Máquina de soporte vectorial con optimización mínima secuencial

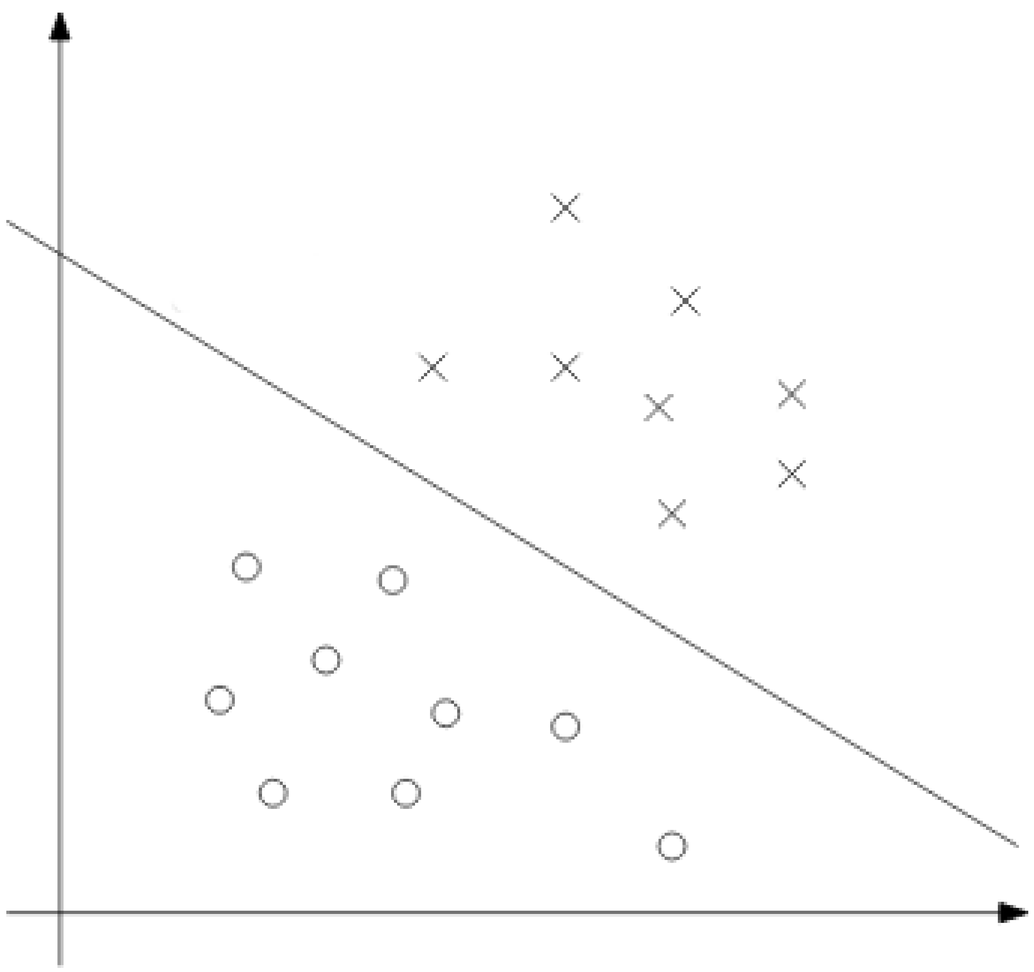
Omar Trejo Navarro

Instituto Tecnológico Autónomo de México

Problema de clasificación binaria

Imagine cualquier problema en el que tiene un cierto número de objetos, de los que conoce algunas características, y se clasifican en dos clases distintos. Además suponga que conoce a qué clase pertenece cada uno de estos objetos. Usted, entonces, puede aprender qué características presentan los objetos de las distintas clases y usar ese conocimiento para predecir a qué clase pertenece un nuevo objeto cuya clase real desconoce.

Figura 1 : Dos grupos separabales de objetos



La **máquina de soporte vectorial (MSV)** es una técnica para modelar el problema anterior, y un método rápido y eficiente para resolverla es la **optimización mínima secuencial** (OMS).

Importancia del problema

El problema de clasificación es un problema fundamental en el área de **aprendizaje automático**, que de forma general busca extraer conocimiento de manera automática para posteriormente realizar predicciones.

La MSV es una técnica de **aprendizaje supervisado** (se conoce a qué grupo pertenence cada objeto inicial) que se puede aplicar en general a problemas cuyo objetivo es predecir la pertenencia de un objeto a un grupo con base en datos de objetos similares observados previamente.

Existen muchos problemas que se ajustan a estas características. Algunos ejemplos generales son

- Clasificación de objetos (ej. vinos)
- Clasificación de eventos (ej. incendios)
- Predicción de enfermedades (ej. cáncer)

Naturalmente entre más objetos se puedan analizar y más información se tenga de esos objetos, la clasificación podrá ser mejor. Retos importantes surgen, por los recursos computacionales disponibles, al llevar estás técnicas a **grandes escalas de datos** (más de 100,000 observaciones, por ejemplo, pero pueden ser escalas mucho mayores). La optimización mínima secuencial contiene ideas mejoran el desempeño respecto a este tipo de retos.

Máquina de soporte vectorial con optimización mínima secuencial

Omar Trejo Navarro

Instituto Tecnológico Autónomo de México

Máquina de soporte vectorial

Queremos resolver el problema de **clasificación binaria**: predecir a cuál de las dos clases pertenece un objeto que no se había observado anteriormente. El objetivo es desarrollar un algoritmo eficiente para resolver este problema.

Conjunto de datos

y_i representa la clase a la que pertenece el i-ésimo objeto (ej. enfermo vs sano) y x_i son las características (atributos) que se conocen o se han medido del objeto (ej. altura, peso, triglicéridos, etc.).

$$\mathbb{D} = \{(x_i, y_i) | x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}_{i=1}^{n}$$

Hiperplano separador

Buscamos conocer w y b para encontrar el hiperplano separador que mejor clasificará los nuevos objetos (i.e. separe las clases 1 y -1).

$$h_{w,b}(x_i) = w^T x_i + b$$

Mecanismo de clasificación

El lado del hiperplano en el que se encuentra una observación indica la clase a la que se predice que pertenece un objeto.

$$c_{w,b}(x_i) = signo(w^T x_i + b)$$

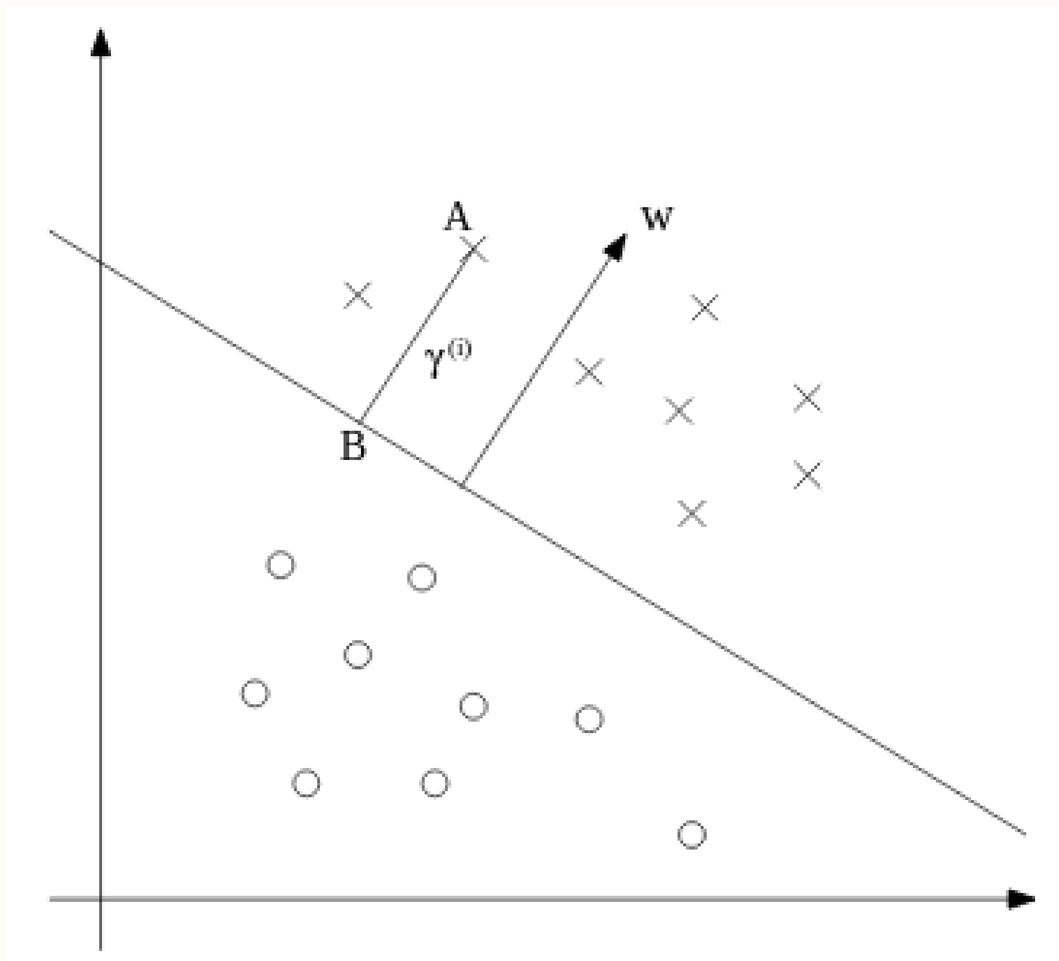
Margen geométrico

$$\gamma_i = y_i \left(\left(\frac{w}{||w||}^T \right) x_i + \frac{b}{||w||} \right)$$

$$\gamma_i > 0 \implies \checkmark$$

$$\gamma_i < 0 \implies \times$$

Figura 2 : Margen respecto al hiperlano separador



Máquina de soporte vectorial con optimización mínima secuencial

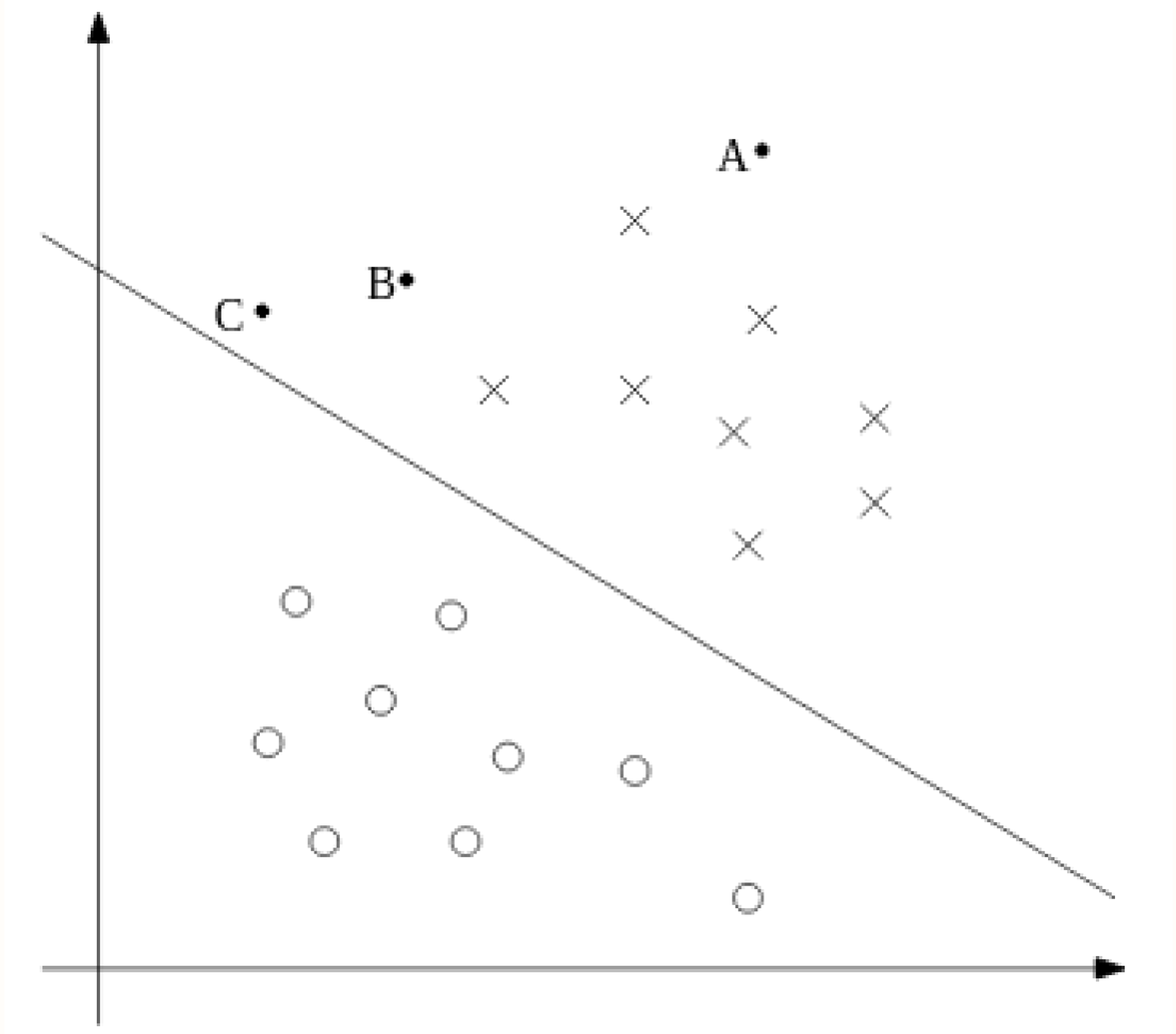
Omar Trejo Navarro

Instituto Tecnológico Autónomo de México

El mejor hiperplano separador

Si los datos son separables, existen infinitos hiperplanos separadores. Queremos en encontrar el mejor (no hacerlo puede tener errores muy costosos, ej. mal diagnóstico de cáncer).

Figura 3 : Sensibilidad respecto al hiperplano



Para hacerlo buscamos dos hiperplanos, uno para cada clase y maximizamos la distancia entre ellos (i.e. $2/||w||$).

$$w^T x - b = 1 \qquad w^T x - b = -1$$

Además debemos asegurar que los objetos se mantengan bien clasificados, por lo tanto

$$\forall i \ni y_i = 1 : w^T x_i - b \geq 1$$

$$\forall i \ni y_i = -1 : w^T x_i - b \leq -1$$

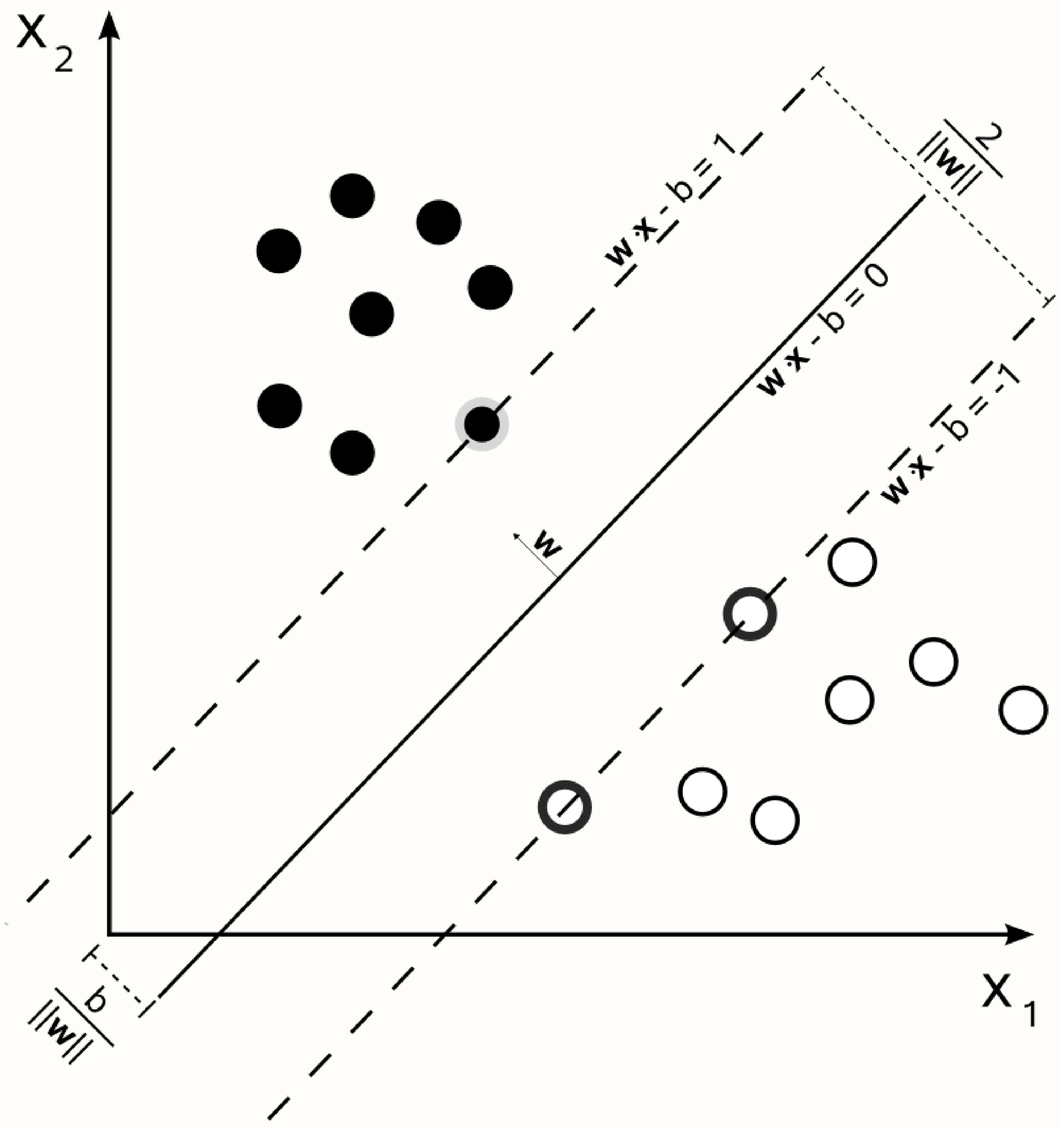
Problema primal

$$\min_{w,b} \frac{||w||^2}{2}$$

$$\text{s.a. } y_i(w^T x + b) \geq 1$$

$$\forall i \in \{1, \ldots, n\}$$

Figura 4 : Ilustración del problema primal



Usando los **multiplicadoers Karush-Kuhn-Tucker** expresamos el problema como

$$\min_{w,b} \max_{\alpha \geq 0} \left\{ \frac{||w||^2}{2} - \sum_{i=n}^{n} \alpha_i [y_i(w^T x_i - b) - 1] \right\}$$

Máquina de soporte vectorial con optimización mínima secuencial

Omar Trejo Navarro

Instituto Tecnológico Autónomo de México

Vectores de soporte

Todos aquellos $\alpha_i = 0$ representan obervaciones que se encuentran "bien" clasificados (i.e. tienen márgenes "grandes"). Aquellos $\alpha_i > 0$ representan observaciones cuya x_i está sobre la frontera.

$$\alpha_i > 0 \implies \text{en el margen}$$

$$\alpha_i = 0 \implies \text{bien clasificado}$$

La complejidad del problema depende de cuántos vectores de soporte existan. En problemas no degenerados normalmente hay proporcionalmente un número pequeño.

Problema dual

Este problema es mucho más fácil de resolver computacionalmente que el problema primal. Los algoritmos de solución trabajan con este problema.

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} y_i \alpha_i x_i y_j \alpha_j x_j$$

$$\text{s.a. } 0 \leq \alpha_i \quad \forall i \in \{1, \ldots, n\}$$

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

Margen suave

En 1995, Cortes y Vapnik sugieren un un *margen suave* para permitir objetos mal clasificados si las clases no son separables. Las ideas son las mismas, y el hiperplano separador óptimo seguirá buscando hacer la mejor separación posible.

Para lograrlo se introducen **variables de holgura** $\xi_i \geq 0$ que miden el grado de error que se introduce en el problema. Ahora las restricciones son

$$y_i(w^T x - b) \geq 1 - \xi_i$$

El problema primal se convierte en

$$\min_{w,b} \frac{||w||^2}{2} + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.a. } y_i(w^T x + b) \geq 1 - \xi_i$$

$$\wedge \xi_i > 0 \quad \forall i \in \{1, \ldots, n\}$$

El problema dual se convierte en

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} y_i \alpha_i x_i y_j \alpha_j x_j$$

$$\text{s.a. } 0 \leq \alpha_i \leq C \quad \forall i \in \{1, \ldots, n\}$$

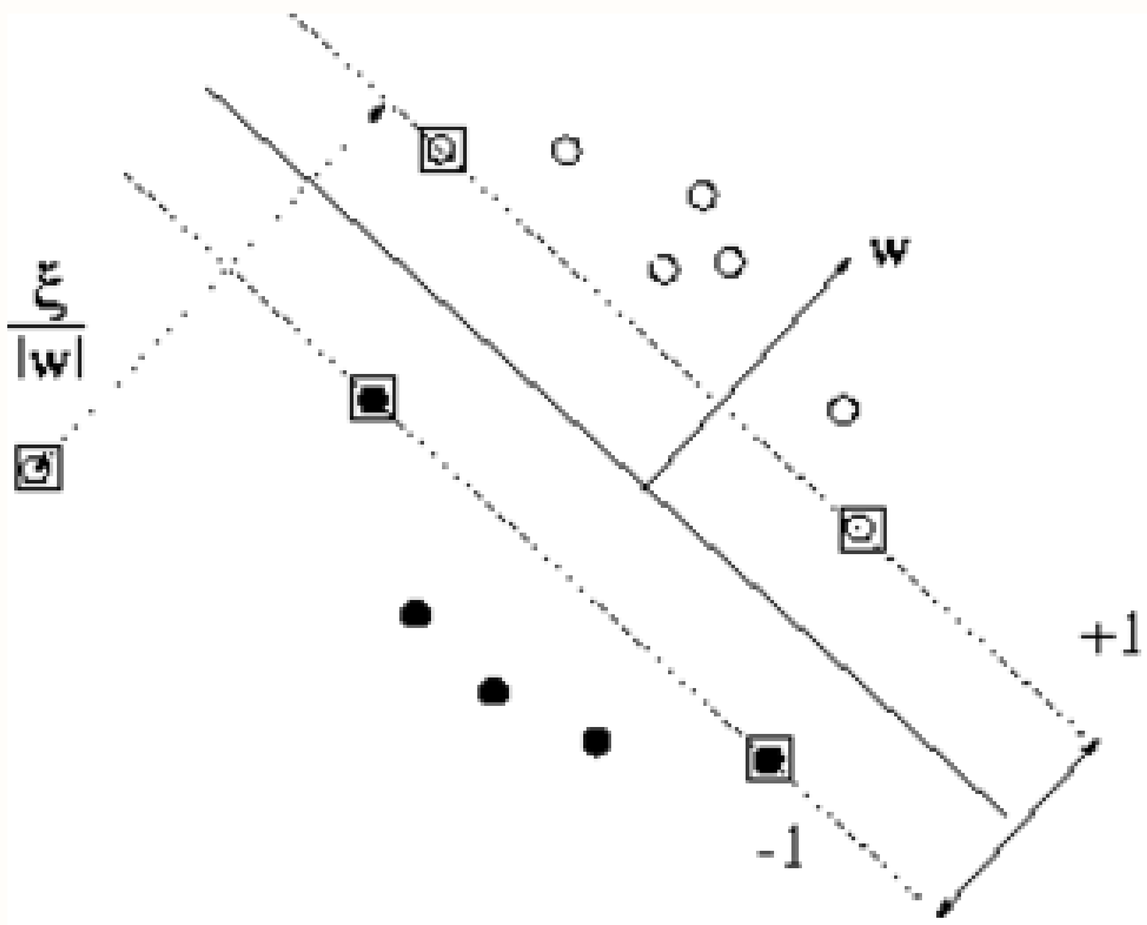
$$\sum_{i=1}^{n} y_i \alpha_i = 0$$

Máquina de soporte vectorial con optimización mínima secuencial

Omar Trejo Navarro

Instituto Tecnológico Autónomo de México

Figura 5 : MSV con margen suave



La solución se puede expresar como

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

Clasificación no-lineal con kernels

Si remplazamos los productos punto entre los atributos por un **kernel**, podemos "llevar" los datos a un **espacio de dimensión superior**, donde puede ser más fácil hacer la separación lineal. Algunos ejemplos de kernels:

- Lineal: $k(x_i, x_j) = x_i x_j$
- Polinomal: $k(x_i, x_j) = (x_i x_j)^d$
- Gaussiano: $exp(-\gamma||x_i - x_j||^2)$

Todas las ecuaciones son simialres y lo único que cambia es que cuando se tenía una multiplicación entre los vectores de atributos, ahora se utiliza alguno de estos kernels.

Figura 6 : Frontera lineal en un espacio superior

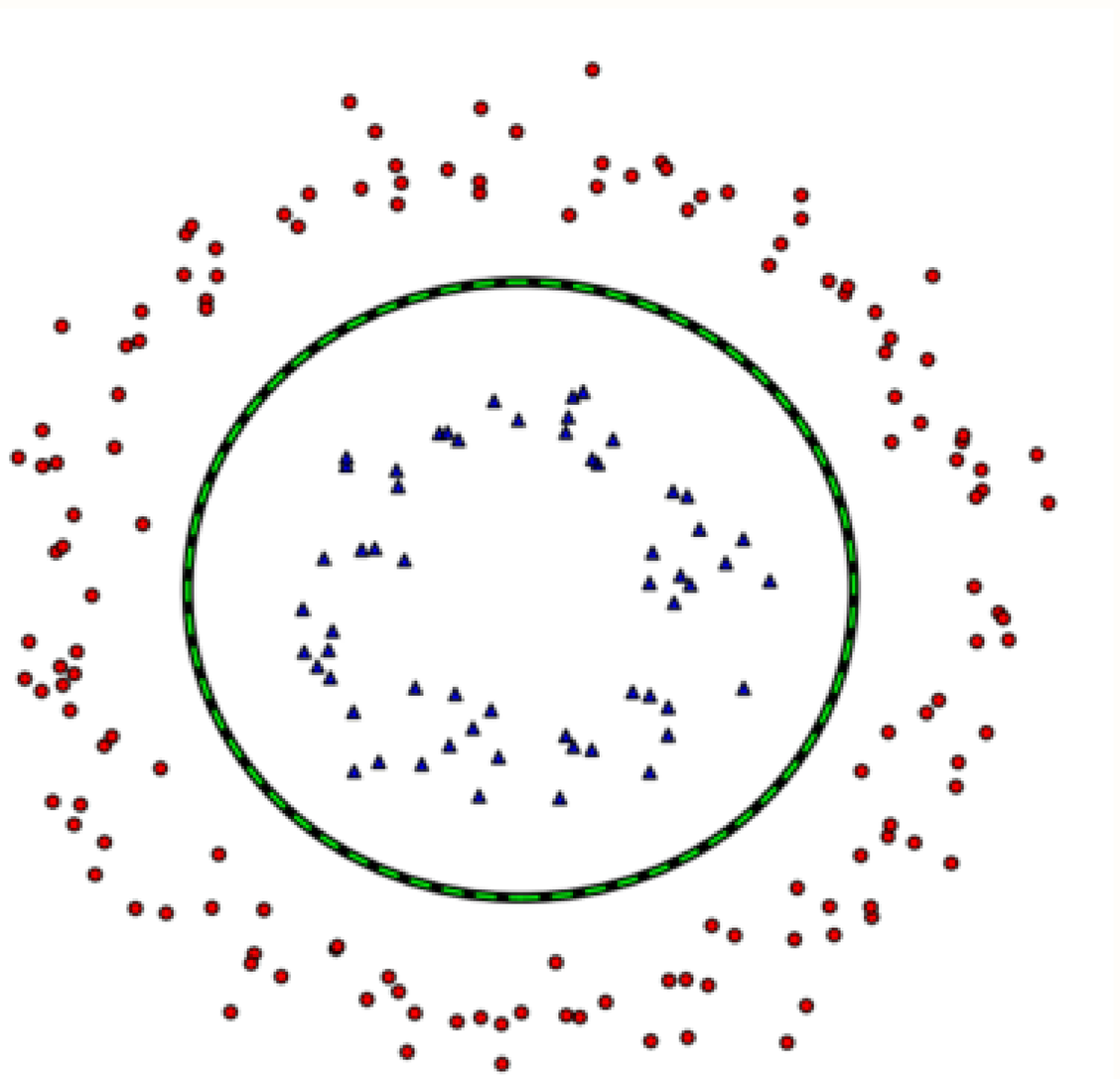
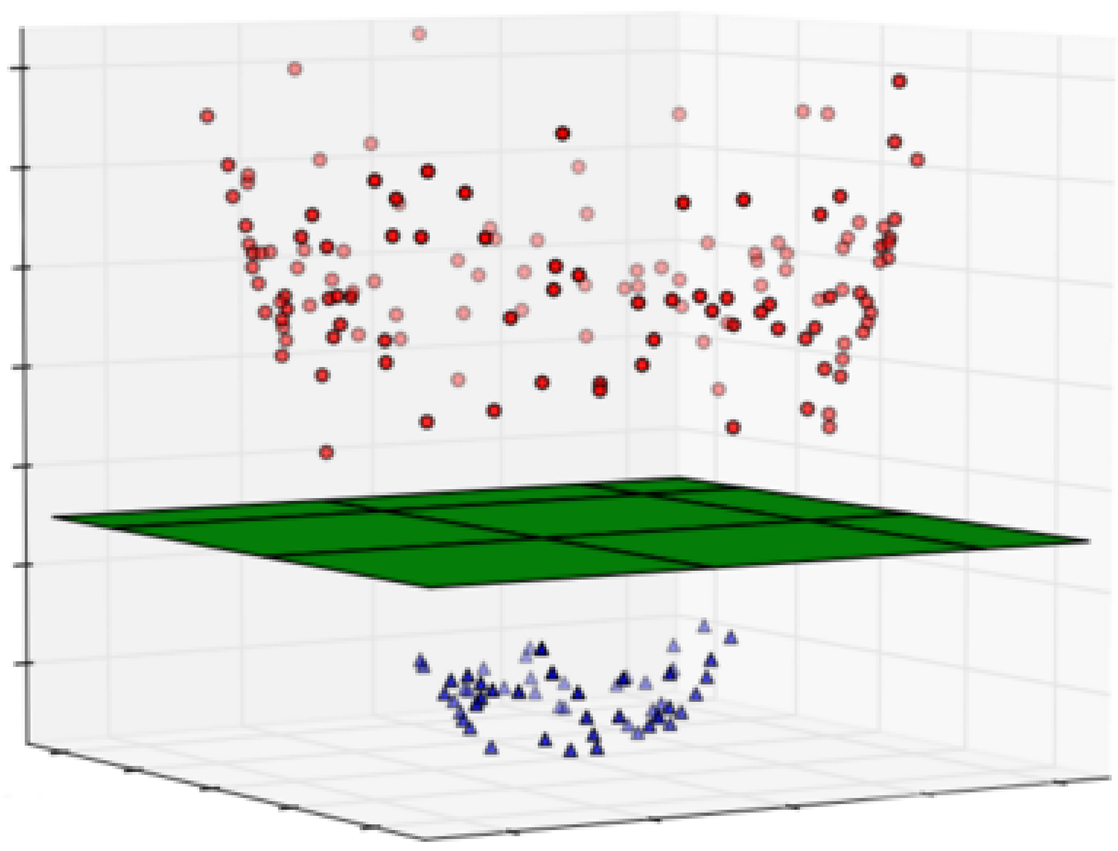


Figura 7 : Representación en un espacio superior



Máquina de soporte vectorial con optimización mínima secuencial

Omar Trejo Navarro

Instituto Tecnológico Autónomo de México

Hiperplano separador óptimo

$$h^*_{w,b}(\hat{x}) = w^{*T}\hat{x} + b^*$$
$$= \left(\sum_{i=1}^{n} y_i \alpha_i^* x_i\right)^T \hat{x} + b^*$$

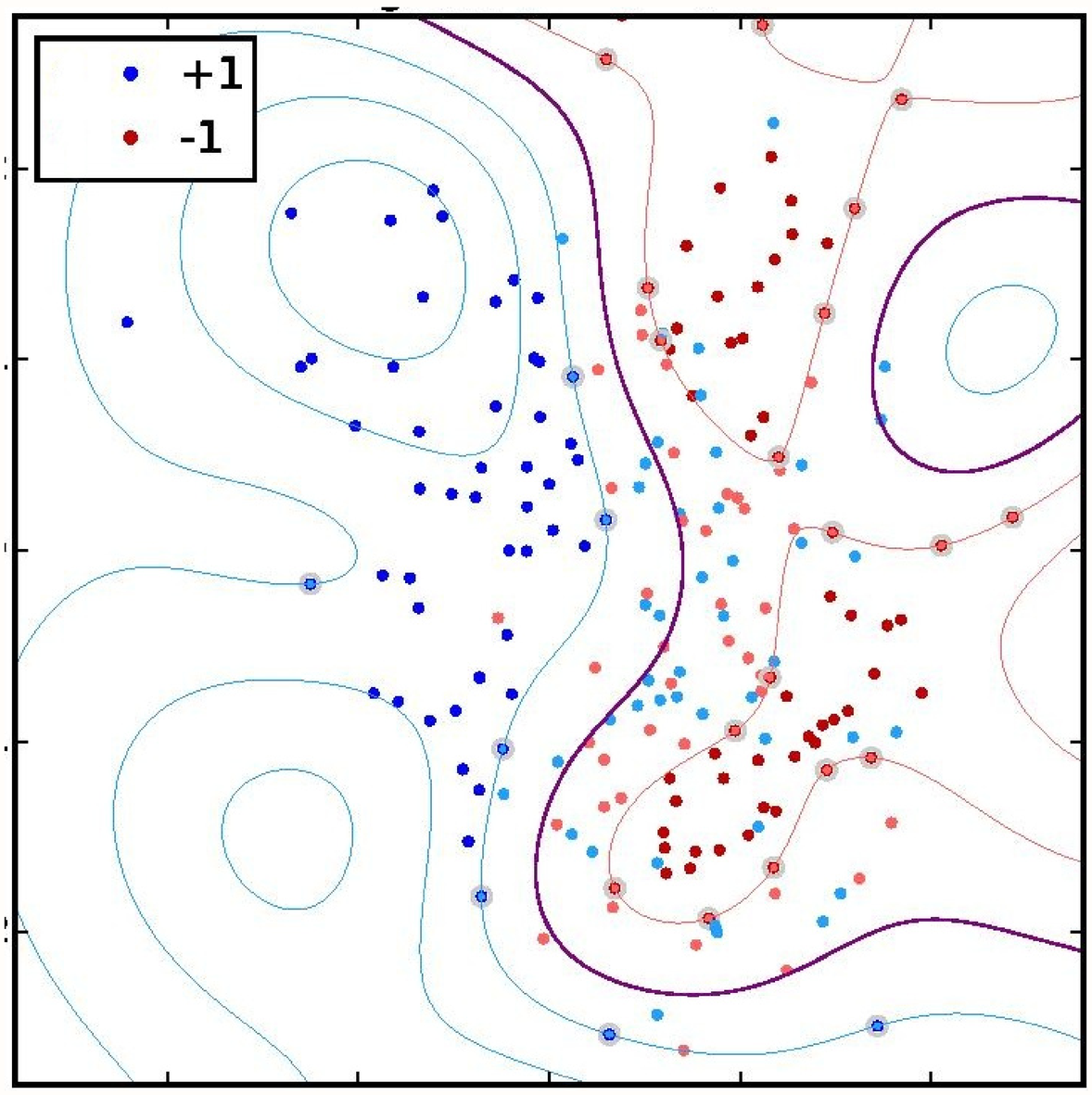
Selección de parámetros

Cuando se va a "entrenar" a la MSV, los **datos deben estar normalizados** y la selección de parámetros se hace por **validación cruzada**, y normalmente se prueba con los valores

$$C \in \{2^{-5}, 2^{-3}, ..., 2^{13}, 2^{15}\}$$
$$\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{1}, 2^{3}\}$$

Esto se puede hacer un poco más eficiente si se utilizan técnicas bayesianas para detectar qué parámetros se deben de probar.

Figura 8 : SVM con margen suave y kernel Gaussiano



Optimización mínima secuencial

Buscan resolver el problema de MSV de manera eficiente, ya que calcular el kernel es costoso e innecesario, y posiblemente no quepa en la memoria cuando se quiera resolver problemas de gran escala.

Dos desventajas de las MSV son:

- Requiere conocer todas las clases
- Parámetros difíciles de interpretar

Máquina de soporte vectorial con optimización mínima secuencial

Omar Trejo Navarro

Instituto Tecnológico Autónomo de México

Máquina de soporte vectorial con optimización mínima secuencial

Omar Trejo Navarro

Instituto Tecnológico Autónomo de México