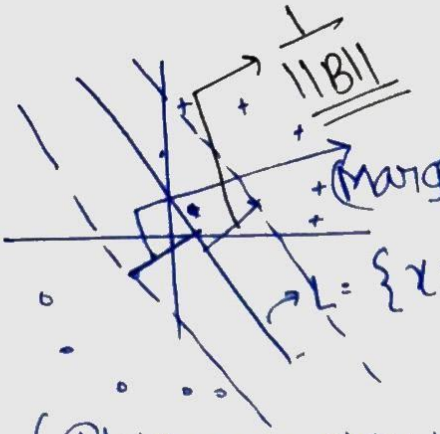
SVM FORMULATION Lecture - 21



$f(x) = \beta_0 + \beta^T x$

Optimal seperating hyperplane
Maximize the distance of closest point to hyperplane.

(Thickness should be same for both hyperplanes)

β_0, β_1, max M
($||\beta|| = 1$)
$\downarrow$
β, should n't be larger.

subject to $y_i(x_i^T\beta + \beta_0) \geq M$ for all $i = 1, \dots N$
$\downarrow$
We need to maximise M, so that there is a maximum seperation between class.

Distance of any x_i from hyperplane

$$\frac{f(x)}{||f'(x)||} = \frac{(x_i^T\beta + \beta_0)}{||\beta||}$$

$\to$ signed distance
$\to$ multiply by y^i to make positive

Need n't worry about $||\beta|| = 1$

$$\frac{y_i(x_i^T\beta + \beta_0)}{||\beta||} \geq M \Rightarrow y_i(x_i^T\beta + \beta_0) \geq M||\beta||$$

Just make sure $||\beta|| = 1/M$

$$\Downarrow$$

$$\boxed{y_i(x_i^T\beta + \beta_0) \geq 1}$$

Max M only change in direction
$\Downarrow$
Min β

$$\left(\begin{array}{l} \min ||\beta|| \\ \text{s.t } \; y_i(x_i^T\beta + \beta_0) \geq 1 \end{array}\right)$$

SVM - interpretation & Analysis

$\min \|\beta\|$ s.t. $y_i(x_i^T\beta+\beta_0) \geq 1$ $\Rightarrow \left\{ \min \underbrace{\frac{1}{2}\|\beta\|^2}_{\text{by our choise}} \right\}$ so that it is easily differentiable

<u>Lagrangian:</u> $L_P = \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i(x_i^T\beta+\beta_0) - 1 \right]$

setting derivatives to '0'.

$$\frac{\partial L_P}{\partial \beta} = 0 \Rightarrow \frac{1}{\cancel{2}}\cancel{2}\cdot\beta - \sum_{i=1}^{N}\alpha_i [y_i x_i] = 0 \quad -(1)$$

$$\beta = \sum_{i=1}^{N} \alpha_i x_i y_i$$

$$\frac{\partial L_P}{\partial \beta_0} = 0 - \sum_{i=1}^{N} \alpha_i y_i = 0 \quad -(2)$$

<u>Dual:</u>

$$L_D := \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{N}\alpha_i\alpha_k y_i y_k x_i^T x_k \quad -(3)$$

Subject to $\alpha_i \geq 0$, $\forall i$

$$\alpha_i [y_i(x_i^T\beta+\beta_0) - 1] = 0 \quad \forall i \cdot - (4)$$

$\alpha_i = 0$ For non-marginal points → which are very far from hyperplane.

↓ will not Contribut to calculating 'β'.

$\alpha_i \neq 0$, x_i is on the margin. → (support vectors)

↳ only points on the margin plane Contribu. to β.

↓ points which lies on margin plane.

$\boxed{\text{Calculate } \beta_0 \text{ from here.}}$

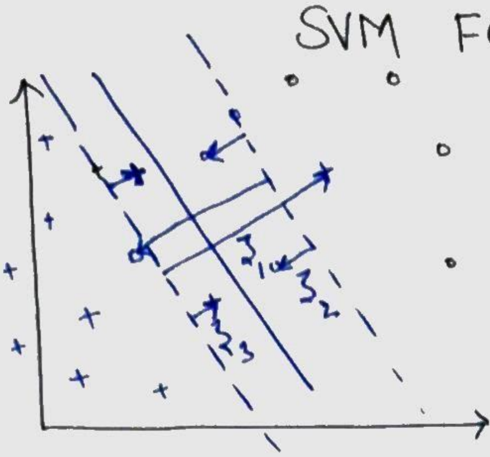
$$\boxed{\begin{aligned} f(x) &= x^T\hat{\beta} + \hat{\beta}_0 \\ &= \sum \alpha_i y_i x_i^T x_i + \beta_0 \end{aligned}}$$

$\boxed{\text{SVM is stable}}$ → $\boxed{\text{less variance}}$

points outs the margin not affect the SVM. only support vectors will affect SVM

Lecture-23

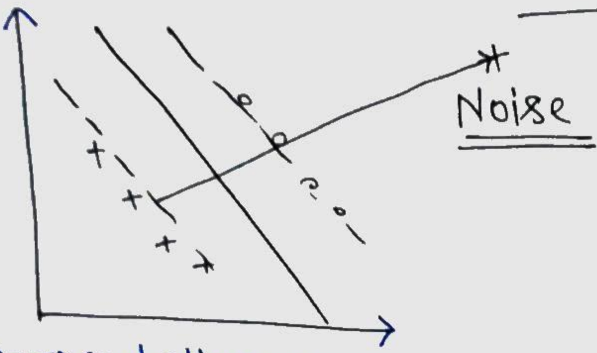
SVM FOR linear Non-seperable data



1. Minimize these distances.

$\xi_1, \xi_2, \xi_3 \rightarrow$ misclassified samples

Qe? why not minimize the maximum distance rather than minimizing sum of distances.



Ans: classifier will try to minimize the noise, and hyperplane will get shifted towards Noise. like.

Formulation:

$$y_i(x_i^T\beta + \beta_0) \geq M(1-\xi_i)$$

$\xi_i \rightarrow$ slack variable

1. Ideally we want to maximise M. and $\xi_i = 0$
2. but I want some sort of relaxation in maximimising M.

why not $(M-\xi_i)$ Ans: Non-Convex optimization ~~loss~~ problem.

ξ_i: ~~what~~ fraction of margin (M)

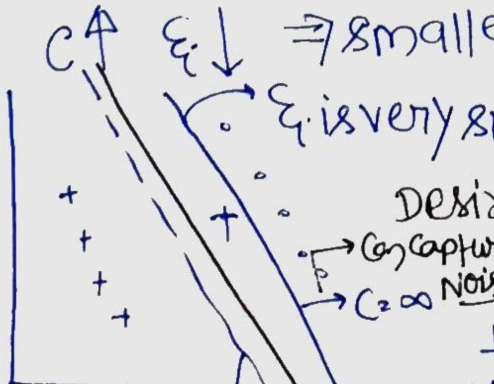
① $\forall_i \ \xi_i \geq 0$, ② $\sum_{i=1}^{N} \xi_i \leq$ Constant

↑ we don't want ξ_i to be very large.

$$\min \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i, \quad \text{sub. to } y_i(x_i^T\beta + \beta_0) \geq (1-\xi_i)$$

since $M = \frac{1}{\|\beta\|}$ like the previous formulation

$C\uparrow \ \xi_i\downarrow \Rightarrow$ smaller margin. $\quad \xi_i \geq 0$



ξ_i is very small. $\rightarrow$ less robust

Desirable ξ_i should be more.

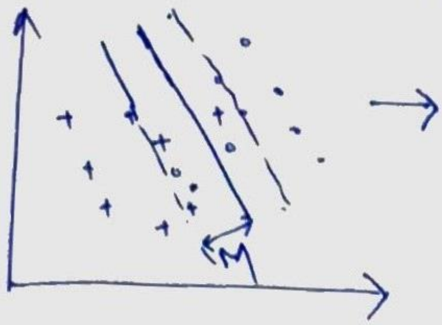
C should be less.

tradoff between increasing C and ξ_i.

$\boxed{C\uparrow \ \xi_i\downarrow, \ C\downarrow \ \xi_i\uparrow}$

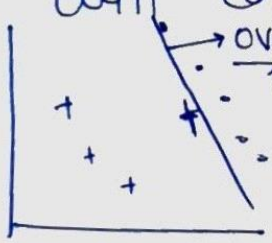
NOTE:

introducing ξ_i makes SVM more stable and less prone to noisy data. Suppose: we put $\xi_i = 0$, then we will try to maximimize M. as shown in figure.



→ We should provide some slackness, because M is calculated via support vectors which lies on the boundary. It doesn't care about the points which lies within the margin. so, we need to provide some slackness in variable. M. by introducing ξ_i which is distance of points which are within the margin from margin plane. If ~~make~~ we need sum of ξ_i to be constant so it shouldn't overshoot.

If we make $C = \infty$, it means all ξ_i are very small, since classifier will try to minimize loss, in doing so, it has to make all $\xi_i = 0$, ⇒ ~~the~~ margin will be very small. overfitting will capture Noise also as shown below



→ overfit due to one noise.

If $C = 0$, ξ_i will be larger, more biased SVM classifier.

underfitting

Lecture-24 SVM KERNELS

$f(x) = x_i^T \beta + \beta_0 \Rightarrow \sum_{i=1}^{N} \alpha_i y_i \boxed{x_i^T x} + \beta_0$

$\rightarrow$ inner product.

is there any good way to compute inner product.

$x \rightarrow h(x)$ "Transformation on x".

$L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle h(x_i), h(x_j) \rangle$

similarily

$f(x) = \sum_{i=1}^{N} \alpha_i y_i \langle h(x), h(x_i) \rangle + \beta_0$

$\langle h(x), h(x_i) \rangle = K(x, x_i)$ - similarity measures

Kernel - $\boxed{\text{symmetric}}$ $\quad x^T A x \geq 0$

(+ve) semi definite $\quad$ semi definite

"popular choice of K"

poly : $(1 + \langle x, x' \rangle)^d$

RBF : $\exp(-\gamma \|x - x'\|^2)$

ANN : $\tanh(k_1 \langle x, x' \rangle + k_2)$

$k_1 \rightarrow$ Constant $\quad k_2 \rightarrow$ Constant.

Poly

$(1 + \langle x_i, x \rangle)^d = (1 + x_1 x_1' + x_2 x_2')^2, \quad d = 2$

$= \underset{1}{1} + \underset{2}{2x_1 x_1'} + \underset{3}{2x_2 x_2'} + \underset{4}{(x_1 x_1')^2} + \underset{5}{(x_2 x_2')^2} + \underset{6}{2x_1 x_1' x_2 x_2'}$

$\boxed{X = (x_1, x_2)}$

$h_1(x) = 1, \quad h_2(x) = \sqrt{2} x_1, \quad h_3(x) = \sqrt{2} x_2$

$h_4(x) = x_1^2, \quad h_5(x) = x_2^2, \quad h_6(x) = \sqrt{2} x_1 x_2$

(6-dimension space)

Primal

$$L_p = \frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\left(\alpha_i\left[y_i(x_i^T\beta+\beta_0)-(1-\xi_i)\right]\right) - \sum_{i=1}^{N}\mu_i\xi_i$$

$\alpha_i, \mu_i \geq 0$

Setting derivatives to '0'.

$\beta = \sum \alpha_i y_i x_i$ — ① $0 = \sum_{i=1}^{N}\alpha_i y_i$ — ②, we don't Need $\sum \xi_i \leq$ constant

$\alpha_i = C - \mu_i$ — ③

↓ It's taken care by 'C' and optimization equation

Dual:

$$L_D = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

Subject to $\boxed{0 \leq \alpha_i \leq C \;\&\; \sum_{i=1}^{N}\alpha_i y_i = 0}$

KKT Condition

$C\xi_i = 0$, — ④ $\alpha_i[y_i(x_i^T\beta+\beta_0)-(1-\xi_i)] = 0$ — ⑤, $\mu_i\xi_i = 0$ — ⑥

$y_i(x_i^T\beta+\beta)-(1-\xi_i) \geq 0$

① If $\alpha_i = 0 \Rightarrow x_i$ far away since $\xi_i = 0$ for farther points.

② $0 < \alpha_i < \infty$, ~~$\alpha_i \neq 0$~~, $\xi_i = 0 \Rightarrow y_i(x_i^T\beta+\beta) = (1-\xi_i)$ → support vector

③ $\boxed{y_i(x_i^T\beta+\beta_0)} - (1-\xi_i) < 1$ → Misclassified

if $\xi_i \neq 0$ ⇓ $\mu_i = 0$ ⇓ $\boxed{\alpha_i = C}$

if $y_i(x_i^T\beta+\beta_0) > 1$, $\xi_i = 0 \Rightarrow \alpha = 0$

if $y_i(x_i^T\beta+\beta_0) = 1 \Rightarrow 0 < \alpha_i < C$ ($\xi_i = 0$)

if $y_i(x_i^T\beta+\beta_0) < 1 \Rightarrow \xi_i > 0, \Rightarrow \mu_i = 0, \Rightarrow \alpha_i = C$

→ in both cases $\alpha_i \neq 0 \Rightarrow$ those x_i's are support vectors.

Lecture-24 SVM kernel

using kernels, we don't need to compute the basis function to transform vectors.

like in previous example, we don't need to comput $h(x)$.

$\boxed{\text{C-SVM}}$ → we call it C-SVM - whatever we learnt so far.

Different type of SVM's are possible where people use different optimisation constraint.

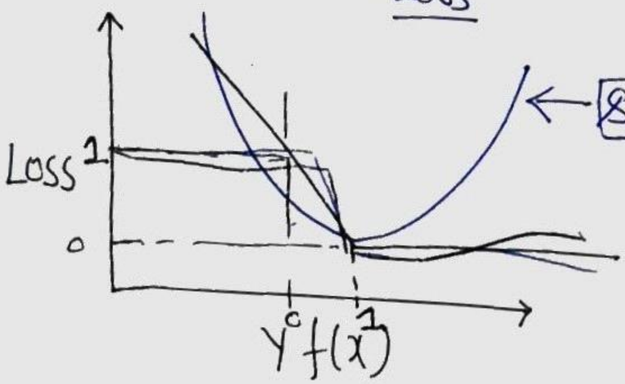
Lecture-25 - Hinge Loss Formulation

$$L_p = \cancel{\sum_{i=1}^{N} \alpha_i} - \frac{1}{2}\|\beta\|^2 - \sum_{i=1}^{N} \alpha_i \left[y_i (x_i^T \beta + \beta_0) - 1 \right]$$

$$= \min_{\beta, \beta_0} \sum_{i=1}^{N} \underbrace{[1 - y_i f(x_i)]_+}_{\text{Count only when It's positive.}} + \frac{1}{2}\|\beta\|^2$$

Loss

Penalty



← $\boxed{\text{squared error}}$

$(1 - y f(x_i))^2$