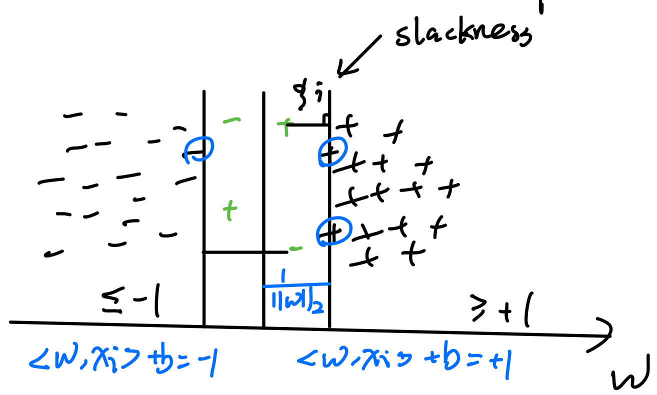
Wrap up SVM

General case: Non-separable case



Previously, $\min \ \frac{1}{2}\|w\|_2^2$

s.t. $y_i(\langle w, x_i \rangle + b) \geq 1$, $i = 1, \cdots, n$

Now, relax the constraints:

$$\min \ \frac{1}{2}\|w\|_2^2 + \underbrace{C \sum_{i=1}^{n} \xi_i}_{\text{penalty}}$$

C → tuning constant

s.t. $y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i$ ... $\alpha_i \geq 0$ (Primal)

$\xi_i \geq 0$, $i = 1, \cdots, n$ ... $\mu_i \geq 0$

$$\text{primal} \Longrightarrow \min \max L \Longrightarrow \max_{(\alpha,\mu)} \min_{(w,b,\xi)} L \xrightarrow{\text{Solved in close form}} \max Q \ (\text{Dual})$$

Primal parameters: w, b, ξ

Dual parameters: α, μ (Lagrange multipliers)

$$L(w, b, \xi, \alpha, \mu) = \underbrace{\frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n}\xi_i}_{\text{Primal objective}} + \underbrace{\sum_{i=1}^{n}\alpha_i\left(1 - \xi_i - y_i\langle w, x_i\rangle + b\right) + \sum_{i=1}^{n}\mu_i(-\xi_i)}_{\text{free the constraints}}$$

≤ 0 in original form

Still concave - convex

$$\min_{w,b,\xi} L \longrightarrow Q(\alpha)$$

Searching for the saddle point

(same for all 3 rounds)

Representer

$$\boxed{\frac{\partial L}{\partial w} = 0} \Longrightarrow \hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Alternative : $\|w - \hat{w}\|_2^2$

$$\boxed{\frac{\partial L}{\partial b} = 0} \Longrightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \text{(caused by } b\text{)}$$

Alter : $+\infty / -\infty$ trick

$$\boxed{\frac{\partial L}{\partial \xi_i} = 0} \Longrightarrow \alpha_i = C - \mu_i \leq C, \ i = 1, \cdots, n \qquad \text{(caused by non-separability)}$$

(box constraints)

$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left\| \sum_{i=1}^{n} \alpha_i y_i x_i \right\|_2^2 \qquad \text{(Same as that in the 1st round)}$$

Kernelize :

$$\hat{w} = \sum_{i=1}^{n} \alpha_i y_i x_i$$

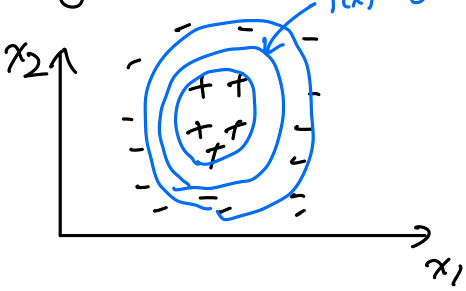
interpolation

$$f(x) = \langle \hat{w}, x \rangle + b = \sum_{i=1}^{n} \alpha_i y_i \underbrace{\langle x_i, x \rangle}_{k(x_i, x)} + b \qquad \text{(learned classifier)}$$

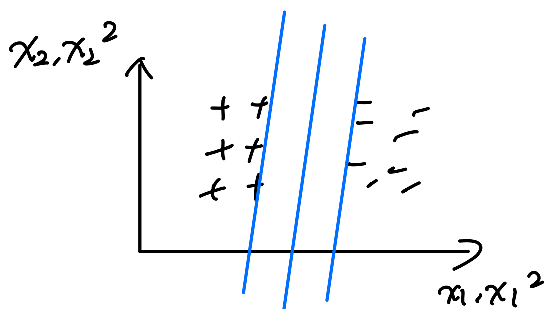
$$Q(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \underbrace{\langle x_i, x_j \rangle}_{k(x_i, x_j)} \qquad \text{(training)}$$

$$x \longrightarrow \phi(x) \qquad \langle x, x' \rangle \longrightarrow \langle \phi(x), \phi(x') \rangle = k(x, x')$$

e.g.



Non-separable



Dual optimization :

$$\max_{\substack{\alpha_i \in [0,c], \forall i \\ \sum_i \alpha_i y_i = 0}} Q(\alpha)$$

If remove $\sum_{i=1}^{n} \alpha_i y_i = 0$ (assuming $b=0$)

$$\max_{\alpha_i \in [0,c]} Q(\alpha)$$

(Coordinate ascent to solve it)

Each time update one α_i

However, if we allow b, one more constraint : $\sum_{i=1}^{n} \alpha_i y_i = 0$

Sol : Each time, change 2 (α_i, α_j) (Sequential minimal optimization)

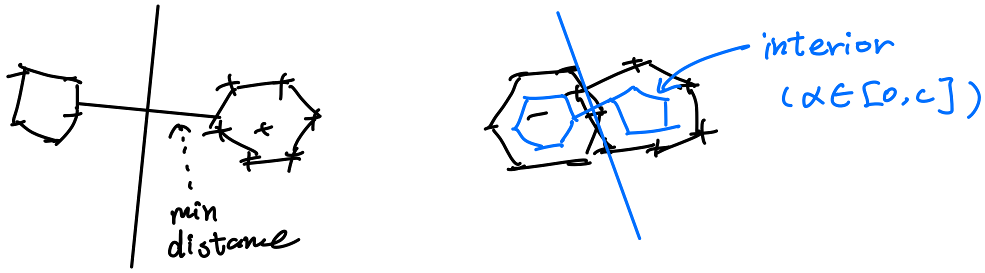
SMO (Linear case, b important)

$\updownarrow$

Kernel case, b negligible

(Dual Coordinate ascent)

Geometry of Dual optimization :



Separable

Non-separable

Primal : Max margin

Dual : Min distance

Stats aspect / Connected to logistic regression

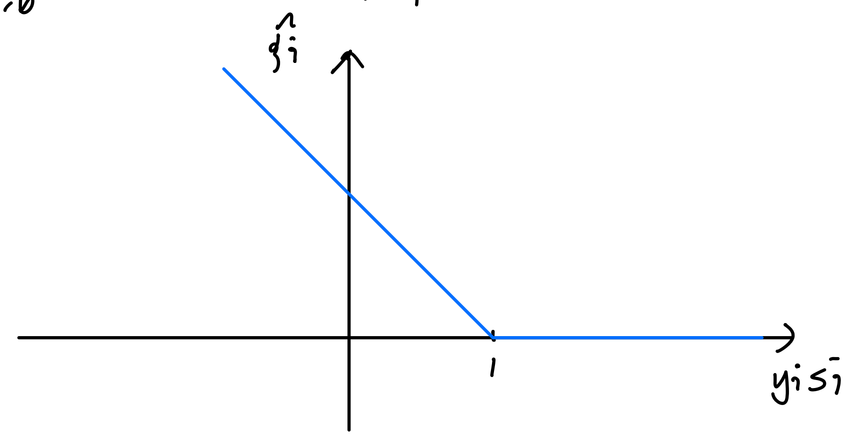
Another way to translate constraints :

$$y_i(\langle w, x_i\rangle + b) \geq 1 \implies \hat{q}_i = 0 \quad \leftarrow \text{penalty}$$

$$\ldots < 1 \implies \hat{q}_i = 1 - y_i(\langle w, x_i\rangle + b)$$

$$\hat{q}_i = \max(0, 1 - y_i \underbrace{(\langle w, x_i\rangle + b)}_{s_i})$$ hinge loss

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{n} \max(0, 1 - y_i(\langle w, x_i\rangle + b))$$



Logistic regression :

$$s_i = \langle w, x_i\rangle + b$$

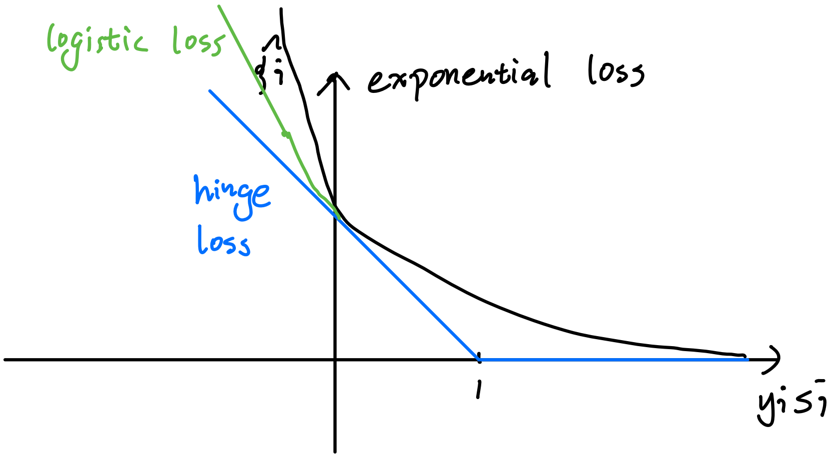
$$p_i = \frac{e^{s_i}}{1+e^{s_i}} = Pr(y_i = +1 \mid s_i), \quad 1 - p_i = \frac{1}{1+e^{s_i}} = Pr(y_i = -1 \mid s_i)$$

Note : y_i is 0/1

$$P(y_i \mid s_i) = \frac{1}{1+e^{-y_i s_i}}$$

$$\text{Loss} = -\log P(y_i \mid s_i) = \log(1+e^{-y_i s_i})$$ logistic loss

$$= \begin{cases} -y_i s_i, & y_i s_i \to -\infty \quad (\text{Similar to } 1 - y_i s_i, \text{ SVM}) \\ e^{-y_i s_i}, & y_i s_i \to +\infty \quad (\text{exponential loss, Adaboost}) \end{cases}$$



similar to lasso forward selection

Rethink

SVM : $\sum_{i=1}^{n} \max(0, 1 - y_i \langle w, x_i \rangle + b) + \underbrace{\frac{1}{2C}\|w\|_2^2}$ margin consideration

logistic : $\sum_{i=1}^{n} \log(1 + \exp(\langle w, x_i \rangle + b)) + \underbrace{\lambda \|w\|_2^2}$ shrinkage consideration

ridge logistic regression