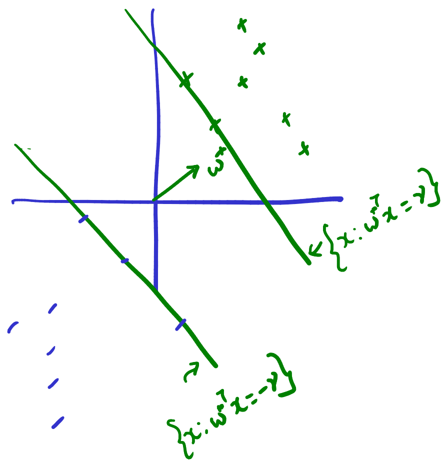


PERCEPTRON

$$\#\text{ mistakes} \le \frac{R^2}{\gamma^2}$$

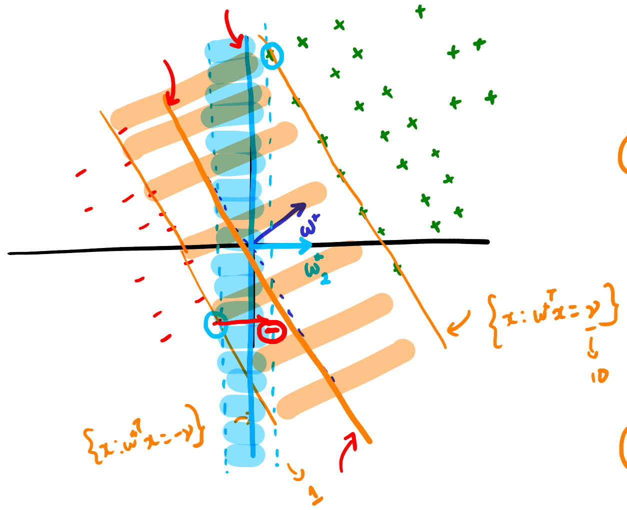
$\|x_i\|^2 \le R^2$

Dataset - L.S with margin γ

$(w^{*T}x_i)y_i \ge \gamma \quad \forall i$

$\gamma > 0$

"QUALITY" OF FINAL SOLUTION



Question

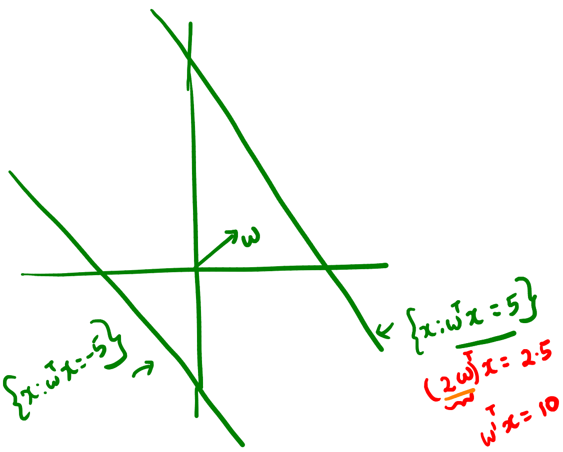
Given that we prefer classifies with large margin, can we directly find them?

Observation

1. # mistakes depends on the best possible w^*'s margin.

2. w_{Perc} need not necessarily be w^*. It could be w_2^* also (blue line)

Goal: To come up with a formulation that maximizes "margin"



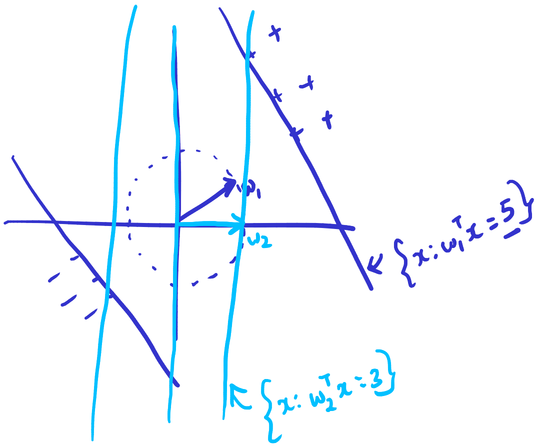
$$\max_{w,\gamma} \ \gamma$$

such that

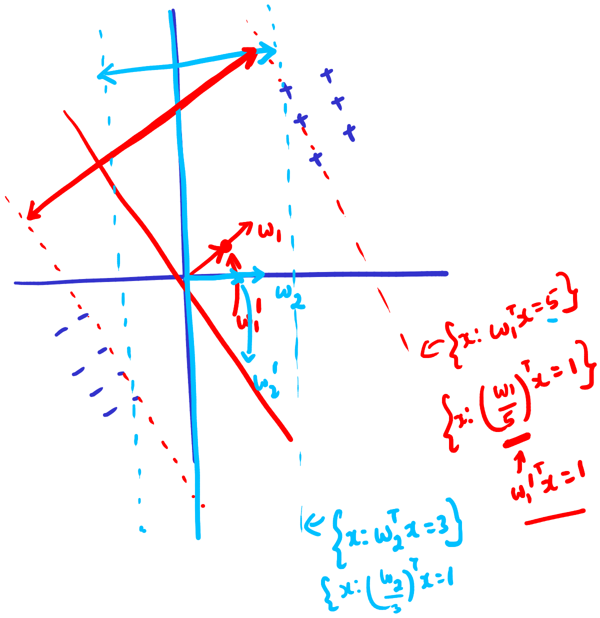
$$(w^Tx_i)y_i \ge \gamma \quad \forall i$$

Issue: Can scale w arbitrarily.

$$\boxed{\begin{aligned} &\max_{w,\gamma} \ \gamma \\ &(w^Tx_i)y_i \ge \gamma \quad \forall i \\ &\|w\|^2 = 1 \end{aligned}}$$

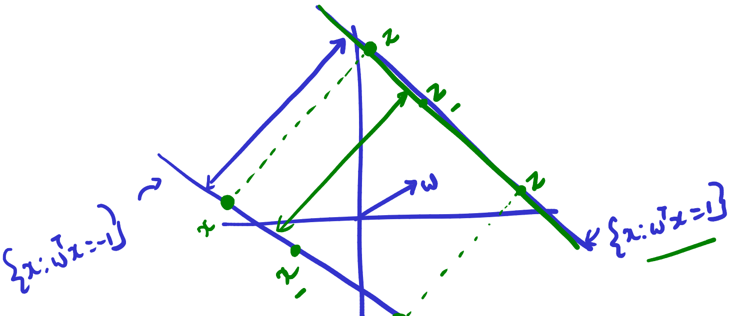


$\Rightarrow$



$$\max_{w} \quad \text{width}(w)$$
$$\text{s.t.} \quad (w^T x_i) y_i \geq 1 \quad \forall i$$

What is width(w) ?

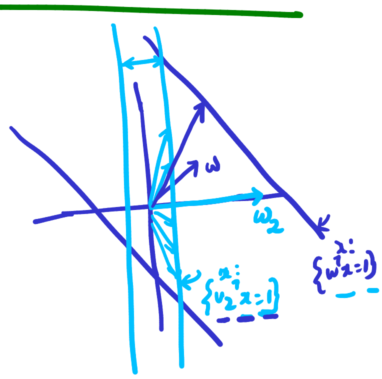


$$\boxed{w^T x = -1}$$

$$\min_{z} \quad \boxed{\frac{1}{2} \| x - z \|^2}$$
$$\text{s.t.} \quad w^T z = +1$$

Solution : $\quad \text{width}(w) = \boxed{\frac{2}{\|w\|^2}}$

$$\max_{w} \quad \frac{2}{\|w\|^2}$$
$$\text{s.t.} \quad (w^T x_i) y_i \geq 1$$



Equivalently

$$\boxed{\begin{aligned} &\min_{w} \quad \frac{1}{2}\|w\|^2 \\ &\text{s.t.} \quad (w^T x_i) y_i \geq 1 \quad \forall i \end{aligned}}$$

$$\min_{w \in \mathbb{R}^d} \ \underbrace{\frac{1}{2}\|w\|^2}$$

$$\text{s.t.} \ \underline{\forall i} \quad (w^T x_i) y_i \geq 1$$

— (A)

DETOUR

$$\min_{w} \ f(w) \qquad \text{s.t.} \ g(w) \leq 0$$

$$\mathcal{L}(w, \alpha) = f(w) + \alpha\, g(w)$$

Fix any w.

Consider $\underbrace{\max_{\alpha \geq 0} \mathcal{L}(w, \alpha)} = \underbrace{\max_{\underline{\alpha \geq 0}} \ \underbrace{f(w)} + \underbrace{\alpha\, g(w)}}$

$w = [1\ 2\ 3\ 4]$: $f(w) = -100$, $g(w) = 5$

$$\max_{\underline{\alpha \geq 0}} \ \underbrace{-100 + \alpha 5} \uparrow$$

$\alpha = 1$: -95

$\alpha = 10$: -50

$\alpha = 100$: 400

$\vdots$

$w = [3\ 4\ 5\ 1]$: $\boxed{f(w)} = \boxed{100}$, $g(w) = -5$

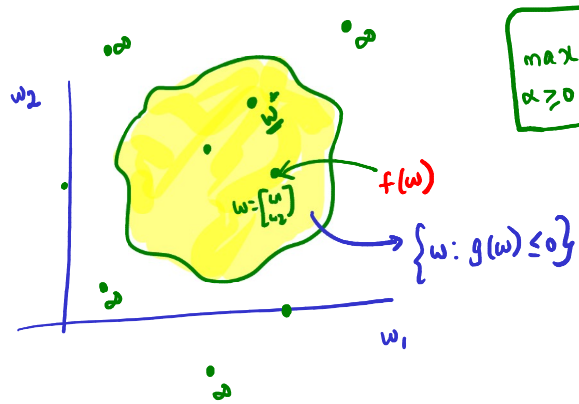
$100 - 5\boxed{\alpha}$

$\alpha = 1 \Rightarrow 95$

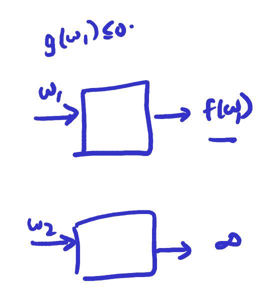
$\alpha = 10 \Rightarrow 50 \downarrow$

$\alpha = 0 \Rightarrow \boxed{100}$

$$= \begin{cases} \underline{\infty} & \underline{g(w) > 0} \\ \underline{f(w)} & \underline{g(w) \leq 0} \end{cases}$$



$$\boxed{\max_{\alpha \geq 0} \ f(w) + \alpha\, g(w)}$$



$$\boxed{\min_{w} \ f(w)} \equiv \min_{w} \left[\max_{\alpha \geq 0} \ f(w) + \alpha\, g(w) \right]$$

- Can we swap min and max in (B)?

- In general, No! But if f and g are "nice" functions [convex functions], then yes!
  ↳ [Quadratic / Linear]

$$\min_{w} \left[\max_{\alpha \geq 0} f(w) + \alpha g(w) \right] \equiv \max_{\alpha \geq 0} \left[\min_{w} f(w) + \alpha g(w) \right]$$

For convex f and g.

For multiple constraints

$$\boxed{\begin{aligned} &\min_{w} f(w) \\ &\text{s.t. } g_i(w) \leq 0 \quad \forall i = 1 \ldots k \end{aligned}} \equiv \min_{w} \left[\max_{\substack{\alpha_1, \ldots, \alpha_k \\ \geq 0 \; \geq 0 \; \geq 0}} f(w) + \alpha_1 g_1(w) + \alpha_2 g_2(w) + \cdots + \alpha_k g_k(w) \right]$$

$$\equiv$$

$$\max_{\alpha_1 \geq 0, \ldots, \alpha_k \geq 0} \left[\min_{w} \left[f(w) + \alpha_1 g_1(w) + \cdots + \alpha_k g_k(w) \right] \right]$$

$$\min_{w} \frac{1}{2} \|w\|^2 \leftarrow \text{Quadratic in } w$$

$$\text{s.t. } (w^T x_i) y_i \geq 1 \quad \forall i = 1, \ldots, n \leftarrow \text{Linear in } w$$

$$\equiv \underbrace{1 - (w^T x_i) y_i}_{g_i(w)} \leq 0 \quad \forall i = 1, \ldots n$$

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}$$

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

$$\min_{w} \max_{\alpha \geq 0} \left[\frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right] \equiv \max_{\alpha \geq 0} \left[\min_{w} \left[\frac{1}{2} \|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right) \right] \right]$$

Fix some $\alpha \geq 0$. $\quad \alpha = \begin{bmatrix} 1 \\ 5 \\ 7 \end{bmatrix}$

$$\min_{w} \underbrace{\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)}$$

$$w_\alpha^* + \sum_{i=1}^{n} \alpha_i (-x_i y_i) = 0$$

$$\boxed{w_\alpha^* = \sum_{i=1}^{n} \alpha_i x_i y_i}$$

$\in \mathbb{R}^d$ $\quad \geq 0 \quad \{\pm 1\}$

→ Substitute back value of w_α^* in the objective

$$w_\alpha^* = X y \alpha$$

$$X = \underbrace{\begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}_{d \times n} \begin{bmatrix} y_1 & & 0 \\ & \ddots & \\ 0 & & y_n \end{bmatrix}_{n \times n}} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}_{n \times 1}$$

$$\frac{1}{2}\|w\|^2 + \sum_{i=1}^{n} \alpha_i \left(1 - (w^T x_i) y_i\right)$$

substitute $w_\alpha^* = X y \alpha$ into

on simplification

$$= \alpha^T \mathbf{1} - \frac{1}{2} (X y \alpha)^T (X y \alpha)$$

$[\vdots]_{n \times 1}$

DUAL PROBLEM

$$\boxed{\max_{\boxed{\alpha \geq 0}} \underbrace{\alpha^T \mathbf{1} - \frac{1}{2} \alpha^T y^T \boxed{X^T X} y \alpha}}$$

$\mathbb{R}^{n \times n}$ → kernel K

PRIMAL

$$\min_{w \in \mathbb{R}^d} \frac{1}{2}\|w\|^2$$

s.t. $\boxed{(w^T x_i) y_i \geq 1}$

What have we gained?

- Dual variable dimension in $\mathbb{R}_+^n$ while primal problem dimension is $\mathbb{R}^d$

- Dual constraints are "easier"

- More Importantly dual depends on $x^T x$ and so can be "KERNELIZED"!

$$\underbrace{w^*_{\alpha^*} = \sum_{i=1}^{n} \alpha_i^* x_i y_i}$$

→ This says optimal w^* is a linear combination of the data points where importance of a datapoint is given by α_i^* (for i^{th} data point)

→ Question: Where are the "IMPORTANT" points? (i.e., points for which $\alpha_i^* > 0$)

REVISITING THE LAGRANGIAN

$$\min_{w} \underbrace{\left[\max_{\alpha \geq 0} f(w) + \alpha g(w) \right]}_{\text{Primal}} \equiv \max_{\alpha \geq 0} \underbrace{\left[\min_{w} f(w) + \alpha g(w) \right]}_{\text{Dual}}$$

w^* is the primal solution

α^* is the dual solution

$$\max f(w^*) + \alpha g(w^*) = \min_{w} f(w) + \alpha^* g(w)$$

$$f(w^*) = \min_w f(w) + \alpha^* g(w)$$
$$\leq f(w^*) + \alpha^* g(w^*)$$

$$\Rightarrow \cancel{f(w^*)} \leq \cancel{f(w^*)} + \alpha^* g(w^*)$$

$$\Rightarrow \boxed{\alpha^* g(w^*) \geq 0} \quad -(1)$$

But we already know $\underline{\alpha^* \geq 0}$ & $\underline{g(w^*) \leq 0}$

$$\Rightarrow \boxed{\alpha^* g(w^*) \leq 0} \quad -(2)$$

$$(1) \ \& \ (2) \Rightarrow \boxed{\alpha^* g(w^*) = 0} \rightarrow$$ COMPLEMENTARY SLACKNESS

For multiple constraints,

$$\boxed{\alpha_i^* g_i(w^*) = 0 \quad \forall i}$$

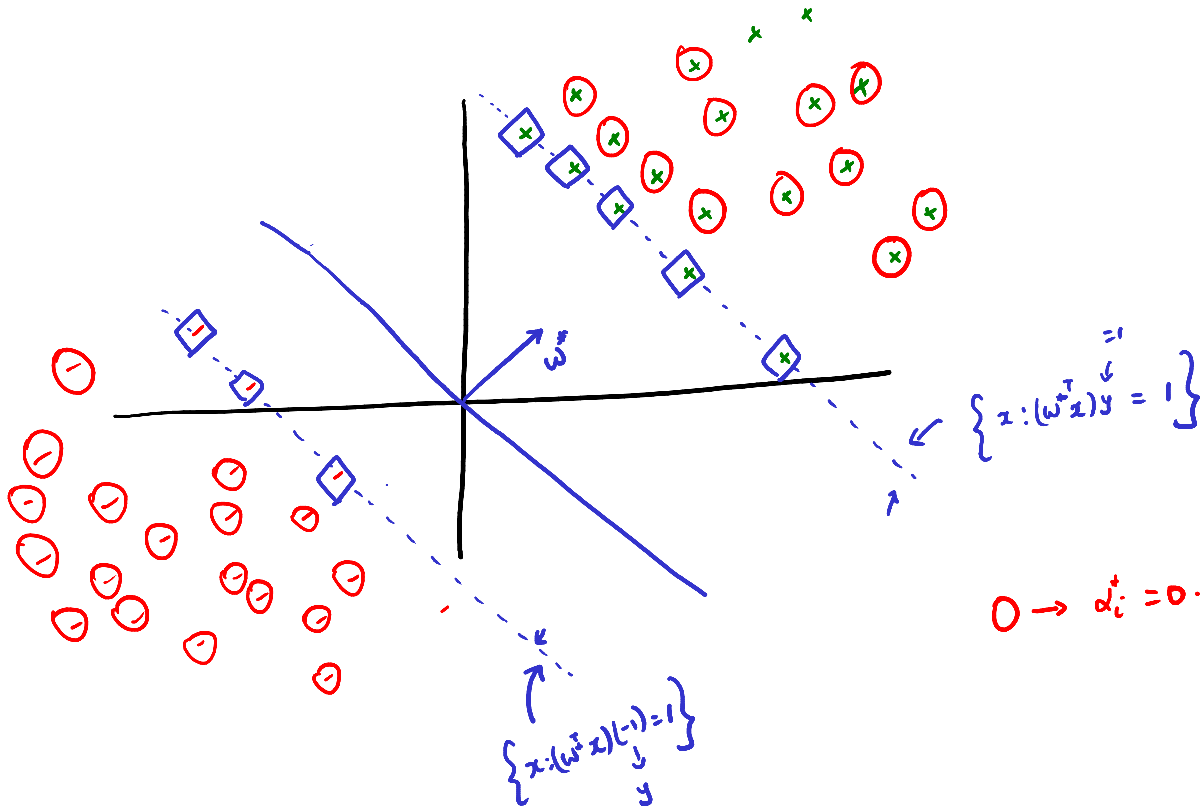
In our problem

$$\underline{\alpha_i^*}\big(\underbrace{1 - (w^{*T}x_i) y_i}_{g_i(w^*)}\big) = 0 \quad \forall i$$ [by complementary slackness]

$$\Rightarrow \text{If } \underline{\alpha_i^* > 0} \overset{C.S}{\Rightarrow} 1 - (w^{*T}x_i) y_i = 0$$

$$\Downarrow$$

$$\boxed{(w^{*T}x_i) y_i = 1}$$



- Only the points that are on the "SUPPORTING" hyperplane can contribute to w^*
- These special points are called "SUPPORT VECTORS"
- ALGORITHM $\rightarrow$ SUPPORT VECTOR MACHINE (SVM) [Vapnik et. al]
- w^* is a sparse linear combination of the data points.

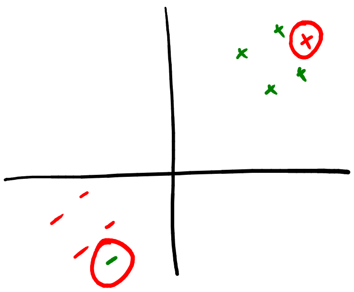
Given x_{test} = $\quad w^{*T} x_{test} = \left(\sum_{i=1}^{n} \alpha_i^* x_i y_i\right)^T x_{test}$

$$= \sum_{i=1}^{n} \alpha_i^* y_i (x_i^T x_{test})$$

x_{test} $\quad w^{*T} \phi(x_{test}) = \sum_{i=1}^{n} \alpha_i^* y_i \, k(x_i, x_{test})$

QUESTIONS

- How to adapt the SVM algorithm when data has outliers.



- KERNELS can help but is not the right way to solve this!

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

s.t. $(w^T x_i) y_i \geq 1 \quad \forall i$

Insight: Make every w feasible.

- Fix any w. w classifies some points correctly and misclassifies some points
- The incorrectly classified points "pay bribe" to go to the "correct" side!

MODIFIED FORMULATION

$\geq 0 \to$ HYPER PARAMETER.

SOFT MARGIN PRIMAL FORMULATION $\leftarrow$

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

s.t. $(w^T x_i) y_i + \xi_i \geq 1 \quad \forall i$

$\xi_i \geq 0 \quad \forall i$

$C = 0$

$\Rightarrow$ Bribes don't cost

$\Rightarrow$ $w = 0 \in \mathbb{R}^d$ is the solution

$C = \infty$

$\Rightarrow$ Linear separable case.