SUPPORT VECTOR MACHINES

- CLASSIFICATION

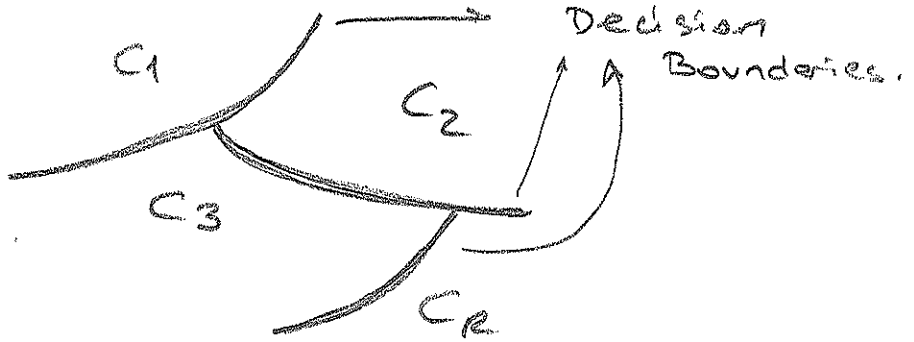
Let $x \in \mathbb{R}^n$, assume there are R classes $C_1, C_2, \ldots, C_m$

Find / build a machine / algorithm :

$x \Rightarrow$ [Classifier] $\Rightarrow i_0(x) \Rightarrow$ class information.

$i_0(x)$: class membership function.

Assume that input space ($\mathbb{R}^n$) is divided into regions.



Finding decision boundaries solves classification problem.

$\Rightarrow$ DISCRIMINANT FUNCTIONS

Assume that we can find R functions

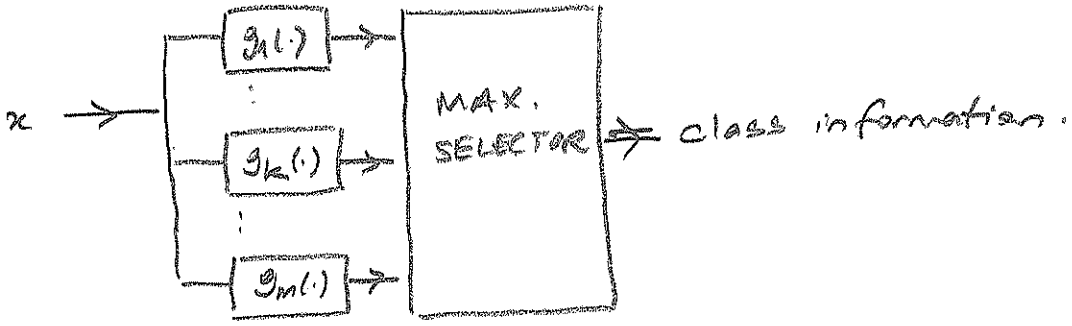
$g_i : \mathbb{R}^n \to \mathbb{R} \qquad i = 1, 2, \ldots$

$\Rightarrow \quad x \in C_k \iff g_k(x) > g_j(x) \qquad j = 1, 2, \ldots, m, \; j \neq k$

$\Rightarrow$ in the region of C_k, the relevant discriminant function $g_k(\cdot)$ is maximum.

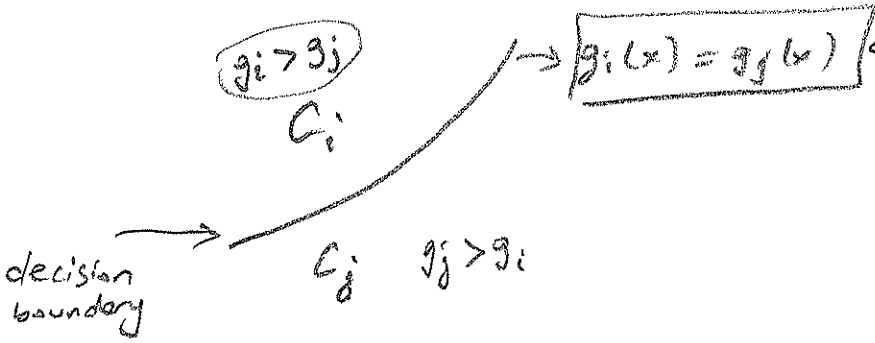
Advantages: If we can find such discriminant fn.s

1°) A machine can be built:



2°) If $g_k(\cdot)$ are LINEAR ⇒ a standard NN structure

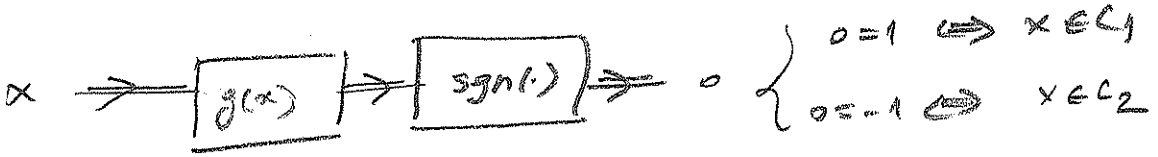
3°) Decision boundaries can be computed.



SPECIAL CASE: 2 class

$$x \in C_1 \Leftrightarrow g_1(x) > g_2(x) \Leftrightarrow g(x) = g_1(x) - g_2(x) > 0$$
$$x \in C_2 \Leftrightarrow g_2(x) > g_1(x) \Leftrightarrow g(x) = g_1(x) - g_2(x) < 0$$

⇒ Decision boundary ≡ $g(x) = 0$

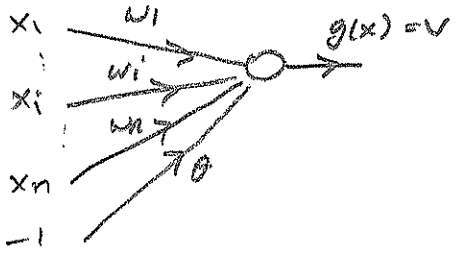


Called DICHOTOMIZER

— LINEAR DISCRIMINANT FUNCTIONS

$$g(x) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n - \theta$$

⇒ related to NN

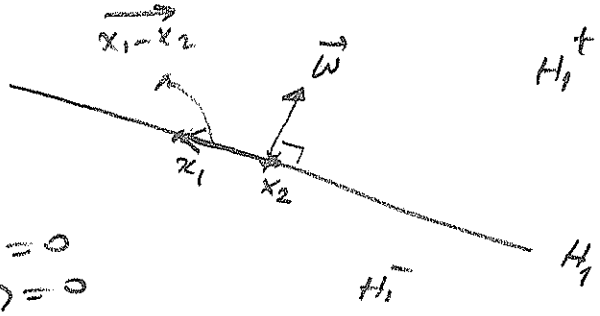


$$w = \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \qquad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

$$\boxed{g(x) = \underset{\sim}{w}^T \underset{\sim}{x} - \theta}$$

1°) $g(x)=0$ is called a HYPERPLANE H_1

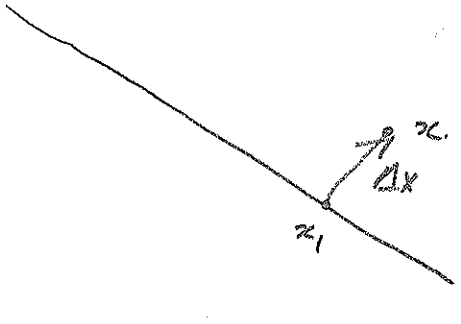
2°) $w \perp H_1$.



Let $x_1 \in H_1 \Leftrightarrow g(x_1) = 0$
$x_2 \in H_1 \quad g(x_2) = 0$

$$\Leftrightarrow \begin{cases} w^T x_1 - \theta = 0 \\ w^T x_2 - \theta = 0 \end{cases} \quad \boxed{w^T (x_1 - x_2) = 0}$$

3°) w, as a vector points towards + side H_1^+



$x_1 \in H_1 \Rightarrow g(x_1) = 0$

$\Rightarrow w^T x_1 - \theta = 0$

$x \in H_1^+$

$\Rightarrow g(x) = w^T x - \theta > 0$

write $x = x_1 + \Delta x \Rightarrow g(x) = w^T x_1 - \theta + w^T \Delta x > 0$

$\Rightarrow w^T \Delta x > 0$ for any Δx pointing to H_1^+

$\Rightarrow |\measuredangle(w, \Delta x)| < 90^\circ$

$w^T \Delta x = \|w\| \cdot \|\Delta x\| \cos \measuredangle(w, \Delta x) > 0$

$\Rightarrow$ w points at + side

4°) Let x_1^* be on H_1, x be a generic points on H_1

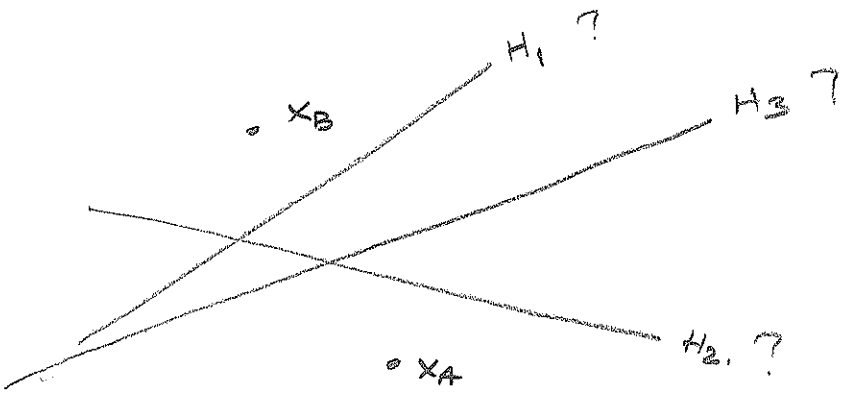
$\Rightarrow$ $g(x) = w^T x - \theta \Rightarrow$
$g(x_1) = w^T x_1 - \theta = 0$

$\boxed{w^T (x - x_1) = 0}$ Line formula

$\Rightarrow g(x) = w^T x - \theta \Leftrightarrow g(x) = w^T x - w^T x_1 \Leftrightarrow \theta = w^T x_1$

(xlote that if $x_1 \in H_1$, $x_2 \in H_1 \Rightarrow \boxed{w^T x_1 = w^T x_2}$

- MINIMUM DISTANCE CLASSIFIER

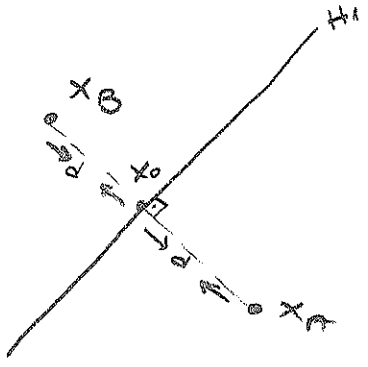


$C_1 = \{x_B\}$, $C_2 = \{x_A\}$.

Question: what is the BEST hyperplane separating C_1 & C_2 ?

Answer: the line which has EQUAL DISTANCE to both x_A and x_B is the BEST ⇒ May tolerate larger perturbations.

PROBLEM: Given x_B and x_A, find the line H_1 which bisects them



x_0 : midpoint ⇔ on H_1 ⇔ $x_0 = \frac{1}{2}(x_A + x_B)$

normal to H_1 ⇒ $x_B - x_A$ ⇒ $\boxed{w = (x_B - x_A)}$

Line formula $\boxed{w^T(x - x_0) = 0}$

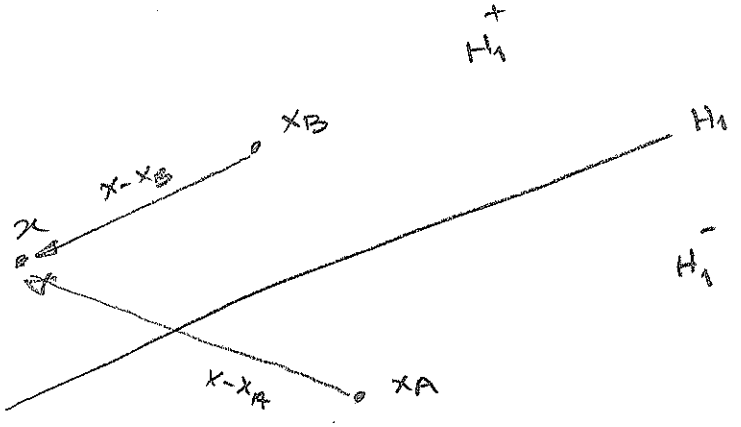
$$\boxed{(x_B - x_A)^T x - \frac{1}{2}(x_B - x_A)^T (x_B + x_A) = 0}$$

$$(x_B - x_A)^T (x_B + x_A) = x_B^T x_B - x_A^T x_A = \|x_B\|^2 - \|x_A\|^2$$

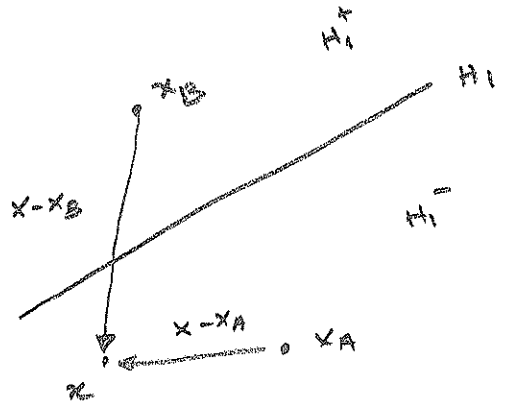
$$\Rightarrow \quad \boxed{g(x) = w^T x - \theta}$$

$$\boxed{w = (x_B - x_A) \qquad \theta = \|x_B\|^2 - \|x_A\|^2}$$

GEOMETRICAL IDEA



$$x \in H_1^+ \Rightarrow \quad \|x - x_B\| < \|x - x_A\|$$



$$x \in H_1^- \Rightarrow \quad \|x - x_A\| < \|x - x_B\|$$

$\Rightarrow$ $x \in H_1^+ \Rightarrow$ minimum distance to x_B

$x \in H_1^- \Rightarrow$ minimum distance to x_A

— RELATION WITH DISCRIMINANT FUNCTIONS

$\Rightarrow \quad g_1(x) = \|x - x_B\| \quad$ or $\quad g_1(x) = \|x - x_B\|^2$

$= (x - x_B)^T (x - x_B)$

$g_2(x) = \|x - x_A\| \quad$ or $\quad g_2(x) = \|x - x_A\|^2$

$= (x - x_A)^T (x - x_A)$

$\Rightarrow \quad x \in C_1 \Rightarrow \quad g_1(x) < g_2(x)$ (minimum distance)

$x \in C_2 \Rightarrow \quad g_2(x) < g_1(x)$

$\Rightarrow$ opposite of discriminant functions

$\Rightarrow$ Actually $-g_1(x)$ and $-g_2(x)$ are discriminant functions.

DECISION BOUNDARY :

$g_1(x) = g_2(x)$

$\Rightarrow \quad \|x - x_B\|^2 = \|x - x_A\|^2$

$\Rightarrow \quad (x - x_B)^T (x - x_B) = (x - x_A)^T (x - x_A)$

$\Rightarrow \quad \|x\|^2 - 2 x_B^T x + \|x_B\|^2 = \|x\|^2 - 2 x_A^T x + \|x_A\|^2$

$\Rightarrow \quad 2 (x_B - x_A)^T x - (\|x_B\|^2 - \|x_A\|^2) = 0$

$\Rightarrow \quad \boxed{(x_B - x_A)^T x - \frac{1}{2} (\|x_B\|^2 - \|x_A\|^2) = 0}$

what if we have more than one member in classes?

$C_1 = \{x_1, \dots, x_P\}$ $\quad C_2 = \{x_{P+1}, \dots, x_N\}$

SOLUTION: Choose a MEANINGFUL class representative point for each classes.

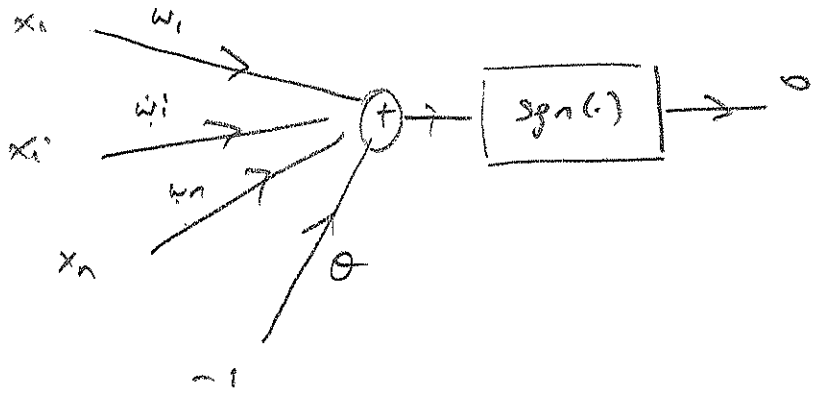
say x_B represents C_1, x_A represents C_2

$\Rightarrow$ use min. distance classifier.

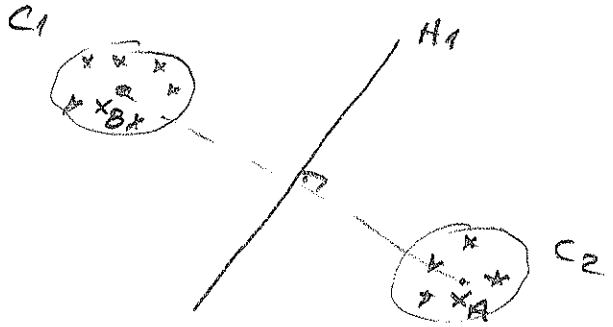
Typically. x_B = cluster center for $C_1 = \frac{x_1 + \dots + x_P}{P}$

x_A = cluster center for $C_2 = \frac{x_{P+1} + \dots + x_N}{N-P}$

$\Rightarrow \quad w = x_B - x_A \quad , \quad \theta = \frac{1}{2}(\|x_B\|^2 - \|x_A\|^2)$



Works fine if C_1 and C_2 are WELL SEPARATED



- MULTICATEGORY CASE

$C_1 = \{x_1, \dots, x_5\} \rightarrow$ class representative $\underset{\sim}{x}^1$

$C_2 = \{x_{5+1}, \dots, x_9\} \rightarrow$ class representative $\underset{\sim}{x}^2$

$\vdots$

$C_i = \{x_{t+1}, \dots, x_r\} \rightarrow$ class representative $\underset{\sim}{x}^i$

$\vdots$

$C_m = \{x_{w+1}, \dots, x_N\} \rightarrow$ class representative $\underset{\sim}{x}^m$

NOTE: One may choose class representatives as cluster centers.

$x \in C_i$ if $\|x - x^i\|^2 < \|x - x^j\|^2 \quad j \neq i$

$\|x - x^i\|^2 = (x - x^i)^T (x - x^i) = \|x\|^2 - 2x_i^T x + \|x^i\|^2$

$\Rightarrow x \in C_i$ if $\|x\|^2 - 2x_i^T x + \|x^i\|^2 < \|x\|^2 - 2x_j^T x + \|x^j\|^2$

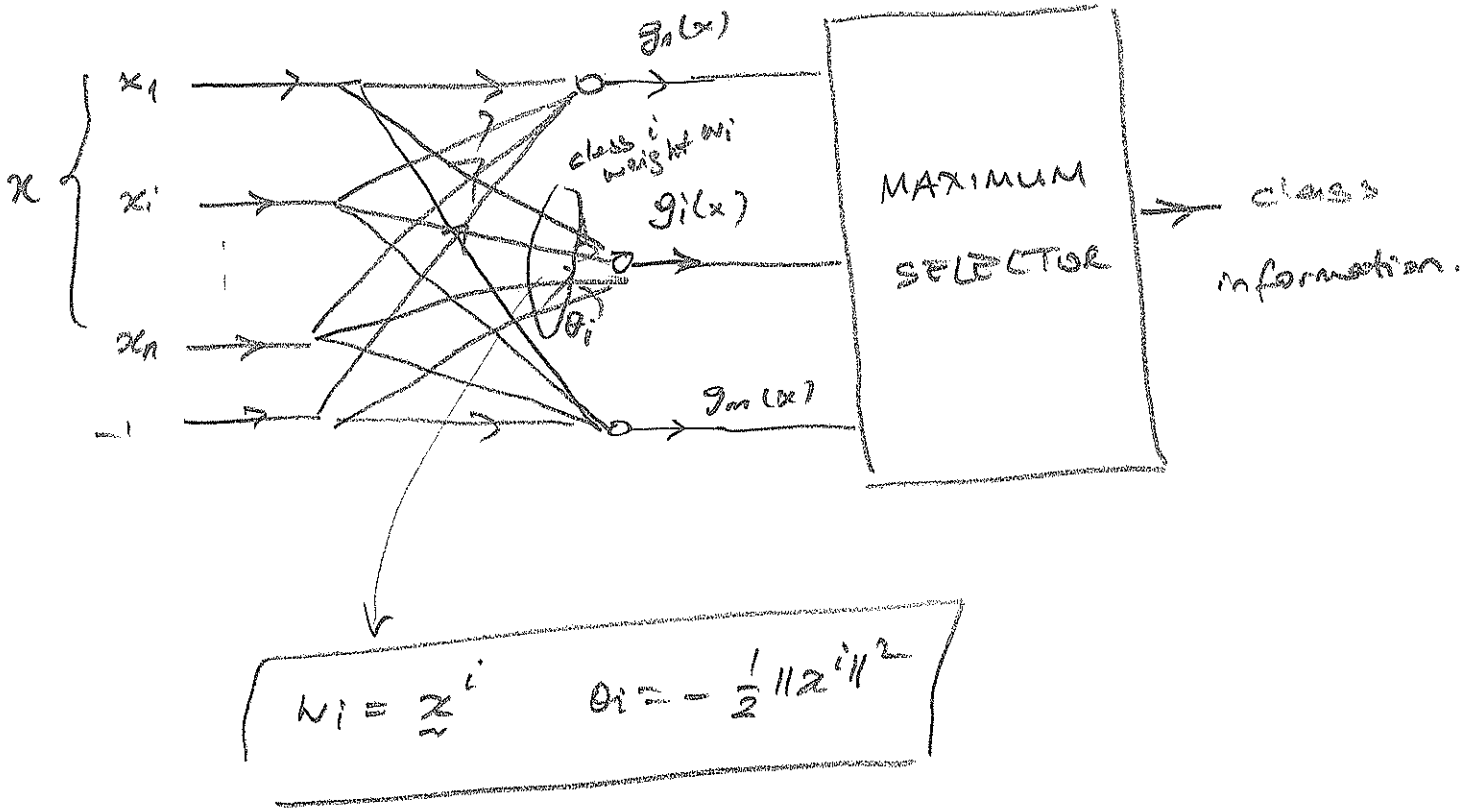
$\Rightarrow$ if $2x_i^T x - \|x^i\|^2 > 2x_j^T x - \|x^j\|^2$

$\Rightarrow$ Discriminant function for $C_i \Rightarrow g_i(x) = x_i^T x - \frac{1}{2}\|x^i\|^2$

$x \in C_i \Leftrightarrow g_i(x) > g_j(x) \quad j \neq i$

$g_i(x) = x_i^T x - \frac{1}{2}\|x^i\|^2$

CLASSIFIER



$$w_i = x^i \qquad \theta_i = -\frac{1}{2}\|x^i\|^2$$

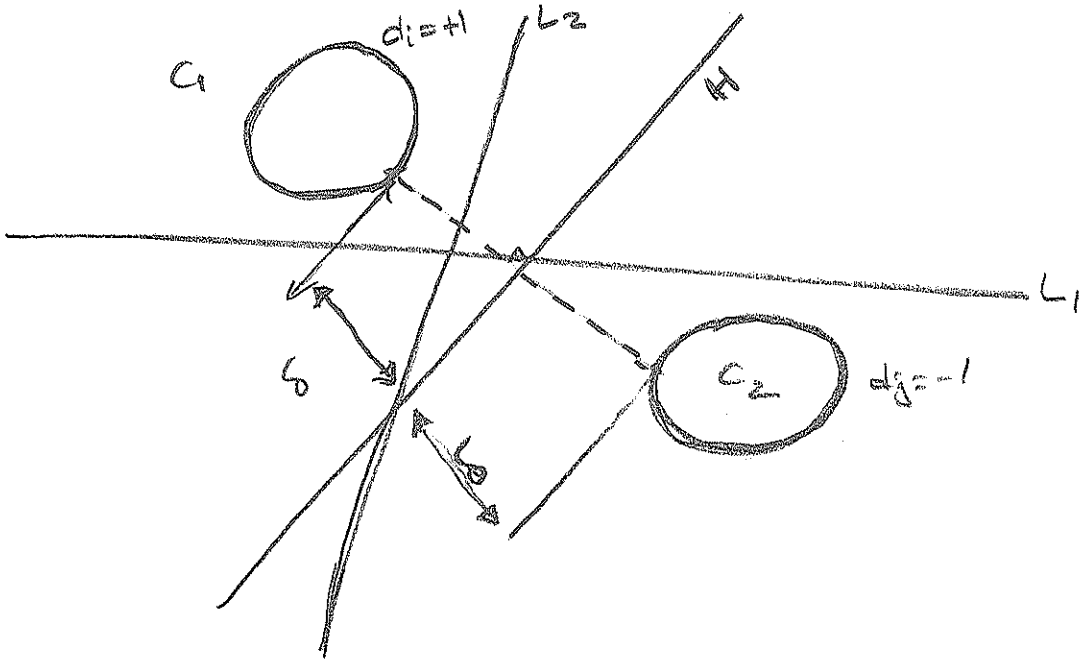
SUPPORT VECTOR MACHINES (SVM)

- Extension of Minimum Distance classifier.
- Linear case (LSVM)
- Assume that we have 2 classes

$$C_1 = \{x_1, \ldots, x_i, \ldots x_m\} \Rightarrow d_i = +1 \quad \text{for } i = 1, 2, \ldots, m$$

$$C_2 = \{x_{m+1}, \ldots, x_j, \ldots x_N\} \Rightarrow d_j = -1 \quad \text{for } j = m+1, \ldots, N$$

- Assume that the classes are linearly seperable.
- By using perceptron algorithm, a seperating hyperplane can be found.
- Linearly seperable case $\Rightarrow$ infinitely many solutions.



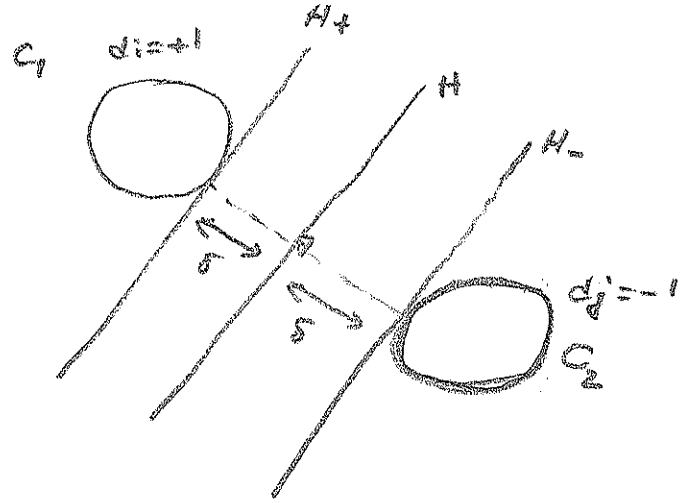
Among infinitely many solutions, which one is "the best"?

→ Preferably H_1. why?

- SVM is based on finding such a hyperplane H

H : $\boxed{g(x) = w^T x - \theta = 0}$

- Should be on the midpoint between C_1 & C_2



$\boxed{\begin{aligned} H_+ \Rightarrow \quad & g_+(x) = w^T x - \theta = b \\ H_- \Rightarrow \quad & g_-(x) = w^T x - \theta = -b \end{aligned}}$

$b > 0$ is a scaler.

PROPERTY 1: For simplicity, we may normalize $b \Rightarrow 1$

Why? $H^+ \quad w^T x - \theta = b \iff \frac{1}{b} w^T x - \frac{1}{b}\theta = 1$

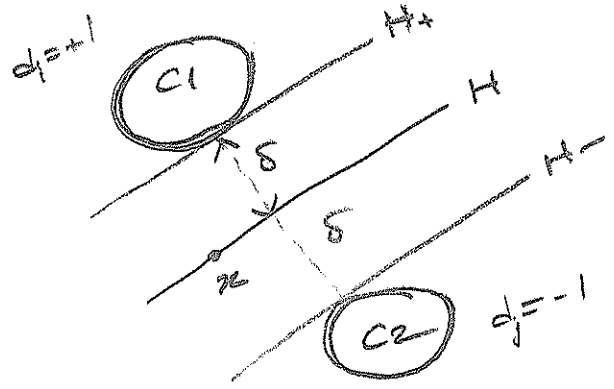
$H^- \quad w^T x - \theta = -b \iff \frac{1}{b} w^T x - \frac{1}{b}\theta = -1$

Define $\hat{w} = \frac{1}{b} w \qquad \hat{\theta} = \frac{1}{b}\theta$

$\Rightarrow$ $\boxed{\begin{aligned} H : \; & g(x) \Rightarrow \hat{w}^T x - \hat{\theta} = 0 \\ H_+ : \; & g_+(x) \Rightarrow \hat{w}^T x - \hat{\theta} = +1 \\ H_- : \; & g_-(x) = \hat{w}^T x - \hat{\theta} = -1 \end{aligned}}$

- What is the distance δ to C_1 & C_2 ?

- Let x be on $H \Rightarrow \boxed{w^T x - \theta = 0}$



$\vec{w}$: normal to H

$\Rightarrow \quad y = x + \delta w$ should be on H_+

$\Rightarrow \quad w^T y - \theta = 1 \Rightarrow \boxed{w^T x + \delta w^T w - \theta = 1}$

But $x \in H \Rightarrow \boxed{w^T x - \theta = 0} \Rightarrow \delta w^T w = 1 \Rightarrow \boxed{\delta = \frac{1}{\|w\|^2}}$

FACT: Distance δ is inversely proportional to $\|w\|$.

AIM: Maximize the distance $\delta \Rightarrow$ minimize $\|w\|$

PROBLEM 1 :

$$\text{minimize} \quad \frac{1}{2}\|w\|^2$$

$$\text{subject to} \quad w^T x_i - \theta \geq 1 \quad \text{if } x_i \in C_1 \ (d_i = +1)$$

$$w^T x_j - \theta \leq -1 \quad \text{if } x_j \in C_2 \ (d_j = -1)$$

$\Rightarrow$ QUADRATIC OPTIMIZATION PROBLEM with linear constraints.

$\Rightarrow$ CONVEX $\Rightarrow$ has a GLOBAL MIN. SOLUTION (UNIQUE!)

- SECOND FORM :

P2 :

$$\text{minimize} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad d_i(w^T x_i - \theta) \geq 1 \qquad i = 1, 2, \ldots, N$$

- Various solvers solve this kind of problems.

- One analytic method : LAGRANGE MULTIPLIERS

Solution of P2 is an EXTREMUM POINT of

P3

$$L(w, \theta, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{N} \alpha_i \left[d_i (w^T x_i - \theta) - 1 \right]$$

Here, $\alpha_i \geq 0$ are called LAGRANGE MULTIPLIERS

NOTE: Actually, we should use min. wrt w, θ and max wrt α_i!

P4 :

$$\min_{w,\theta} \max_{\alpha} L(w, \theta, \alpha)$$

Why? Otherwise choose $\alpha_i = \infty \Rightarrow$ get $-\infty$, which is minimum.

SOLUTION: Karush-Kuhn-Tucker (KKT) conditions.

$$\frac{\partial L}{\partial w} = 0 \Rightarrow \boxed{w - \sum_{i=1}^{N} \alpha_i d_i x_i = 0} \quad (1)$$

$$\frac{\partial L}{\partial \theta} = 0 \Rightarrow \boxed{\sum_{i=1}^{N} \alpha_i d_i = 0} \quad (2)$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \Rightarrow \boxed{d_i (w^T x_i - \theta) - 1 = 0} \quad (3)$$

+ Remember: $\boxed{\alpha_i \geq 0}$ (4)

(1) $\Rightarrow$ $\boxed{w = \sum_{i=1}^{N} \alpha_i d_i x_i}$ (**)

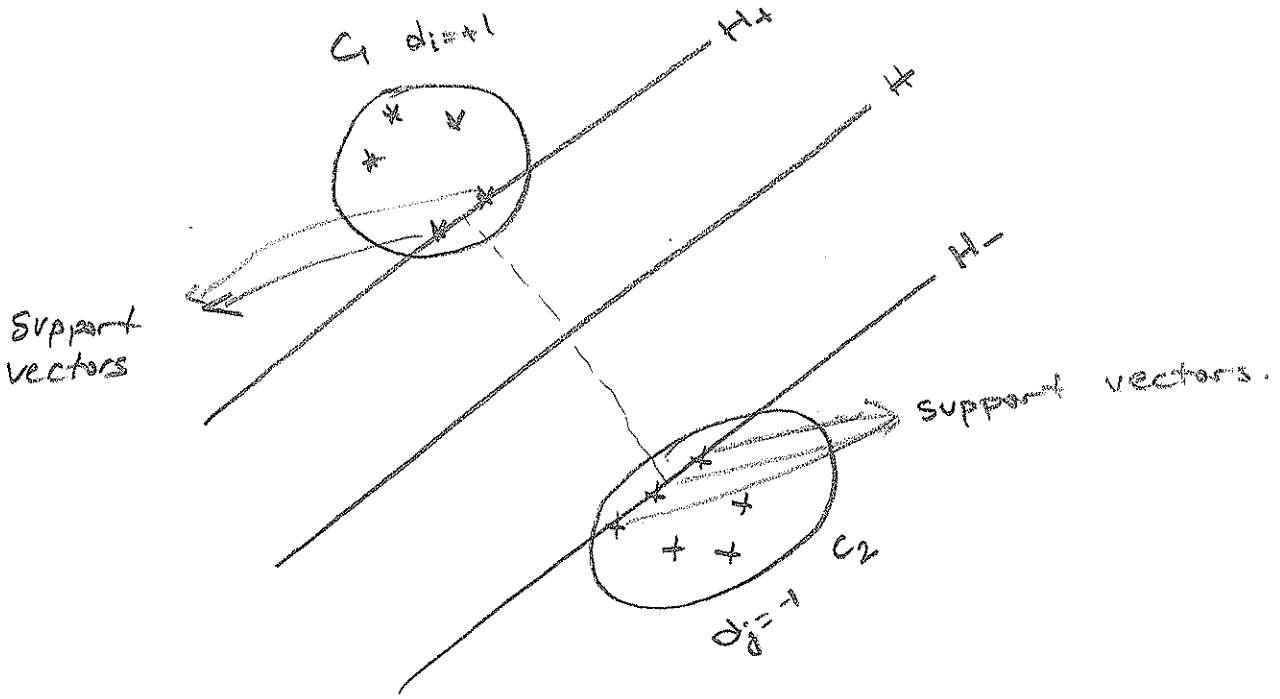
FACT: w is a linear combination of some training patterns x_i? Which ones?

(3) $\Rightarrow$ $d_i (w^T x_i - \theta) - 1 = 0$

$d_i = +1 \Rightarrow w^T x_i - \theta = 1 \Rightarrow x_i$ is ON H_+

$d_i = -1 \Rightarrow w^T x_i - \theta = -1 \Rightarrow x_i$ is ON H_-

$\Rightarrow$ The only relevant x_i's on (**) are the patterns which are on H_+ and H_-. They are called SUPPORT VECTORS



CONCLUSION: $\alpha_i > 0$ for support vectors, and $\alpha_i = 0$ for all other patterns.

CONCLUSION: Optimal weight w is determined by only SUPPORT VECTORS. The threshold term θ can be found easily:

if x_i is a support vector

$$\Rightarrow \quad d_i [w^T x_i - \theta] - 1 = 0 \quad \Rightarrow \quad \boxed{w^T x_i - \theta = 1/d_i}$$

But $d_i = \pm 1 \Rightarrow 1/d_i = d_i \quad \Rightarrow \quad w^T x_i - \theta = d_i$

$$\Rightarrow \quad \boxed{\theta = w^T x_i - d_i}$$

Sometimes an average over the support vectors is used to find θ for ROBUSTNESS, i.e. let N_s denote the # of support vectors

$$\boxed{\theta_{ave} = \frac{1}{N_s} \sum_{i=1}^{N_s} (w^T x_i - d_i) \qquad x_i \text{ is SV}}$$

- DUAL FORM OF OPTIMIZATION

- Since $\omega = \sum_{i=1}^{N} \alpha_i d_i x_i$, $\sum_{i=1}^{N} \alpha_i d_i = 0$

$\Rightarrow$ use it on $L(\omega, \theta, \alpha)$

$$\Rightarrow L(\omega,\theta,\alpha) = \frac{1}{2}\omega^T\omega - \sum_{i=1}^{N} \alpha_i \left[d_i(\omega^T x_i - \theta) - 1 \right]$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j - \sum_{i=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j + \theta \sum_{i=1}^{N} \alpha_i d_i + \sum_{i=1}^{N} \alpha_i$$

$$\Rightarrow \boxed{L(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i=1}^{N} \alpha_i \alpha_j d_i d_j x_i^T x_j}$$

DUAL PROBLEM:

$$\text{D.P.} \quad \boxed{\begin{array}{ll} \max\limits_{\alpha} & L(\alpha) \\ \text{s.t} & \alpha_i \geq 0 \end{array}}$$

$\Rightarrow$ much simpler in form $\Rightarrow$ easy to solve.

$\Rightarrow$ various software packages.

NOTE: In the solution of D.P., some $\alpha_i = 0$, some $\alpha_j > 0$.

The ones $\alpha_j > 0$ corresponds to SUPPORT VECTORS.

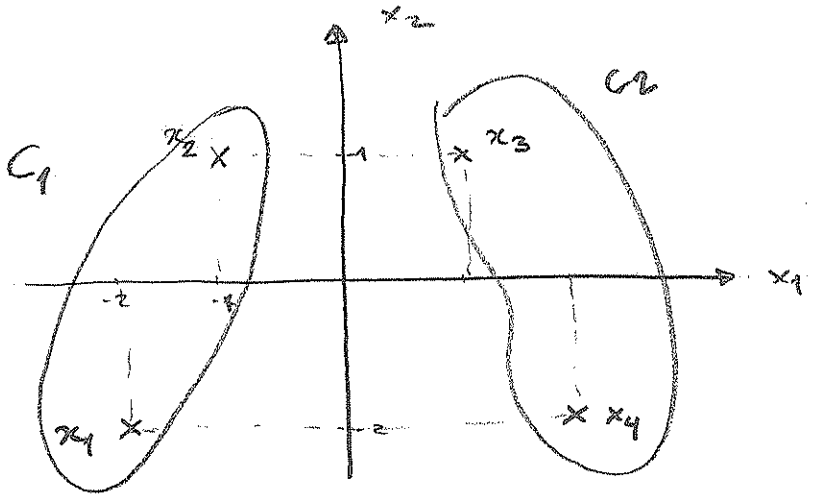
$\Rightarrow$ $\boxed{\omega = \sum_{j=1}^{N_s} \alpha_j d_j x_j}$ x_j support vector; N_s : # of support vectors.

θ can be found as before.

– AN EXAMPLE

$$C_1 = \left\{ x_1 = \begin{pmatrix} -2 \\ -2 \end{pmatrix}, \; x_2 = \begin{pmatrix} -1 \\ +1 \end{pmatrix} \right\} \quad d_1 = d_2 = +1$$

$$C_2 = \left\{ x_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \; x_4 = \begin{pmatrix} 2 \\ -2 \end{pmatrix} \right\} \quad d_3 = d_4 = -1$$



ANALYTIC SOLUTION FOR SVM

$$\text{Dual Prob} \Rightarrow L(\alpha) = \sum_{i=1}^{4} \alpha_i - \frac{1}{2} \sum_{i=1}^{4} \sum_{j=1}^{4} \alpha_i \alpha_j d_i d_j x_i^T x_j$$

$$\Rightarrow L(\alpha) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - 4\alpha_1^2 - \alpha_2^2 - \alpha_3^2 - 4\alpha_4^2 - 4\alpha_1\alpha_3 - 4\alpha_2\alpha_4$$

PROBLEM:

$$\max \; L(\alpha)$$
$$\text{s.t.} \quad \alpha_i \geq 0$$

— Should use quadratic optimization packages

— Or use K-T-H conditions use Linear Programming

— For This special case :

$$L(\alpha) = \underbrace{[\alpha_1 + \alpha_3 - 4\alpha_1^2 - \alpha_3^2 - 4\alpha_1\alpha_3]}_{L_1} + \underbrace{[\alpha_2 + \alpha_4 - \alpha_2^2 - 4\alpha_4^2 - 4\alpha_2\alpha_4]}_{L_2}$$

$L_1(\alpha_1, \alpha_3) \Rightarrow$ max. separately

$L_2(\alpha_2, \alpha_4) \Rightarrow$ max. separately

$$\frac{\partial L_1}{\partial \alpha_1} = 1 - 8\alpha_1 - 4\alpha_3 = 0 \quad \Rightarrow \quad 2\alpha_1 + \alpha_3 = 1/4$$

$$\frac{\partial L_1}{\partial \alpha_3} = 1 - 2\alpha_3 - 4\alpha_1 = 0 \quad \Rightarrow \quad 2\alpha_1 + \alpha_3 = 1/2$$

} conflict!

Choose first $\Rightarrow \quad \alpha_3 = \frac{1}{4} - 2\alpha_1$

$\Rightarrow$ put in $L_1 \rightarrow \quad L_1 = \alpha_1 + \left(\frac{1}{4} - 2\alpha_1\right) - 4\alpha_1^2 - \left(\frac{1}{4} - 2\alpha_1\right)^2 - 4\alpha_1\left(\frac{1}{4} - 2\alpha_1\right)$

$$= \frac{3}{16} - \alpha_1$$

For L_1 max $\Rightarrow \alpha_1 = 0 \Rightarrow \alpha_3 = \frac{1}{4} \Rightarrow L_1 = 3/16$

Choose second $\Rightarrow \quad \alpha_3 = \frac{1}{2} - 2\alpha_1$

$\Rightarrow$ put in $L_1 \quad L_1 = \alpha_1 + \left(\frac{1}{2} - 2\alpha_1\right) - 4\alpha_1^2 - \left(\frac{1}{2} - 2\alpha_1\right)^2 - 4\alpha_1\left(\frac{1}{2} - 2\alpha_1\right)$

$$= \frac{1}{4} - \alpha_1$$

For L_1 max $\Rightarrow \alpha_1 = 0 \Rightarrow \alpha_3 = 1/2 \Rightarrow \boxed{L_1 = \frac{1}{4}}$ ***

Second one is bigger $\Rightarrow$ $\boxed{\begin{array}{l}\alpha_1 = 0 \\ \alpha_3 = 1/2\end{array}}$

For the $L_2 \rightarrow$ by symmetry $\alpha_1 \Leftrightarrow \alpha_4$, $\alpha_2 \Leftrightarrow \alpha_3$

$$\Rightarrow \boxed{\alpha_4 = 0, \quad \alpha_2 = 1/2}$$

RESULT 1: $\boxed{x_2 \text{ and } x_3 \text{ are SUPPORT VECTORS!}}$

RESULT 2:

$$w = \alpha_2 d_2 x_2 + \alpha_3 d_3 x_3$$

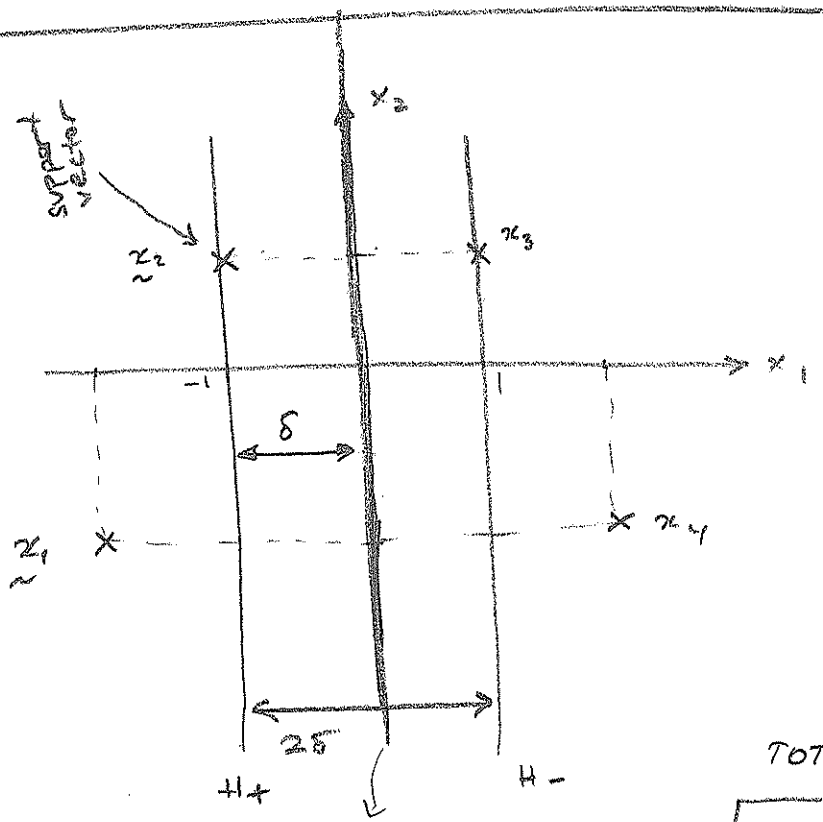
$$= \frac{1}{2}(x_2 - x_3) = \frac{1}{2}\left(\begin{bmatrix}-1\\1\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix}-2\\0\end{bmatrix} = \begin{bmatrix}-1\\0\end{bmatrix}$$

$$\theta = w^T x_2 - d_2 = w^T x_3 - d_3 = 0$$

$$\Rightarrow \boxed{g(x) = \underset{\sim}{w}^T \underset{\sim}{x} - \theta = -x_1 = 0 \Rightarrow \text{VERTICAL LINE}}$$

$$\boxed{H_+: \quad w^T x - \theta = 1 \Rightarrow -x_1 = 1 \Rightarrow \boxed{x_1 = -1}}$$

$$\boxed{H_-: \quad w^T x - \theta = -1 \Rightarrow -x_1 = -1 \Rightarrow \boxed{x_1 = 1}}$$



$$\boxed{\delta = \frac{1}{\|w\|} = \frac{1}{1} = 1}$$

TOTAL DISTANCE

$$\boxed{\delta_T = 2\delta = \frac{2}{\|w\|} = 2}$$