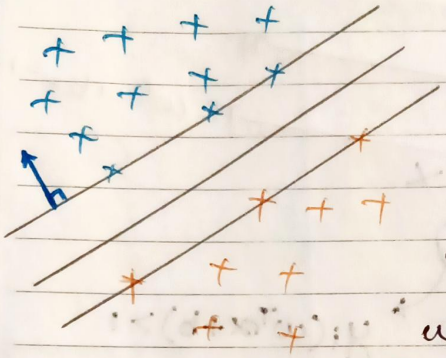
IN-DEPTH SVM

SVM

Mathematical formulation of SVM.



Objective: To find a hyper-plane that does margin maximization.

we need to find a hyperplane Π :- margin-maximization

Π :- $w^T x + b = 0$ (w not necessarily unit vectos).

Π_+ : $w^T x + b = 1$

Π_- : $w^T x + b = -1$

By simple co-ordinate concepts we can get

$$d = \frac{2}{\|w\|}$$

So, we have to maximize 'd' with the constraint that all (+ve) points will lie above Π_+ & all (-ve) pts below Π_-.

So, the optimization function ~~been~~ becomes

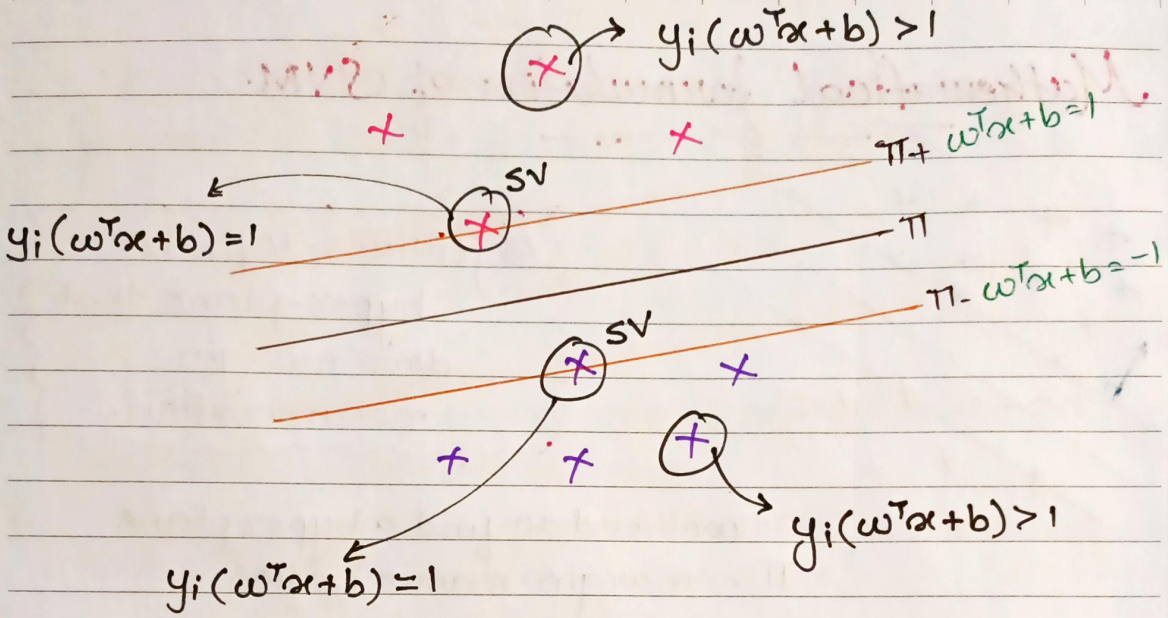
Hard-margin

$$(w^*, b^*) = \arg\max_{w,b} \frac{2}{\|w\|}$$

s.t.

$$\forall i,\ y_i(w^T x_i + b) \geq 1$$

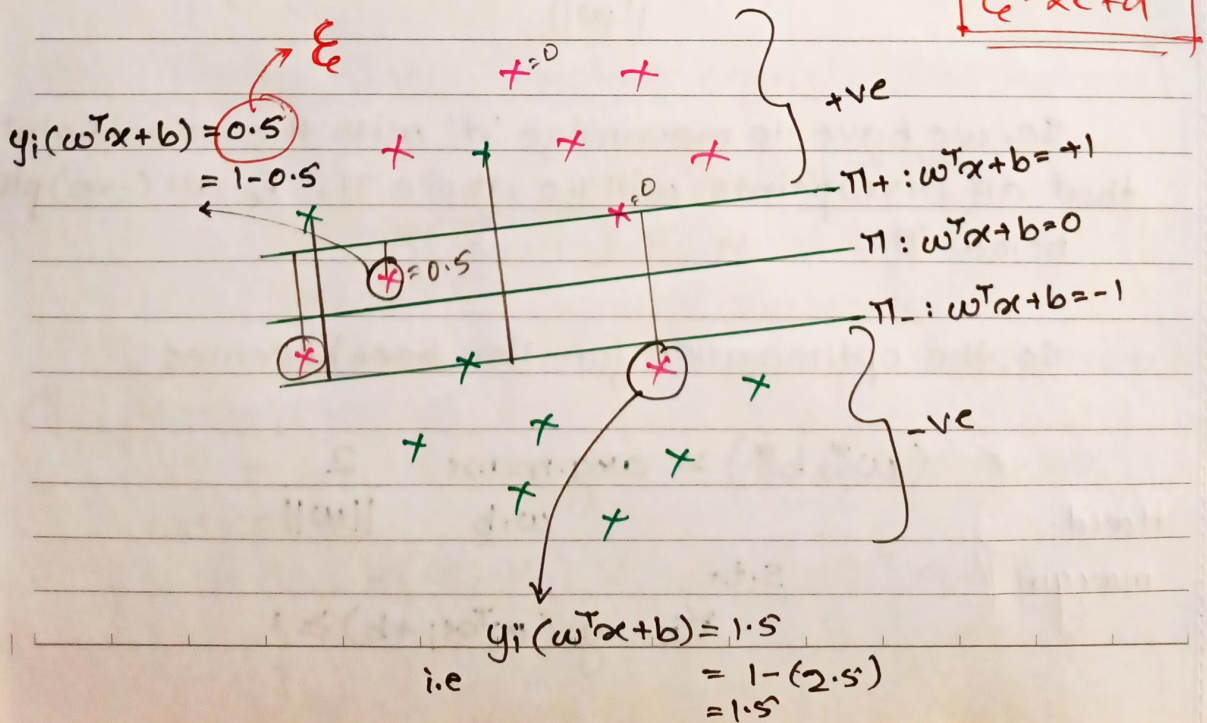
SV: support vectors



But what if we have a dataset that's not linearly separable but almost linearly seperable?

→ we formulate an alternative soft-margin constraint that allows some errors to persist.

ξ = zeta



$x_i \rightarrow \xi_i$ ξ = zeta

$\xi_i = 0$ if $y_i(w^T x + b) \geq 1$
(i.e. for correctly classified points)

$\xi_i > 0$ and it is equal to some units of distn away from the correct hyperplane in the incorrect direction.

Now, we want to minimize these errors.

So, our optimization function becomes

$$\arg\min_{w,b} \frac{\|w\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

$\frac{1}{\text{margin}}$ → avg. distance of misclassified points. (Hinge loss)

s.t. → defn of zeta

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

C → hyperparameter

C↑ → tendency to make mistakes↓
→ overfit (High variance)

C↓ → underfit (High Bias)

$\forall i$
- correctly classified points $\xi_i = 0$
- incorrectly classified pts $\xi_i > 0$

maximizing $\frac{2}{\|w\|}$ same as minimizing $\frac{\|w\|}{2}$.

soft margin of SVM:

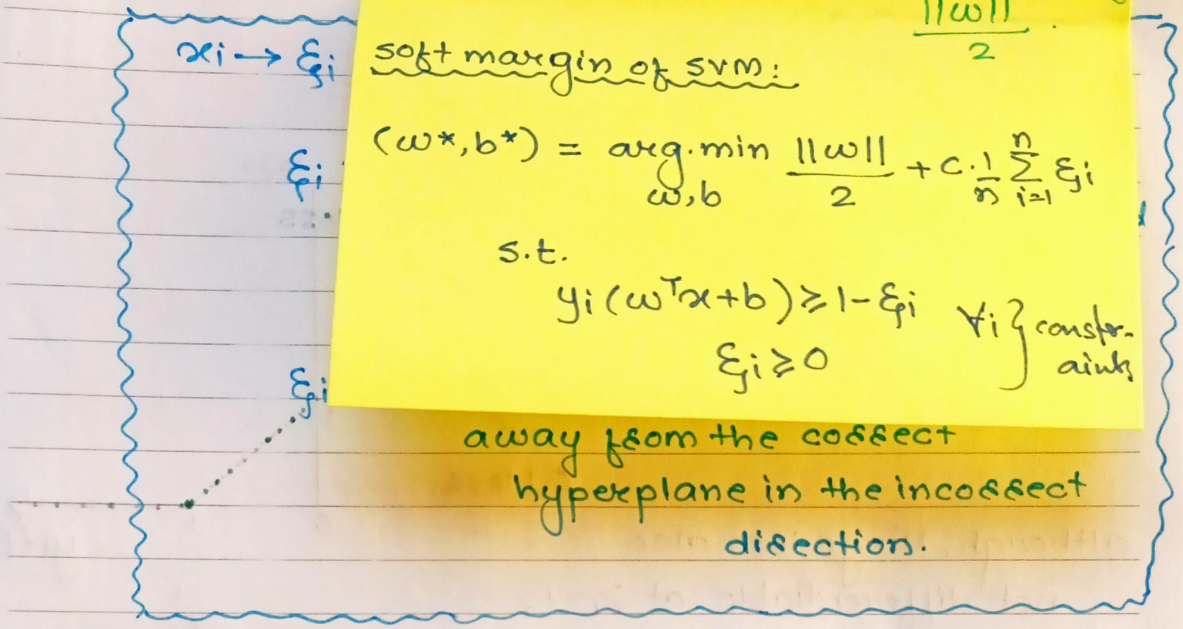
$$(w^*, b^*) = \arg\min_{w,b} \frac{\|w\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

s.t.

$$y_i(w^T x + b) \geq 1 - \xi_i \quad \forall i$$

$$\xi_i \geq 0$$

} constraints



Now, we want to minimize these errors.

So, our optimization function becomes

$$\arg\min_{w,b} \frac{\|w\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

$\frac{1}{\text{margin}}$

$\frac{1}{n}\sum_{i=1}^{n} \xi_i$ → avg. distance of misclassified points. (Hinge loss)

s.t.

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

→ defn of zeta

C → hyperparameter

C↑ → tendency to make mistakes↓
→ overfit (High variance)

C↓ → underfit (High Bias)

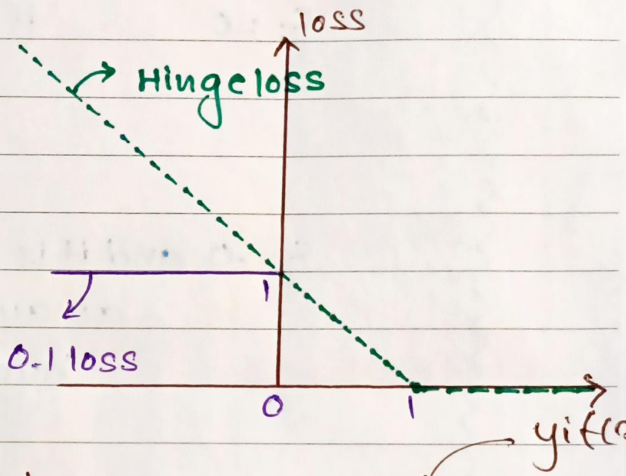
$\forall i$

correctly classified points $\xi_i = 0$

incorrectly classified pts $\xi_i > 0$

Loss minimization : Hinge Loss

Hinge loss + Regularization gives us SVM.



although, Hinge loss also not differentiable at $z_i = 1$, but since it is continuous unlike 0-1 loss, we can work around hacks to work with it.

$y_i f(x_i) = y_i(w^T x + b) = z_i$

$z_i > 0$: x_i is correctly classified

$z_i < 0$: x_i is incorrectly classified.

$$\text{Hinge loss} = \begin{cases} z_i \geq 1 \; ; & \text{Hinge loss} = 0 \\ z_i < 1 \; ; & \text{Hinge loss} = 1 - z_i \end{cases}$$

alternatively,

$$\text{Hinge loss} = \max(0, 1 - z_i)$$

— Andrew Ng

Lagrange Duality

consider a problem of the following form:

$$\min_w \quad f(w)$$
$$\text{s.t.} \quad h_i(w) = 0, \quad i = 1, \ldots, l.$$

we define the Lagrangian to be

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Here, β_i's = Lagrange multipliers.

To find, set $\mathcal{L}$'s partial derivatives to 0;

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0,$$

and solve for w and β.

consider the following, called primal optimization problem:

$$\min_w \quad f(w)$$
$$\text{s.t.} \quad g_i(w) \leq 0, \quad i = 1, \ldots, k$$
$$h_i(w) \leq 0, \quad i = 1, \ldots, l.$$

To solve it, we start by defining the generalized Lagrangian.

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{\ell} \beta_i h_i(w)$$

Here, the α_i's and β_i's are the Lagrange multipliers.

Consider the quantity,

$$\theta_p(w) = \max_{\alpha, \beta; \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$$

Here, p = "primal".

Let some w be given. If w violates any of the primal ~~constraints~~ constraints

(i.e. if either $g_i(w) > 0$ or $h_i(w) \neq 0$ for some i)

then,

$$\theta_p(w) = \max_{\alpha, \beta; \alpha_i \geq 0} f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{\ell} \beta_i h_i(w)$$

$$= \infty \quad (1)$$

Conversely, if the constraints are indeed satisfied for a particular value of w,

then $\theta_p(w) = f(w)$. Hence

$$\theta_p(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise} \end{cases}$$

Hence, if we consider the minimization problem

$$\min_w \theta_p(w) = \min_w \max_{\alpha,\beta;\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta)$$

same as, primal problem.

Optimal value of the objective to be $p^* = \min_w \theta_p(w)$

value of the primal problem.

Now, we define

$$\theta_D(\alpha,\beta) = \min_w \mathcal{L}(w,\alpha,\beta)$$

where, D = "dual"

Note: In defn θ_p we were optimizing (maximizing) ~~w.r.t~~ w.r.t α, β here we are minimizing w.r.t. w.

∴ Dual optimization problem:

$$\max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \min_{\omega} \mathcal{L}(\omega,\alpha,\beta)$$

Optimal value of the dual problem's objective :

$$d^* = \max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\omega)$$

How are the primal and the dual problems related ?

$$d^* = \max_{\alpha,\beta:\alpha_i \geq 0} \min_{\omega} \mathcal{L}(\omega,\alpha,\beta) \leq \min_{\omega} \max_{\alpha,\beta:\alpha_i \geq 0} \mathcal{L}(\omega,\alpha,\beta) = p^*$$

always remember :

"~~min~~ max min" of a function always being less than or equal to the "min max".

for certain conditions,

$$d^* = p^*$$

i.e. there exists a_i, b_i, so that
$h_i(w) = a_i^T w + b_i$

affline → linear, except contains extra intercept term b_i.

Suppose f and the g_i's are convex, and the h_i's are affine.

Suppose that the constraints g_i are feasible; means there exists some w so that $g_i(w) < 0$ for all i.

under above assumptions, there must exist w^*, α^*, β^* so that w^* is the solution to the primal problem α^*, β^* are the solution to the dual problem, and moreover $p^* = d^* = \mathcal{L}(w^*, \alpha^*, \beta^*)$

moreover, w^*, α^* and β^* satisfy the Karush-Kuhn-Tucker (KKT) conditions

$$\frac{\partial}{\partial w_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, d \quad \text{---(3)}$$

$$\frac{\partial}{\partial \beta_i} \mathcal{L}(w^*, \alpha^*, \beta^*) = 0, \quad i = 1, \ldots, l \quad \text{---(4)}$$

$$\alpha_i^* g_i(w^*) = 0, \quad i = 1, \ldots, k \quad \text{---(5)}$$

$$g_i(w^*) \leq 0, \quad i = 1, \ldots, k \quad \text{---(6)}$$

$$\alpha^* \geq 0, \quad i = 1, \ldots, k \quad \text{---(7)}$$

moreover, if some w^*, α^*, β^* satisfy the KKT conditions, then it is also a solution to the primal and dual problems.

Eqn ⑤, which is called the KKT dual complementarity condition.

Specifically, it implies that if $\alpha_i^* > 0$, then $g_i(w^*) = 0$.

i.e. "$g_i(w) \leq 0$" constraint is active, meaning it holds with equality rather than with inequality.

This will be key for showing that SVM has only a small number of "support vectors".

Dual Form of SVM

Following (primal) optimization problem for finding the optimal margin classifier:

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$

s.t.

$$y_i(w^T x_i + b) \geq 1, \quad i = 1, \ldots, n \qquad \text{—(8)}$$

constraints as

$$g_i(w) = -y_i(w^T x_i + b) + 1 \leq 0$$

when we construct ~~lang~~ Lagrangian for optimization problem we have:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i (w^T x_i + b) - 1 \right] \qquad \text{—(9)}$$

Note: there's only "α_i" but no "β_i" Lagrange multipliers, since the problem has only inequality constraints.

To find the dual form, first minimize $\mathcal{L}(w,b,\alpha)$ w.r.t w and b (for fixed α) to get Θ_D.

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{n} \alpha_i y_i x_i = 0$$

$$\therefore \boxed{w = \sum_{i=1}^{n} \alpha_i y_i x_i} \quad \text{---(10)}$$

derivative w.r.t b,

$$\frac{\partial}{\partial b} \mathcal{L}(w,b,\alpha) = \sum_{i=1}^{n} \alpha_i y_i = 0 \quad \text{---(11)}$$

If we take the defn of w in eqn (10) and put it into the Lagrangian eqn (9), we get

$$\mathcal{L}(w,b,\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j (x_i)^T x_j - b \sum_{i=1}^{n} \alpha_i y_i$$

but from eqn (11), last term = 0

$$\therefore \mathcal{L}(w,b,\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j x_i^T x_j$$

∴ with the constraint $\alpha_i \geq 0$, we obtain

dual optimization problem

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle x_i, x_j \rangle$$

$$\text{s.t. } \alpha_i \geq 0, \quad i = 1, \ldots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

Important Observations:

1. for every x_i, we have an α_i
2. x_i's only occur in the form of $x_i^T x_j$

on solving we get

$$f(x_q) = \sum_{i=1}^{n} \alpha_i y_i x_i^T x_q + b$$

for non-sv, $\alpha_i = 0$
so, for $f(q) \rightarrow$ only the support vector matters.

3. $\alpha_i > 0$ only for support vectors else 0.

since x_i always occurs only as $x_i^T x_j \rightarrow x_i \cdot x_j$

cosine_sim$\langle x_i, x_j \rangle$

so, basically cosine similarity value are only required to solve the optimization and ultimately get the model.

we can replace cosine-similarity with any other kind of similarity which makes it more powerful.

Generalized dual form-

$$\max_{\alpha_i} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \underbrace{k(x_i, x_j)}_{\text{called Kernel Function.}}$$

(can be any kind of similarity b/w x_i & x_j)

Even during evaluation x_i occurs only as $x_i^T x_q$

$$f(x_q) = \sum_{i=1}^{n} \alpha_i y_i \underbrace{x_i^T x_q}_{k(x_i, x_q)} + b$$

Kernel trick → replacing $x_i^T x_j$ with generalized similarity function $k(x_i, x_j)$.

Quadratic Programming.

The hard margin and soft margin problems are both convex quadratic optimization problems with linear constraints.

Such problems are known as Quadratic Programming (QP) problems