

IN-DEPTH SVM

Mathematical formulation of SVM.



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

we need to find a hyperplane
 Π is margin-maximization

$\Pi: \omega^T x + b = 0$ (ω not necessarily unit vector).

$$\Pi_+: \omega^T x + b = 1$$

$$\Pi_-: \omega^T x + b = -1$$

By simple coordinate concepts we can get

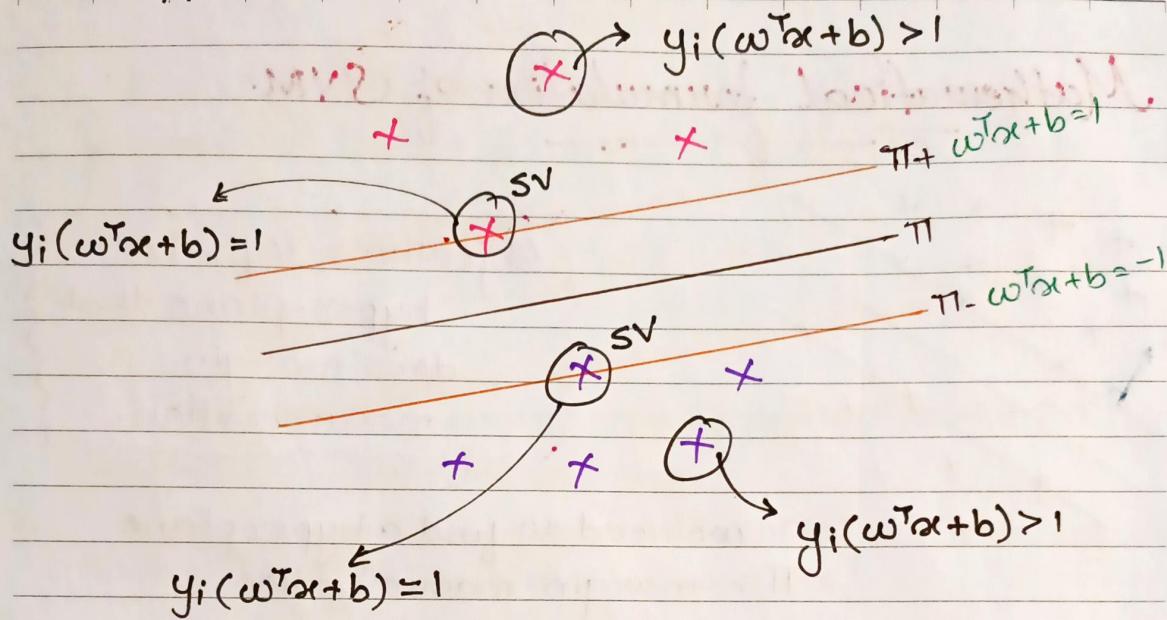
$$d = \frac{2}{\|\omega\|}$$

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 becomes

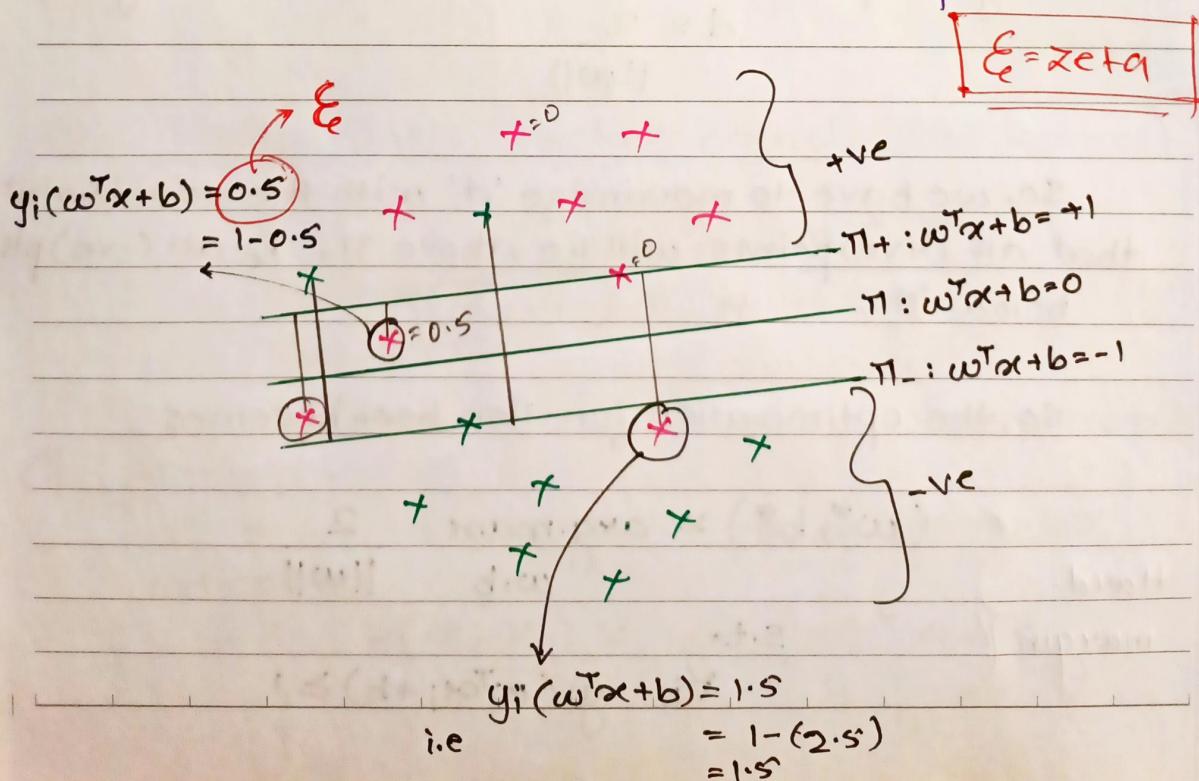
$$\left. \begin{array}{l} (\omega^*, b^*) = \arg \max_{\omega, b} \frac{2}{\|\omega\|} \\ \text{Hard-margin} \end{array} \right\} \quad \text{s.t.} \quad \forall i, y_i(\omega^T x_i + b) \geq 1$$

SV: support vectors



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

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



$$x_i \rightarrow \xi_i$$

$\zeta = \xi_i$

$$\xi_i = 0$$

if $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$

(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

$$\underset{\mathbf{w}, b}{\text{arg. min}} \frac{\|\mathbf{w}\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^n \xi_i$$

margin

s.t.

avg. distance of misclassified points (Hinge loss)

$$y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$$

$C \leftrightarrow$ hyperparameters

$$\xi_i \geq 0$$

$C \uparrow \rightarrow$ tendency to make mistakes \downarrow

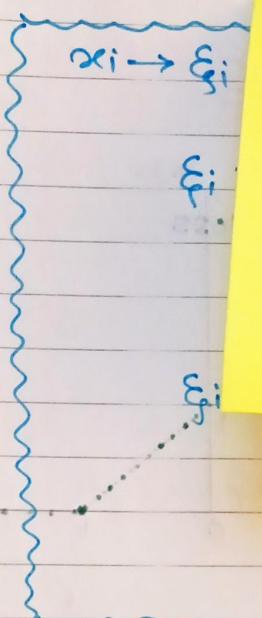
\rightarrow overfit (High variance)

$C \uparrow \rightarrow$ underfit (High Bias)

y_i } correctly classified points
 $\xi_i = 0$

} incorrectly classified pts $\xi_i > 0$

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



soft margin of SVM:

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

s.t.

$$y_i(\omega^T x_i + b^*) \geq 1 - \epsilon_i \quad \forall i \quad \begin{matrix} \text{const.} \\ \epsilon_i \geq 0 \end{matrix}$$

away from the correct
hyperplane in the incorrect
direction.

Now, we want to minimize these errors.

So, our optimization function becomes

$$\underset{\substack{\text{margin} \\ \omega, b}}{\arg \min} \frac{\|\omega\|}{2} + C \cdot \frac{1}{n} \sum_{i=1}^n \epsilon_i \quad \begin{matrix} \rightarrow \text{avg. distance} \\ \text{of misclassified} \\ \text{points.} \\ (\text{Hinge loss}) \end{matrix}$$

s.t.

$$y_i(\omega^T x_i + b) \geq 1 - \epsilon_i$$

$C \rightarrow$ hyperparameter

$$\epsilon_i \geq 0$$

$C \uparrow \rightarrow$ tendency to make

mistakes \downarrow

\rightarrow overfit (High variance)

y_i } correctly

classified points

$$\epsilon_i = 0$$

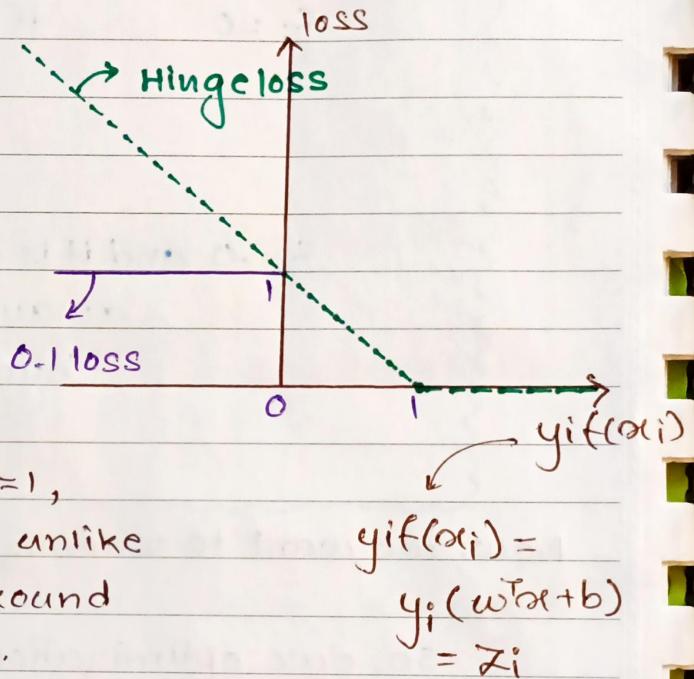
$C \downarrow \rightarrow$ underfit (High Bias)

incorrectly

classified pts $\epsilon_i > 0$

Loss minimization : Hinge Loss

Hinge loss + Regularization
gives us
SVM.



$z_i > 0$: 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:

$$\min_{\omega} f(\omega)$$

$$\text{s.t. } b_i(\omega) = 0, i=1, \dots, l.$$

we define the Lagrangian to be

$$L(\omega, \beta) = f(\omega) + \sum_{i=1}^l \beta_i b_i(\omega)$$

Here, β_i 's = Lagrange multipliers.

To find, set L 's partial derivatives to 0;

$$\frac{\partial L}{\partial \omega_i} = 0; \quad \frac{\partial L}{\partial \beta_i}$$

and solve for ω and β .

consider the following, called **primal optimization problem**:

$$\min_{\omega} f(\omega)$$

$$\text{s.t. } g_i(\omega) \leq 0, i=1, \dots, k$$

$$b_i(\omega) \leq 0, i=1, \dots, l.$$

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

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

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

Consider the quantity,

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

Here, p = "primal".

Let some ω be given. If ω violates any of the primal constraints, then

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

then,

$$\begin{aligned} \theta_p(\omega) &= \max_{\alpha, \beta: \alpha_i \geq 0} f(\omega) + \sum_{i=1}^k \alpha_i g_i(\omega) \\ &\quad + \sum_{i=1}^l \beta_i h_i(\omega) \\ &= \infty \end{aligned}$$

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

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

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

Hence, if we consider the minimization problem

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

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

$$\Theta_D(\alpha, \beta) = \min_w L(w, \alpha, \beta)$$

where, D = "dual"

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

∴ Dual optimization problem:

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

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?

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

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

$$h_i(\omega) = a_i^T \omega + b_i$$

affine \rightarrow linear, except contains extra intercept term b_i .

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

Suppose that the constraints g_i are feasible; means there exists some ω so that

$$g_i(\omega) < 0 \text{ for all } i.$$

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

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

$$\frac{\partial}{\partial \omega_i} L(\omega^*, \alpha^*, \beta^*) = 0, \quad i=1, \dots, d \quad \rightarrow ③$$

$$\frac{\partial}{\partial \beta_i} L(\omega^*, \alpha^*, \beta^*) = 0, \quad i=1, \dots, l \quad \rightarrow ④$$

$$\alpha_i^* g_i(\omega^*) = 0, \quad i=1, \dots, k \quad ⑤$$

$$g_i(\omega^*) \leq 0, \quad i=1, \dots, k \quad ⑥$$

$$\alpha^* \geq 0, \quad i=1, \dots, k \quad \rightarrow ⑦$$

moreover, if some $\omega^*, \alpha^*, \beta^*$ 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 $g_i(\omega^*) = 0$.

i.e. " $g_i(\omega) \leq 0$ " 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:

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

s.t.

$$y_i(w^T x_i + b) \geq 1, i=1, \dots, n$$

constraints as

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

when we construct Langrangian for optimization problem we have:

$$L(w, b; \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i [y_i(w^T x_i + b) - 1]$$

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

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

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

$$\therefore \left[\omega = \sum_{i=1}^n \alpha_i y_i x_i \right] \quad \text{--- (10)}$$

derivative w.r.t b ,

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

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

$$\begin{aligned} \mathcal{L}(\omega, b, \alpha) &= \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n y_i y_j \alpha_i \alpha_j (\omega_i)^T \omega_j \\ &\quad - b \sum_{i=1}^n \alpha_i y_i \end{aligned}$$

but from eqn (11), last term = 0

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

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

dual optimization problem

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

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

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

Important Observations:

1. for every x_i , we have an α_i

2. x_i 's only occur in the form of $\underline{\alpha_i^T x_i}$

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 vectors 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-

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

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 \rightarrow 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