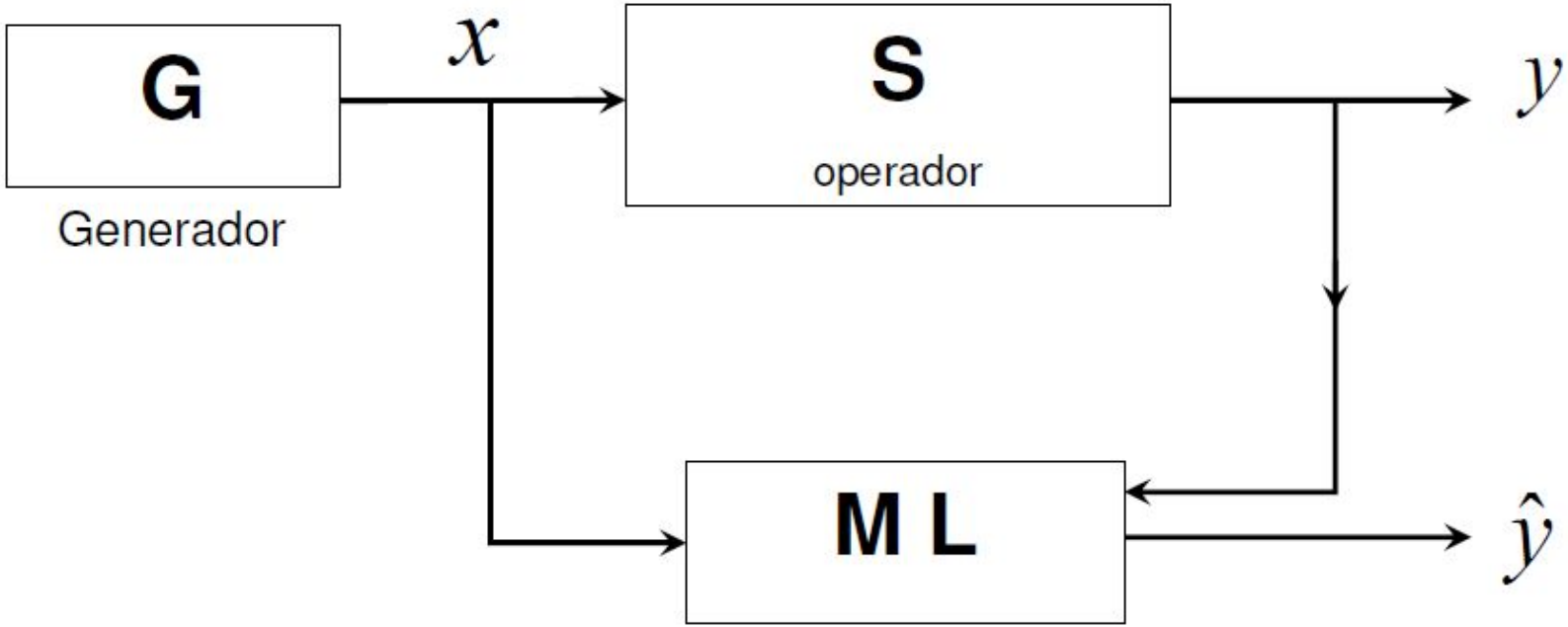
Introducción SVM

Erick López Ovando

El modelo de aprendizaje



Aprendizaje Supervisado

- Riesgo Funcional

$$R(\omega) = \int L(y, f(x, \omega))\, p(x, y)$$

- La solución que minimiza el funcional de Riesgo

$$f(x, \omega^*) = \arg\min_{\omega \in \Omega} R(\omega)$$

Aprendizaje Supervisado

- Riesgo Empírico

$$R_{emp}(\omega) = \frac{1}{l}\sum_{p=1}^{l} L(y_p, f(x_p, \omega))$$

- Riesgo Estructural

$$R_{srm} = R_{emp}(\omega) + \Phi(h_k, l)$$

Riesgo Estructural

- La función $\Phi(h_k, l)$ es directamente proporcional a la dimensión VC, que es una medida de la amplitud del espacio de aproximación.

$$R(\omega) \leq R_{emp}(\omega) + \Phi(h_k, l)$$

- La cota superior del Funcional de Riesgo (dado un espacio de hipótesis) es

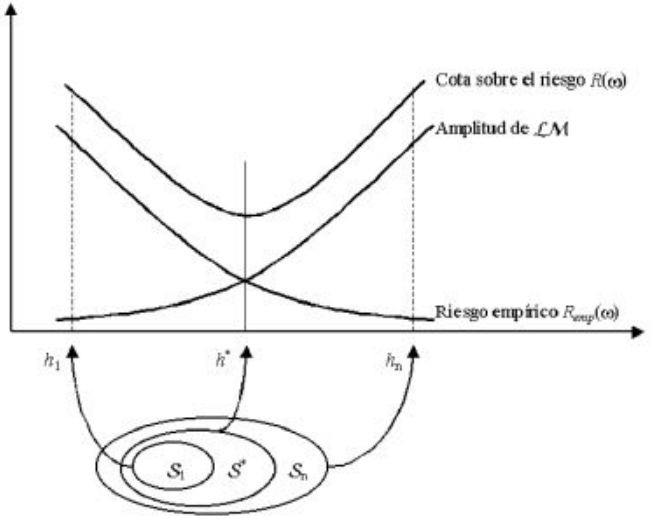
$$R_{emp}(\omega) + \sqrt{\frac{h(\log(2l/h)+1) - \log(\eta/4)}{l}}$$

con probabilidad $1-\eta$, donde h es la dimensión VC y l el tamaño de la muestra de entrenamiento.

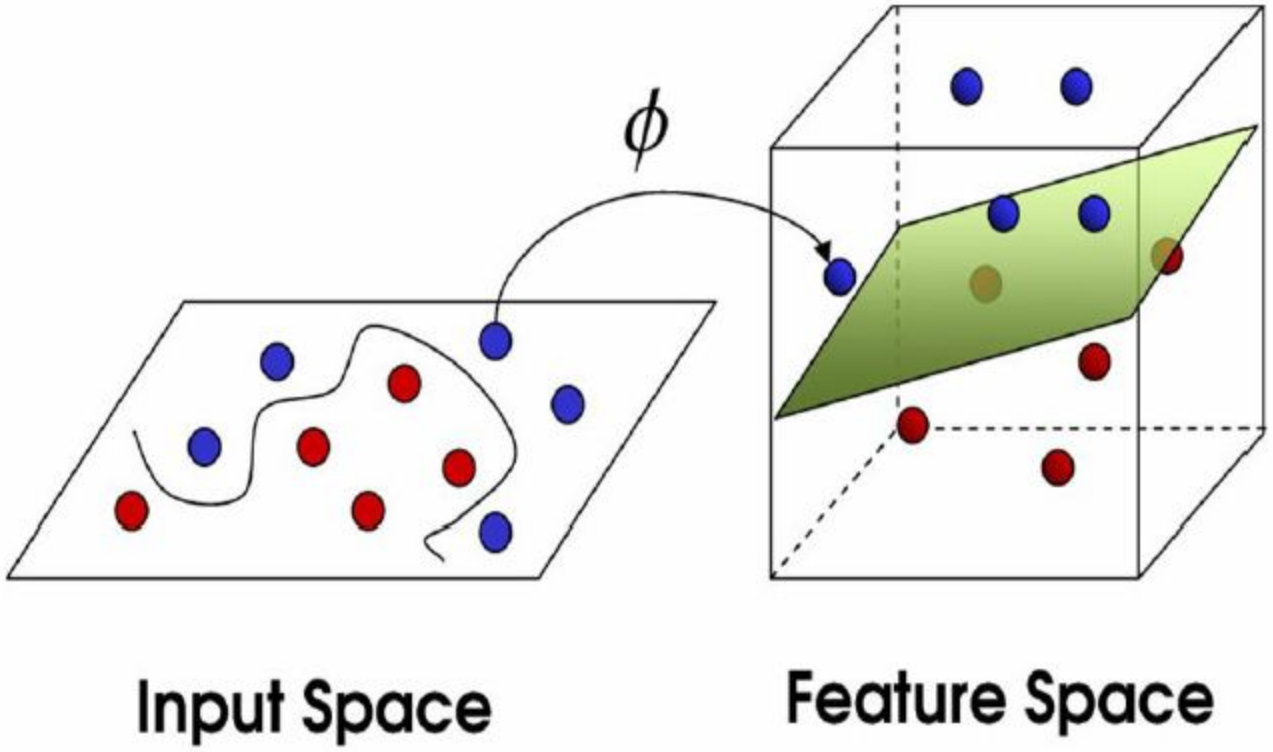
Minimización del Riesgo Estructural

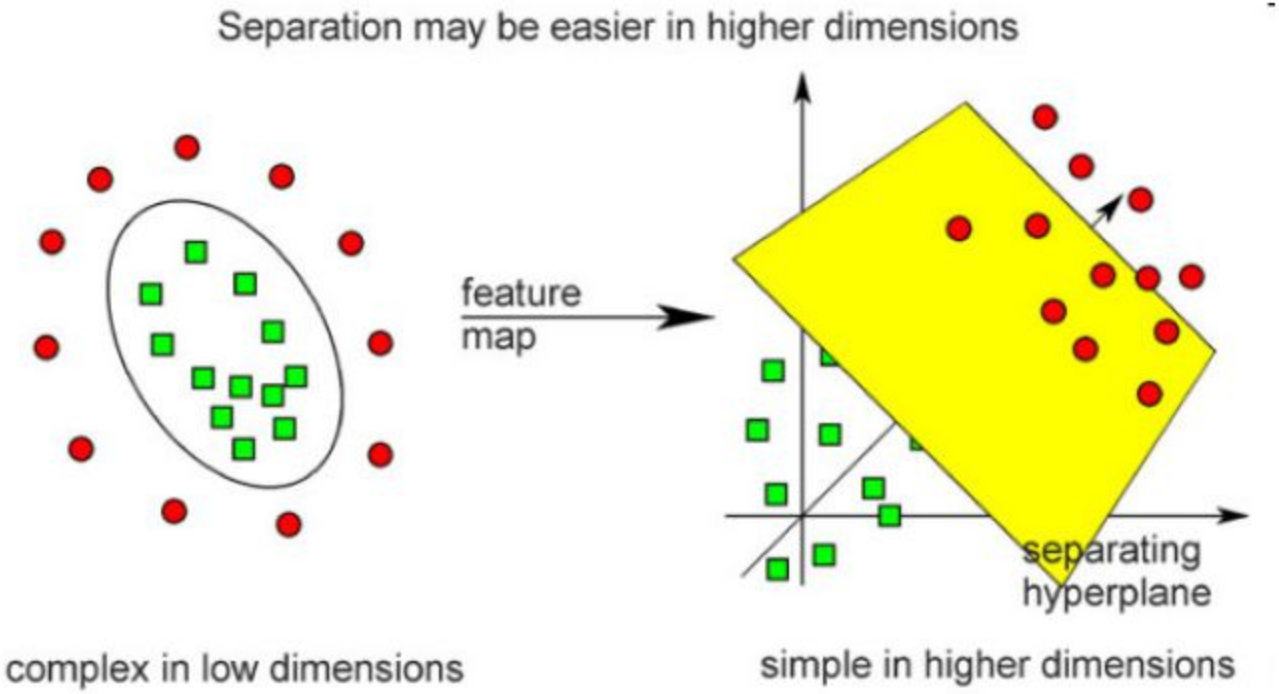
- Minimización del Riesgo Empírico (ERM): podemos interpretarlos como sistemas que tratan de reducir el error empírico

- Minimización del Riesgo Estructural (SRM): estudian el riesgo estructural en el espacio de hipótesis



Principle of Support Vector Machines (SVM)




Separation may be easier in higher dimensions
feature
map
separating
hyperplane
complex in low dimensions
simple in higher dimensions

Potencialidades

- Un gran potencial en problemas de clasificación
  - Muchas aplicaciones con excelentes resultados
  - Superando en varias ocasiones a un gran numero de maquinas, incluyendo a las Redes Neuronales Artificiales.
- Una de las grandes ventajas que tiene con el resto, es que no necesita usar todos los datos de la muestra
  - Tiene un gran desempeño incluso con pocos datos.

Potencialidades

- Se ha extendido al pronostico, a problemas de regresión
  - Al igual que en clasificación, con excelentes resultados.
- En la actualidad hay muchas propuestas para clasificación multi-clase, con buenos desempeños.
- La aplicabilidad que tiene en la actualidad abarca una gran gama de problemas en múltiples áreas.

Desventajas

- Existe un problema de optimización a resolver que en ocasiones el motor del equipo usado no soporta la ejecución.

- Hay que sintonizar parámetros que en ocasiones se complica más de lo esperado
  - En el caso de regresión, puede sobre-ajustarse con facilidad
  - Una SVM sólo sirve para el problema en particular que fue entrenada.

Maquinas de Vectores de Soporte

"No hay nada más practico que una buena teoría"

- Introducidas en los 90 por Vapnik
- Se basan en la Minimización del Riesgo Estructural (SRM)
- 92 – aparece el concepto de maximización del margen y uso de Kernels
- 95 – se propone el margen blando

Representación de los Datos

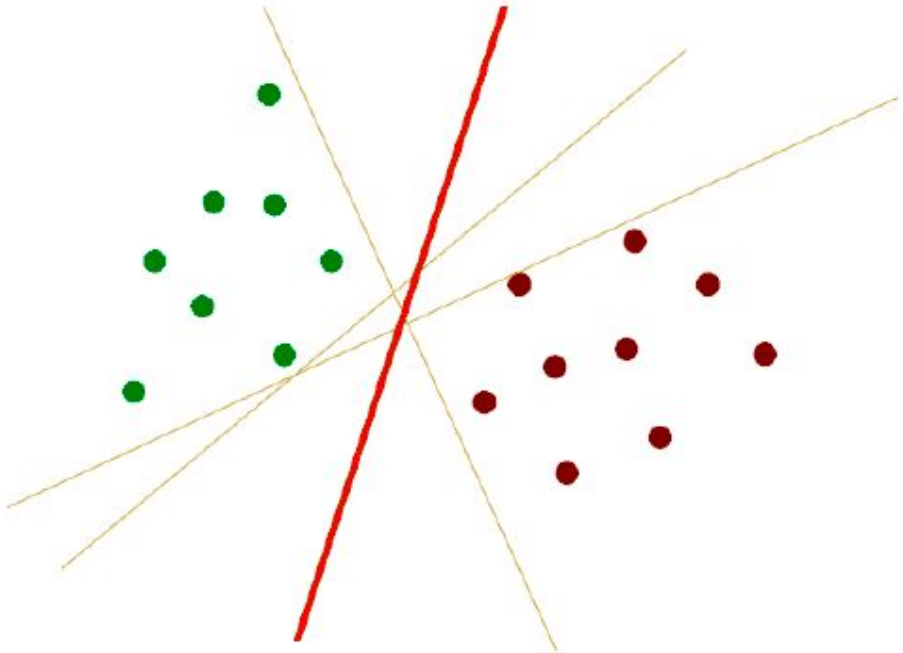
- Vector de D+1 dimensional (X, Y) donde

$$X = (x_1, x_2, \ldots, x_D)$$

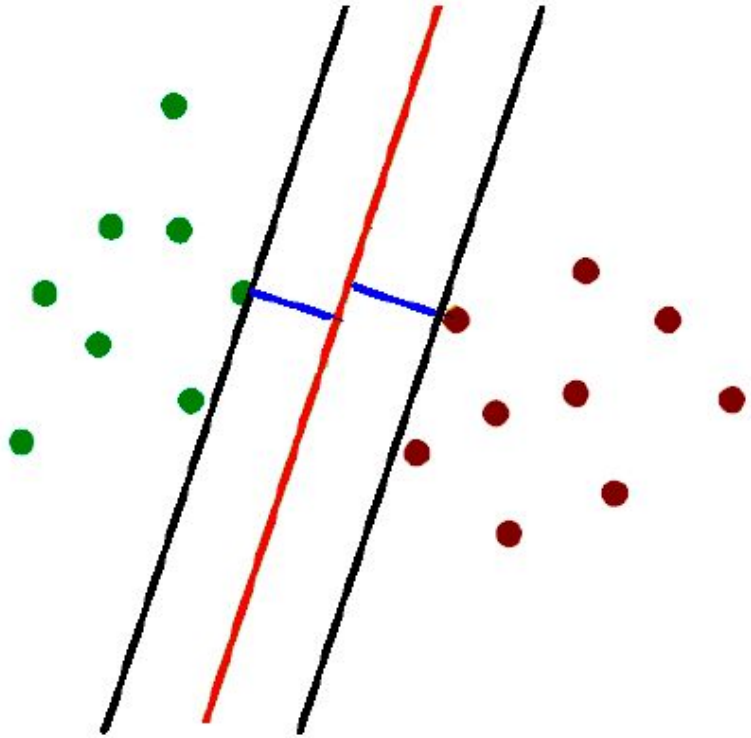
$$Y = \{\pm 1\}$$

- Un ***X*** tendrá asociado un label, la clase a la que pertenece

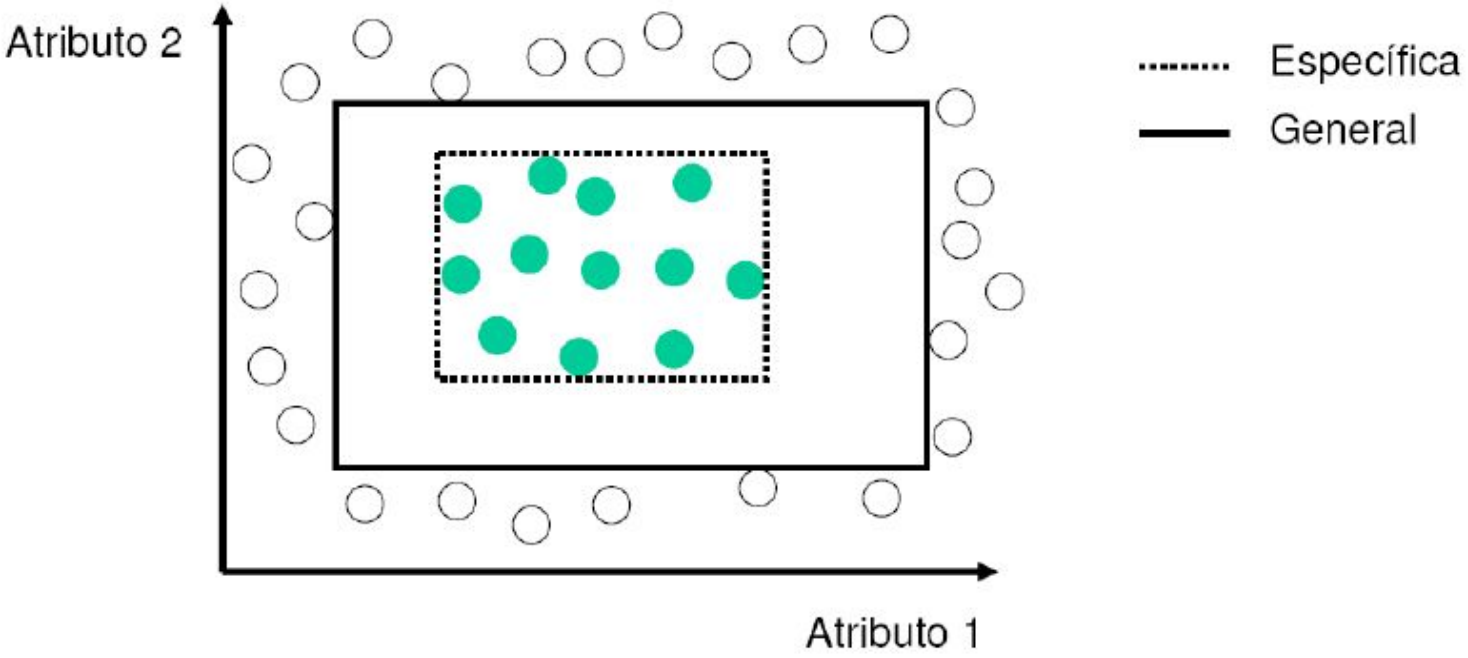
Separar los datos



Separación optima



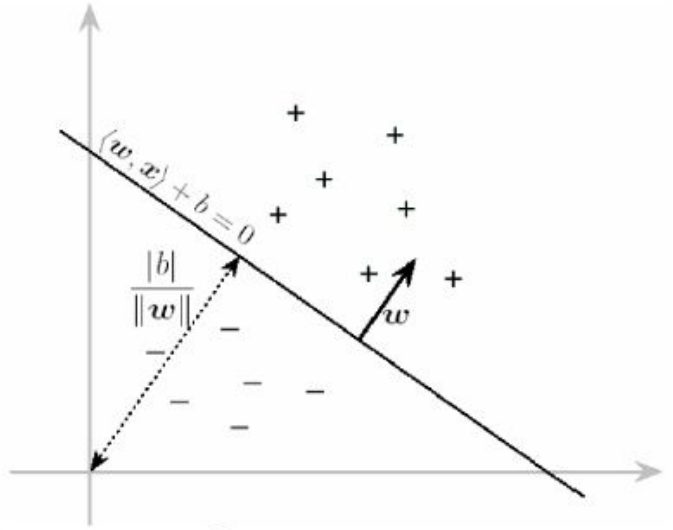
Margen General/Especifico



Planteamiento

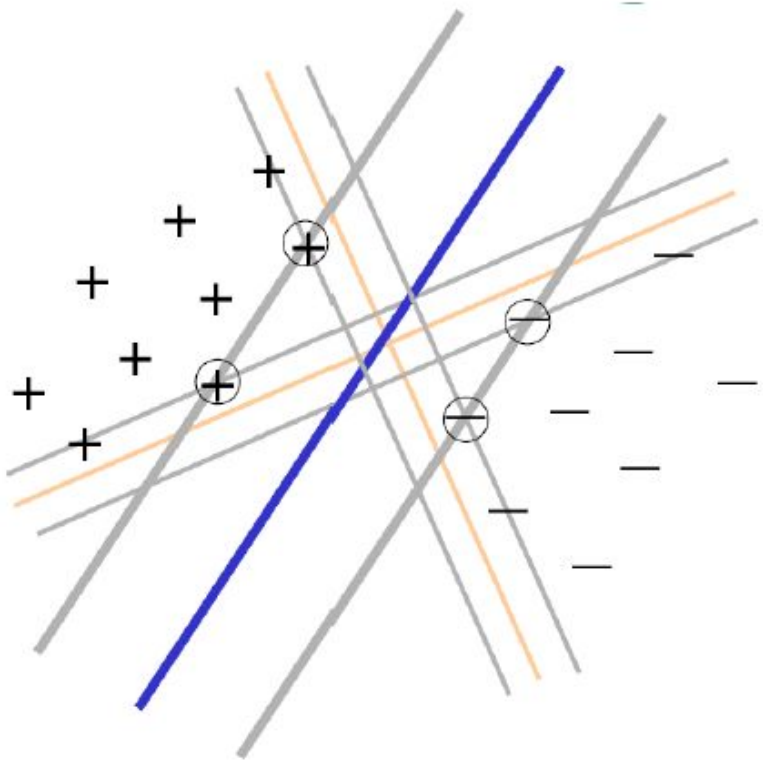
$$S = \{(x_1, y_1), \ldots, (x_n, y_n)\} \qquad x_i \in X,\ y_i \in \{-1, +1\}$$

$$f : X \to \Re \qquad f(x) = \langle \omega, x \rangle + b \qquad f(x_i) = \begin{cases} \geq 0 & \text{si } y_i \in +1 \\ < 0 & \text{si } y_i \in -1 \end{cases}$$



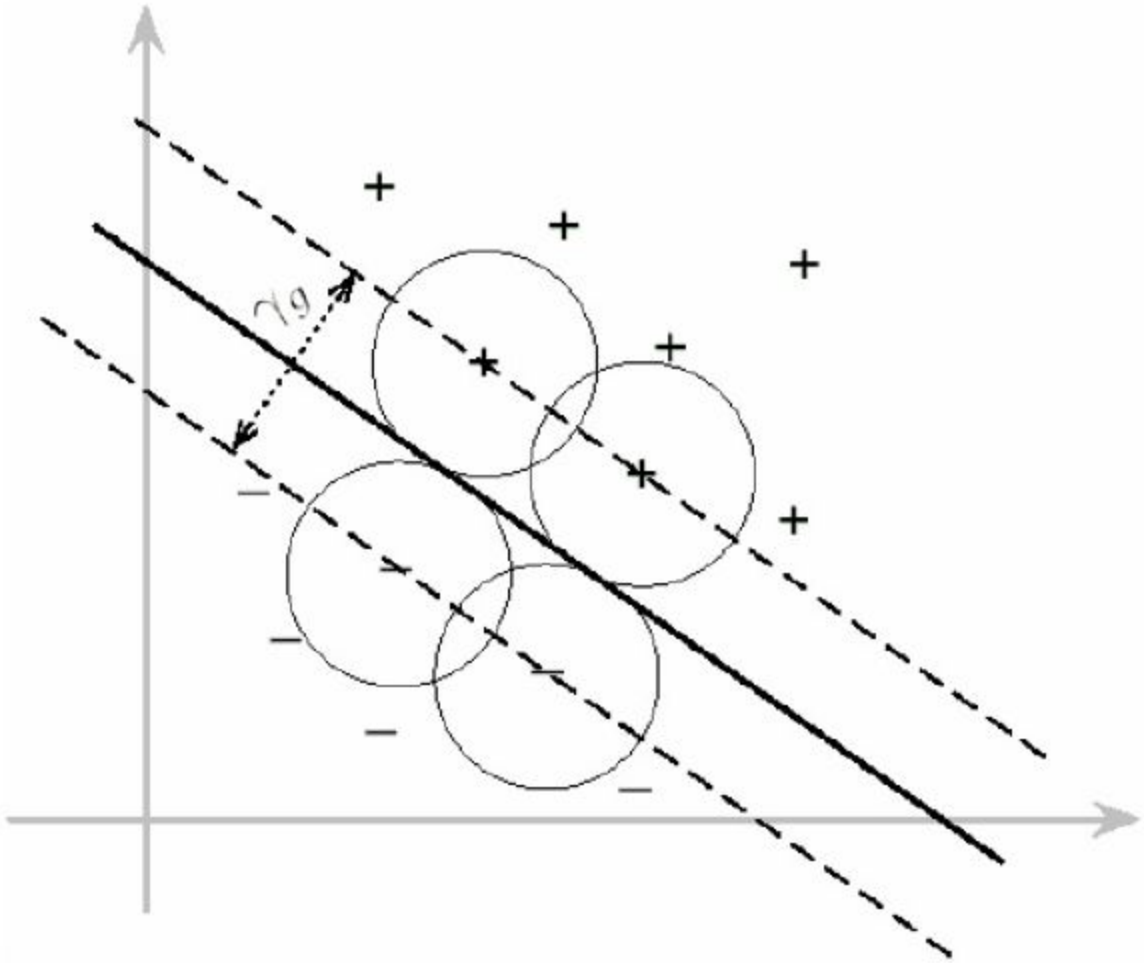
$$h(x) = \text{sgn}(\langle \omega, x \rangle + b)$$

Maximización del Margen



¿Por qué maximizar el margen?

- Resistencia al ruido en los datos de entrada
- Resistencia al error en el cálculo de la función de clasificación
- Propiedades matemáticas que permiten acotar de manera razonable el error de generalización


2δ

Margen funcional y geométrico

- **Margen funcional** es la menor diferencia entre aplicar la función a los ejemplos de la clase positiva y negativa

$$\gamma_f = \min_+(h(x_+)) - \max_-(h(x_-))$$

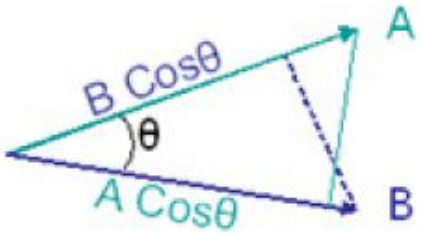
- **Margen geométrico** es la distancia entre los ejemplos de ambas clases, es decir, la suma de la distancia del hiperplano al ejemplo más próximo de cada clase

$$\gamma_g = \min_+(d(\omega, b; x_+)) + \min_-(d(\omega, b; x_-))$$

- Para maximizar uno de ellos debemos mantener fijo el otro. Si mantenemos fijo el funcional ($\|\omega\| = 1$) , podemos maximizar el geométrico

recordatorio

- Recordando algebra lineal, la distancia de cualquier punto hacia una recta (plano, n-dimensional) es el largo del vector que une el punto con la recta de forma perpendicularmente.



- Para obtener el largo de este vector se proyecta el vector del punto sobre el vector ortonormal de la recta, mediante el producto escalar de los vectores dividido por la norma del vector normal.

$$A \cdot B = |A||B|Cos\theta$$

- Si recuerdan la definición del producto escalar por medio del coseno, resulta mas fácil ver como ocurre la proyección.

$$\Pr oy_A B = \frac{A \cdot B}{|A|} \qquad \Pr oy_B A = \frac{A \cdot B}{|B|}$$

Maximizando el margen geométrico

$$\min_+(\langle \omega, x_+ \rangle + b) = +1$$

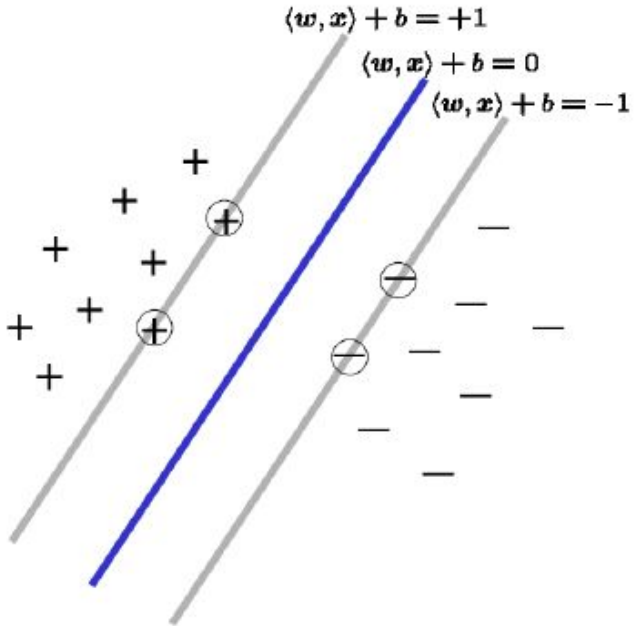
$$\max_-(\langle \omega, x_- \rangle + b) = -1$$

$$y_i(\langle \omega, x_i \rangle + b) \geq +1 \ , \quad i = 1, \ldots, n$$

$$\gamma_g = \min_+(d(\omega, b; x_+)) + \min_-(d(\omega, b; x_-))$$

$$= \min_+\left(\frac{|\langle \omega, x_+ \rangle + b|}{\|\omega\|}\right) + \min_-\left(\frac{|\langle \omega, x_- \rangle + b|}{\|\omega\|}\right)$$

$$= \frac{|+1|}{\|\omega\|} + \frac{|-1|}{\|\omega\|} = \frac{2}{\|\omega\|}$$



Maximizar el margen → minimizar la norma

- Resolviendo el problema a continuación, obtendremos el hiperplano de margen geométrico máximo que clasifica correctamente todos los ejemplos. Para resolverlo aplicaremos métodos conocidos de optimización de funciones

$$\min \quad \|\omega\|$$

$$s.a. \quad (\langle \omega, x_i \rangle + b) \geq +1 \qquad \forall x_i \in +1$$

$$(\langle \omega, x_i \rangle + b) \leq -1 \qquad \forall x_i \in -1$$

Optimización de funciones

- Problema primal

$$\begin{aligned} \min \quad & f(\omega) && \omega \in \Omega \\ s.a. \quad & g_i(\omega) \le 0 && i = 1, \ldots, k \end{aligned}$$

el objetivo es obtener los valores de las variables primales ω que minimizan la función objetivo f. Las solución está sujeta a que dichos valores respeten las restricciones de desigualdad g_i

- Programación Lineal : f y g_i lineales
- Programación Cuadrática : f cuadrática y g_i lineales
- Conjunto Admisible : todos los puntos del dominio que cumplen las restricciones
- Óptimo ω^* : $f(\omega^*) \le f(\omega)$ para otro ω del conjunto admisible

Convexidad: óptimos globales

- **Def 1:** Un dominio es convexo si y solo si el segmento de la recta que une cualquier par de puntos del dominio también está incluido en el dominio

$$\forall u, v \in \Omega, \forall \theta \in (0,1) \quad \text{entonces} \quad \theta u + (1-\theta) v \in \Omega$$

si el dominio es convexo, las restricciones lineales no eliminan la convexidad del conjunto admisible

- **Def 2:** Una función es convexa si

$$f(\theta v + (1-\theta) u) \leq \theta f(v) + (1-\theta) f(u), \qquad \theta \in (0,1)$$

- **Def 3:** Una función doblemente diferenciable es convexa si su matriz Hessiana es semidefinida positiva

- **Def 4:** Si tanto el dominio, como la función objetivo y las restricciones son convexas, entonces el problema se dice que es convexo

Convexidad: óptimos globales

- **Prop 1:** Si una función es convexa, entonces cualquier mínimo local es también global

Demo: Para cualquier $v \neq \omega^*$, por definición de mínimo local, existirá un θ suficientemente cerca de 1 tal que,

$$
\begin{aligned}
f(\omega^*) &\leq f(\theta\omega^* + (1-\theta)v) \\
f(\omega^*) &\leq \theta f(\omega^*) + (1-\theta) f(v) \\
(1-\theta) f(\omega^*) &\leq (1-\theta) f(v)
\end{aligned}
$$

$f(\omega^*) \leq f(v)$ para cualquier v y por tanto ω^* mínimo global

Teoría de Lagrange – función lagrangiano

- Dado el problema de optimización

$$\begin{aligned} \min \quad & f(\omega) && \omega \in \Omega \\ \text{s.a.} \quad & g_i(\omega) \le 0 && i = 1, \ldots, k \end{aligned}$$

se define la función lagrangiano como

$$L(\omega, \alpha) = f(\omega) + \sum_{i=1}^{k} \alpha_i g_i(\omega)$$

donde los α_i se denominan multiplicadores de Lagrange (o variables duales) y deben tener un valor no negativo. Indican la importancia de cada restricción.

Dualidad

- **Def 5:** El problema dual del problema primal planteado es:

$$\max \quad W(\alpha) = \inf_{\omega \in \Omega} L(\omega, \alpha) = \inf_{\omega \in \Omega} \; f(\omega) + \sum_{i=1}^{k} \alpha_i g_i(\omega)$$

$$\text{s.a.} \quad \alpha_i \geq 0$$

bajo ciertas condiciones, al resolver el problema dual (restricciones más simples) obtenemos también la solución del problema primal asociado.

Dualidad

- **Teorema :** Sea ω una solución admisible del problema primal y del dual, entonces $W(\alpha) \leq f(\omega)$

$$W(\alpha) = \inf_{\omega \in \Omega} L(\omega, \alpha) \leq L(\omega, \alpha) = f(\omega) + \sum \alpha_i g_i(\omega) \leq f(\omega)$$

- El valor del problema dual está acotado superiormente por el primal

$$\sup\{W(\alpha) : \alpha_i \geq 0\} \quad \leq \quad \inf\{f(\omega) : g(\omega) \leq 0\}$$

- Si $f(\omega^*) = W(\alpha^*)$ respetándose las restricciones, entonces ω^* y α^* son, respectivamente, las soluciones del primal y dual.

Condiciones de Karush-Kuhn-Tucker (KKT)

- **Teorema :** Dado el problema de optimización primal planteado, si es convexo, las condiciones necesarias y suficientes para que ω^* sea óptimo es que exista α^* tal que

$$\frac{\partial L(\omega^*, \alpha^*)}{\partial \omega} = 0$$

$$\alpha_i^* g_i(\omega^*) = 0, \qquad i = 1, \ldots, k$$

$$g_i(\omega^*) \leq 0, \qquad i = 1, \ldots, k$$

$$\alpha_i^* \geq 0, \qquad i = 1, \ldots, k$$

Condiciones KKT

- Los valores de las variables primales y duales que alcanzan los óptimos están relacionadas por las ecuaciones de las condiciones KKT
  - Las derivadas parciales del lagrangiano respecto a las variables primarias han de ser cero
  - Condición complementaria: las restricciones activas, aquellas que valen exactamente cero, su multiplicador de Lagrange podrá ser mayor o igual que cero. Sin embargo, para las condiciones inactivas, las que valgan estrictamente menos que cero, el multiplicador asociado debe ser cero (dispersión de la solución)
  - Estos valores han de cumplir las restricciones del primal y el dual

- **Consecuencia (KKT).** Se puede solucionar el problema primal a través de una solución del problema dual. Este punto de vista es a veces interesante cuando el problema dual es más fácil de resolver que el primal

Clasificación: problema primal

$$\begin{aligned} \min \quad & \|\omega\| \\ \text{s.a.} \quad & (\langle \omega, x_i \rangle + b) \geq +1 \quad \forall x_i \in +1 \\ & (\langle \omega, x_i \rangle + b) \leq -1 \quad \forall x_i \in -1 \end{aligned}$$

dado que $\|\omega\|^2 = \langle \omega, \omega \rangle$ podemos cambiar esta versión directa por otra equivalente más operativa para calcular sus derivadas

$$\begin{aligned} \min \quad & \frac{1}{2}\langle \omega, \omega \rangle \\ \text{s.a.} \quad & y_i(\langle \omega, x_i \rangle + b) \geq 1 \qquad i = 1,\ldots,n \end{aligned}$$

Clasificación: KKT

- Todas las funciones que intervienen son convexas y diferenciales. Se puede aplicar las condiciones de KKT

$$L(\omega,b,\alpha)=\frac{1}{2}\langle\omega,\omega\rangle-\sum_{i=1}^{n}\alpha_i\left[y_i\left(\langle\omega,x_i\rangle+b\right)-1\right]$$

$$\frac{\partial L(\omega,b,\alpha)}{\partial\omega}=\omega-\sum_{i=1}^{n}\alpha_i y_i x_i=0 \qquad \rightarrow \quad \omega=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L(\omega,b,\alpha)}{\partial b}=\sum_{i=1}^{n}\alpha_i y_i=0$$

Clasificación: lagrangiano

$$L(\omega,b,\alpha)=\frac{1}{2}\langle\omega,\omega\rangle-\sum_{i=1}^{n}\alpha\left[y_i\left(\langle\omega,x_i\rangle+b\right)-1\right]$$

$$=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left\langle x_i,x_j\right\rangle-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left\langle x_i,x_j\right\rangle-\sum_{i=1}^{n}\alpha_i y_i b+\sum_{i=1}^{n}\alpha_i$$

$$=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j\left\langle x_i,x_j\right\rangle+\sum_{i=1}^{n}\alpha_i$$

Clasificación: problema dual

$$\max \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i, x_j \right\rangle + \sum_{i=1}^{n}\alpha_i$$

$$\text{s.a.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\alpha_i \geq 0 \qquad i = 1, \ldots, n$$

Clasificación: análisis

- La solución ω^* es una combinación lineal de los ejemplos de entrenamiento

$$\omega^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

- No intervienen todos (**dispersión**), sólo los que tienen un multiplicador de Lagrange distinto de cero (vectores de soporte)

- En el caso separable, los vectores de soporte son los ejemplos que estén justo en el margen de cada clase (condición KKT)

$$\alpha_i^* \left[y_i \left(\left\langle \omega, x_i^* \right\rangle + b^* \right) - 1 \right] = 0 \qquad i = 1, \ldots, n$$

- La variable primal b no aparece en el problema dual, se debe calcular a partir de

$$b^* = -\frac{\max_{-} \left(\left\langle \omega^*, x_{-} \right\rangle \right) + \min_{+} \left(\left\langle \omega^*, x_{+} \right\rangle \right)}{2}$$

Parámetro b

- Obtener el parámetro b, sabemos que para un $\alpha_j > 0$

$$\sum_{i=1}^{n} \alpha_i y_i < x_i, x_j > + b = y_j$$

- Entonces promediando sobre todos los vectores de soporte

$$b = \frac{\sum_{j \in VS} \left(y_j - \sum_{i=1}^{n} \alpha_i y_i < x_i, x_j > \right)}{\#VS}$$

Clasificación: conclusiones

- **Margen:** Obtenemos la solución que, desde un punto de vista estructural, tiene menor posibilidad de cometer errores futuros

- **Convexidad:** La solución se obtiene resolviendo un programa de optimización cuadrática, convexo, sin mínimos locales y resoluble en tiempo polinomial

- **Dualidad y kernels:** El problema dual depende de productos escalares entre los ejemplos. Podremos sustituirlo por el producto escalar en un espacio de características mediante un kernel

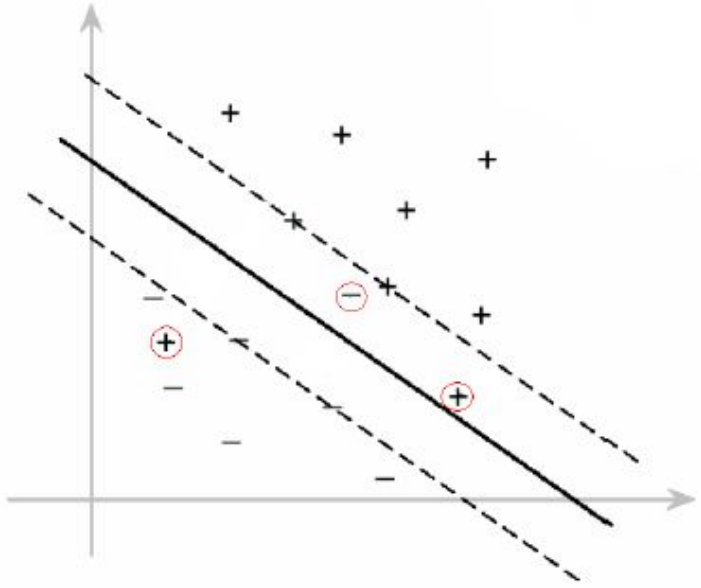
$$\max \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i, x_j) + \sum_{i=1}^{n}\alpha_i$$

- **Dispersión:** La solución depende de los vectores de soporte

$$h(x) = \operatorname{sgn}\left(\sum_{i \in VS}\alpha_i^* y_i k(x_i, x) + b^*\right)$$

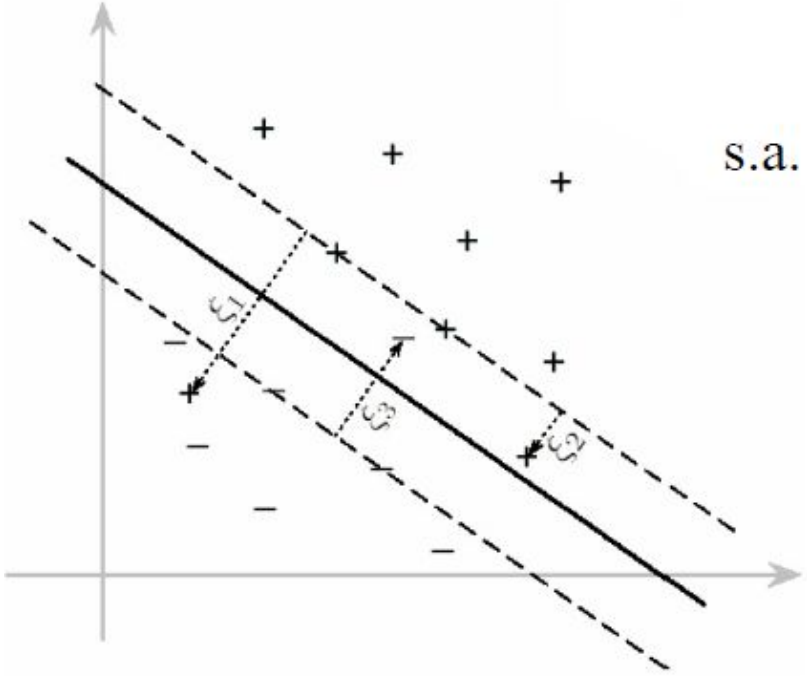
Soft Margin

- Considere el caso en que existan objetos de la clase A dentro de la clase B.
- No existe una separación lineal.
- Relajemos la clasificación, permitiendo errores.



C-SVM (primal)

$$\min \quad \frac{1}{2}\langle \omega, \omega \rangle + C \sum_{i=1}^{n} \xi_i$$



$$\text{s.a.} \quad y_i \left(\langle \omega, x_i \rangle + b \right) \geq 1 - \xi_i \qquad i = 1, \ldots, n$$

$$\xi_i \geq 0 \qquad i = 1, \ldots, n$$

C-SVM (lagrangiano)

$$L(\omega,b,\xi,\alpha,\beta)=\frac{1}{2}\langle\omega,\omega\rangle+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left[y_i\left(\langle\omega,x_i\rangle+b\right)-1+\xi_i\right]-\sum_{i=1}^{n}\beta_i\xi_i$$

$$\frac{\partial L(\omega,b,\xi,\alpha,\beta)}{\partial\omega}=\omega-\sum_{i=1}^{n}\alpha_i y_i x_i=0 \qquad \rightarrow \quad \omega=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L(\omega,b,\xi,\alpha,\beta)}{\partial b}=\sum_{i=1}^{n}\alpha_i y_i=0$$

$$\frac{\partial L(\omega,b,\xi,\alpha,\beta)}{\partial\xi_i}=C-\alpha_i-\beta_i=0 \qquad \rightarrow \quad C=\alpha_i+\beta_i$$

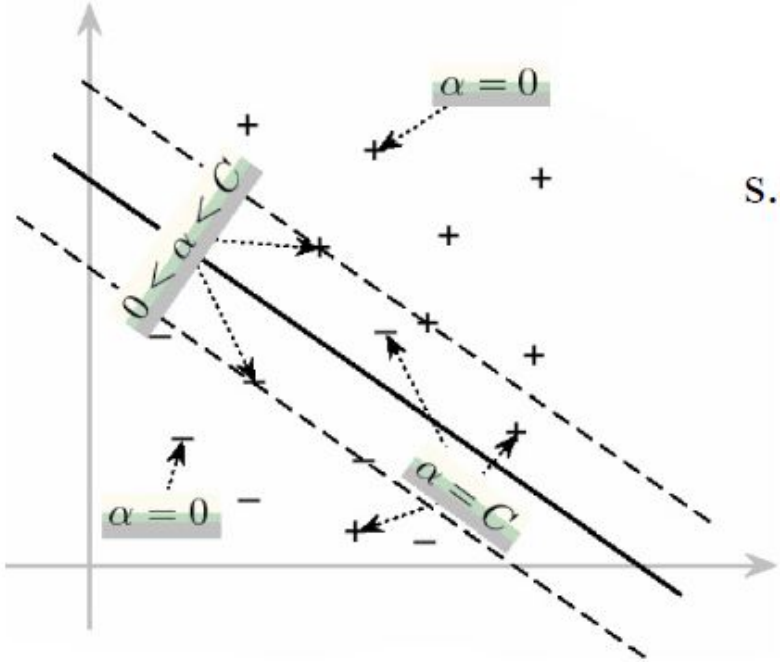
C-SVM (lagrangiano)

$$L(\omega,b,\xi,\alpha,\beta) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i, x_j \right\rangle + \sum_{i=1}^{n}(\alpha_i+\beta_i)\xi_i$$
$$-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i, x_j \right\rangle - \sum_{i=1}^{n}\alpha_i y_i b + \sum_{i=1}^{n}\alpha_i$$
$$-\sum_{i=1}^{n}\alpha_i\xi_i - \sum_{i=1}^{n}\beta_i\xi_i$$

$$= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i, x_j \right\rangle + \sum_{i=1}^{n}\alpha_i$$

C-SVM (dual)

$$\max \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i, x_j \right\rangle + \sum_{i=1}^{n}\alpha_i$$



$$\text{s.a.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C \qquad i = 1, \ldots, n$$

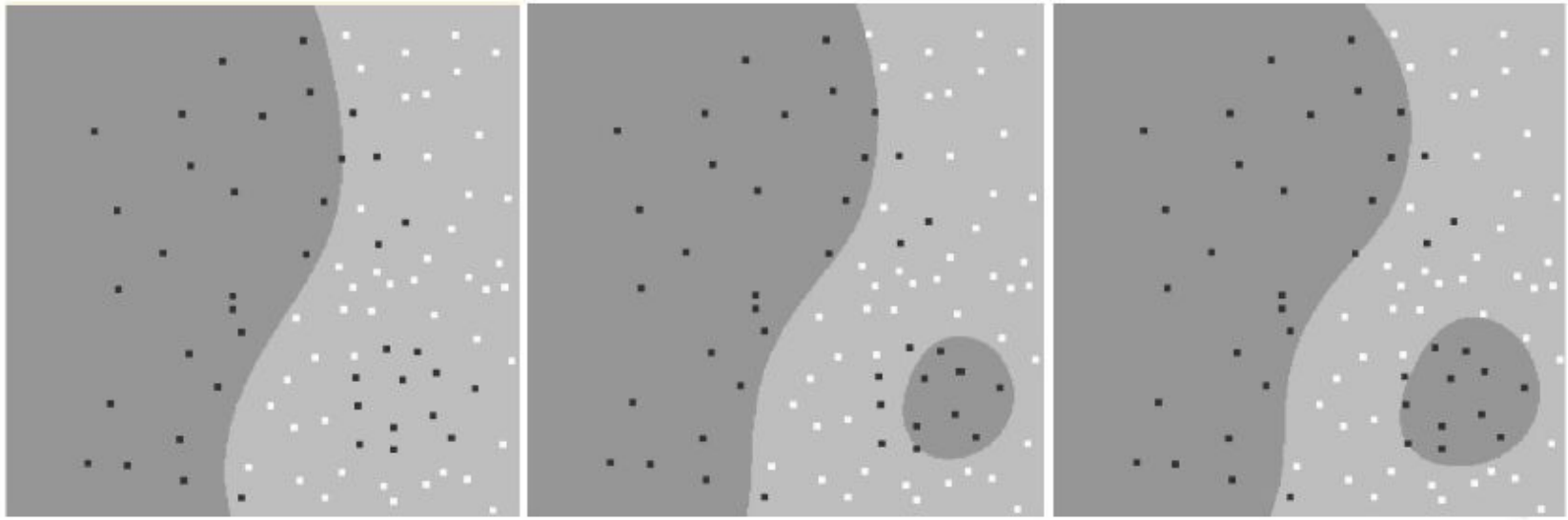
Regularización

- Las SVM's seleccionan la función que cumple la siguiente condición:

$$\min_{h \in H_k} \quad \|h\|_{H_k} + C\ L(h, S)$$

- El primer sumando representa la "complejidad" de la hipótesis elegida, se prefiere la más simple
- El segundo sumando sirve para controlar el **coste** de la hipótesis elegida, medido sobre los datos de entrenamiento utilizados.
- La constante C es la que nos permite regular la solución de compromiso entre ambos términos, complejidad y coste.
- La determinación del valor adecuado para en una aplicación real, es quizás más difícil que decidir el kernel a emplear, ya que éste en muchos casos puede venir dado por los datos.

Regularización



más bajo intermedio más alto

valor de C

Desventaja de la C-SVM

- El parámetro C es una constante que determina el trade-off entre dos objetivos:
  - Minimizar el error de entrenamiento
  - Maximizar el Margen

- Desafortunadamente C es un parámetro intuitivo
  - En la practica existen algunos criterios por donde buscar el valor de C (rango)

- Se propone reemplazar C por el parámetro υ, el cual controla el margen de error y los vectores de soporte

υ-SVM (primal)

$$\min \quad \frac{1}{2}\|\omega\|^2 - \upsilon\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i$$

$$\text{s.a.} \quad y_i(\langle \omega, x_i \rangle + b) \geq \rho - \xi_i \qquad i = 1,\ldots,n$$

$$\xi_i \geq 0 \qquad i = 1,\ldots,n$$

$$\rho \geq 0$$

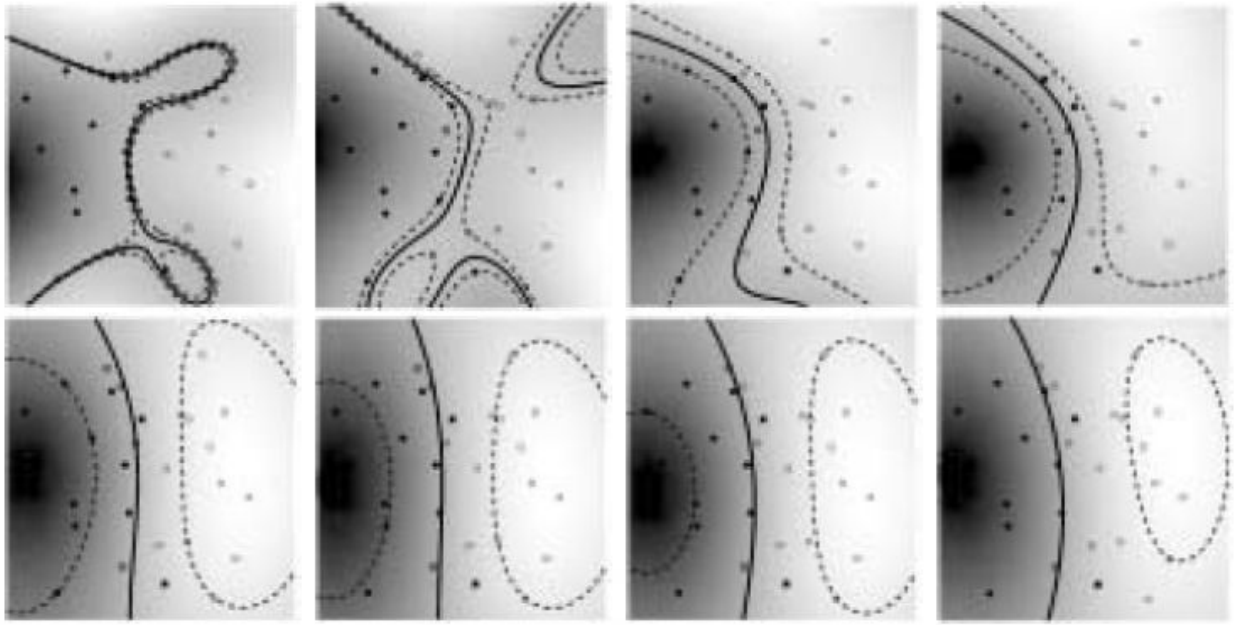


Figure Toy problem (task: separate circles from disks) solved using ν-SV classification, with parameter values ranging from $\nu = 0.1$ (top left) to $\nu = 0.8$ (bottom right). The larger we make ν, the more points are allowed to lie inside the margin (depicted by dotted lines). Results are shown for a Gaussian kernel, $k(x, x') = \exp(-\|x - x'\|^2)$.

ν	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
fraction of errors	0.00	0.07	0.25	0.32	0.39	0.50	0.61	0.71
fraction of SVs	0.29	0.36	0.43	0.46	0.57	0.68	0.79	0.86
margin $\rho/\|\mathbf{w}\|$	0.005	0.018	0.115	0.156	0.364	0.419	0.461	0.546

υ-SVM (lagrangiano)

$$L(\omega,b,\xi,\rho,\alpha,\beta,\delta)=\frac{1}{2}\langle\omega,\omega\rangle-\upsilon\rho+\frac{1}{n}\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left[y_i\left(\langle\omega,x_i\rangle+b\right)-\rho+\xi_i\right]-\sum_{i=1}^{n}\beta_i\xi_i-\delta\rho$$

$$\frac{\partial L(\omega,b,\xi,\rho,\alpha,\beta,\delta)}{\partial\omega}=\omega-\sum_{i=1}^{n}\alpha_i y_i x_i=0 \quad\rightarrow\quad \omega=\sum_{i=1}^{n}\alpha_i y_i x_i$$

$$\frac{\partial L(\omega,b,\xi,\rho,\alpha,\beta,\delta)}{\partial b}=\sum_{i=1}^{n}\alpha_i y_i=0$$

$$\frac{\partial L(\omega,b,\xi,\rho,\alpha,\beta,\delta)}{\partial\xi_i}=\frac{1}{n}-\alpha_i-\beta_i=0 \quad\rightarrow\quad \frac{1}{n}=\alpha_i+\beta_i$$

$$\frac{\partial L(\omega,b,\xi,\rho,\alpha,\beta,\delta)}{\partial\rho}=-\upsilon+\sum_{i=1}^{n}\alpha_i-\delta=0 \quad\rightarrow\quad \upsilon=\sum_{i=1}^{n}\alpha_i-\delta$$

υ-SVM (lagrangiano)

$$L(\omega,b,\xi,\rho,\alpha,\beta,\delta)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i,x_j\right\rangle-\left(\sum_{i=1}^{n}\alpha_i-\delta\right)\rho+\sum_{i=1}^{n}(\alpha_i+\beta_i)\xi_i$$

$$-\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i,x_j\right\rangle-\sum_{i=1}^{n}\alpha_i y_i b+\sum_{i=1}^{n}\alpha_i\rho-\sum_{i=1}^{n}\alpha_i\xi_i$$

$$-\sum_{i=1}^{n}\beta_i\xi_i-\delta\rho$$

$$=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i,x_j\right\rangle$$

υ-SVM (dual)

$$\max \quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \left\langle x_i, x_j \right\rangle$$

$$\text{s.a.} \quad \sum_{i=1}^{n}\alpha_i y_i = 0$$

$$\sum_{i=1}^{n}\alpha_i \geq \upsilon$$

$$0 \leq \alpha_i \leq \frac{1}{n} \qquad i = 1,\ldots,n$$

Función de decisión

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b \right)$$

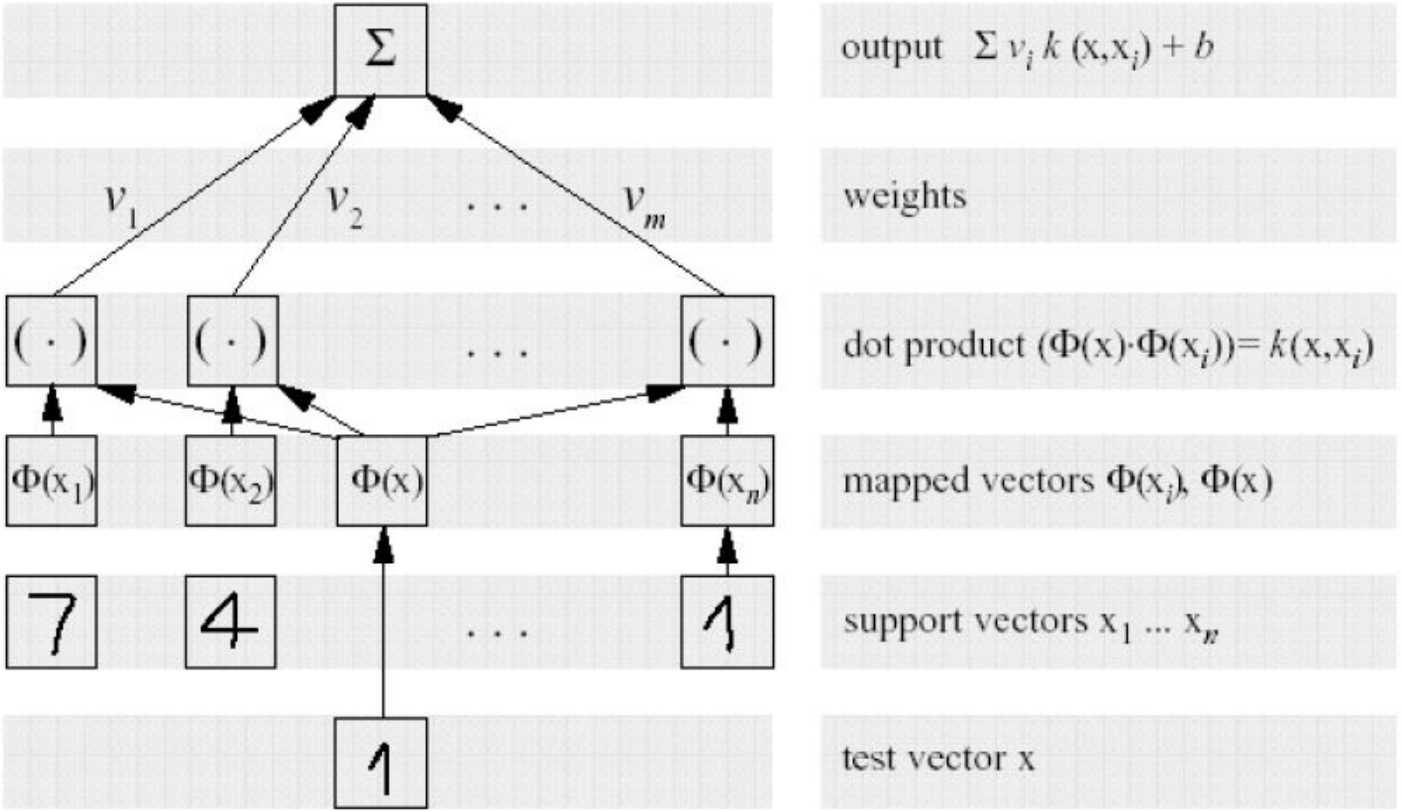
Parámetro b y ρ

- Para obtener b y ρ, se considerara 2 conjuntos $S_{\pm}$, de igual tamaño $s > 0$, que contenga los vectores de soporte x_i con $0 < \alpha_i < 1$ y $y_i = \pm 1$ respectivamente. Igualando $\xi_i = 0$ y según las condiciones KKT

$$b = -\frac{1}{2s} \sum_{x \in S_+ \cup S_-} \sum_j \alpha_j y_j \left\langle x, x_j \right\rangle$$

$$\rho = \frac{1}{2s} \left(\sum_{x \in S_+} \sum_j \alpha_j y_j \left\langle x, x_j \right\rangle - \sum_{x \in S_-} \sum_j \alpha_j y_j \left\langle x, x_j \right\rangle \right)$$

Arquitectura de la SVM



Robusto?

- La arquitectura de la maquina tiene implícitamente la cualidad de ser Robusto
  - La clasificación depende solamente de los vectores de soporte
  - La función de perdida es discreta finita (2 clases)
  - Los errores son penalizados y su influencia es acotada (cota sobre el multiplicador de lagrange asociado)

Idea

- Trabajar con kernel ofrece una solución para proyectar un conjunto de datos dentro de un espacio de característica altamente dimensional que incrementa la posibilidad de encontrar un hiperplano separador.

Truco del Kernel

- Todo algoritmo lineal, que dependa de un producto interno, puede transformarse fácilmente en un algoritmo no lineal.

- Este algoritmo no lineal (en el espacio de entrada) es equivalente al algoritmo lineal (en el espacio de características).

- El espacio de característica puede ser infinito-dimensional (donde se asegura la linealidad).

Truco del Kernel

- El mapeo al espacio de característica es muy costoso y más aun trabajar en él.
- Por el teorema de mercer, no se necesita trabajar en el nuevo espacio, y más aun, no se necesita mapear los inputs
- Sólo necesito el resultado del kernel, es decir, la similaridad (basado en producto interno), entre dos vectores.

Algunos Kernel

Sea $x = (x_1, x_2, \ldots, x_m) \in X$

- Kernel Lineal : $k(x, x') = \sum_{i=1}^{m} x_i \cdot x'_i$
- Kernel Polinomial : $k_p(x, y) = \left(x^T y + 1\right)^p$
- Kernel Gaussiano : $k(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}}$
- Sigmoidal : $k(x, y) = \tanh(cx^T \cdot y + \theta)$

Propiedades

- Si k_1 y k_2 son kernels, y $\alpha_1, \alpha_2 \geq 0$ entonces

$$\alpha_1 k_1 + \alpha_2 k_2$$

es un kernel

- Si $k_1, k_2, \ldots$ son kernels, y $k(x, x') = \lim_{n \to \infty} k_n(x, x')$

existe para todos los x, x', entonces k es un kernel

- Si k_1 y k_2 son kernels, entonces $k_1 k_2$, definido por

$$(k_1 k_2)(x, x') = k_1(x, x') k_2(x, x')$$

es un kernel

Propiedades

- Si k_1 y k_2 son kernels definidos respectivamente sobre $X_1 \times X_1$ y $X_2 \times X_2$, entonces el producto tensorial

$$(k_1 \otimes k_2)(x_1, x_2, x_1', x_2') = k_1(x_1, x_1')k_2(x_2, x_2')$$

  es un kernel sobre $(X_1 \times X_2) \times (X_1 \times X_2)$.

- Si k_1 y k_2 son kernels definido respectivamente sobre $X_1 \times X_1$ y $X_2 \times X_2$, entonces la suma directa

$$(k_1 \oplus k_2)(x_1, x_2, x_1', x_2') = k_1(x_1, x_1') + k_2(x_2, x_2')$$

  es un kernel sobre $(X_1 \times X_2) \times (X_1 \times X_2)$.

- En ambos casos $x_1, x_1' \in X_1$ y $x_2, x_2' \in X_2$

Kernel Compuestos

- Diferentes kernels proporcionan diferentes características a los estimadores.
- Los datos pueden transformarse hacia varios espacios de Hilbert (con diferentes kernels) a la vez.
- Alternativamente, diferentes fragmentos de cada patrón se pueden transformar a diferentes espacios de Hilbert

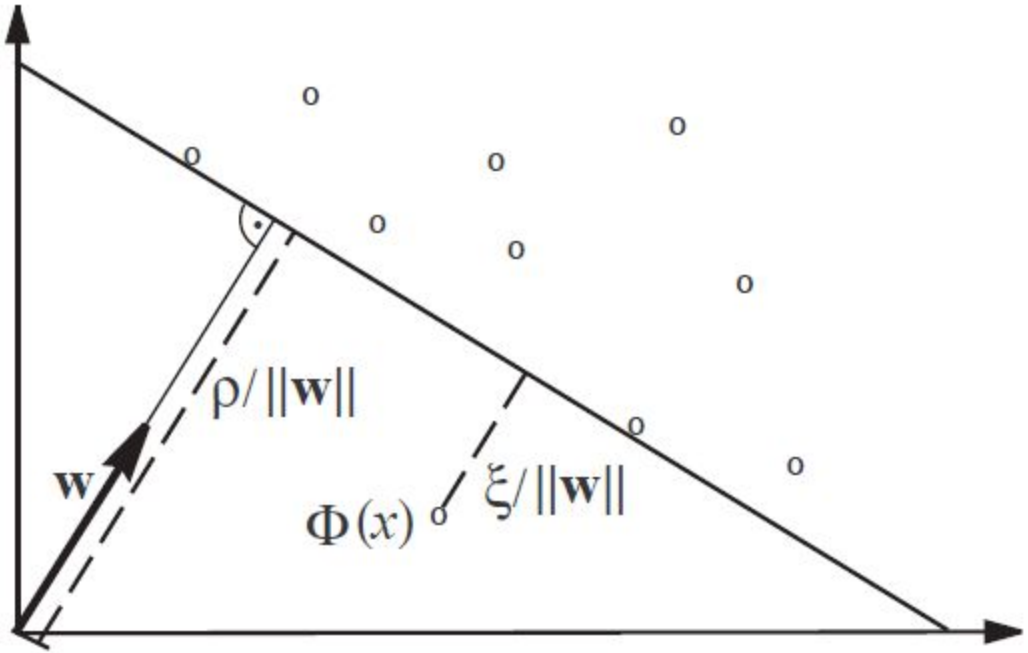
- La concatenación de estos espacios de Hilbert es un nuevo espacio de Hilbert.

$$\underset{\mathbf{w}\in\mathcal{H},\boldsymbol{\xi}\in\mathbb{R}^m,\rho\in\mathbb{R}}{\text{minimize}} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu m}\sum_i \xi_i - \rho,$$
$$\text{subject to } \langle \mathbf{w}, \Phi(x_i)\rangle \geq \rho - \xi_i, \quad \xi_i \geq 0.$$

$$f(x) = \operatorname{sgn}\left(\langle \mathbf{w}, \Phi(x)\rangle - \rho\right),$$

$$L(\mathbf{w}, \boldsymbol{\xi}, \rho, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{1}{\nu m}\sum_i \xi_i - \rho - \sum_i \alpha_i(\langle \mathbf{w}, \Phi(x_i)\rangle - \rho + \xi_i) - \sum_i \beta_i \xi_i,$$

$$\mathbf{w} = \sum_i \alpha_i \Phi(x_i),$$


ρ/ ||w||
w
Φ(x)
ξ/ ||w||

$$\alpha_i = \frac{1}{\nu m} - \beta_i \leq \frac{1}{\nu m}, \quad \sum_i \alpha_i = 1.$$

$$f(x) = \operatorname{sgn}\left(\sum_i \alpha_i k(x_i, x) - \rho\right).$$

$$\begin{aligned} \underset{\boldsymbol{\alpha} \in \mathbb{R}^m}{\text{minimize}} \; & \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(x_i, x_j), \\ \text{subject to} \; & 0 \leq \alpha_i \leq \frac{1}{\nu m}, \\ & \sum_i \alpha_i = 1. \end{aligned}$$

$$\underset{R\in\mathbb{R},\boldsymbol{\xi}\in\mathbb{R}^m,c\in\mathcal{H}}{\text{minimize}} \; R^2 + \frac{1}{\nu m}\sum_i \xi_i \text{ for } 0 < \nu \leq 1$$

$$\text{subject to } \|\Phi(x_i) - c\|^2 \leq R^2 + \xi_i \text{ and } \xi_i \geq 0 \text{ for } i \in [m].$$

This leads to the dual,

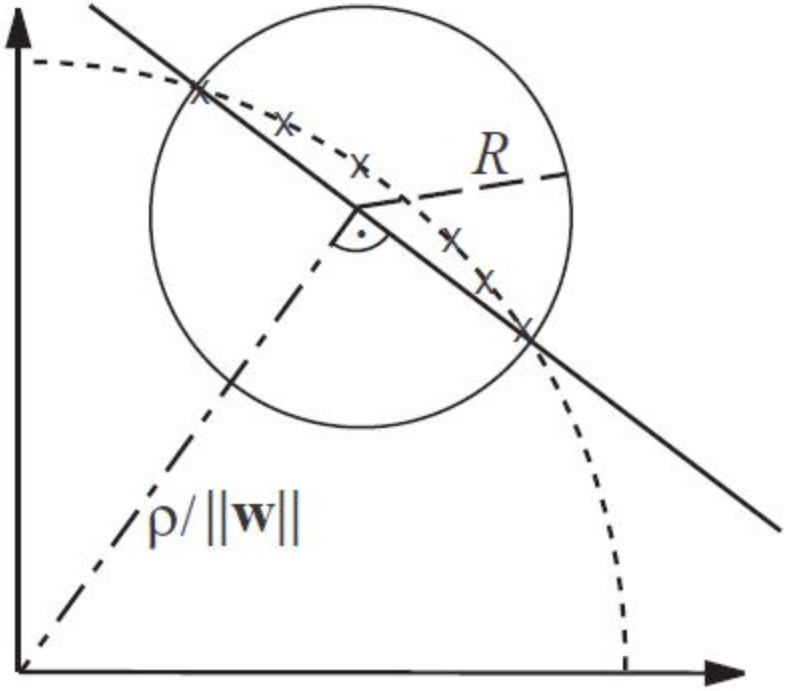
$$\underset{\boldsymbol{\alpha}}{\text{minimize}} \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) - \sum_i \alpha_i k(x_i, x_i),$$

$$\text{subject to } \; 0 \leq \alpha_i \leq \frac{1}{\nu m} \text{ and } \sum_i \alpha_i = 1,$$

and the solution

$$c = \sum_i \alpha_i \Phi(x_i),$$

$$f(x) = \operatorname{sgn}\left(R^2 - \sum_{ij} \alpha_i \alpha_j k(x_i, x_j) + 2\sum_i \alpha_i k(x_i, x) - k(x, x)\right).$$


R
ρ/ ||w||