UNIVERSIDADE
DE VIGO

Departamento de Estatística
e Investigación Operativa

Minería de Datos, 2008-2009
Support Vector Machines para Clasificación y Regresión

Desarrollo Histórico

1. El origen de las **support vector machines** (SVM), es una generalización del **hiperplano separador de máximo margen** (definición 1, pág. 2) para una muestra de datos en un problema de clasificación, [Vapnik 1982], [Vapnik 1998], [Schölkopff y Smola 2002], [Cristianini y Shawe-Taylor 2000], [Cristianini y Shawe-Taylor 2004], [Vert *et al.* 2004].

2. Los hiperplanos separadores sufren de dos debilidades en muchos problemas de aplicación: por una parte, exigen la separabilidad perfecta de la muestra y, por otra, poseen carácter lineal.

3. Las redes **support vector machines** se desarrollan para evitar estas limitaciones, siendo sus hitos históricos fundamentales, los siguientes:

   a) La creación del algoritmo *Soft-Margin* [Cortes y Vapnik 1995], para enfrentarse a muestras no separables o a problemas donde no interesa la separabilidad perfecta (p. ej. ruido en las observaciones).

   b) La transformación previa del espacio de entrada $\mathcal{X} \subset \mathbb{R}^d$ en un nuevo espacio (*feature espace*) de mayor dimensión, $\mathcal{U} = \phi(\mathcal{X}) \subset \mathbb{R}^s$, en el que las fronteras lineales de los hiperplanos separadores dan lugar, mediante la transformación inversa, a fronteras no lineales en el espacio de entrada, (*Kernel Trick* [Boser *et al.* 1992]).

   c) La generalización de las SVM a problemas de regresión mediante la utilización de la pérdida ε**-insensible** de Vapnik, [Drucker *et al.* 1997].

Hiperplano Separador de Máximo Margen

1. Nos centramos en el problema de clasificación con dos clases con codificación $\mathcal{Y} = \{-1, 1\}$ que se pretende resolver mediante:

   a) Una función discriminante lineal $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \bar{\mathbf{w}}^T\bar{\mathbf{x}}$ que define un hiperplano separador[1] $\tau_{\mathbf{w},b} \equiv \{\mathbf{x} \in \mathcal{X} : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$

   b) Una regla de clasificación $g_{f_{\bar{\mathbf{w}}}}(\mathbf{x}) = \text{signo}(f_{\bar{\mathbf{w}}}(\mathbf{x})) \in \{-1, 1\}$.

[1] Si $\mathbf{w}, \mathbf{x} \in \mathbb{R}^d$,el producto interno entre $\mathbf{w}$ y $\mathbf{x}$ se denota indistintamente por $\mathbf{w}^T\mathbf{x}$ o por $\langle \mathbf{w}, \mathbf{x} \rangle$. Al tratar los hiperplanos de máximo margen (y más adelante las SVM) es común utilizar la notación $\langle \mathbf{w}, \mathbf{x} \rangle$ con objeto de manejar mejor sus propiedades de operador lineal: $\left\langle \sum_{i=1}^n \alpha_i \mathbf{x}_i, \mathbf{x} \right\rangle = \sum_{i=1}^n \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle$.

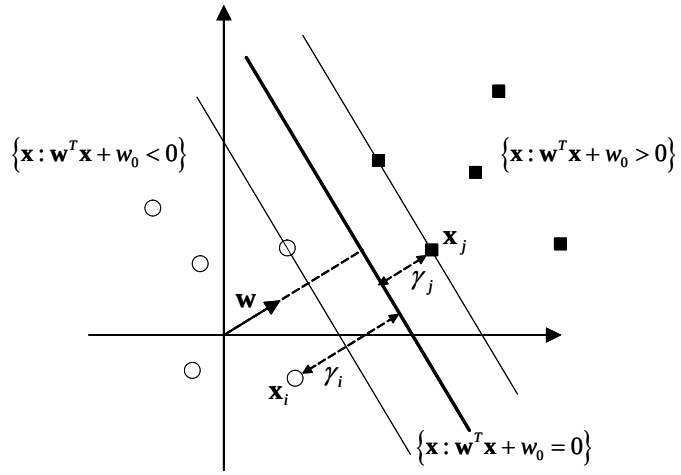

Figura 1: Muestra separable, hiperplano separador y margen geométrico γ_i de una observación $(\mathbf{x}_i, y_i)$

Margen

1. **Distancia de un punto a un hiperplano**.

   a) Si $\mathbf{w}^T\mathbf{x}_i + b = 0$, entonces el hiperplano $\tau_{\mathbf{w},b} \equiv \{\mathbf{x} \in \mathcal{X} : \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}$ pasa por $\mathbf{x}_i$.

   b) La distancia desde $\mathbf{x}_i$ al hiperplano $\{\mathbf{w}^T\mathbf{x} + b = 0\}$ será la distancia entre $\mathbf{x}_i$ y el punto $\mathbf{x} = \mathbf{x}_i + \lambda\mathbf{w}$ del hiperplano que, por tanto, verifica:

$$\begin{aligned} 0 &= \mathbf{w}^T\mathbf{x} + b = \mathbf{w}^T(\mathbf{x}_i + \lambda\mathbf{w}) + b = \mathbf{w}^T\mathbf{x}_i + \lambda\mathbf{w}^T\mathbf{w} + b \\ \iff \lambda &= \frac{-(\mathbf{w}^T\mathbf{x}_i + b)}{\mathbf{w}^T\mathbf{w}} = \frac{-(\mathbf{w}^T\mathbf{x}_i + b)}{\|\mathbf{w}\|^2} \end{aligned}$$

   Sustituyendo ese valor de λ, dicha distancia será:

$$\mathsf{d}(\mathbf{x}_i, \mathbf{x}) = \|\mathbf{x} - \mathbf{x}_i\| = \|\lambda\mathbf{w}\| = |\lambda| \, \|\mathbf{w}\| = \frac{1}{\|\mathbf{w}\|} \left|\mathbf{w}^T\mathbf{x}_i + b\right|$$

   Definición 1 *Una muestra* $\mathbf{z}^n$ *se dice* ***linealmente separable*** *si verifica:* $y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i) \geq 0, \forall i \in \{1 : n\}$ *siendo* $f_{\bar{\mathbf{w}}}(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$*, en cuyo caso, el hiperplano* $\{\mathbf{w}^T\mathbf{x} + b = 0\}$ *se denomina* ***hiperplano separador****. En este marco, se denomina:*

   a) ***Margen funcional*** *de la observación* $(\mathbf{x}_i, y_i)$ *con respecto al hiperplano* $\{\mathbf{w}^T\mathbf{x}_i + b = 0\}$*, a la cantidad:*

$$\chi_i = \left|\mathbf{w}^T\mathbf{x}_i + b\right| = y_i(\mathbf{w}^T\mathbf{x}_i + b) = y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i)$$

   b) ***Margen geométrico*** *de la observación* $(\mathbf{x}_i, y_i)$ *con respecto al hiperplano* $\{\mathbf{w}^T\mathbf{x}_i + b = 0\}$*, a la distancia euclídea entre* $\mathbf{x}_i$ *y el hiperplano, es decir:*

$$\gamma_i = \frac{1}{\|\mathbf{w}\|} \left|\mathbf{w}^T\mathbf{x}_i + b\right| = \frac{1}{\|\mathbf{w}\|} y_i(\mathbf{w}^T\mathbf{x}_i + b) = \frac{1}{\|\mathbf{w}\|} y_i f_{\bar{\mathbf{w}}}(\mathbf{x}_i) = \frac{1}{\|\mathbf{w}\|}\chi_i$$

2. Así, si el vector $\mathbf{w}$ está normalizado ($\|\mathbf{w}\| = 1$), el margen funcional coincide con el margen geométrico.

Definición 2 *El **margen** (o **margen geométrico**) del hiperplano $\tau_{\mathbf{w},b}$ **con respecto a la muestra** $\mathbf{z}^n$, se define como:*

$$\gamma(\tau_{\mathbf{w},b}) = \min_{i\in\{1:n\}} \frac{1}{\|\mathbf{w}\|} |\langle \mathbf{w}, \mathbf{x}_i\rangle + b| = \frac{1}{\|\mathbf{w}\|} \min_{i\in\{1:n\}} |\langle \mathbf{w}, \mathbf{x}_i\rangle + b| = \frac{1}{\|\mathbf{w}\|}\chi(\tau_{\mathbf{w},b}) \tag{1}$$

*donde $\chi(\tau_{\mathbf{w},b})$ se denomina **margen funcional**, que depende del tamaño de $\mathbf{w}$.*

Algoritmo de Entrenamiento del Hiperplano Separador de Máximo Margen

1. El hiperplano separador de máximo margen se obtiene como solución del siguiente programa de optimización:
$$\max_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \left\{ \gamma(\tau_{\mathbf{w},b}) = \min_{i\in\{1:n\}} |\langle \mathbf{w}, \mathbf{x}_i\rangle + b| \right\}$$
$$\text{sujeto a: } \|\mathbf{w}\| = 1$$

2. Alternativamente, en lugar de normalizar el vector $\mathbf{w}$ en la expresión (1), puede normalizarse el margen funcional, $\chi(\tau_{\mathbf{w},b}) = 1$ (en cuyo caso se dice que el hiperplano está en **forma canónica**), con lo que el problema de optimización persigue maximizar $1/\|\mathbf{w}\|$ con la restricción $\chi(\tau_{\mathbf{w},b}) = 1$, es decir:
$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 \tag{2}$$
$$\text{sujeto a: } y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) \geq 1, i = 1, \ldots, n$$

3. Ahora el margen en la muestra es:
$$\gamma(\tau_{\mathbf{w},b}) = \frac{1}{2}\left(\frac{\langle \mathbf{w}, \mathbf{x}_i^+\rangle + b}{\|\mathbf{w}\|} - \frac{\langle \mathbf{w}, \mathbf{x}_i^-\rangle + b}{\|\mathbf{w}\|} \right) = \frac{1}{\|\mathbf{w}\|}$$
donde $\mathbf{x}_i^+$ son los puntos de la clase $+1$ y $\mathbf{x}_i^-$ los de la clase -1 que saturan la restricción del problema anterior: $y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) = 1$

4. El problema anterior es cuadrático con restricciones lineales (convexo) con lo que **existe solución única** y las condiciones de óptimo de (Kuhn-Tucker) son necesarias y suficientes.

5. La solución puede obtenerse a través del **problema dual**, y resulta combinación lineal de un subconjunto SV de puntos de la muestra, denominados **vectores soporte** (*support vectors*) pues la solución se expresa en términos de dichos puntos, ([Schölkopff y Smola 2002], [Cristianini y Shawe-Taylor 2000], [Vapnik 1998]):
$$\begin{aligned} \mathbf{w} &= \sum_{\mathbf{x}_i\in \text{SV}} \beta_i \mathbf{x}_i \\ &\Rightarrow \quad f_{\mathbf{w},b}(\mathbf{x}) = \sum_{\mathbf{x}_i\in \text{SV}} \beta_i \langle \mathbf{x}_i, \mathbf{x}\rangle + b \end{aligned} \tag{3}$$

6. Con ello, dado un nuevo punto $\mathbf{x}$ a clasificar, la solución se obtiene calculando el producto interno con los puntos de la muestra.

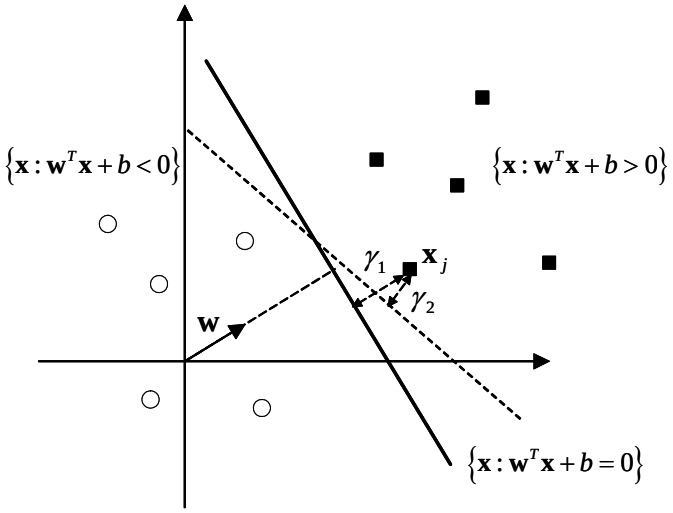

Figura 2: El hiperplano en línea continua produce un mayor margen γ_1 en la muestra que el representado con linea discontinua con margen γ_2.

Hiperplano Separador Soft-Margin

1. Cuando la muestra no es linealmente separable, o no interesa su separación perfecta, (lo que es frecuente en la práctica, p. ej. debido a que la población no lo sea –solapamiento de las clases– o si los datos están sujetos a ruido), es preciso admitir un conjunto de observaciones mal clasificadas, (enfoque *soft margin* en contraposición del anterior, denominado *hard margin*).

2. Esto se logra formulando variables *slack*, $\xi_i \geq 0$, $i = 1 : n$, que permiten la violación de las restricciones:
$$y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \xi_i \geq 1,\ i = 1 : n$$

3. En este punto, aparece un conflicto de intereses entre un número mínimo de observaciones mal clasificadas y una clasificación lo más robusta posible. Este conflicto puede formularse como [Cortes y Vapnik 1995] (compárese con el problema (2)):
$$\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} \{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i\} \\ \left\{ \begin{array}{ll} y_i(\langle \mathbf{w}, \mathbf{x} \rangle + b) \geq 1 - \xi_i & \\ \xi_i \geq 0 & i = 1, \ldots, n \end{array} \right. \end{array} \tag{4}$$
donde el parámetro C expresa la importancia asignada a los casos mal clasificados.

4. Nótese que el problema (4) no es una formulación extraña pues equivale al problema:
$$\min_{\mathbf{w},b,\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \ell(y_i, f_{\mathbf{w},b}(\mathbf{x}_i)) \right\} \equiv \min_{\mathbf{w},b,\boldsymbol{\xi}} \left\{ \sum_{i=1}^{n} \ell(y_i, f_{\mathbf{w},b}(\mathbf{x}_i)) + \frac{1}{2C}\|\mathbf{w}\|^2 \right\}$$
que equivale a un principio inductivo basado en el riesgo empírico regularizado, utilizando como pérdida:
$$\ell(y, f(\mathbf{x})) = [1 - yf(\mathbf{x})]_+ \doteq \left\{ \begin{array}{ll} 1 - yf(\mathbf{x}) & \text{si } 1 - yf(\mathbf{x}) \geq 0 \\ 0 & \text{en caso contrario} \end{array} \right. \tag{5}$$

5. De nuevo, el programa (4) es un programa cuadrático con restricciones lineales, cuya solución puede obtenerse mediante el programa dual, y es de forma análoga a (3).

6. La selección de C es un problema de selección del modelo, y se puede realizar por ejemplo mediante validación cruzada.

Support Vector Machines

El Método del Núcleo (*The Kernel Trick*)

1. El problema del hiperplano separador de máximo margen es que posee carácter lineal cuando, en la práctica, este tipo de modelo puede no ser suficiente.

   Una táctica para conseguir fronteras no lineales utilizando técnicas lineales es utilizar una transformación no lineal del espacio de entrada (al modo de la regresión polinómica o de las redes neuronales con capa oculta), para aplicar métodos lineales en el espacio transformado (**espacio de características** o *features*) de tal forma que, sus soluciones en dicho espacio den lugar, mediante la transformación inversa, a soluciones no lineales en el espacio original.

2. Las redes neuronales basan su potencial en la misma táctica. En dichas redes, es usual que $h < d$ (el número de nuevas variables suele ser menor que el número de variables originales), sin embargo, en el caso de las SVM es normal lo contrario pudiendo alcanzar el espacio de *features* dimensión incluso infinita.

3. Teniendo en cuenta que la solución de los hiperplanos separadores es de la forma (3):

$$f_{\mathbf{w},b}(\mathbf{x}) = \sum_{\mathbf{x}_i \in \mathrm{SV}} \beta_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b$$

   (que requiere calcular el producto interno con los puntos de la muestra), y que el producto interno es $\langle \mathbf{x}_i, \mathbf{x} \rangle = \sum_{j=1}^{d} (\mathbf{x}_i)_j (\mathbf{x})_j$ podría pensarse que la utilización de espacios intermedios de dimensión infinita impedirá la utilización de este tipo de técnicas.

4. La solución a este problema [Boser *et al.* 1992] reside en elegir una transformación:

$$\boldsymbol{\phi} : \mathcal{X} \subset \mathbb{R}^d \to \mathcal{U}$$

   con $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x}) \in \mathcal{U}$ posiblemente de dimensión infinita, de tal manera que su producto interno esté definido mediante una función κ definida positiva:

$$\left\langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \right\rangle = \sum_{i=1}^{\infty} \phi_i(\mathbf{x}) \phi_i(\mathbf{x}') = \kappa(\mathbf{x}, \mathbf{x}') \tag{6}$$

   de tal forma que dicho producto interno esté bien definido, donde $\phi_i(\mathbf{x})$ son las componentes de $\boldsymbol{\phi}(\mathbf{x})$.

Support Vector Machines para Clasificación

1. Utilizando el método anterior, planteamos el problema de clasificación en el nuevo espacio de mayor dimensión, donde ahora la muestra es $\{(y_i, \boldsymbol{\phi}_i)\}_{i=1}^{n}$ con $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\mathbf{x}_i)$, $i = 1 : n$.

$\kappa(\mathbf{x}, \mathbf{x}')$	**Observaciones**
$\langle \mathbf{x}, \mathbf{x}' \rangle$	$\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$. Modelo lineal
$\cos(x - x')$	
$\exp(-(x + x')^p), 0 < p \leq 1$	
$\exp(-\left\|\mathbf{x} - \mathbf{x}'\right\|^p), 0 < p \leq 2$	Gausiana, Laplaciana,...
$(c^2 + \left\|\mathbf{x} - \mathbf{x}'\right\|^2)^{-p}, p > 0$	Multicuádrica inversa
$\langle \mathbf{x}, \mathbf{x}' \rangle^p, p \in \mathbb{N}$	Polinómico homogéneo
$(\langle \mathbf{x}, \mathbf{x}' \rangle + c)^p, p \in \mathbb{N}$	Polinómico no homogéneo
$\cos(\langle \mathbf{x}, \mathbf{x}' \rangle)$	Coseno con producto interno

Cuadro 1: Diversas funciones definidas positivas que pueden utilizarse como núcleos en las SVM.

2. La obtención del hiperplano *soft-margin* (o el hiperplano de máximo margen, si se pretende una separación perfecta) en el nuevo espacio con la nueva muestra, la solución resulta:

$$\begin{aligned} \mathbf{w} &= \sum_{\mathbf{x}_i \in \text{SV}} \beta_i \boldsymbol{\phi}(\mathbf{x}_i) \\ &\Rightarrow \quad f_{\mathbf{w},b}(\boldsymbol{\phi}(\mathbf{x})) = \sum_{\mathbf{x}_i \in \text{SV}} \beta_i \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}) \rangle + b = \sum_{\mathbf{x}_i \in \text{SV}} \beta_i \kappa(\mathbf{x}_i, \mathbf{x}) + b \end{aligned} \tag{7}$$

que es la expresión general de las **support vector machines** para clasificación tanto en el caso *hard margin*, como en el caso *soft margin*, que proviene de la resolución de los problemas (2) o (4) respectivamente, en el espacio de características.

3. Por ejemplo, el programa del caso *soft margin* posee ahora la forma:

$$\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} \{\frac{1}{2} \left\|\mathbf{w}\right\|^2 + C \sum\limits_{i=1}^{n} \xi_i\} \\ \left\{ \begin{array}{l} y_i(\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i) \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{array} \right| \; i = 1, \ldots, n \end{array} \tag{8}$$

4. Con el método del núcleo, se resuelve el problema de la dimensionalidad del nuevo espacio sin necesidad de conocer específicamente dicho espacio o la transformación $\boldsymbol{\phi} : \mathcal{X} \subset \mathbb{R}^d \rightarrow \mathcal{U}$, pues basta el conocimiento de κ para construir la solución.

5. La tarea de modelización puede realizarse seleccionando directamente κ o bien la transformación $\boldsymbol{\phi}$, aunque lo primero es lo más usual pues un mismo núcleo κ puede resultar de diferentes transformaciones verificando la condición (6).

6. La tabla 1 muestra núcleos utilizados frecuentemente en las support vector machines pero es válida cualquier función definida positiva[2].

Support Vector Machines para Regresión

1. Utilizando el método del núcleo expuesto en el apartado anterior, planteamos ya el problema de regresión en el espacio de características.

[2]Si y sólo si la matriz $\mathbf{K}$ es semidefinida positiva con $(\mathbf{K})_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ para cualquier conjunto $\{\mathbf{x}_1, ...\mathbf{x}_n\}$. (Nota: una función de covarianza o covariograma $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \text{Cov}(\mathbf{Y}(\mathbf{x}_i), \mathbf{Y}(\mathbf{x}_j))$ satisface la misma condición).

2. En dicho espacio, la extensión al problema de regresión de la idea de *soft margin* equivale a permitir soluciones que no se ajusten exactamente a todos los puntos $\{(\boldsymbol{\phi}(\mathbf{x}_i), y_i)\}_{i=1}^n$ de la muestra transformada.

3. Esto puede formularse generalizando el problema (8) de la siguiente forma [Drucker *et al.* 1997]:

$$\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}'} \{\frac{1}{2}\|\mathbf{w}\|_{\mathcal{U}}^2 + C\sum_{i=1}^n(\xi_i+\xi_i')\} \\ \left\{ \begin{array}{c} \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + b - y_i \le \varepsilon + \xi_i \\ y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + b) \le \varepsilon + \xi_i' \\ \xi_i, \xi_i' \ge 0 \end{array} \right| \; i = 1, \ldots, n \end{array} \tag{9}$$

   donde $\mathbf{w} \in \mathcal{U}$, y $\xi_i, \xi_i' \in \mathbb{R}$ son variables *slack* que evitan que la solución tenga que contener en la banda de radio ε todos los puntos $(\mathbf{x}_i, y_i)$ de la muestra, como medio de defenderse de posibles *outliers* y evitar el sobreajuste.

4. Esta formulación equivale a la utilización de la **función de pérdida ε-insensible** ([Vapnik 1995], [Vapnik 1998]) con $p = 1, 2$:

$$\ell(y, f(\mathbf{x})) = |y - f(\mathbf{x})|_\varepsilon^p = \text{máx}\{0, (|y - f(\mathbf{x})| - \varepsilon)^p\} \tag{10}$$

   que permite ver mejor el principio inductivo utilizado, basado en el riesgo empírico regularizado. Para $p = 1$:

$$\min_{\mathbf{w},b}\{\frac{1}{2}\|\mathbf{w}\|_{\mathcal{U}}^2 + C\sum_{i=1}^n |y_i - f_{\mathbf{w},b}(\mathbf{x}_i)|_\varepsilon\} \tag{11}$$

   donde $f_{\mathbf{w},b}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle_{\mathcal{U}} + b$. (Pero este problema es más difícil de resolver que (9), debido a la función $|\cdot|_\varepsilon$).

5. La solución al problema (9) se obtiene formulando el Lagrangiano con respecto a las variables primales $\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}'$ y duales $\boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\rho}, \boldsymbol{\rho}'$:

$$\begin{aligned} L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}', \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\rho}, \boldsymbol{\rho}') &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n(\xi_i+\xi_i') + \sum_{i=1}^n \alpha_i[\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b - y_i - \varepsilon - \xi_i] \\ &\quad + \sum_{i=1}^n \alpha_i'[y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b) - \varepsilon - \xi_i'] - \sum_{i=1}^n \rho_i\xi_i - \sum_{i=1}^n \rho_i'\xi_i' \end{aligned}$$

6. Las condiciones de óptimo para las variables primales son:

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^n(\alpha_i' - \alpha_i)\boldsymbol{\phi}(\mathbf{x}_i) = 0 \\ \frac{\partial L}{\partial b} &= -\sum_{i=1}^n(\alpha_i' - \alpha_i) = 0 \\ \frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \rho_i = 0, \; i = 1:n \\ \frac{\partial L}{\partial \xi_i'} &= C - \alpha_i' - \rho_i' = 0, \; i = 1:n \end{aligned}$$

que, sustituyéndolas en el Lagrangiano, producen el siguiente problema dual:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\alpha}'\in\mathbb{R}^n}\{-\tfrac{1}{2}\sum_{i,j=1}^{n}(\alpha_i'-\alpha_j)(\alpha_j'-\alpha_j)\kappa(\mathbf{x}_i,\mathbf{x}_j)-\varepsilon\sum_{i=1}^{n}(\alpha_i'+\alpha_i)+\textstyle\sum_{i=1}^{n}(\alpha_i'-\alpha_i)y_i \quad \left\{\begin{array}{c}\displaystyle\sum_{i=1}^{n}(\alpha_i'-\alpha_i)=0\\ 0\leq\alpha_i,\alpha_i'\leq C,\ i=1:n\end{array}\right. \tag{12}$$

con las siguientes condiciones adicionales de complementariedad de Kuhn-Tucker:

$$\left.\begin{array}{l}\alpha_i[\langle\mathbf{w},\boldsymbol{\phi}(\mathbf{x}_i)\rangle+b-y_i-\varepsilon-\xi_i]=0\\ \alpha_i'[y_i-(\langle\mathbf{w},\boldsymbol{\phi}(\mathbf{x}_i)\rangle+b)-\varepsilon-\xi_i']=0\\ \xi_i\xi_i'=0,\ \alpha_i\alpha_i'=0\\ (C-\alpha_i)\xi_i=0,\ (C-\alpha_i')\xi_i'=0\end{array}\right|, i=1:n \tag{13}$$

pues $0=\xi_i\rho_i=\xi_i(C-\alpha_i)$ y $0=\xi_i'\rho_i'=\xi_i'(C-\alpha_i')$.

7. **Observaciones**. Sobre la solución del programa (12)-(13), destacamos lo siguiente haciendo $\beta_i=\alpha_i'-\alpha_i$, (el problema de clasificación en sus versiones *hard* y *soft margin* admite observaciones similares):

   a) El óptimo $\mathbf{w}$ posee la forma: $\mathbf{w}=\sum_{i=1}^{n}\beta_i\boldsymbol{\phi}_i$ con $\boldsymbol{\phi}_i=\boldsymbol{\phi}(\mathbf{x}_i)$, $i=1:n$ al igual que en el problema de clasificación (ahora en el espacio de características).

   b) Los vectores que satisfacen $\beta_i=\alpha_i'-\alpha_i\neq 0$, ($\alpha_i>0$ ó $\alpha_i'>0$), es decir, los que saturan las dos primeras restricciones en (13), son los únicos que participan en la expresión de la solución.

      1) Por ello, se denominan **vectores soporte** (*support vectors*) y son los vectores realmente relevantes o *difíciles* para el problema).
      2) De ellos, los que verifican $0<\alpha_i<C$ o $0<\alpha_i'<C$ están ubicados justo en la frontera de la banda de radio ε, y el resto, $\alpha_i=C$ o $\alpha_i'=C$, quedan fuera de esa banda contabilizados como errores por la función de pérdida.

   c) Con ello, el óptimo es combinación lineal de los vectores soporte $\mathbf{w}=\sum_{\text{s.v.}}\beta_i\boldsymbol{\phi}_i$ y la solución al problema (12) viene dada por:

   $$f_{\mathbf{w},b}(\mathbf{x})=\langle\mathbf{w},\boldsymbol{\phi}(\mathbf{x})\rangle+b=\sum_{\text{s.v.}}\beta_i\,\langle\boldsymbol{\phi}(\mathbf{x}_i),\boldsymbol{\phi}(\mathbf{x})\rangle_{\mathcal{U}}+b=\sum_{\text{s.v.}}\beta_i\kappa(\mathbf{x}_i,\mathbf{x})+b \tag{14}$$

   que es la forma general de las SVM para regresión.

   d) La solución obtenida se expresa únicamente en términos de los vectores realmente relevantes o *difíciles* para el problema (los vectores soporte): depende de la dificultad intrínseca del problema a resolver y no, por ejemplo, del tamaño de la muestra o de la dimensión del espacio de entrada.

   e) Una ventaja de lo anterior, es que la solución sería la misma si se eliminan de la muestra todos los puntos que no resultan relevantes, lo que permite adoptar tácticas de entrenamiento progresivo añadiendo datos sucesivamente y conservando únicamente los vectores soporte de las etapas anteriores.

8. La figura 3 muestra el ajuste de cuatro redes SVM gausianas en un problema de regresión. En cada subfigura puede observarse la banda $\pm\varepsilon$ centrada en la estimación proporcionada por cada red, así como los vectores soporte de la misma, identificados con estrellas.

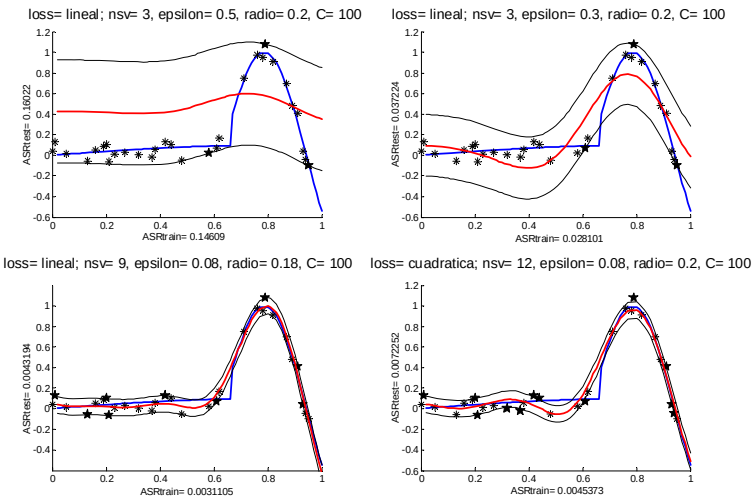

Figura 3: De izquierda a derecha y de arriba a abajo: redes SVM entrenadas con pérdida ε-insensible lineal para $\varepsilon = 0{,}5$, $0{,}3$ y $0{,}08$, respectivamente, y red SVM entrenada con pérdida ε-insensible cuadrática con $\varepsilon = 0{,}08$ y demás parámetros iguales a la tercera de las anteriores.

a) Puede observarse la influencia del parámetro ε en la complejidad del estimador final: las tres primeras redes se entrenaron utilizando la pérdida ε-insensible lineal con $\varepsilon = 0{,}5$, $0{,}3$ y $0{,}08$ (este último fué el valor óptimo sobre una muestra de validación), obteniéndose respectivamente 3, 3 y 9 vectores soporte, lo que muestra la influencia inversa de ε sobre dicho número.

b) La última red se entrenó con pérdida ε-insensible cuadrática y los demás parámetros iguales a la última de las anteriores ($\varepsilon = 0{,}08$), obteniéndose 12 vectores soporte. Por tanto, esta red resultó menos parsimoniosa que la correspondiente obtenida con pérdida ε-insensible lineal, lo que suele ser el caso en general.

9. Finalmente, la figura 4 muestra la influencia de los hiperparámetros ε y σ (escala del núcleo), sobre el error en una muestra de validación (gráfico de la izquierda) y el número de vectores soporte de la solución (gráfico de la derecha).

   a) Al respecto del gráfico de la izquierda:

      1) La zona de mayores errores es la zona de infraajuste de la red que se corresponde con valores grandes de ε. La zona de sobreajuste no se muestra debido a su rápido crecimiento y se obtiene para valores muy pequeños de ambos parámetros, sobre todo de ε. En un rango de valores moderados de ambos parámetros, el parámetro ε posee mucha mayor influencia en el grado de complejidad de la red.
      2) A medida que ε es mayor, el valor óptimo del parámetro de escala del núcleo también se va haciendo mayor pues no es necesaria tanta complejidad en el modelo.

   b) El gráfico de la derecha de la figura 4 (nótese que la escala de los ejes del plano horizontal está invertida con respecto al gráfico de la derecha), muestra que:

      1) Los valores pequeños de ε producen un gran incremento en el número de vectores soporte debido al estrechamiento de la banda $\pm\varepsilon$. Si este parámetro es suficientemente grande, no habría vectores soporte.

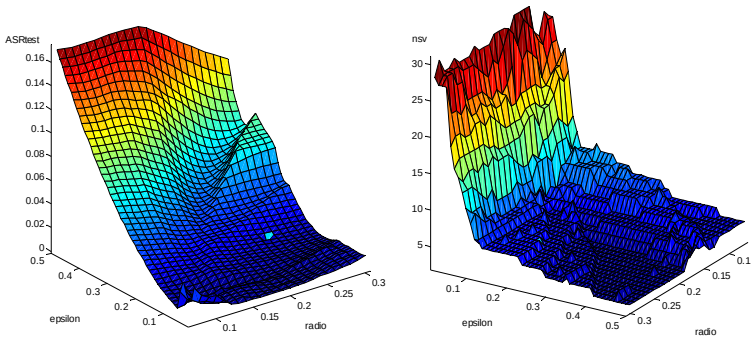

Figura 4: Izquierda: error en una muestra de validación en función del parámetro ε (izqda.) y del parámetro de escala del núcleo (dcha.). Los valores extremos, especialmente de ε, empeoran la capacidad de generalización. Derecha: número de vectores soporte de la solución en función de dichos parámetros. Ahora, es sobre todo el incremento de ε el que provoca el incremento de dicho valor. (Nótese que la escala de los ejes del plano horizontal de la segunda figura, está invertida con respecto a la primera (para mejor visualización)).

2) Para cada valor fijo de ε, se obtiene un mayor número de vectores soporte cuanto menor es el parámetro de escala del núcleo pues esto implica mayor complejidad en el modelo y con ello más violaciones de la banda $\pm\varepsilon$.

Casos Particulares

1. **Splines** (**Redes RBF de regularización**): si $p = 2$ y $\varepsilon = 0$ se tiene el problema:

$$\left.\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum\limits_{i=1}^{n}\xi_i^2 \right\} \\ y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b) = \xi_i,\ i = 1:n \end{array}\right\} \equiv \min_{\mathbf{w},b}\left\{ \sum_{i=1}^{n}\left(y_i - (\langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}_i)\rangle + b)\right)^2 + \frac{1}{2C}\|\mathbf{w}\|^2 \right\}$$

cuya solución es:

$$\begin{aligned} \hat{\mathbf{w}} &= \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \frac{1}{2C}\mathbf{I}_n\right)^{-1}\boldsymbol{\Phi}^T\mathbf{y} \Longleftrightarrow \boldsymbol{\Phi}^T\mathbf{y} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \frac{1}{2C}\mathbf{I}_n\right)\hat{\mathbf{w}} = \boldsymbol{\Phi}^T\boldsymbol{\Phi}\hat{\mathbf{w}} + \frac{1}{2C}\hat{\mathbf{w}} \\ \hat{\mathbf{w}} &= 2C\cdot\boldsymbol{\Phi}^T(\mathbf{y} - \boldsymbol{\Phi}\hat{\mathbf{w}}) = \boldsymbol{\Phi}^T\boldsymbol{\beta} = \sum_{\mathbf{i=1}}^{\mathbf{n}}\beta_i\boldsymbol{\phi}(\mathbf{x}_i) \text{ con } \boldsymbol{\beta} = 2C\cdot(\mathbf{y} - \boldsymbol{\Phi}\hat{\mathbf{w}}) \\ &\Rightarrow \quad \hat{f}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x})\rangle + b = \sum_{i=1}^{n}\beta_i\kappa(\mathbf{x},\mathbf{x}_i) + b \end{aligned}$$

donde todos los vectores son vectores soporte (debido a que $\varepsilon = 0$).

2. **Kriging** y **Procesos Gausianos**. De lo anterior se deduce:

$$\begin{aligned} \frac{1}{2C}\boldsymbol{\beta} &= (\mathbf{y} - \boldsymbol{\Phi}\hat{\mathbf{w}}) = (\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\beta}) = \mathbf{y} - \mathbf{K}\boldsymbol{\beta} \\ (\mathbf{K}+\frac{1}{2C}\mathbf{I}_n)\boldsymbol{\beta} &= \mathbf{y} \Longleftrightarrow \boldsymbol{\beta} = (\mathbf{K}+\frac{1}{2C}\mathbf{I}_n)^{-1}\mathbf{y} \end{aligned}$$

donde $\boldsymbol{\Phi}\boldsymbol{\Phi}^T = \mathbf{K}$ es la **matriz Gram**. Así, si $b = 0$:

$$\hat{y}(\mathbf{x}) = \hat{f}(\mathbf{x}) = \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle = \sum_{i=1}^{n} \beta_i \kappa(\mathbf{x}, \mathbf{x}_i) = \boldsymbol{\kappa}_{\mathbf{x}}^T \boldsymbol{\beta} = \boldsymbol{\kappa}_{\mathbf{x}}^T (\mathbf{K}+\frac{1}{2C}\mathbf{I}_n)^{-1}\mathbf{y}$$

que coincide con la predicción del **kriging simple** cuando el núcleo κ coincide con el covariograma de la función aleatoria.

   a) Con $b \neq 0$, se tiene la expresión del **kriging ordinario**.

   b) Si en lugar de un término b, se incluye un término polinómico en el modelo (SVM semi-paramétricas de [Smola *et al.* 1999]) se obtiene el kriging universal.

3. **Regresión lineal por mínimos cuadrados regularizados (*ridge regression*)**: Si $\kappa(\mathbf{x}, \mathbf{x}_i) = \langle \mathbf{x}, \mathbf{x}_i \rangle$, es decir la transformación es la identidad $\boldsymbol{\phi}(\mathbf{x}) = \mathbf{x}$, y además $p = 2$ y $\varepsilon = 0$, el problema (9) resulta:

$$\left.\begin{array}{c} \min\limits_{\mathbf{w},b,\boldsymbol{\xi}} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i^2 \right\} \\ y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = \xi_i',\ i = 1:n \end{array}\right\} \equiv \min_{\mathbf{w},b} \left\{ \sum_{i=1}^{n} (y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b))^2 + \frac{1}{2C}\|\mathbf{w}\|^2 \right\}$$

4. **Regresión lineal por mínimos cuadrados**. Si $C = \infty$ se tiene la regresión usual por mínimos cuadrados, que también admite una expresión en términos del producto interno con los puntos de la muestra y donde todos los vectores son vectores soporte (debido a que $\varepsilon = 0$):

$$\begin{aligned} \hat{\mathbf{w}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-2}\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\boldsymbol{\beta} = \sum_{i=1}^{n}\beta_i\mathbf{x}_i \\ \Rightarrow\ \hat{f}(\mathbf{x}) &= \langle \mathbf{w}, \boldsymbol{\phi}(\mathbf{x}) \rangle + b = \sum_{i=1}^{n}\beta_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b \end{aligned}$$

Bibliografía

Boser B E, Guyon I M y Vapnik V [1992]. A Training Algorithm for Optimal Margin Classifiers, *en* D Haussler, ed., *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, ACM Press, Pittsburgh, PA, pp. 144–152.

Cortes C y Vapnik V [1995]. Support Vector Networks, *Machine Learning* **20**, 273–297.

Cristianini N y Shawe-Taylor J [2000]. *Support Vector Machines*, Cambridge University Press.

Cristianini N y Shawe-Taylor J [2004]. *Kernel Methods for Pattern Analysis*, Cambridge University Press.

Drucker H, Burges C, Kaufman L, Smola A y Vapnik V [1997]. Support Vector Regression Machines, *en* M Mozer, M Jordan y T Petsche, eds, *Advances in Neural Information Processing Systems*,, MIT Press, pp. 155–161.

Schölkopff B y Smola A J [2002]. *Learning with Kernels*, The MIT Press.

Smola A, Frieß J y Scholköpf B [1999]. Semiparametric Support Vector and Linear Programming Machines, *en* M S Kearns, S A Solla y D A Cohn, eds, *Advances in Neural Information Processing Systems*, MIT Press, pp. 585–591.

Vapnik V [1982]. *Estimation of Dependences Based on Empirical Data*, Springer.

Vapnik V [1995]. *The Nature of Statistical Learning Theory*, Springer, N.Y.

Vapnik V [1998]. *Statistical Learning Theory*, John Wiley.

Vert J P, Tsuda K y Schölkopf B [2004]. MIT Press, chapter A primer on kernel methods, pp. 35–70.