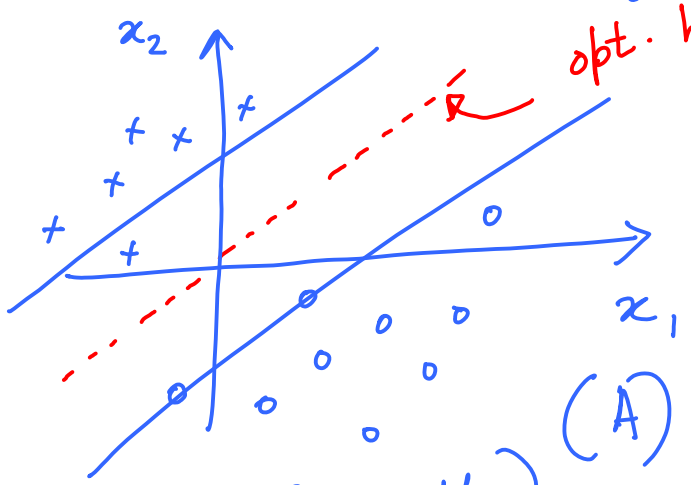
Optimal hyperplane for non separable patterns

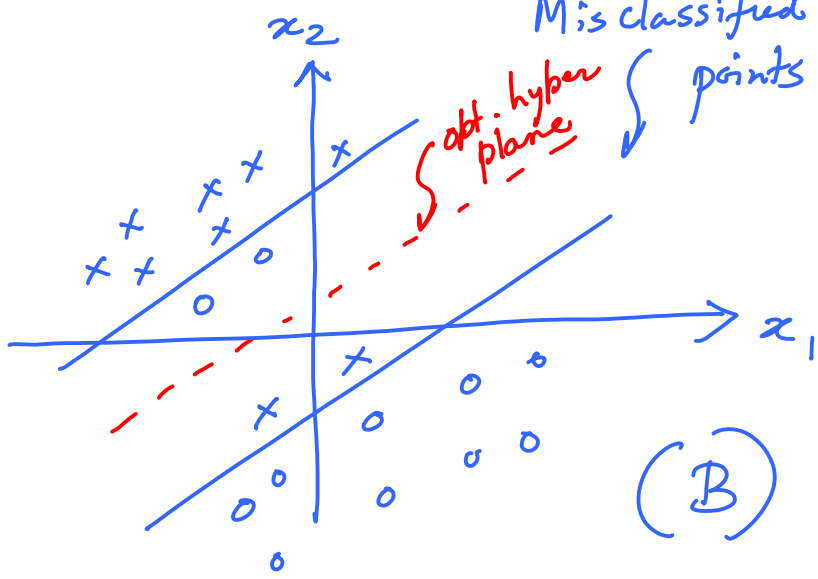
Consider the following cases



(A)

(Linearly Separable)

The points are on the correct side of the decision boundary



(B)

Some of the points are inside the region of separation but on the incorrect side

Let us introduce a new set of non negative scalar variables $\{\xi_i\}_{i=1}^{N}$ into the defn. hyperplane

$$d_i\left(\underline{w}^T \underline{x}_i + b\right) \geq 1 - \xi_i ; \quad i = 1, \ldots, N \qquad \text{(I)}$$

ξ_i : slack variables

(Gives us a measure of deviation from ideal conditions of pattern separability)

For $0 \leq \xi_i \leq 1$ ——— Case (A)

For $\xi_i > 1$, it corresponds to Case (B)

The support vectors are those that satisfy the equality (precisely) even if $\xi_i > 0$ in (I)

GOAL: Find a separating hyperplane for which the error is minimized when averaged over the training set

Formulate $\phi(\xi) = \sum_{i=1}^{N} I(\xi_i - 1)$

Indication function

$$I(\xi) = \begin{cases} 0 & \text{if } \xi \leq 0 \\ 1 & \text{else} \end{cases}$$

To make the problem tractable,

$$\phi(\xi) = \sum_{i=1}^{N} \xi_i$$

$$\phi(\underline{w}, \underline{\xi}) = \frac{1}{2} \underline{w}^T w + C \sum_{i=1}^{N} \xi_i$$

$\xi = (\xi_1 \dots \xi_N)$

objective function

trade off in the complexity of the machine & the # of non-separable patterns

Constraint

GOAL:

We need to optimize $\phi(\underline{w}, \underline{\xi})$ w.r.t $\underline{w}$ and $\{\xi_i\}_{i=1}^{N}$

Let us formulate the primal and dual problems

PRIMAL :

Given the training set $\{\underline{x}_i, d_i\}_{i=1}^N$

find the opt. values of $\underline{w}$ and b /

$$d_i(\underline{w}^T \underline{x}_i + b) \geq 1 - \xi_i \quad , \forall i = 1, \ldots, N$$

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

We need to choose the wt. vector $\underline{w}$ and slack variables $\{\xi_i\}_{i=1}^N$ that minimize the cost functional

$$\Phi(\underline{w}, \underline{\xi}) = \frac{1}{2} \underline{w}^T \underline{w} + C \sum_{i=1}^N \xi_i$$

↑ user specified param.

Dual Problem : Find Lagrange multipliers $\{\alpha_i\}_{i=1}^{N}$

that maximizes

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i d_i \alpha_j d_j \underline{x}_i^T \underline{x}_j$$

Subject to

(1) $\sum_{i=1}^{N} \alpha_i d_i = 0$

(2) $0 \leq \alpha_i \leq C$

$i = 1, \ldots, N$

Observe the change: In case of linear separability $\alpha_i \geq 0$

Note that neither slack variables ξ_i nor the Lagrange multipliers appear in the dual problem.

After proceeding with the opt. steps

N_s ← # support vectors

$$\underline{w}_{opt} = \sum_{i=1}^{N_s} \alpha_{opt,i}\, d_i\, \underline{x}_i$$

$$\alpha_i \left[d_i \left(\underline{w}^T \underline{x}_i + b \right) - \left(1 - \xi_i \right) \right] = 0 \qquad i = 1, \ldots, N$$

$$\mu_i \xi_i = 0 \; ; \qquad i = 1, \ldots, N$$

← Lagrange multipliers to ensure that slack variables ξ_i are non negative

At the saddle point, for the primal problem

$$\frac{\partial J(\cdot)}{\partial \xi_i} = 0$$

$$\Rightarrow \quad \alpha_i + \mu_i = C$$

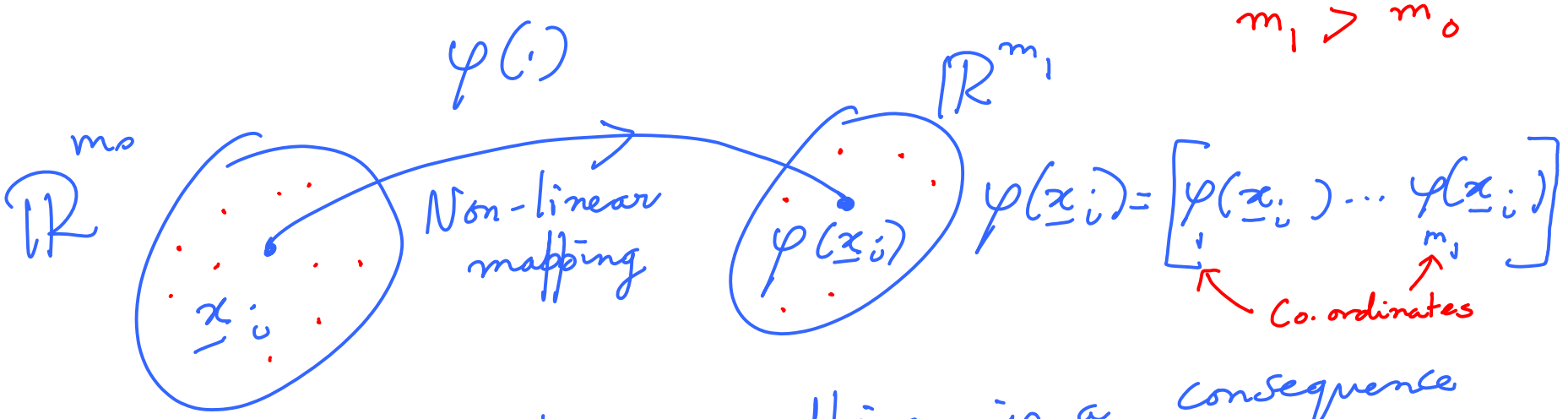
$$\therefore \quad \xi_i = 0 \quad \text{if} \quad \alpha_i < C$$

Building SVMs for pattern recognition

Idea behind SVMs hinges on

(1) Non-linear mapping of an input vector into a higher dimensional feature space hidden from both i/p and o/p.

(2) Construction of an optimal hyperplane in lieu of (1) above.



The notion of non-linear mapping is a consequence of Cover's Theorem that guarantees linear separability of patterns with high probability when input patterns are non-linearly transformed into a higher dim. features space

Inner Product Kernel

Let $\underline{x}$ denote a vector drawn from an input space, say of dim. m_0.

Let $\{\varphi_j(\underline{x})\}_{j=1}^{m_1}$ denote the set of non-linear transformations from the input to the feature space

m_1 : dim. of feature space

$\varphi_j(\underline{x})$ is defined a priori $\forall j$

$\varphi(\underline{x}) = [\varphi_1(\underline{x}) \ldots \varphi_j(\underline{x}) \ldots \varphi_{m_1}(\underline{x})]^T$

Given a set of non-linear transformations, we can construct a hyperplane acting as a decision surface as follows:

$$\sum_{j=1}^{m_1} w_j \varphi_j(\underline{x}) + b = 0$$

(bias)

where $\{w_j\}_{j=1}^{m_1}$ denotes the set of linear weights connecting the feature space to the o/p space along with bias 'b'.

Folding the bias

$$\sum_{j=0}^{m_1} w_j \varphi_j(\underline{x}) = 0 \qquad \underline{w}^T \underline{\varphi} = 0$$

so that $w_0 = b$; $\varphi_0(\underline{x}) = 1$

Let $\underline{\varphi}(\underline{x}) = \left[\varphi_0(\underline{x}) \;\; \varphi_1(\underline{x}) \;\; \cdots \;\; \varphi_{m_1}(\underline{x})\right]^T$ $\forall \underline{x}$

$\varphi_0(\underline{x}) = 1$

↖ vector in $\mathbb{R}^{m_1+1}$

In effect, the vector $\varphi(\underline{x})$ is the 'image' of $\underline{x}$ induced in the feature space.

$$\underline{w}^T \varphi(\underline{x}) = 0 \qquad (1)$$

Following Condition 1 in the derivative of Lagrangian w.r.t. $\underline{w}$ with the transformation $\varphi(\cdot)$ on $\underline{x}$

$$\underline{w} = \sum_{i=1}^{N} \alpha_i d_i \varphi(\underline{x}_i) \qquad (2)$$

↑ Notice the detail

$$\therefore \left[\sum_{i=1}^{N} \alpha_i d_i \varphi^T(\underline{x}_i) \right] \varphi(\underline{x}) = 0$$

(Using ① in ②) ——— 2(4)

Now $\underbrace{\varphi^T(\underline{x}_i)\,\varphi(\underline{x})}$ represents the inner product of 2 vectors induced in the feature space by i/p $\underline{x}$ and pattern $\underline{x}_i$ in the training set.

↙ $\langle \varphi(\underline{x}_i), \varphi(\underline{x}) \rangle$

Let us define the kernel $K(\cdot)$

$$K(\underline{x}, \underline{x}_i) \triangleq \varphi^T(\underline{x})\, \varphi(\underline{x}_i)$$

$$= \sum_{j=0}^{m_1} \varphi_j(\underline{x})\, \varphi_j(\underline{x}_i)$$

$\varphi_j(\underline{x})$: j-th coordinate of $\varphi(\underline{x})$

$\forall\, i = 1, \ldots, N$

This 'kernel' is symmetric in the arguments

i.e., $K(\underline{x}, \underline{x}_i) = K(\underline{x}_i, \underline{x}) \quad \forall\, i$

In terms of the kernel, the opt. hyper plane is now given by

$$\sum_{i=1}^{N} \alpha_i d_i K(\underline{x}, \underline{x}_i) = 0 \qquad (3)$$

Kernel

MERCER'S THEOREM (Mercer, 1908)

Let $K(x, x')$ be a continuous symmetric kernel defined over the interval $a \leq x \leq b$ and $a \leq x' \leq b$.

The kernel $K(\cdot)$ can be expanded as

$$K(x, x') = \sum_{i=0}^{\infty} \lambda_i \varphi_i(x) \varphi_i(x')$$

$\lambda_i > 0 \quad \forall i$ (4)

For the expansion (4) to be valid,

and for it to be absolutely and

uniformly convergent,

$$\int_a^b \int_a^b K(x, x') \psi(x) \psi(x') \, dx \, dx' \geq 0$$

and

$$\int_a^b \psi^2(x) \, dx < \infty$$

Here, one can think of

$$\psi(x) \triangleq \varphi^T(x)\,\varphi(x) = \langle \varphi(x), \varphi(x) \rangle$$

where $\varphi(x) = \left[\varphi_0(x) \quad \cdots \quad \varphi_\infty(x) \right]^T$

Example: $\psi(x) = (1+x)^2$ over $[0, 1]$

$$\psi(x) = 1 + 2x + x^2$$

$$= (1 \;\; \sqrt{2}\,x \;\; x)(1 \;\; \sqrt{2x} \;\; x)^T$$

The functions $\{\varphi_i(x)\}$ are called the eigen functions of the expansion and λ_i s are the eigenvalues

$\lambda_i > 0 \Longrightarrow K(x, x')$ is +ve definite

Let us explore this for a $\underline{x}$ i.e; a vector $\underline{x}$

$(1 + x)^2$; $(1 + \underline{x}^T \underline{x})^2$ $\underline{x} \in \mathbb{R}^d$

1) In light of Mercer's theorem,

For $\lambda_i \neq 1$, the i^{th} image $\sqrt{\lambda_i}\,\varphi_i(\underline{x})$ induced in the feature space by the i^{th} Coord. in $\varphi(\underline{x})$ is an eigen function of the expansion.

2) In theory, the dim. of feature space can be ∞!

Mercer's theorem tells us if a candidate Kernel is actually an inner product kernel in some space $\Rightarrow$ admissible within the SVM design frame work

Examples of I.P. kernels

Machine	I.P. Kernel	Remarks
Polynomial learning machine	$(\underline{x}^T \underline{x}_i + 1)^p$	Power 'p' specified apriori
Radial basis function	$\exp\left(\frac{-1}{2\sigma^2} \|\underline{x} - \underline{x}_i\|^2\right)$	Width σ^2 common to all kernels, specified apriori or estimated adaptively

Optimum Design of a SVM

The expansion of the inner product kernel $K(\underline{x}, \underline{x}_i)$ allows us to construct a decision surface which is non-linear in the i/p space, but with image in the feature space 'linear'.

GOAL: Set up the optimization problem using $K(\cdot)$

$$\underline{\alpha} = (\alpha_1, \ldots \alpha_N)$$

Construct the optimization problem

$$\max_{\underline{\alpha}} \; Q(\underline{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j K(\underline{x}_i, \underline{x}_j)$$

Observe the $K(\cdot,\cdot)$ as against $\underline{x}_i^T \underline{x}_j$ in linearly separable situation in the i/p space.

Subject to

(1) $\sum_{i=1}^{N} \alpha_i d_i = 0$

(2) $0 \le \alpha_i \le C \qquad i = 1, \ldots, N$

$b = w_0$ for $\varphi_0(\underline{x}) = 1$

The inner product $\underline{x}_i^T \underline{x}_j$ we dealt with under SVM design for linear separability is replaced with $K(\underline{x}_i, \underline{x}_j)$

Define a symmetric matrix

$$K := \left[K(\underline{x}_i, \underline{x}_j) \right]_{i,j=1}^{N}$$

$$\begin{bmatrix} K(\underline{x}_1, \underline{x}_1) & \cdots & K(\underline{x}_1, \underline{x}_N) \\ \vdots & & \\ K(\underline{x}_N, \underline{x}_1) & \cdots & K(\underline{x}_N, \underline{x}_N) \end{bmatrix}$$

Gram matrix

After having found opt. values for the Lagrange multipliers $\alpha_{opt,i}$, we computes opt. $\underline{w}$ as

$$\underline{w}_{opt} = \sum_{i=1}^{N} \alpha_{opt,i} \, d_i \, \varphi(\underline{x}_i)$$

Image induced in the feature space

ε- insensitive loss function

GOAL: Come up with a robust estimator for a regression problem insensitive to small changes in the model.

Problems seen: A least squares (L.S) estimator is sensitive to the presence of outliers & performs poorly over distributions with long tails.

Idea

Minimize max. degradation.

Construct a loss function

$$L(d, y) = |d - y|$$

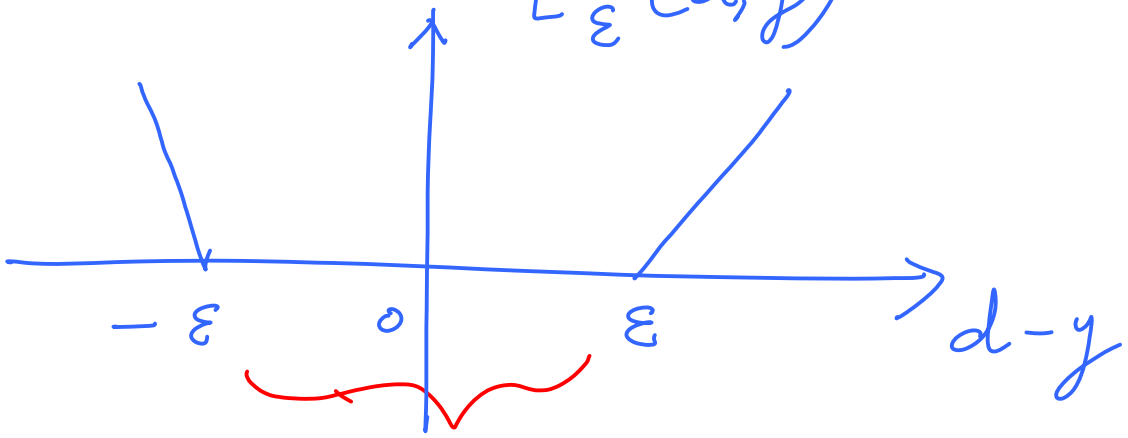
d: desired response; y: estimator o/p

Define $L_\varepsilon(d, y) = \begin{cases} |d - y| - \varepsilon, & |d-y| \geq \varepsilon \\ 0, & \text{else} \end{cases}$

'ε' is a prescribed parameter

$L_\varepsilon(d, y)$

ε - insensitive loss function



insensitive upto an ε

SVMs for non linear regression

Consider a nonlinear regression model

$$d = f(\underline{x}) + v$$

$f(\cdot)$ is unknown, statistics of v is unknown

We have the training set $\left\{ \underline{x}_i, d_i \right\}_{i=1}^{N}$

GOAL: Get an estimate of the dependence of d on $\underline{x}$

Let 'd' be estimated by 'y' using a linear combination of some non linear basis functions $\{\varphi_j(\underline{x})\}_{j=0}^{m_1}$

$$y = \sum_{j=0}^{m_1} w_j \varphi_j(\underline{x}) = w^T \varphi(\underline{x}) = \langle \underline{w}, \varphi(\underline{x}) \rangle$$

where $\varphi(\underline{x}) = [\varphi_0(\underline{x}) \cdots \varphi_{m_1}(\underline{x})]^T$; $\varphi_0(\underline{x}) = 1$

$w = [w_0 \cdots w_{m_1}]^T$; $w_0 = b$

Minimize the empirical risk

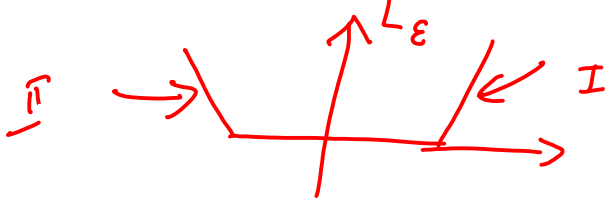
$$R_{emp} = \frac{1}{N} \sum_{i=1}^{N} L_{\varepsilon}(d_i, y_i)$$

Subject to $\|\underline{w}\|^2 \leq C_0$, C_0 is a constant

Let us formulate it as a constrained opt. problem by introducing 2 sets of slack variables

$$\{\xi_i\}_{i=1}^{N} \quad \text{and} \quad \{\xi_i'\}_{i=1}^{N}$$

$$\begin{cases} d_i - \underline{w}^T \varphi(\underline{x}_i) \leq \varepsilon + \xi_i, & i = 1, \dots, N \\ \underline{w}^T \varphi(\underline{x}_i) - d_i \leq \varepsilon + \xi_i', & i = 1, \dots, N \\ \xi_i \geq 0 \\ \xi_i' \geq 0 \end{cases}$$



The constrained opt. problem is equivalent to minimizing the cost functional

$$\phi(\underline{w}, \underline{\xi}, \underline{\xi}') = \underbrace{C \sum_{i=1}^{N} (\xi_i + \xi_i')}_{\text{misclassified points}} + \underbrace{\frac{1}{2} \underline{w}^T \underline{w}}_{\text{norm}^2 \text{ of the } \underline{w}}$$

Forming the Lagrangian $J(\underline{w}, \underline{\xi}, \underline{\xi}', \underline{\alpha}, \underline{\alpha}', \underline{\gamma}, \underline{\gamma}')$ in terms of the cost & the constraints using $\alpha, \alpha', \gamma, \gamma'$ as Lagrange multipliers

$$J(\cdot) = \underbrace{C \sum_{i=1}^{N} (\xi_i + \xi_i') + \frac{1}{2} \underline{w}^T \underline{w}}_{\text{Cost}}$$

$$- \sum_{i=1}^{N} \alpha_i \left(\underline{w}^T \varphi(\underline{x}_i) - d_i + \varepsilon + \xi_i \right)$$

$$- \sum_{i=1}^{N} \alpha_i' \left(d_i - \underline{w}^T \varphi(\underline{x}_i) + \varepsilon + \xi_i' \right)$$

} Conditions for the constraints over $L_\varepsilon(\cdot)$

$$- \sum_{i=1}^{N} \gamma_i \xi_i - \sum_{i=1}^{N} \gamma_i' \xi_i'$$

} Conditions for slack variables

Taking $\frac{\partial J(\cdot)}{\partial \underline{w}^T} = 0$

$$\Rightarrow \underline{w} = \sum_{i=1}^{N} (\alpha_i - \alpha_i') \varphi(\underline{x}_i)$$

Verify: $\gamma_i = C - \alpha_i$
$\gamma_i = C - \alpha_i'$
} Putting a constraint on α_i, α_i'

One can formulate the dual problem towards maximization

$$Q(\underline{\alpha}_i, \underline{\alpha}_i') = \sum_{i=1}^{N} d_i (\alpha_i - \alpha_i') - \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i') - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\alpha_i - \alpha_i')(\alpha_j - \alpha_j') K(\underline{x}_i, \underline{x}_j)$$

Inner product kernel

where $K(\cdot,\cdot) = \varphi^T(\underline{x}_i)\varphi(\underline{x}_j)$ is an I.P. kernel $\langle \varphi(\underline{x}_i), \varphi(\underline{x}_j) \rangle$

NOTE:

Training Set

Given $\{x_i, d_i\}_{i=1}^{N}$, find $\{\alpha_i\}_{i=1}^{N}$ and $\{\alpha_i'\}_{i=1}^{N}$ that max. $Q(\alpha_i, \alpha_i')$

Subject to (1) $\sum_{i=1}^{N} \alpha_i - \alpha_i' = 0$

(2) $0 \le \alpha_i, \alpha_i' \le C$ $\quad i = 1, \ldots, N$

Now, with $\underline{\varepsilon}$ and $\underline{C}$ as our free parameters

$$F(\underline{x}, \underline{w}) = \underline{w}^T \varphi(\underline{x})$$

$$= \sum_{i=1}^{N} (\alpha_i - \alpha_i') K(\underline{x}, \underline{x}_i)$$

XOR Problem Revisited

Let us consider the XOR problem using SVMs (Cherkassy, 1998)

Start with a kernel

$$K(\underline{x}, \underline{x}_i) = (1 + \underline{x}^T \underline{x}_i)^2$$

$$\underline{x} = [x_1 \; x_2]^T \qquad \underline{x}_i = [x_{i1} \; x_{i2}]^T$$

$$K(\underline{x}, \underline{x}_i) = \left(1 + (x_1\ x_2) \cdot \begin{bmatrix} x_{i1} \\ x_{i2} \end{bmatrix}\right)^2$$

$$= \left(1 + x_1 x_{i1} + x_2 x_{i2}\right)^2$$

$$= 1 \cdot 1 + x_1^2 x_{i1}^2 + 2 x_1 x_2 x_{i1} x_{i2} + 2 x_1 x_{i1} + 2 x_2 x_{i2} + x_2^2 x_{i2}^2$$

Let us express $K(\cdot,\cdot) = \langle \varphi(\underline{x}), \varphi(\underline{x}_i) \rangle$

$$\varphi(x) = \begin{bmatrix} 1 & x_1^2 & \sqrt{2}\, x_1 x_2 & x_2^2 & \sqrt{2}\, x_1 & \sqrt{2}\, x_2 \end{bmatrix}^T$$

$$\varphi(\underline{x}_i) = \begin{bmatrix} 1 & x_{i_1}^2 & \sqrt{2}\, x_{i_1} x_{i_2} & x_{i_2}^2 & \sqrt{2}\, x_{i_1} & \sqrt{2}\, x_{i_2} \end{bmatrix}^T$$

For the XOR problem,

$(-1, -1) \longrightarrow -1$

$(-1, +1) \longrightarrow +1$

$(+1, -1) \longrightarrow +1$

$(+1, +1) \longrightarrow -1$

$$K := \left[K(\underline{x}_i, \underline{x}_j) \right]$$

$$K(\underline{x}_i, \underline{x}_j) = \varphi^T(\underline{x}_i)\,\varphi(\underline{x}_j)$$

Each $\underline{x}_i, \underline{x}_j$

$$K = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}_{4\times 4}$$

From the dual problem, the objective $Q(\underline{\alpha})$

$$Q(\underline{\alpha}) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - \frac{1}{2}\Big(9\alpha_1^2 - 2\alpha_1\alpha_2 - 2\alpha_1\alpha_3 + 2\alpha_1\alpha_4 + 9\alpha_2^2 + 2\alpha_2\alpha_3 - 2\alpha_2\alpha_4 + 9\alpha_3^2 - 2\alpha_3\alpha_4 + 9\alpha_4^2 \Big)$$

$$\sum_{i=1}^{4}\sum_{j=1}^{4} \alpha_i \alpha_j d_i d_j K(\underline{x}_i, \underline{x}_j)$$

$$\frac{\partial Q(\underline{\alpha})}{\partial \alpha_i} = 0 \qquad i = 1, \dots, 4$$

Doing this yields

$$\begin{cases} 9\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 1 \\ -\alpha_1 + 9\alpha_2 + \alpha_3 - \alpha_4 = 1 \\ -\alpha_1 + \alpha_2 + 9\alpha_3 - \alpha_4 = 1 \\ \alpha_1 - \alpha_2 - \alpha_3 + 9\alpha_4 = 1 \end{cases}$$

Solving this set of eqns

$$\alpha_{opt,i} = \frac{1}{8} \qquad i = 1, \dots, 4$$

Now, all 4 inputs $\{\underline{x}_i\}_{i=1}^{4}$ are support vectors

$$Q(\alpha) = \frac{1}{4}$$

$$\frac{1}{2}\|\underline{w}_o\|^2 = \frac{1}{4} \implies \|\underline{w}_o\| = \frac{1}{\sqrt{2}}$$

$$\underline{w}_o = \sum_{i=1}^{4} \alpha_{o,i}\, d_i\, \varphi(\underline{x}_i)$$

$$= \frac{1}{8}\left[-\varphi(\underline{x}_1) + \varphi(\underline{x}_2) + \varphi(\underline{x}_3) - \varphi(\underline{x}_4)\right]$$

$$\underline{w}_o = \left[0 \quad 0 \quad -\frac{1}{\sqrt{2}} \quad 0 \quad 0 \quad 0 \right]^T$$

The opt. hyperplane is given by

$$\underline{w}_o^T \varphi(\underline{x}) = 0$$

Linear frame work

$$\left[0 \quad 0 \quad -\frac{1}{\sqrt{2}} \quad 0 \quad 0 \quad 0 \right] \begin{bmatrix} 1 \\ x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \\ \sqrt{2}\, x_1 \\ \sqrt{2}\, x_2 \end{bmatrix} = 0$$

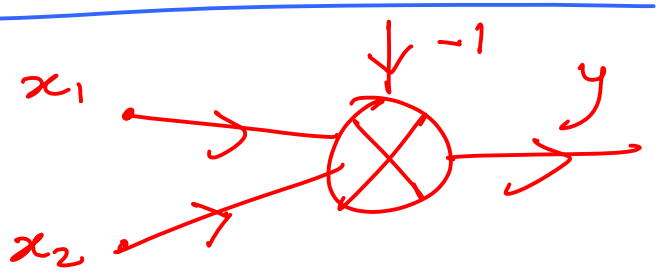
$$\Rightarrow \quad x_1 x_2 = 0$$

is the decision boundary

$y = -x_1 x_2$ ← It is nonlinear

$\underline{x}$	$y = -x_1 x_2$
$(-1, -1)$	-1
$(-1, +1)$	$+1$
$(+1, -1)$	$+1$
$(+1, +1)$	-1

as desired!



Hilbert Space

Let $\{\underline{x}_k\}_{k=1}^{\infty}$ be an orthonormal basis for an inner product space $\mathcal{F}$ (possibly of ∞ dimensions)

$$\langle \underline{x}_j, \underline{x}_k \rangle = \underline{x}_j^T \underline{x}_k = \begin{cases} 1 & j = k \\ 0 & \text{else} \end{cases}$$

Let $\mathcal{H}$ be the largest and most inclusive space of vectors for which the set $\{\underline{x}_k\}_{k=1}^{\infty}$ is a basis.

Then any vector $\underline{x}$ not necessarily lying in $\mathcal{F}$ can be written as $\underline{x} = \sum_{k=1}^{\infty} a_k \underline{x}_k$

$\underline{x}^* = \lim_{n\to\infty} \underline{x}^{(n)} \notin \mathcal{F}$

↑ Converging vector (subtle issue)

Define a new vector

$$\underline{y}_n = \sum_{k=1}^{n} a_k \underline{x}_k$$

$$\underline{y}_m = \sum_{k=1}^{m} a_k \underline{x}_k$$

For $n > m$

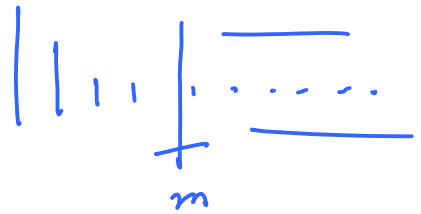
$$\|\underline{y}_n - \underline{y}_m\|^2 = \left\| \sum_{k=1}^{n} a_k \underline{x}_k - \sum_{k=1}^{m} a_k \underline{x}_k \right\|^2$$

$$= \left\| \sum_{k=m+1}^{n} a_k \underline{x}_k \right\|^2$$

$$\leq \sum_{k=m+1}^{n} a_k^2 \underbrace{\|\underline{x}_k\|^2}_{\text{norm. 1}} = \sum_{k=m+1}^{n} a_k^2$$

1) $\sum_{k=m+1}^{n} a_k^2 \longrightarrow 0$ as $m, n \to \infty$

2) $\sum_{k=1}^{n} a_k^2 < \infty$



Pick an $\varepsilon > 0$ and $m \gg$ such that $\sum_{k=m+1}^{\infty} a_k^2 < \varepsilon$

Since $\sum_{k=1}^{\infty} a_k^2 = \sum_{k=1}^{m} a_k^2 + \sum_{k=m+1}^{\infty} a_k^2$ ——— (1)

$\sum_{k=1}^{\infty} a_k^2 < \infty$

A sequence of vectors in a normed space, for which the Euclidean distance

$$\| \underline{y}_n - \underline{y}_m \| < \varepsilon \quad \text{for any } \varepsilon > 0 \quad \text{and } m, n > M$$

is a convergent sequence called a Cauchy sequence

$$\{ \underline{y}_n \}_{n=1}^{\infty} \quad \text{is Cauchy}$$

$$\| \underline{y}_n - \underline{y}_m \| \xrightarrow[m,n \to \infty]{} 0$$

(Hilbert space) $\mathcal{H}$ is more complete than $\mathcal{I}$ (I.P)

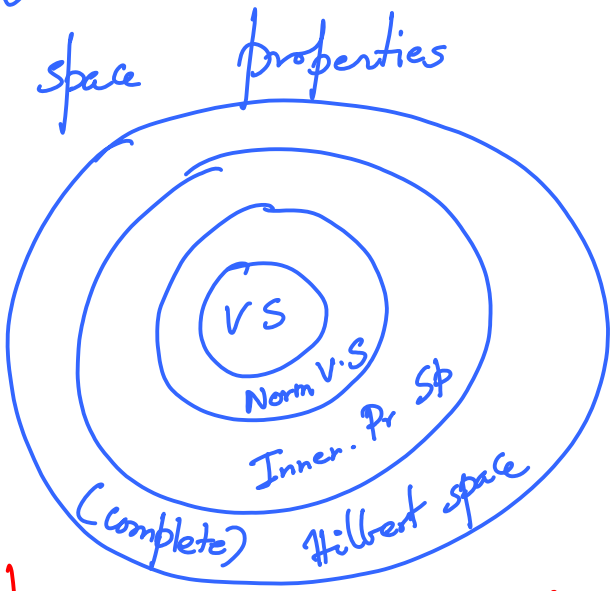
1) Every Cauchy sequence of vectors taken from $\mathcal{H}$ Converges to a limit in $\mathcal{H}$.

2) $\mathcal{H}$ inherits the inner product space properties

NOTE: Are all Cauchy sequences Convergent? No

$$\sum_{k \geq 0} \frac{1}{k!} \longrightarrow e \notin \mathbb{Q}$$

over all rational nos $\mathbb{Q}$



All convergent sequences are Cauchy
But, all Cauchy sequences are not convergent