Support Vector Machines

Jesús Manuel de la Cruz García, Matilde Santos Peña
Dpto. Arquitectura de Computadores y Automática.
Universidad Complutense

jmcruz@fis.ucm.es, msantos@dacya.ucm.es
http://www.dacya.ucm.es/area-isa/index.php?page=home

Sao Paulo, Brasil, October 2010

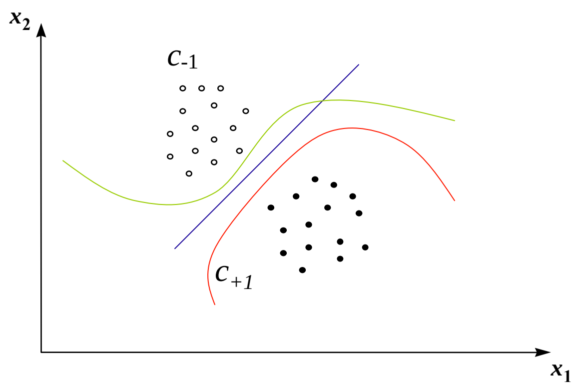
Presentation

- Two Learning Problems
- Linearly Separable Classes: Linear SVM
- Nonlinearly Separable Classes: Nonlinear SVM
- Multiclass SVM
- SVM for regression
- Applications

A Classification Learning Problem

Set of data $\{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots, n\}$

Two different classes $y_i = \in \{-1, +1\}$



Problem: find $f(\mathbf{x};\ \theta)$ such that

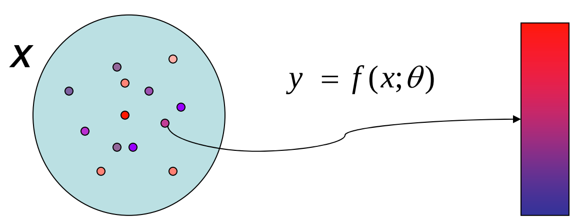
$$y = f(\mathbf{x}; \theta)$$

is a classifier

A regression Learning Problem

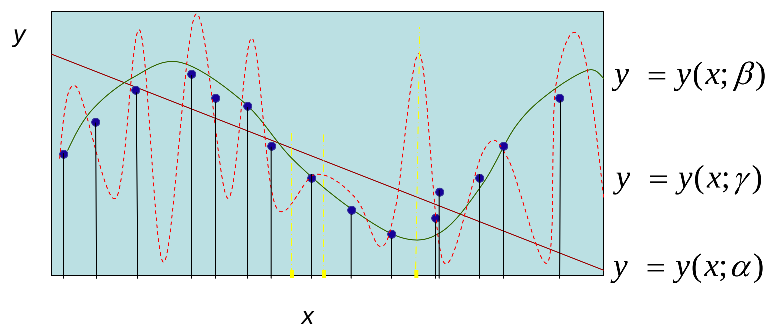
Set of data $\{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots, n\}$

y_i continuous domain



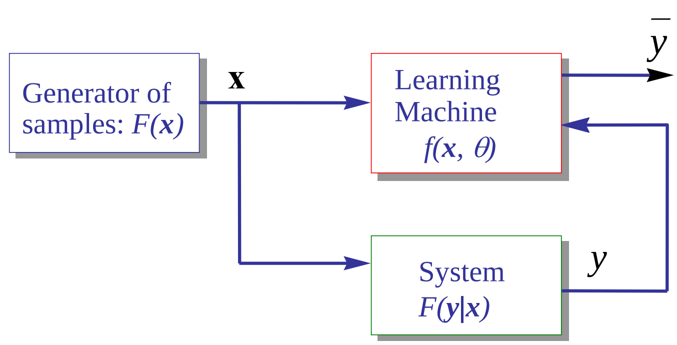
If we are given a finite number of samples
What is the number we assign to a new sample?

Model Complexity Problem



- There exists infinite ways of approaching the points of a curve.
- It does not suffice to have a good approach, the complexity of the model needs to be controlled.
- A measure of the complexity of the model must be introduced.

General Learning Machine

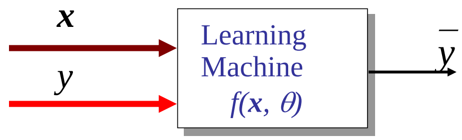


- ***G**:* $\mathbf{x} \in R^m$, drawn independently from a fixed probability function $F(\mathbf{x})$
- ***S**:* $y = F(y|\mathbf{x})$, system output. $F(y|\mathbf{x})$ unknown

$$F(\mathbf{x},y) = F(y \mid \mathbf{x})\, F(\mathbf{x})$$

- ***L.M.**:* implements a set of functions $f(\mathbf{x},\theta)$, $\theta \in \Lambda$

Learning Problem: Theoretical Formulation



For a given θ we *measure* the discrepancy between y and $f(\mathbf{x}, \theta)$:
$Q(y, f(\mathbf{x}, \theta))$

How do we choose θ, or equivalently $f(\mathbf{x}, \theta)$?

We choose the function that gives the minimum mean value of the discrepancy

$$R(\theta) = \int Q(y, f(x, \theta))\, dF(x, y)$$

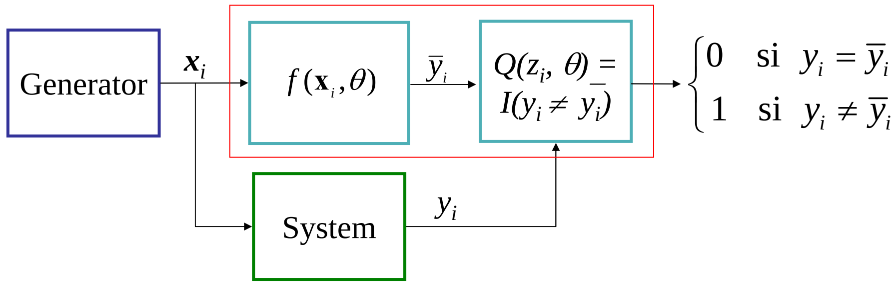
Learning Problem: Empirical Risk Minimization Principle

- We have a finite number of samples:
  $(\mathbf{x}_1, y_1) = \mathbf{z}_1, (\mathbf{x}_2, y_2) = \mathbf{z}_2, , \ldots, (\mathbf{x}_N, y_N) = \mathbf{z}_N,$
- Fix a family of functions: $f(\mathbf{x}, \theta)$
- Measure the discrepancy between y_i and $f(\mathbf{x}_i, \theta)$: $Q(z_i, \theta)$
- Calculate the mean value of the discrepancies

$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N} Q(z_i, \theta)$$

- Choose the value of θ that minimizes the Empirical Risk: R_{emp}

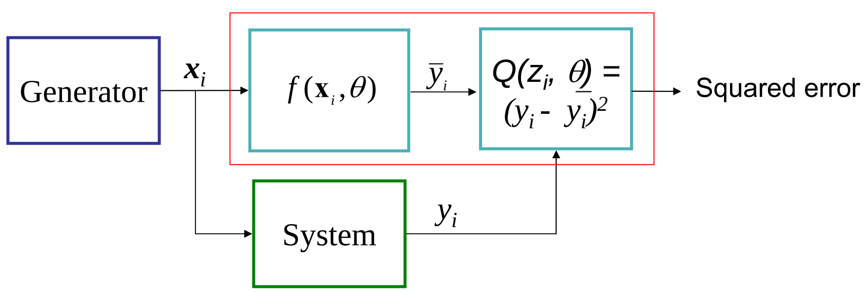
Classification as a Learnin Machine



The best possible choice of $f(\mathbf{x}, \theta)$ is the one that minimizes the empirical risk:

$$R_{emp}(\theta) = \frac{1}{N} \sum_{i=1}^{N} I(y_i \neq \bar{y}_i)$$

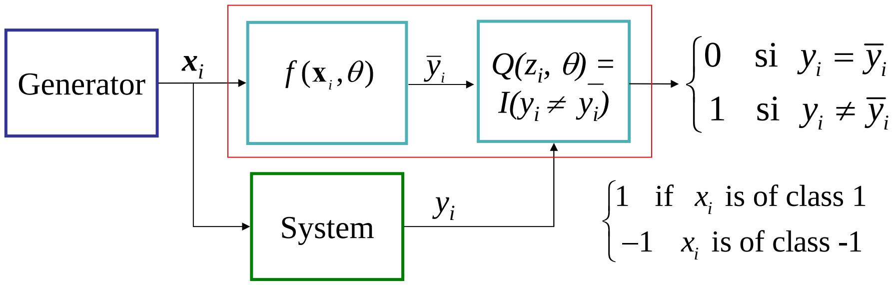
Regression as a Learning Machine



The best possible choice of $f(\mathbf{x}, \theta)$ is the one that minimizes the empirical risk:

$$R_{emp}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}_i)^2$$

Classification: Linearly Separable Case

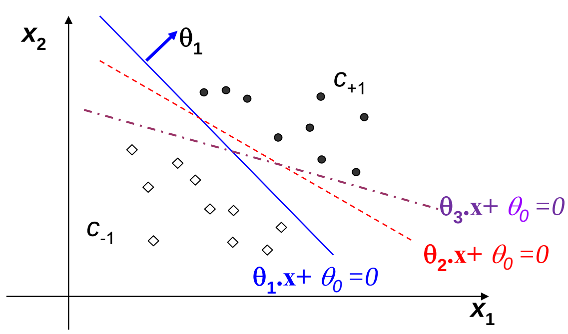


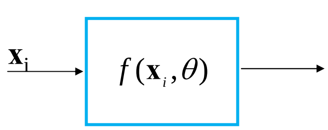
We consider a separating hyperplane: $D(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\theta} \rangle + \theta_0$

$y_i = f(\mathbf{x}_i, \boldsymbol{\theta}) = \text{sign } D(\mathbf{x}_i)$

Linearly Separable Case

How do we compute the hyperplane ?



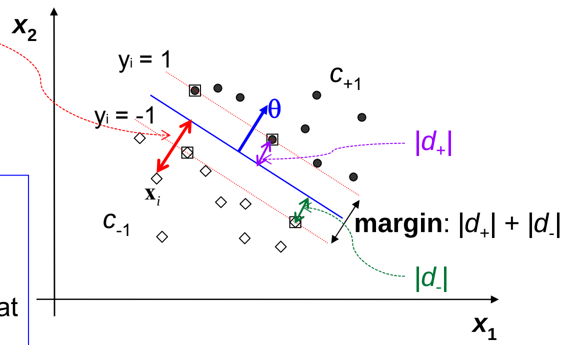


$y_i = 1$, if $\mathbf{x}_i \in c_{+1}$,
$y_i = -1$, if $\mathbf{x}_i \in c_{-1}$

Linearly Separable Case

Optimal separating hyperplane: the one maximizing the margin → maximal margin classifier

distance: $|d_i| = \dfrac{|D(\mathbf{x}_i)|}{\|\boldsymbol{\theta}\|}$



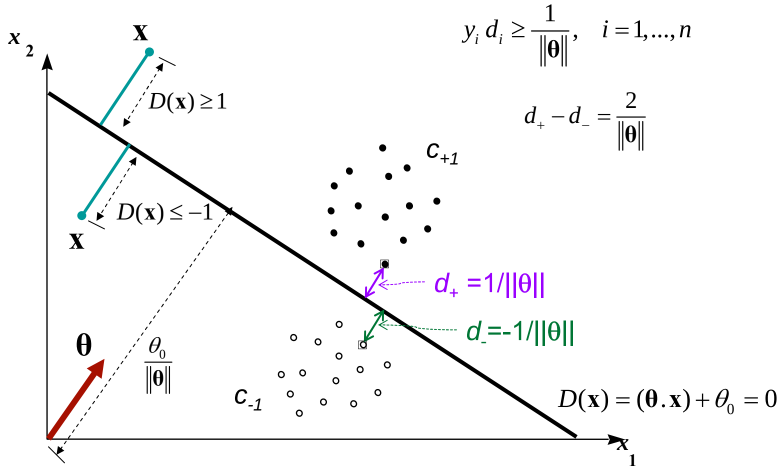
Since: $D(\mathbf{x}) = 0$ holds if
we take $k\boldsymbol{\theta}$, $k\theta_0$
We choose $\boldsymbol{\theta}$, θ_0 such that
$min_{\,i=1,..,m}\ |\ D(\mathbf{x})\ | = 1$

Linearly Separable Case

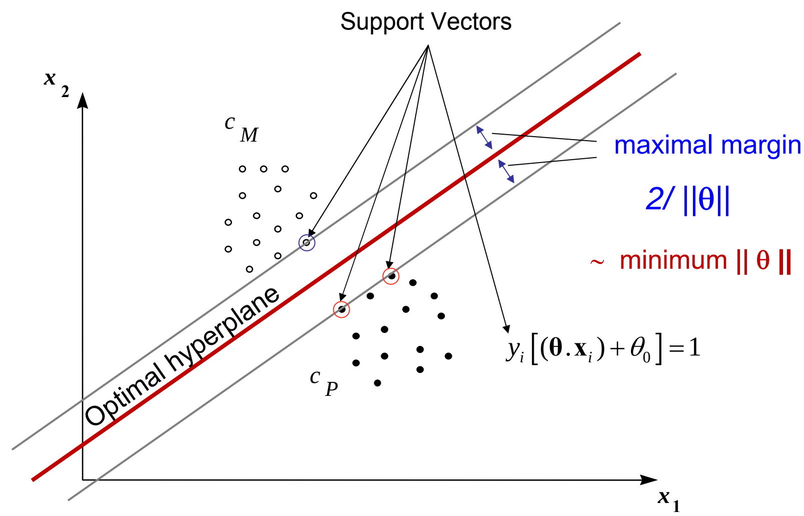
We choose θ, θ_0 such that $\quad y_i\, D(\mathbf{x}_i) \geq 1, \quad i = 1,\ldots,n$

$$y_i\, d_i \geq \frac{1}{\|\boldsymbol{\theta}\|}, \quad i = 1,\ldots,n$$

$$d_+ - d_- = \frac{2}{\|\boldsymbol{\theta}\|}$$



Computation of the Optimal Hyperplane



Optimal Separating Hyperplane

With the election of θ that generates the given straight line, it is verified

$$R_{emp}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \bar{y}_i) = 0$$

And the learning process has been done.

The prediction is done through the relationship

$\mathbf{x}_i$ → **sign [(θ. $\mathbf{x}_i$)+ θ_0]** → $\bar{y}_i$

Optimization Problem

- The problem of finding the optimal separating hyperplane can be formulated as:

$$\min \quad J(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}.\boldsymbol{\theta}$$

subject to:

$$y_i\left[(\boldsymbol{\theta}.\mathbf{x}_i) + \theta_0\right] \geq 1, \quad i = 1,\ldots,n$$

- Convex optimization problem: minimize a quadratic function subject to linear inequalities constraints →
  There exists a global minimum without local minima

Solution of the Optimization Problem

The solution has always the form

$$\boldsymbol{\theta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i$$

α_i are Lagrange multipliers:

$$\alpha_i^*\left[y_i\left((\boldsymbol{\theta}^*.\mathbf{x}_i) + \theta_0^*\right) - 1\right] = 0, \quad i = 1,\ldots,n$$

The $\mathbf{x}_i$ associated to the α_i different from zero are called "support vectors"

$$\boldsymbol{\theta}^* = \sum_{\substack{vectores \\ soporte}} \alpha_i^* y_i \mathbf{x}_i$$

Solution of the Optimization Problem

The estimated bias is:

$$\theta_0^* = \frac{1}{|sv|}\left(\sum_{\substack{\text{sup}\ port\\ vectors}} \frac{1 - y_i\left(\boldsymbol{\theta}^*.\mathbf{x}_i\right)}{y_i}\right), \quad |sv| = \text{ number of support vectors}$$

SVM Decission Function

The classification rule is:

$$D(\mathbf{x}) = \text{sign}\left(\sum_{\substack{\text{support}\\ \text{vectors}}} \alpha_i^* y_i (\mathbf{x}_i\,.\,\mathbf{x}) + \theta_0^*\right)$$

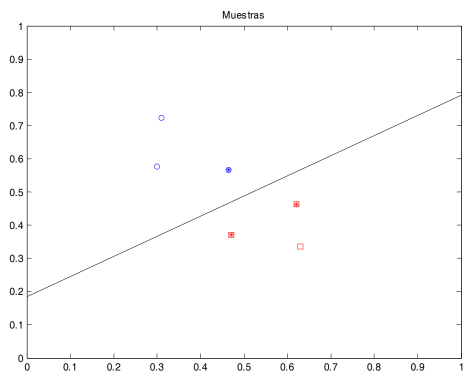
It is necessary to compute the dot product of the element to be classified with the Support Vectors $\mathbf{x}_i$

Example

	x_1	x_2	y_i	α_i	$\boldsymbol{\theta}^*.\mathbf{x}_i+\theta_0$
$\mathbf{x}_1$	0.3107	0.7234	1.0	0	3.5166
$\mathbf{x}_2$	0.2988	0.5754	1.0	0	2.1011
$\mathbf{x}_3$	0.4645	0.5666	1.0	69.2589	1.0000
$\mathbf{x}_4$	0.4704	0.3713	-1.0	31.4601	-1.0000
$\mathbf{x}_5$	0.6302	0.3358	-1.0	0	-2.3335
$\mathbf{x}_6$	0.6213	0.4630	-1.0	37.7988	-1.0000

$$\boldsymbol{\theta}^* = (\text{-}6.1125 \;\; 10.0601)$$

$$\theta_0^* = -1.8592$$



Optimization Problem and Solution

Problem:

$$\min_{\boldsymbol{\theta}} \frac{1}{2}\boldsymbol{\theta}.\boldsymbol{\theta}$$

$$\text{subject to} \;\; y_i(\boldsymbol{\theta}.\mathbf{x}_i + \theta_0) \geq 1, \quad i = 1,\cdots,n$$

Dual formulation:

$$L(\boldsymbol{\theta},\theta_0,\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 - \sum_{i=1}^{n}\alpha_i\left[y_i(\boldsymbol{\theta}.\mathbf{x}_i + \theta_0) - 1\right], \quad \alpha_i \geq 0, i = 1,\dots,n$$

$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)$ Lagrange multipliers vector

Optimal solution is a saddle point of *L:*

- minimum w.r.t. θ, θ_0
- maximum w.r.t. $\boldsymbol{\alpha}$

Optimization Problem and Solution

- Minimum of L w.r.t. θ, θ_0

$$\frac{\partial L}{\partial \theta_0} = \sum_{i=1}^{m} \alpha_i y_i = 0 \quad \text{and} \quad \frac{\partial L}{\partial \boldsymbol{\theta}} = \boldsymbol{\theta} - \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i = 0$$

$$\Downarrow$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad \boldsymbol{\theta} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i \qquad \textbf{(1)}$$

- Maximum of L w.r.t. $\boldsymbol{\alpha}$ using **(1)**

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \mathbf{x}_i . \mathbf{x}_j$$

$$\text{subject to} \quad \alpha_i \geq 0, \; i = 1, \ldots, n \qquad \textbf{(2)}$$

$$\text{and} \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Optimization Problem and Solution

(2) $\equiv$

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = 1_n^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha}$$

Only dot products

$$\text{subject to} \quad \boldsymbol{\alpha} \geq 0, \; \boldsymbol{\alpha}^T \mathbf{y} = 0$$

where

$$\mathbf{H} = \left[H_{ij} \right], H_{ij} = y_i y_j (\mathbf{x}_i . \mathbf{x}_j), i, j = 1, \ldots, n$$

$$\mathbf{y} = (y_1, \cdots, y_n)^T$$

Solution

$$\alpha_i^*, \quad i = 1, \ldots, n \quad \Rightarrow \quad \boldsymbol{\theta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i$$

$$\alpha_i^* \left[y_i (\boldsymbol{\theta}^* . \mathbf{x_i} + \theta_0) - 1 \right] = 0, \quad \alpha_i^* \geq 0, i = 1, \ldots, n$$

Optimization Problem and Solution

$$(2) \equiv \max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = 1_n^T \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{H} \boldsymbol{\alpha}$$

$$\text{subject to} \quad \boldsymbol{\alpha} \geq 0, \;\; \boldsymbol{\alpha}^T \mathbf{y} = 0$$

where $\mathbf{H} = \left[H_{ij}\right], H_{ij} = y_i y_j (\mathbf{x}_i . \mathbf{x}_j), i, j = 1, \ldots, n$

$$\mathbf{y} = (y_1, \cdots, y_n)^T$$

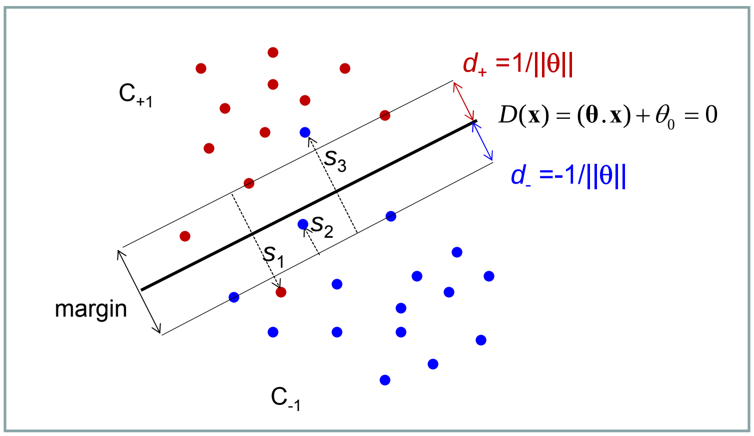
Solution $\alpha_i^*, \quad i = 1, \ldots, n \quad \Rightarrow \quad \boldsymbol{\theta}^* = \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i$

$$\alpha_i^* \left[y_i (\boldsymbol{\theta}^* . \mathbf{x}_i + \theta_0) - 1 \right] = 0, \quad \alpha_i^* \geq 0, i = 1, \ldots, n$$

Linearly Nonseparable Case

We introduce *slack* variables s_i for each observation $(\mathbf{x}_i, y_i)$

$\mathbf{s} = (s_1, \ldots, s_n)^T \geq 0$



Linearly Nonseparable Case

The constraints are now:

$$y_i(\boldsymbol{\theta}.\mathbf{x_i}+\theta_0)\geq 1-s_i,\quad s_i\geq 0,\quad i=1,\cdots,n$$

If $\mathbf{x}_i$ verifies: $y_i(\boldsymbol{\theta}.\mathbf{x_i}+\theta_0)\geq 1$ then $s_i = 0$

Else: $s_i > 0$

We consider a function of the slack variables: $C\sum_{i=1}^{m} s_i$

C is a regularization parameter

Optimization Problem for the Linearly Nonseparable case

- Statement of the new problem

$$\min_{\boldsymbol{\theta}} \quad \frac{1}{2}\boldsymbol{\theta}.\boldsymbol{\theta}+C\sum_{i=1}^{m} s_i$$

$$\text{subject to} \quad y_i(\boldsymbol{\theta}.\mathbf{x_i}+b)\geq 1-s_i,\ i=1,\cdots,n$$

$$s_i\geq 0,\ i=1,\cdots,n$$

- Dual formulation:

$$L(\boldsymbol{\theta},\theta_0,\mathbf{s},\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{2}\boldsymbol{\theta}.\boldsymbol{\theta}+C\sum_{i=1}^{n} s_i-\sum_{i=1}^{n}\alpha_i\left[y_i(\boldsymbol{\theta}.\mathbf{x_i}+\theta_0)-(1-s_i)\right]-\sum_{i=1}^{n}\beta_i s_i$$

$\alpha_i,\beta_i\geq 0, i=1,...,n$ Lagrange multipliers

Optimal solution is a saddle point of *L:*

- minimum w.r.t. θ, θ_0 , **s**
- maximum w.r.t. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$

Optimization Problem for the Linearly Nonseparable case

• Minimum of L w.r.t. θ, θ_0 , **s**

$$\frac{\partial L}{\partial \theta_0} = \sum_{i=1}^{m} \alpha_i \, y_i = 0, \quad \frac{\partial L}{\partial \boldsymbol{\theta}} = \boldsymbol{\theta} - \sum_{i=1}^{m} \alpha_i \, y_i \, \mathbf{x}_i = 0, \quad \frac{\partial L}{\partial s_i} = C - \alpha_i - \beta_i = 0$$

$$\Rightarrow \quad \boldsymbol{\theta}^* = \sum_{i=1}^{n} \alpha_i \, y_i \, \mathbf{x}_i, \quad \sum_{i=1}^{n} \alpha_i \, y_i = 0, \quad \alpha_i = C - \beta_i \qquad \textbf{(1)}$$

• Maximum of L w.r.t. $\boldsymbol{\alpha}$ using **(1)**

$$\max_{\boldsymbol{\alpha}} \quad L(\boldsymbol{\alpha}) = 1_n^T \boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T \mathbf{H} \, \boldsymbol{\alpha}$$

$$\text{subject to} \quad \mathbf{0} \le \boldsymbol{\alpha} \le C 1_n, \; \boldsymbol{\alpha}^T \mathbf{y} = 0$$

Only dot products

If C = ∞ the problem becomes the hard-margin separable case

Optimization Problem for the Linearly Nonseparable case

• Possible values of the optimal Lagrange multipliers $\boldsymbol{\alpha^*}$

$$\alpha_i^* = 0 \quad \Rightarrow \quad y_i(\boldsymbol{\theta}.\mathbf{x_i} + \theta_0) \ge 1 \quad \text{and} \quad s_i = 0$$

$$0 < \alpha_i^* < C \quad \Rightarrow \quad y_i(\boldsymbol{\theta}.\mathbf{x_i} + \theta_0) = 1 \quad \text{and} \quad s_i = 0$$

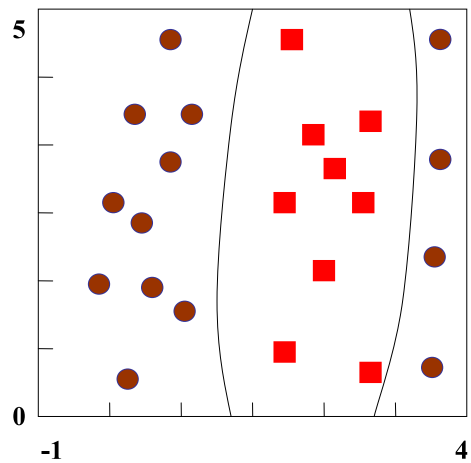
$\mathbf{x}_i$ support vector: margin vector

$$\alpha_i^* = C \quad \Rightarrow \quad y_i(\boldsymbol{\theta}.\mathbf{x_i} + \theta_0) \le 1 \quad \text{and} \quad s_i \ge 0$$

$\mathbf{x}_i$ support vector: error vector

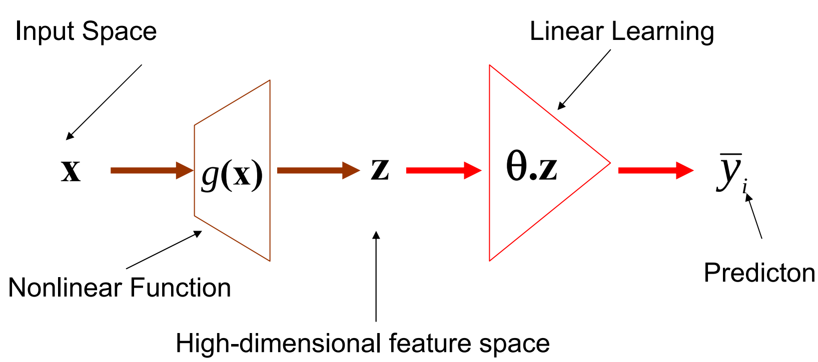
$$\boldsymbol{\theta}^* = \sum_{\substack{\text{support} \\ \text{vectors}}} \alpha_i^* y_i \mathbf{x}_i$$

Nonlinear Classification Problem



Neural Networks provide a solution

Searching for a Hyperplane in a High Dimensional Space



We search for optimal hyperplanes in the feature space **z**

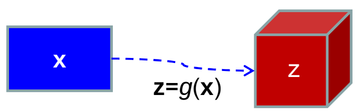
Decission function in the space z

$$D(\mathbf{x}) = \text{signo}\left(\sum_{\substack{\text{support}\\\text{vectors}}} \alpha_i y_i (g(\mathbf{x}_i).g(\mathbf{x})) + \theta_0 \right)$$

We choose nonlinear mappings *g(***x***)* such that:

$$(g(\mathbf{x}_i).g(\mathbf{x})) = K(\mathbf{x}_i.\mathbf{x})$$

*K(***x***,* **y***):* inner product Kernel



$\mathbf{z}_i \; . \; \mathbf{z}_j \; = K(\mathbf{x}_i \; \mathbf{x}_j)$

Searching for a Hyperplane in a High Dimensional Space

- All the necessary relationships to find the optimal hyperplane have a dot product form within the feature space
- It is not necessary to make the transformation to the feature space, but to know how to compute dot products in the feature space

Kernel functions

- Polynomials of degree *q*:

$$H(\mathbf{x},\bar{\mathbf{x}}) = [(\mathbf{x}.\bar{\mathbf{x}})+1]^q$$

- Radial basis functions:

$$H(\mathbf{x},\bar{\mathbf{x}}) = \exp\left\{-\frac{|\mathbf{x}-\bar{\mathbf{x}}|^2}{\sigma^2}\right\}$$

- Neuronal Network (Sigmoid)

$$H(\mathbf{x},\bar{\mathbf{x}}) = \tanh\left[b(\mathbf{x}.\bar{\mathbf{x}})+c\right]$$

Aplications

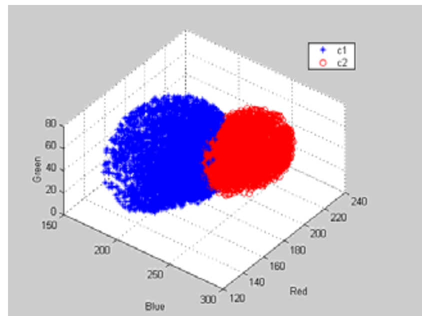
- **Sunflower Classification**
  G. Pajares, E. Besada-Portas, J.M. de la Cruz, "Analysis of support vector machines and Bayesian methods for sunflower classification". Recent Res. Devel. Pattern Rec.; 3[2002]: 1-14, Transworld Research Network.
- **Stereovision**
  G. Pajares, J.M. de la Cruz. "On combining Support Vector Machines and Simulated Annealing in Stereovision Matching". IEEE Trans. on Systems, Man , and Cybernetics-Part B: Cybernetics, vol. 34. August, 2004.
- **Nuclear Fusion**
  S. Dormido-Canto, G. Farias, R. Dormido, J. Vega, J. Sánchez, M. Santos."TJ -II wave forms analysis with wavelets and support vector machines". *Review of Scientific Instruments.* Vol. 75, Issue 10, pp. 4254-4257, October 2004
- **Diagnostic of Brain Tumors**.
  Instituto de Neurocirugía, Hospital Universitario La Paz, Instituto de Investigaciones Biomédicas del CSIC y Universidad Comnplutense.

Clasificación de girasoles

Clasificación de los píxeles de una imagen como pertenecientes a los pétalos, c_1, o a la parte central de los girasoles, c_2.

Cada elemento tiene tres atributos: $\mathbf{x}\ (x_{rojo}, x_{verde}, x_{azul})$

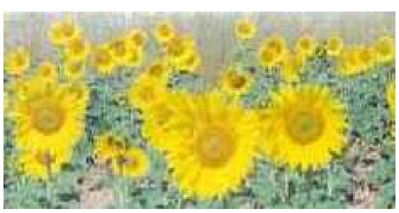
Imagen de entrenamiento

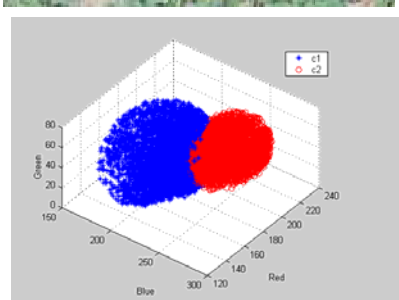


800 puntos de entrenamiento

Clasificación de girasoles

Figuras para clasificación





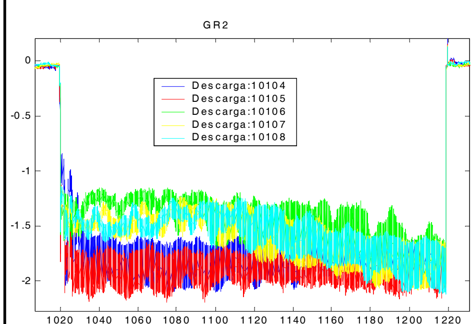
Kernel	Param.	Nº VS	Error %
VP	q=3	10	1.43
RBF	σ=16	11	1.38
TLNN	b=1/4 c=-1	9	1.49

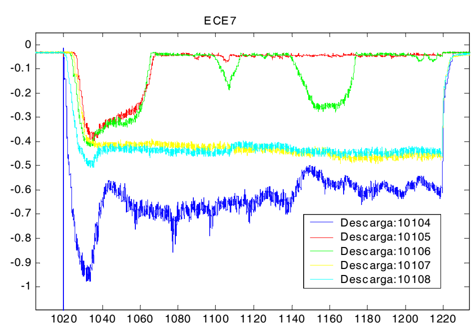
Señales de Fusión Nuclear: TJII

Dispositivo del tipo stellarator
Los plasmas se producen y se calientan con ECRH
Dispone de 940 canales digitales para medidas experimentales.
Necesidad de disponer de mecanismos automáticos
de clasificación y recuperación de las señales

Clases de señales de la base de datos del TJII	
BOL5	Señal de bolometría
ECE7	Emisión electrón-ciclotrón
RX306	Rayos-X blandos
ACTON275	Señal espectroscópica
HALFAC3	Emisión de la línea ALFA de hidrógeno
Densidad2	Densidad electrónica de línea

Ejemplo de señales

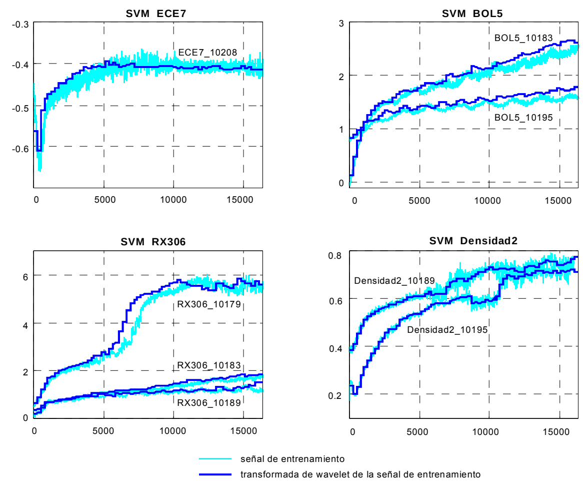




Señales con 16384 muestras reducidas en un factor de 2^8
mediante la Transformada de Wavelets

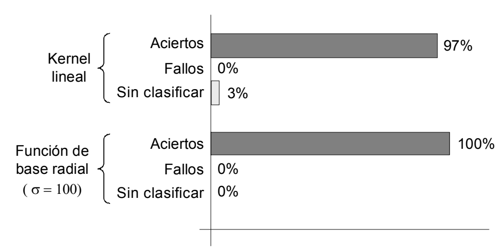
Uso de wavelets y aprendizaje estocástico

Representación de los VS para ECE7, BOL5, RX306 y Densidad2

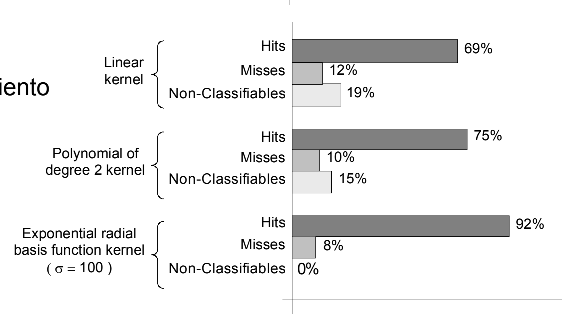


Uso de wavelets y aprendizaje estocástico

Resultados para 4 clases
ECE7, BOL5, RX306 y Densidad2
con 40 señales de entrenamiento
y 32 señales de test



Resultados para 6 clases
con 60 señales de entrenamiento
y 40 señales de test

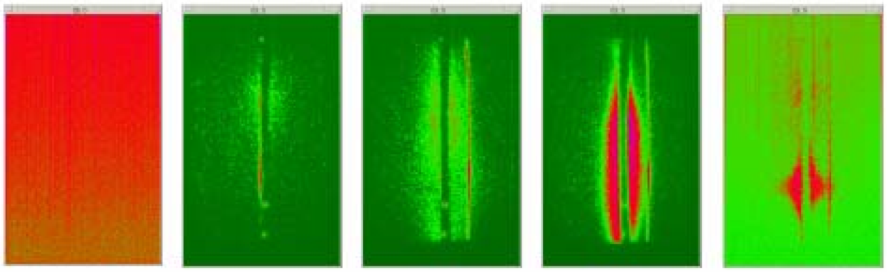


Diagnóstico de Scattering Thompson

- La dispersión Thomson consiste en la reemisión de la radiación incidente en el plasma por parte de los electrones libres
- Se trata de un diagnóstico óptico no perturbativo
- Medible solo usando los láseres más potentes (rubí en el caso del TJ-II)
- La distribución de velocidades de los electrones se traduce en un **ensanchamiento espectral** de la luz dispersa relacionado con la temperatura electrónica (por efecto Doppler)
- El número **total** de fotones dispersos es proporcional a la densidad electrónica
- T.S. <=> espectroscopía con resolución espacial, por lo tanto es (en TJ-II) un sistema espectroscópico bidimensional

Diagnóstico de Scattering Thompson

Patrones de imágenes del TJII Scattering Thompson



(a) (b) (c) (d) (e)

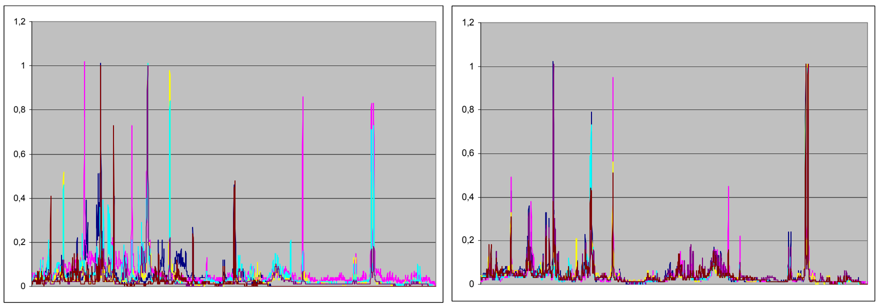
Cada imagen tiene 221760 pixels, que se reducen a 900 atributos utilizando Transformada de Wavelets, y mediante tratamiento de imágenes se reduce a 10 atributos por imagen.

Actúa de manera automática con un 98% de aciertos

Diagnóstico de tumores cerebrales

Espectros de Resonancia Magnética Nuclear de Protones.

- 16384 puntos por muestra
- 10 tipos distintos de patologías
- Mejores resultados conocidos hasta ahora



Conjunto de señales correspondientes a dos tipos de patologías distintas