal{L}(w, b, a) = \|w\|^2 - \sum_{n=1}^{N} a_n \cdot (t_n\, y(x_n) - 1)$$

↑ set of Lagrange multipliers

KKT conditions:

$$\begin{cases} a_n \geq 0 \\ t_n \cdot y(x_n) - 1 \geq 0 \\ a_n (t_n \cdot y(x_n) - 1) = 0 \end{cases}$$

these conditions say that for every point x_n:

$a_n = 0$
$t_n \cdot y(x_n) > 1$
inactive constraint
x_n is NOT a S.V.

OR

$a_n > 0$
$t_n \cdot y(x_n) = 1$
active constraint
x_n is a S.V.

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

optimizing $\mathcal{L}(w, b, a)$ w.r.t. w and b:

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

in some cases, $\phi(x)$ is infinite-dimensional. So we cannot compute, nor store it.

So we will work with the DUAL LAGRANGIAN

Substituting w in the Lagrangian function and making sure that condition $\sum_{n=1}^{N} a_n . t_n = 0$ is met:

THE DUAL LAGRANGIAN is:

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

KERNEL function

where $a_n \geq 0$

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

KERNEL function

$$K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$$
$$(x, y) \longmapsto \phi^T(x) \cdot \phi(y)$$

$$\phi : \mathbb{R}^d \longrightarrow \mathbb{R}^D$$

GRAM MATRIX

① K is $N \times N$, symmetric

$$K = \begin{bmatrix} \phi^T(x_1)\cdot\phi(x_1) & \phi^T(x_1)\cdot\phi(x_2) & \cdots & \phi^T(x_1)\phi(x_n) \\ \phi^T(x_2)\cdot\phi(x_1) & \phi^T(x_2)\cdot\phi^T(x_2) & \cdots & \phi^T(x_2)\cdot\phi(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ \phi^T(x_n)\phi(x_1) & \phi^T(x_n)\phi(x_2) & \cdots & \phi^T(x_n)\phi(x_n) \end{bmatrix}$$

② K is positive semi-definite

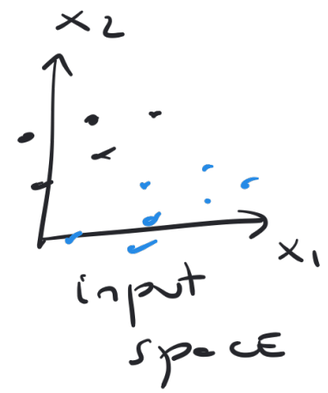
$$z^T K z \geq 0, \quad \forall z \in \mathbb{R}^n \setminus \{\vec{0}\}$$

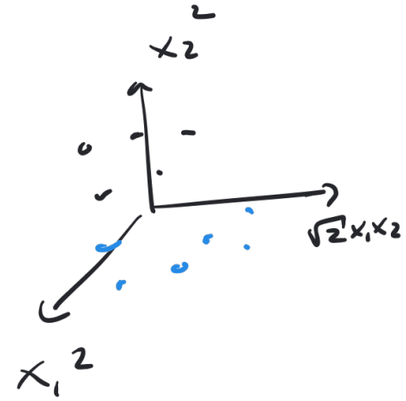
③ all eigenvalues are ≥ 0. which means that some columns / rows ARE colinear.

KERNEL TRICK

$$\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$$

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \longmapsto \begin{bmatrix} x_1^2 \\ \sqrt{2} \cdot x_1 \cdot x_2 \\ x_2^2 \end{bmatrix}$$





$$K(x, y) = \phi^T(x) \cdot \phi(y)$$

$$= \begin{bmatrix} x_1^2, \sqrt{2}\, x_1 \cdot x_2, x_2^2 \end{bmatrix} \begin{bmatrix} y_1^2 \\ \sqrt{2} \cdot y_1 \cdot y_2 \\ y_2^2 \end{bmatrix} \qquad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \; y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$= x_1^2 \cdot y_1^2 + 2 \cdot x_1 \cdot x_2 \cdot y_1 \cdot y_2 + x_2^2 \cdot y_2^2$$

$$= (x_1 \cdot y_1 + x_2 y_2)^2$$

$$= \left(\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right)^2 = (x^T \cdot y)^2$$

RADIAL BASIS FUNCTION (RBF)

it maps the input space to an infinite-dimensional space.

$$\phi_{RBF} : \mathbb{R}^d \longrightarrow \mathbb{R}^\infty$$

$$K_{RBF} : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$$
$$(x, y) \longmapsto e^{-\gamma \cdot \|x-y\|^2}$$

equivalent

$$e^{-\frac{\|x-y\|^2}{\sigma^2}}, \quad \gamma = \frac{1}{\sigma^2}$$

Proof: $\gamma = \frac{1}{2}$ (without loss of generality)

$$K_{RBF}(x, y) = \exp\left(-\frac{1}{2}\|x - y\|^2\right)$$

$$= \exp\left(-\frac{1}{2}(x - y)^T(x - y)\right)$$

$$= \exp\left(-\frac{1}{2}(x^T x - x^T y - y^T x + y^T y)\right)$$

$$= \exp\left(-\frac{1}{2}x^T x - \frac{1}{2}y^T y + x^T y\right)$$

$$= \underbrace{\exp\left(-\frac{1}{2}x^T x - \frac{1}{2}y^T y\right)}_{= c \equiv \text{constant}} \cdot \exp(x^T y)$$

$$= c \cdot \exp(x^T y)$$

$$= c \cdot \sum_{n=0}^{\infty} \frac{(x^T y)^n}{n!} = c \cdot \sum_{n=0}^{\infty} \frac{K_{poly}^{(n)}(x, y)}{n!}$$

Taylor series expansion

Q.E.D.

- The dual Lagrangian takes a form of a QUADRATIC PROGRAMMING optimization. We can use known algorithms to find the solution for a.
  - The solution to a quadratic optimization problem on N variables is $\Theta(N^3)$.
- Going to the dual Lagrangian, we have $D \longrightarrow N$ variables ($N \equiv$ # samples and $D \equiv$ # of features in ϕ), $N >> D$.

$$w = \sum_{n=1}^{N} a_n . t_n . \phi(x_n)$$

and $$y(x) = w^T \phi(x) + b$$

During ~~test~~, we only use the S.V.s to make a decision:

$$y(x) = \sum_{n=1}^{N} a_n . t_n . \phi^T(x_n) . \phi(x) + b$$

$y(x)$: test point

$$= \sum_{m \in S} a_m . t_m . \phi^T(x_m) . \phi(x) + b$$

$$= \sum_{m \in S} a_m . t_m . K(x_m, x) + b$$

$a_n > 0$ and $t_n \cdot y(x_n) = 1$ for x_n that are S.Vs

where $$y(x) = \sum_{n=1}^{N} a_n \cdot t_n \cdot \phi^T(x_n) \cdot \phi(x) + b$$

$$= \sum_{m \in S} a_m \cdot t_m \cdot \phi^T(x_m) \cdot \phi(x) + b$$

Substituting,

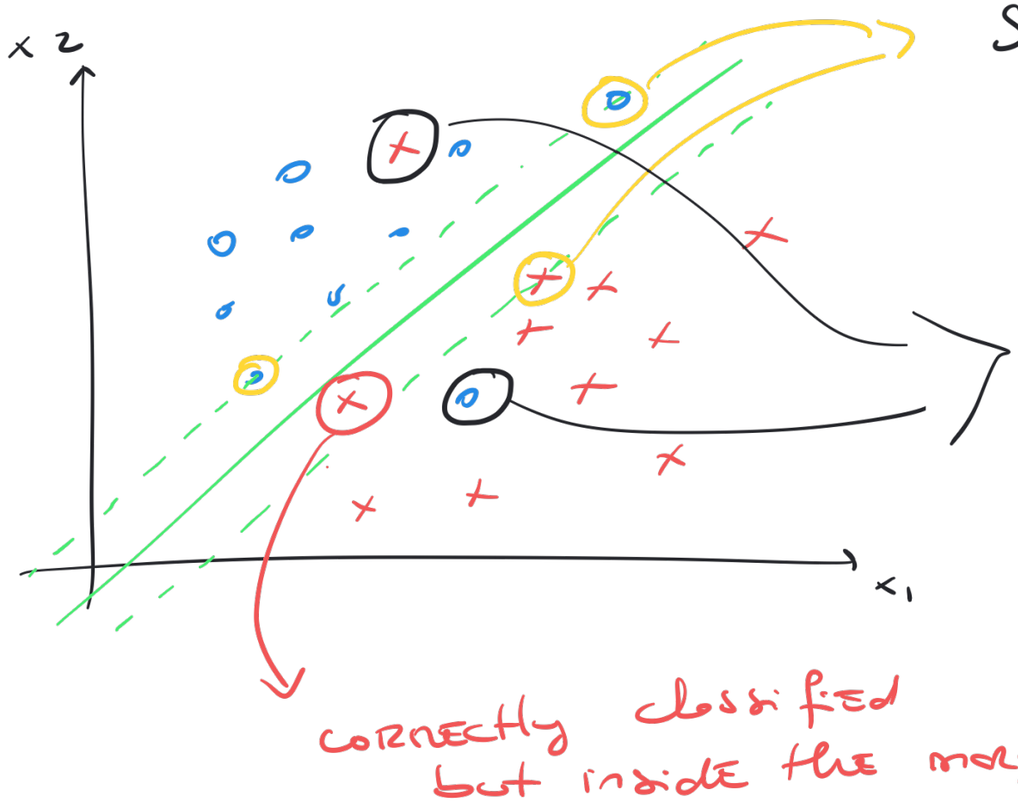
$$t_n \cdot \left(\sum_{m \in S} a_m \cdot t_m \cdot \phi^T(x_m) \cdot \phi(x_n) + b \right) = 1$$

$$b = \frac{1}{N_s} \sum_{m \in S}' \left[t_n - \left(\sum_{n \in S} a_n \cdot t_n \cdot K(x_m, x_n) \right) \right]$$

$N_s \equiv$ # of S.V.s

$S \equiv$ set of all samples that are S.V.s

Soft-Margin SVM



Slack Variable

$\xi_n = 0$ for all x_n that correctly classified and on or outside of margin

$$\xi_n = |t_n - y(x_n)| \geq 0$$