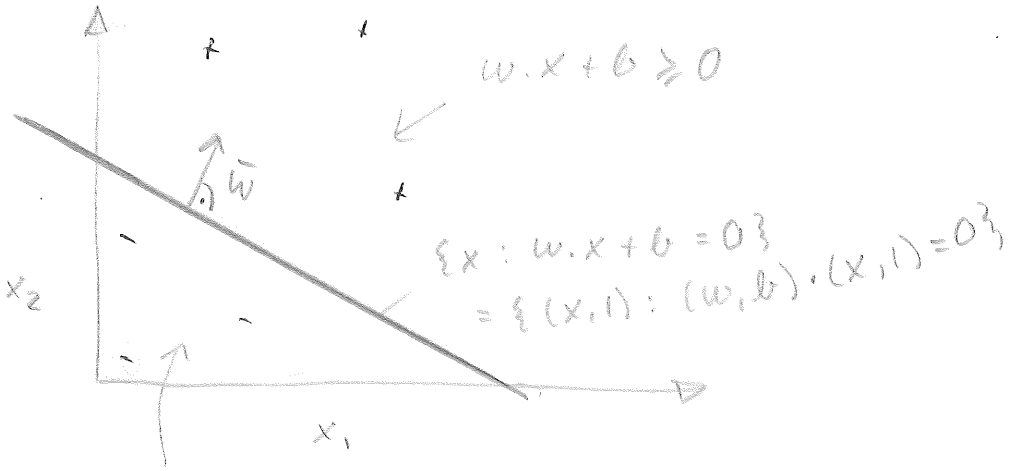
SUPPORT VECTOR MACHINES

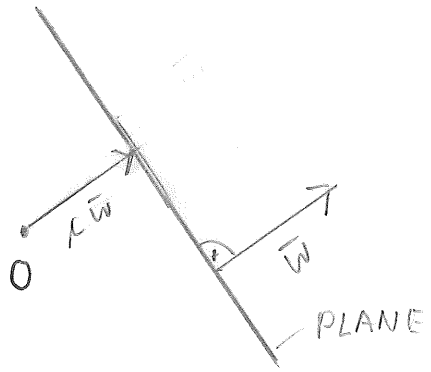
SEPARATE DATA INTO 2 CLASSES BASED ON A HYPERPLANE



IF $b \geq 0$ THEN ORIGIN ON $\geq$ SIDE

WHAT IS DISTANCE TO ORIGIN ?

INTERSECTION BETWEEN LINE $c\bar{w}$ & PLANE :



$$w \cdot (cw) + b = 0$$

$$c = \frac{-b}{\|w\|^2}$$

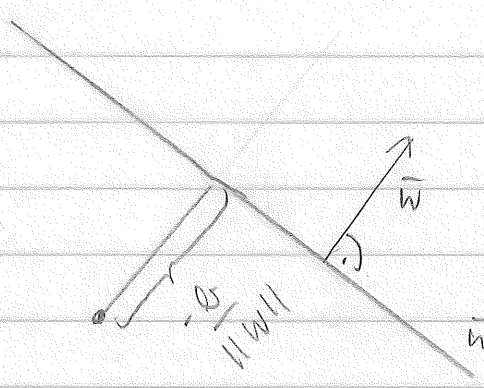
$$\|cw\| = |c| \, \|w\|$$

$$= \frac{|b|}{\|w\|}$$

DISTANCE

THE SIGNED DISTANCE OF ORIGIN TO $\{x : w \cdot x + b = 0\}$ IS $\frac{-b}{\|w\|}$.

DISTANCE IS 0 IFF $b = 0$



NOTE THAT $\bar{w} \cdot \bar{x} + b = 0$

$\iff q\,\bar{w} \cdot \bar{x} + q\,b = 0$ FOR SCALAR $q \neq 0$

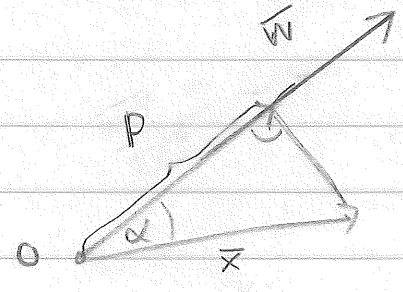
↑ SAME PLANE

$\iff \frac{\bar{w}}{\|\bar{w}\|} \cdot \bar{x} + \frac{b}{\|\bar{w}\|} = 0$

$\frac{\bar{w}}{\|\bar{w}\|}$ ↳ UNIT

DOT PRODUCT: $\bar{w} \cdot \bar{x} = \|w\| \|x\| \cos(\alpha)$

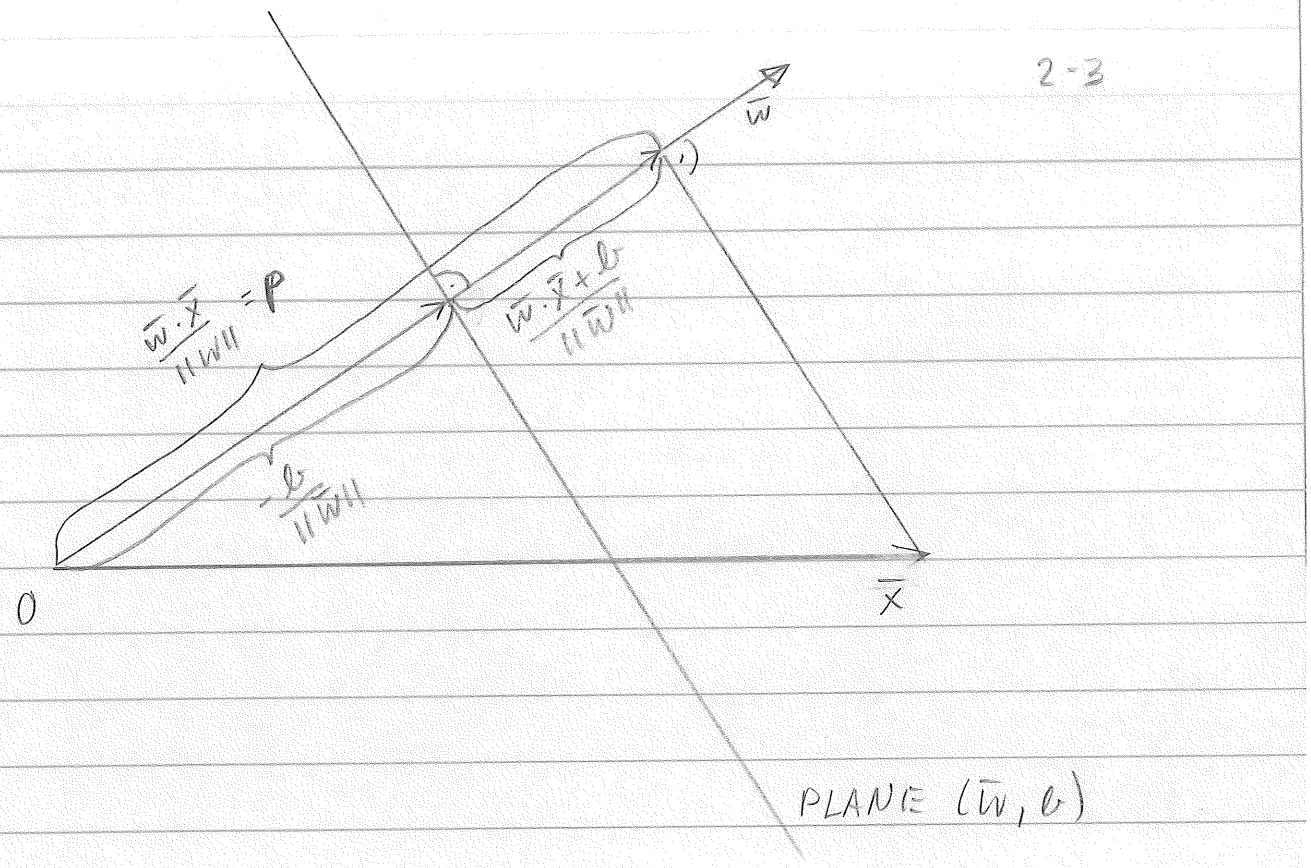
p IS PROJECTION OF x ONTO w



$\cos(\alpha) = \frac{p}{\|\bar{x}\|}$

$p = \|x\| \cos(\alpha)$

$p = \frac{\bar{w} \cdot \bar{x}}{\|w\|}$



$\frac{|\bar{w}\cdot\bar{x} + b|}{\|\bar{w}\|}$ DISTANCE OF $\bar{x}$ TO PLANE

$\frac{\bar{w}\cdot\bar{x} + b}{\|\bar{w}\|}$ SIGNED DISTANCE OF PLANE TO $\bar{x}$

WANT $w\cdot x_i + b > 0$ IF $y_i = +1$

< 0 IF -1

WANT $\underbrace{y_i (w x_i + b)}_{\gamma_i} > 0$

FUNCTIONAL MARGIN OF EXAMPLE $(\vec{x}_i, y_i)$
WRT $(\vec{w}, b)$ IS $\gamma_i = y_i(\vec{w} \cdot \vec{x}_i + b)$

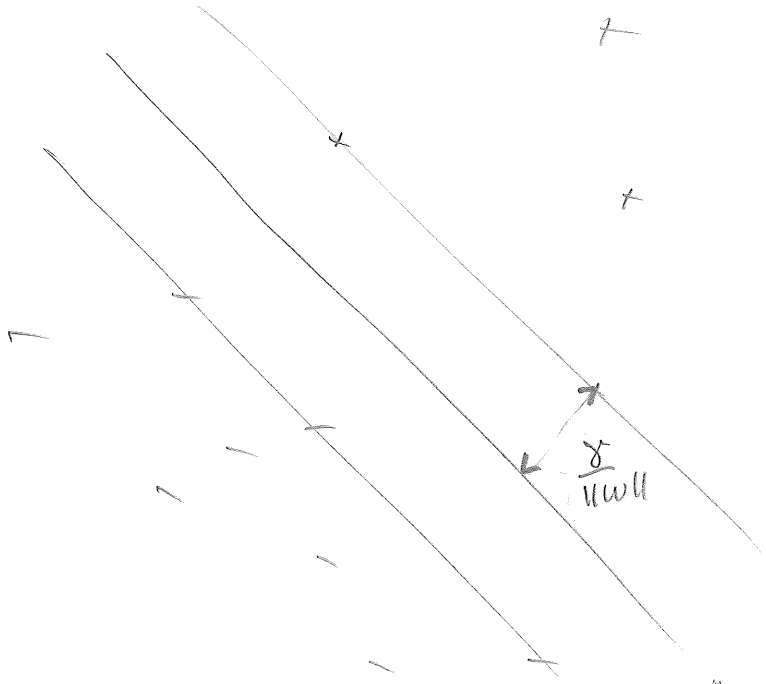
GEOMETRIC MARGIN

... $\frac{\gamma_i}{\|w\|}$

GEOMETRIC MARGINS ARE SCALED DISTANCES

IF $\|w\| = 1$ $\Rightarrow$ NO NORMALIZATION NECC.

WANT PLANE S.T. ALL EXAMPLES HAVE POSITIVE MARGIN



BETTER GOAL:

WANT

MAXIMUM
MARGIN HYPER PLANE

PLANES (w, b) & $(\lambda w, \lambda b)$ FOR $\lambda > 0$ ARE THE SAME

GEOMETRIC MARGIN OF x_i INVARIANT TO SCALING

$$\frac{w}{\|w\|} x_i + \frac{b}{\|w\|}$$

HOW TO DEFINE MAX. MARGIN HYPERPLANE

$\max_{w,b,\gamma} \gamma$ ← FUNCTIONAL

SUBJECT TO $y_i (w \cdot x_i + b) \geq \gamma$

$1 \leq i \leq \ell$

WRONG MARGIN

$= \infty$

$\max_{w,b,\gamma} \gamma$ ← GEOMETRIC

SUBJECT TO $y_i \left(\frac{w}{\|w\|} \cdot x_i + \frac{b}{\|w\|} \right) \geq \gamma$ (*)

$1 \leq i \leq \ell$

GOOD DEF. BUT HARD TO OPTIMIZE

Hyperplanes with functional margin 1
known as canonical hyperplanes

Find w, b

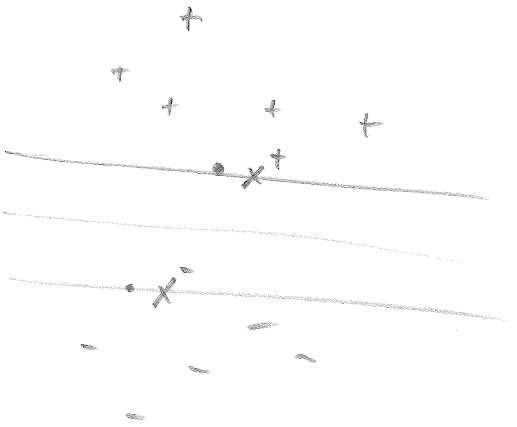
s.t. $\quad y_i (w \cdot x_i + b) \geq 1$

$1 \leq i \leq \ell$

Choose x^+, x^-

s.t. $\quad w \cdot x^+ + b = 1$

$w \cdot x^- + b = -1$



Geometric margin of x^+

$$\frac{w}{\|w\|^2} \cdot x^+ + \frac{b}{\|w\|^2} = \frac{1}{\|w\|^2} \underbrace{(x^+ + b)}_{+1}$$

Geometric margin of x^-

$$\frac{w}{\|w\|^2} x^- + \frac{b}{\|w\|^2} = \frac{1}{\|w\|^2} \underbrace{(x^- + b)}_{-1}$$

EQUIVALENT OPTIMIZATION PROBLEMS

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

MAX MARGIN HYPERPLANE

$$\min \|w\|^2 \quad \text{s.t. } y_i (w x_i + b) \geq 1$$

REALIZES MAX MARGIN HYPERPLANE W. MARGIN $\frac{1}{\|w\|^2}$

DUAL OPTIMIZATION PROBLEM?

$$L(w, b, \alpha) = \frac{1}{2}(\bar{w} \cdot \bar{w}) + \sum_{i=1}^{\ell} \alpha_i \left[y_i \left((\bar{w} \cdot \bar{x}_i + b) - 1 \right) \right]$$

$$\frac{\partial L}{\partial w} = w + \sum_{i=1}^{\ell} \alpha_i y_i \bar{x}_i = 0$$

$$w = -\sum_i y_i \alpha_i \bar{x}_i$$

LINEAR COMB. OF EXAMPLES

$$\frac{\partial L}{\partial b} = \sum_i y_i \alpha_i = 0$$

PLUG IN:

$$L = \frac{1}{2}\left(\sum_i y_i \alpha_i x_i\right) \cdot \left(\sum_j y_j \alpha_j x_j\right) -$$

$$+ \sum_i \alpha_i y_i \left(\sum_j \alpha_j y_j x_j \cdot x_i\right) - \underbrace{\sum_i \alpha_i y_i b}_{0} + \sum_i \alpha_i$$

$$= -\frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j x_i \cdot x_j + \sum_i \alpha_i$$

KEY: - ONLY DOT PRODUCTS MATTER

- KERNELIZABLE !!!

- KEY OBSERVATION OF ORIGINAL SUM PAPER

DUAL PROBLEM

MAXIMISE $$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j x_i \cdot x_j$$

SUBJECT TO $$\sum_i y_i \alpha_i = 0$$

$$\alpha_i \geq 0$$

IF α^* IS SOLUTION TO DUAL

THEN $$w^* = \sum_i y_i \alpha_i^* x_i$$

$$b^* = -\frac{\max_{y_i=-1} w^* \cdot x_i + \min_{y_i=+1} w^* \cdot x_i}{2}$$

GEOMETRIC MARGIN

$$\frac{1}{\|w^*\|_2}$$

KURESH KUHN-TUCKER CONDITION

$$\alpha_i^* \left[y_i \left(w^* \cdot x_i + b^* \right) - 1 \right] = 0$$

$\alpha_i^* > 0$ IFF MARGIN $= 1$

$y_i (w^* \cdot x_i + b^*) - 1 = 0$ IFF MARGIN $= 1$

> 0 IF " > 1

SUPPORT VECTORS :

ALL x_i FOR WHICH $\alpha_i > 0$

REPRESENTING MAX. MARGIN HYPERPLANE IN DUAL

$$w^* \cdot x + b^* = \underbrace{\sum_{i=1}^{\ell} y_i \alpha_i^* x_i}_{w^*} \cdot x + b^*$$

$$= \sum_{i \in SV} y_i \alpha_i^* x_i \cdot x + b^*$$

FOR $j \in SV$: $w^* \cdot x_j + b^* = y_j$ (*)

ALSO : $w^* \cdot w^* = \sum_{i,j} y_i y_j \alpha_i^* \alpha_j^* x_i \cdot x_j$

$$= \sum_{j \in SV} \alpha_j^* y_j \underbrace{\sum_{i \in SV} y_i \alpha_i^* x_i}_{w^*} \cdot x_j$$

(*)

$$= \sum_{j \in SV} \alpha_j^* y_j (y_j - b^*)$$

$y_j \in \{+1, -1\}$

$$= \sum_{j \in SV} \alpha_j^* (1 - y_j b^*)$$

$$= \sum_{j \in SV} \underbrace{\alpha_j^*}_{>0} - \underbrace{\sum_{j \in SV} \alpha_j^* y_j b^*}_{0}$$

MARGIN : $\frac{1}{\|w^*\|_2} = \frac{1}{\sqrt{\sum_j \alpha_j^*}}$

KERNEL TRICK

x_i EXPANDED TO $\phi(x_i)$

$$\phi(x_i) \cdot \phi(x_j) = k(x_i, x_j)$$

$$w^* \cdot \phi(x) + b^* = \sum_{i \in SV} y_i \alpha_i^* \phi(x_i) \cdot \phi(x) + b^*$$

↑ ONE WEIGHT PER FEATURE

$$= \sum_{i \in SV} y_i \alpha_i^* k(x_i, x) + b^*$$

↑ ONE WEIGHT PER SV

$|SV|$ USUALLY $\ll$ DIMENSION OF FEATURE SPACE

GENERALIZATION ERROR GOOD IF

MARGIN LARGE

OR $|SV|$ SMALL

NON-SEPERABLE CASE

MINIMIZE $_{w, \xi, b}$ $w \cdot w + C \sum_i \xi_i$

SUBJECT TO $y_i (w_i \cdot x_i + b) \geq 1 - \xi_i$

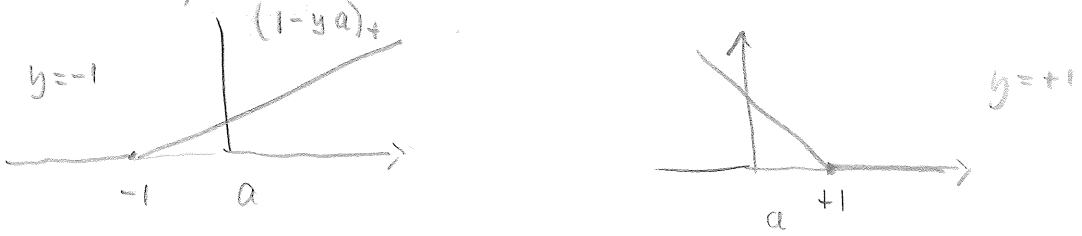
$\xi_i \geq 0$

ONE SLACK VARIABLE PER EXAMPLE

COST IS LINEAR IN SLACK

ALTERNATE DEF.

MINIMIZE $_{w, \xi}$ $w \cdot w + C \sum_i (1 - y_i \overbrace{(w_i \cdot x_i + b)}^{a_i})_+$



(LINEAR) HINGE LOSS

DUAL

MAXIMIZE: $\sum \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j x_i \cdot x_j$

SUBJECT TO: $\sum y_i \alpha_i = 0$

$\alpha_i \in [0, C]$

↑

NEW

BEFORE $C = \infty$

QUADRADIC HINGE

MINIMIZE: $w \cdot w + C \sum_i \xi_i^2$

SUBJ. TO $y_i (w_i \cdot x_i + b) \geq 1 - \xi_i$

$\xi_i \geq 0$

ALTERNATE FORMULATION:

MINIMIZE $w \cdot w + C \sum_i (1 - y_i (w_i \cdot x_i + b))_+^2$

DUAL

MAXIMIZE $\sum \alpha_i - \frac{1}{2} \sum_{ij} y_i y_j \alpha_i \alpha_j x_i \cdot x_j$

$- \frac{1}{2C} \sum_i \alpha_i^2$

SUBJ TO. $\sum_i y_i \alpha_i = 0$

$\alpha_i \geq 0$

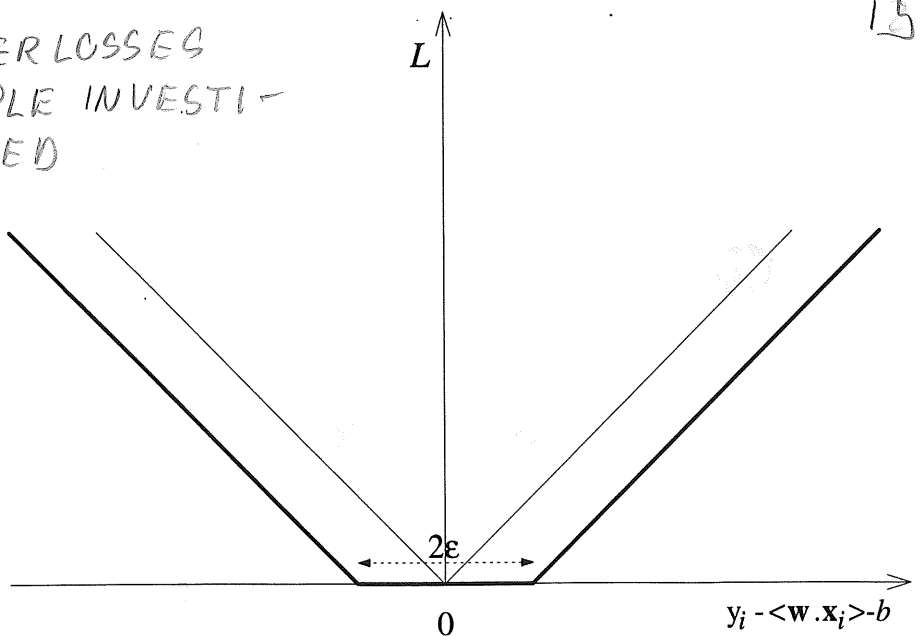


Figure 6.6: The linear ε-insensitive loss for zero and non-zero ε

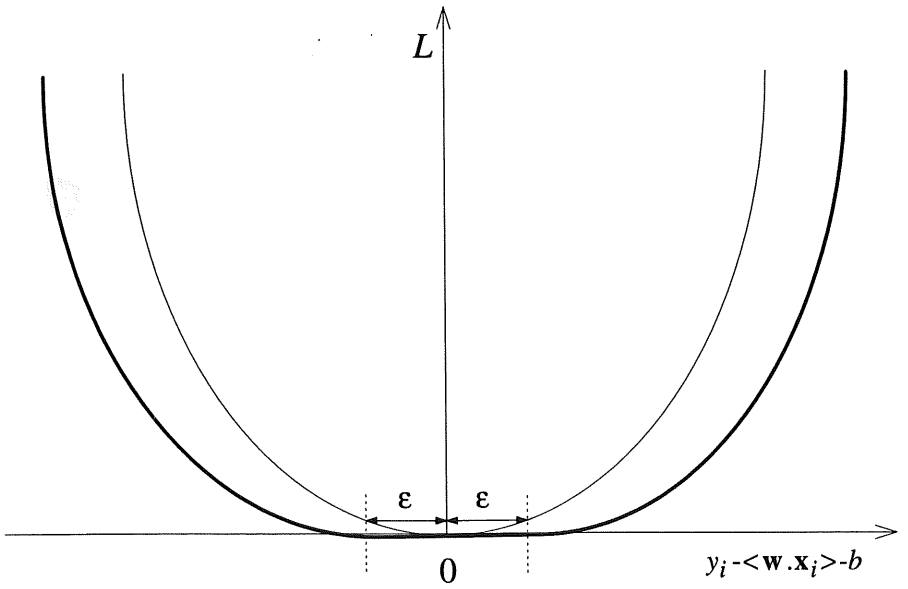


Figure 6.7: The quadratic ε-insensitive loss for zero and non-zero ε

KERNEL METHODS

WHEN APPLICABLE ?

HYPOTHSIS MUST BE DEFINED I.T.O. DOT PRODUCT

$$\hat{y} = h_w(x) = f(w \cdot x)$$

$\hat{y}$: PREDICTION ON NEW x

WHERE w IS LINEAR COMBINATION OF INSTANCES IN TRAINING SET

$$\text{I.E.} : \quad w = \sum_{q=1}^{t} \alpha_q x_q$$

↑ CALLED LINEAR FORM

EXAMPLES :

- WIDROW HOFF
- PERCEPTRON
- BACKPROP
- SUPPORT VECTOR MACHINES

} GD FAMILY

W IN LINEAR FORM WHEN
- PARAMETER DIVERGENCE IS $\|\ \|_2^2$
- LOSS DEPENDS ON DOT PRODUCT

$$w_{t+1} = \underset{w}{\text{ARGINF}} \left(\|w - w_t\|^2 + \eta L_{y_t}(w \cdot x_t) \right)$$

$$\Rightarrow w_{t+1} = \sum_{q=1}^{t} \alpha_q^{t+1} x_q \qquad \text{LINEAR FORM}$$

GD FAMILY

MAIN OTHER FAMILY

$$w_{t+1} = \underset{w \geq 0}{\text{ARGINF}} \left(\sum w_i \ln \frac{w_i}{w_{t,i}} + w_{t,i} - w_i + \eta L_{y_t}(w \cdot x_t) \right)$$

$$\ln w_{t+1,i} = \sum_{q=1}^{t} \gamma_q^{t+1} x_{q,i}$$

$$w_{t+1,i} = e^{\,\cdot\,} \qquad \text{EXPONENTIAL FORM}$$

WINNOW, ...
EGU FAMILY

ASSUME PREDICTION $\hat{y}_t$ IS LINEAR ($\hat{y}_t = w_t \cdot x_t$)
OR LINEAR THRESHOLD FUNCTION ($\hat{y}_t = (w_t \cdot x_t > \theta)$)

LINEAR FORM GOOD BECAUSE
AFTER EXPANDING INSTANCES
STILL EFFICIENT

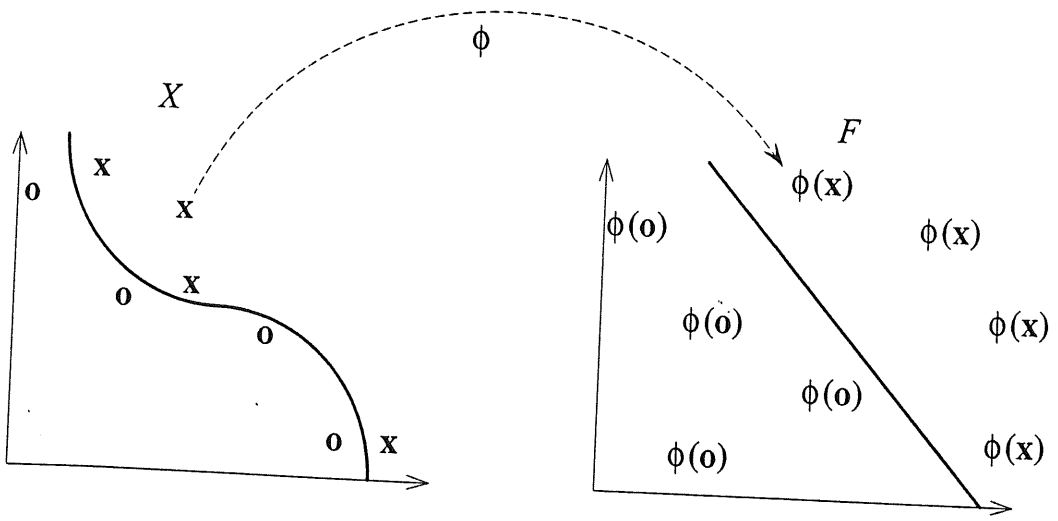


Figure 3.1: A feature map can simplify the classification task

$\bar{X} = (x_1, \ldots, x_n) \longrightarrow \phi(\bar{x}) = (\phi_1(x), \ldots, \phi_N(x))$

$x \in X$ — INPUT SPACE

$\{\phi(x) : x \in X\}$ — FEATURE SPACE

FOR EXAMPLE : $\phi(x_1, x_2) = (x_1^2, x_2^2, x_1 x_2)$

REASONS FOR EXPANSION:

- LINEAR MODELS IN FEATURE SPACE ARE CONVENIENT NON-LINEAR MODELS OF INPUT SPACE

- DIMENSIONALITY REDUCTION (I.E. $N < n$)

- $\left(\underbrace{\sum_{q=1}^{t-1} \alpha_q^t \phi(x_q)}_{\text{WEIGHT VECTOR } w_t \text{ IN LINEAR FORM}} \right) \cdot \phi(x_t)$

$$= \sum_{q=1}^{t-1} \alpha_q^t \phi(x_q) \cdot \phi(x_t)$$

COMPUTE DOT PRODUCTS WITH PAST EXAMPLES

OFTEN EFFICIENT COMPUTATION OF $\phi(x) \cdot \phi(z)$ EVEN WHEN $N \gg n$

x $\qquad\qquad$ $\phi(x)$

$(x_1, x_2, x_3) \longrightarrow (1, x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)$

8 FEATURES

Repeat!

$$\phi(x) \cdot \phi(z) = 1 + x_1z_1 + x_2z_2 + x_3z_3 +$$

$$+ x_1x_2\, z_1z_2 + x_1x_3\, z_1z_3 + x_2x_3\, z_2z_3$$

$$+ x_1x_2x_3\, z_1z_2z_3$$

$$= (1 + x_1z_1) \cdot (1 + x_2z_2) \cdot (1 + x_3z_3)$$

8 TERMS : $\quad 2 \quad \times \quad 2 \quad \times \quad 2$

x $\qquad$ $\phi(x)$

$(x_1, x_2, \ldots, x_n)$ $\qquad$ $(1, x_i, x_i x_j, x_i x_j x_k, \ldots)$

$i \neq$ $\qquad$ all $\neq$

2^n MONOMIALS

$$\phi(x) \cdot \phi(z) = \sum_{I \subseteq \{1,\ldots,n\}} \prod_{i \in I} x_i \prod_{i \in I} z_i$$

$$= \prod_{i=1}^{n} (1 + x_i z_i)$$

$O(n)$ TIME

EFFICIENCY:

$$\underbrace{\left(\sum_{q=1}^{t-1} \alpha_q^t \phi(q) \right)}_{w_t} \cdot \phi(x_t)$$

DIMENSION 2^n

SEEMINGLY TIME $O(2^n)$

$$= \underbrace{\sum_{q=1}^{t-1} \alpha_q^t \underbrace{\phi(q) \cdot \phi(x_t)}_{\text{TIME } O(n)}}_{\text{TIME } O(tn)}$$

MORE EXAMPLES

$$(x_1, \dots, x_n) \longrightarrow \left(\quad x_i x_j \quad \right)_{\substack{1 \le i \le n \\ 1 \le j \le n}}$$

WITH REPEATS

$\bar{x}$ $\qquad\qquad$ $\phi(\bar{x})$

$$\phi(\bar{x}) \cdot \phi(\bar{z}) = \sum_{(i,j)=(1,1)}^{(n,n)} (x_i x_j)(z_i z_j)$$

$$= \left(\sum_{i=1}^{n} x_i z_i \right) \left(\sum_{j=1}^{n} x_j z_j \right)$$

$$= \left(\sum_{i=1}^{n} x_i z_i \right)^2$$

$$= (x \cdot z)^2$$

IMPLICIT MAPPING INTO FEATURE SPACE

A KERNEL IS A FUNCTION k S.T. FOR ALL x, z

$$k(x,z) = \phi(x) \cdot \phi(z)$$

implicit

WHERE ϕ IS A MAPPING FROM THE INPUT SPACE X TO A FEATURE SPACE F WHICH HAS AN INNER PRODUCT

EXAMPLES SO FAR

$$k(x,z) = \prod_{i=1}^{n} (1 + x_i z_i)$$

$$k(x,z) = (x \cdot z)^2$$

ALSO:

$$k(x,z) = (x \cdot z + c)^d$$

t training examples

$w = \sum_{q}^{t} \alpha_q x_q$

hypothesis requires t α's

$w \cdot x = t$ kernal computations

(x: new)

PROPERTIES OF KERNAL FNCTS:

$$k(x,z) = \phi(x) \cdot \phi(z)$$
$$= \phi(z) \cdot \phi(x)$$
$$= k(z,x)$$

SYMMETRY

$$k(x,z)^2 = (\phi(x) \cdot \phi(z))^2$$
$$\leq \|\phi(x)\|^2 \, \|\phi(z)\|^2$$
$$= (\phi(x) \cdot \phi(x))(\phi(z) \cdot \phi(z))$$
$$= k(x,x)\, k(z,z)$$

CAUCHY-SCHWARTZ

CHARECTERIZATION OF KERNALS

MERCER'S TH (FINITE CASE) :

LET $X = \{x_1, \ldots, x_n\}$ BE A FINITE INPUT SPACE. THEN $K(x, z)$ IS A KERNEL FUNCTION IFF

$$K = \left(k(x_i, x_j) \right)_{i,j=1}^{n}$$

IS SYMMETRIC POSITIVE DEFINITE (HAS NON-NEG. EIGEN VALUES).

THERE ARE CONTINUOUS VERSIONS OF THIS THEOREM

WHAT ARE GOOD KERNELS FOR A GIVEN PROBLEM

Special kernal functions

$$k(x,z) = k(x-z)$$

are translation invariant

Example: $k(u) = \sum_{n=1}^{\infty} a_n \cos(nu)$

$$k(x-z) = a_0 + \sum_{n=1}^{\infty} a_n \sin nx \sin nz + \sum_{n=1}^{\infty} a_n \cos(nx)\cos(nz)$$

$$\left(\phi_i(x)\right)_{i=0}^{\infty} = (1, \sin(x), \cos(x), \sin(2x), \cos(2x), \ldots, \sin(nx), \cos(nx)$$

FOURIER COEFFICIENTS

orthogonal features

$$\int_0^{2\pi} \sin(x) \cdot \cos(3x) = 0$$

$$k(x,z) = e^{-(x-z)^2}$$

relating inv. ass Hilb

MAKING KERNELS FROM KERNELS

Proposition 3.12 *Let K_1 and K_2 be kernels over $X \times X$, $X \subseteq \mathbb{R}^n$, $a \in \mathbb{R}^+$, $f(\cdot)$ a real-valued function on X,*

$$\phi : X \longrightarrow \mathbb{R}^m$$

with K_3 a kernel over $\mathbb{R}^m \times \mathbb{R}^m$, and $\mathbf{K}$ a symmetric positive semi-definite $n \times n$ matrix. Then the following functions are kernels:

1. $K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z}) + K_2(\mathbf{x}, \mathbf{z})$,
2. $K(\mathbf{x}, \mathbf{z}) = aK_1(\mathbf{x}, \mathbf{z})$,
3. $K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$,
4. $K(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$,
5. $K(\mathbf{x}, \mathbf{z}) = K_3(\phi(\mathbf{x}), \phi(\mathbf{z}))$,
6. $K(\mathbf{x}, \mathbf{z}) = \mathbf{x}'\mathbf{B}\mathbf{z}$.

Corollary 3.13 Let $K_1(\mathbf{x}, \mathbf{z})$ be a kernel over $X \times X$, $\mathbf{x}, \mathbf{z} \in X$, and $p(x)$ a polynomial with positive coefficients. Then the following functions are also kernels:

1. $K(\mathbf{x}, \mathbf{z}) = p(K_1(\mathbf{x}, \mathbf{z}))$,
2. $K(\mathbf{x}, \mathbf{z}) = \exp(K(\mathbf{x}, \mathbf{z}))$,
3. $K(\mathbf{x}, \mathbf{z}) = \exp(-\|\mathbf{x} - \mathbf{z}\|^2 / \sigma^2)$.