Tecnologías para Inteligencia Artificial (247101009)
Tema 5. Máquinas de vector soporte

Javier Vales Alonso
Máster Universitario en Ingeniería Telemática

2020

Universidad Politécnica de Cartagena

Métodos de kernel

Clasificadores de margen máximo

Clasificadores de margen suave

¿Cómo estudiar esta unidad?

1. Haga una primera lectura de la unidad. Concéntrese en ver las ideas generales y hacer una primera revisión de las matemáticas.
2. Haga una revisión a fondo de las matemáticas y resuelva en el notebook los ejercicios indicados. Intente comprender todos los desarrollos involucrados. En caso de dudas, lea las referencias sugeridas (ver referencias en Tema 0) o contacte con el profesor.
3. Finalmente, debe enviar el notebook a través de AV.

Métodos de kernel

Métodos de kernel

Dada una transformación a un espacio de características fijas denotada por $\boldsymbol{\phi}$, la función $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}')$ se llama **kernel**. El kernel más simple es $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T \boldsymbol{x}'$ y se llama **kernel lineal**.

Los kernels se utilizan para reescribir modelos paramétricos (e.g., regresión lineal o regresión logística) en representaciones *duales*. Éstas, son **no paramétricas** y requieren la evaluación del kernel para el punto $\boldsymbol{x}$ frente a los datos de entrenamiento (o un subconjunto de ellos) a la hora de realizar las predicciones.

Al usar kernels, ya no es necesario calcular transformaciones explícitas desde y hacia un espacio de características. Además los kernels puede estar asociados implicitamente a espacios de características de dimensiones infinitas.

Métodos de kernel (II)

Para ilustrar cómo se obtienen las representaciones duales, considere la función de coste regularizada obtenida para el modelo de regresión lineal (ver Tema 1):

$$\widetilde{J}(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) - t_n)^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w} \tag{1}$$

El $\boldsymbol{w}$ óptimo se obtiene igualando el gradiente respecto a $\boldsymbol{w}$ de la expresión anterior a 0, y puede reescribirse como:

$$\boldsymbol{w} = -\frac{1}{\lambda}\sum_{n=1}^{N}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) - t_n)\boldsymbol{\phi}(\boldsymbol{x}_n) = \sum_{n=1}^{N} a_n\boldsymbol{\phi}(\boldsymbol{x_n}) = \boldsymbol{\Phi}^T\boldsymbol{a}$$

donde $\boldsymbol{a}$ es el vector con elementos $a_n = -\frac{1}{\lambda}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) - t_n)$.

Métodos de kernel (III)

La forma dual de la función de coste se obtiene sustituyendo $\boldsymbol{w}$ por $\boldsymbol{\Phi}^T\boldsymbol{a}$ en la ecuación. (1):

$$\tilde{J}(\boldsymbol{a}) = \frac{1}{2}\boldsymbol{a}^T\boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{t} + \frac{\lambda}{2}\boldsymbol{a}^T\boldsymbol{\Phi}\boldsymbol{\Phi}^T\boldsymbol{a} \quad (2)$$

$\boldsymbol{K}$ es la matriz de *Gram* $\boldsymbol{\Phi}\boldsymbol{\Phi}^T$, es decir, $(\boldsymbol{K})_{nm} = \boldsymbol{\phi}(\boldsymbol{x}_n)^T\boldsymbol{\phi}(\boldsymbol{x}_m) = k(\boldsymbol{x}_n, \boldsymbol{x}_m)$.

Por lo tanto

$$\tilde{J}(\boldsymbol{a}) = \frac{1}{2}\boldsymbol{a}^T\boldsymbol{K}\boldsymbol{K}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{K}\boldsymbol{t} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{t} + \frac{\lambda}{2}\boldsymbol{a}^T\boldsymbol{K}\boldsymbol{a} \quad (3)$$

Métodos de kernel (IV)

Al establecer el gradiente del coste con respecto a $\boldsymbol{a}$ a cero obtenemos $\boldsymbol{a} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{t}$.

En la representación dual las predicciones se calculan como:

$$y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{a}^T \boldsymbol{\Phi} \boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})^T (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1} \boldsymbol{t} \qquad (4)$$

siendo $\boldsymbol{k}(\boldsymbol{x})$ el vector cuyos elementos son $k(\boldsymbol{x}_n, \boldsymbol{x})$, con n=1, . . . , N.

Tenga en cuenta que la ecuación (4) no requiere el uso de los parámetros $\boldsymbol{w}$, pero proporciona la predicción en términos de las evaluaciones del kernel frente a cada dato de entrenamiento.

Métodos de kernel (V)

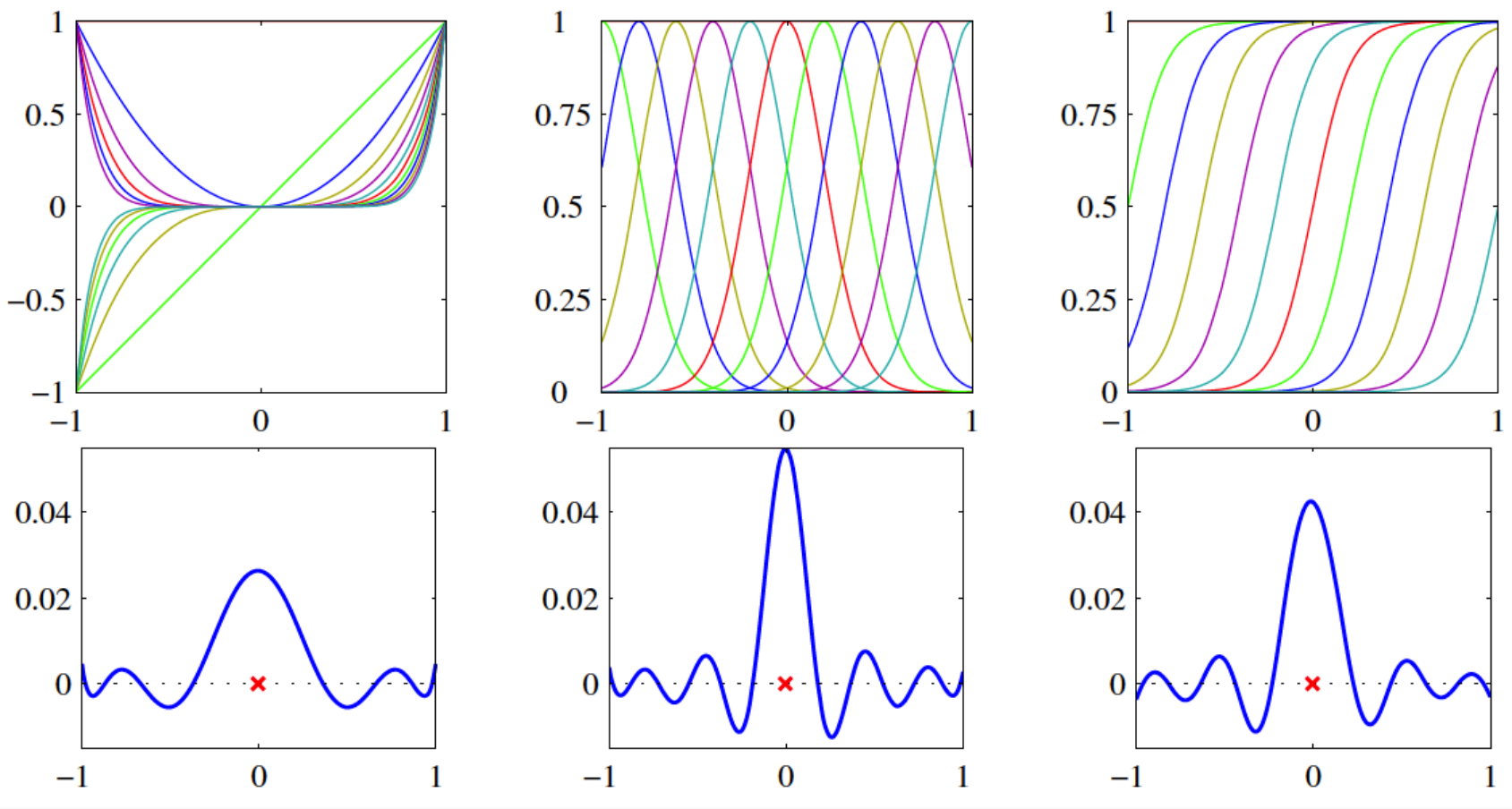
Para evitar la conversión de datos a/desde el espacio de características, se usa el **kernel trick**. Lo ilustramos mediante un ejemplo: consideremos un espacio de entrada de 2 dimensiones y el kernel $k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^T \boldsymbol{z})^2$. Esta función se puede escribir [1] como:

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}\boldsymbol{z})^2 &= (x_1 z_1 + x_2 z_2)^2 = x_1^2 z_1^2 + 2 x_1 z_1 x_2 z_2 + x_2^2 z_2^2 \\ &= (x_1^2, \sqrt{2} x_1 x_2, x_2^2)(z_1^2, \sqrt{2} z_1 z_2, z_2^2)^T = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{z}) \end{aligned}$$

donde el mapeo de características es $\boldsymbol{\phi}(\boldsymbol{x}) = (x_1^2, \sqrt{2} x_1 x_2, x_2^2)^T$. **Los cálculos del kernel se realizan directamente con la entrada, sin transformaciones al espacio de características**.

[1] Por claridad, aquí x_i denota el elemento i-ésimo del vector $\boldsymbol{x}$.

Métodos de kernel (VI)



La figura muestra las evaluaciones de kernel en $k(\boldsymbol{x}, \mathbf{0})$ asociadas a funciones de base de polinomios (izquierda), gaussianas (centro) y sigmoides (derecha).

Métodos de kernel (VII)

Una condición necesaria y suficiente para que una función $k(\boldsymbol{x}, \boldsymbol{x}')$ sea un kernel válido es que su matriz de Gram es semidefinida positiva (todos los autovalores positivos).

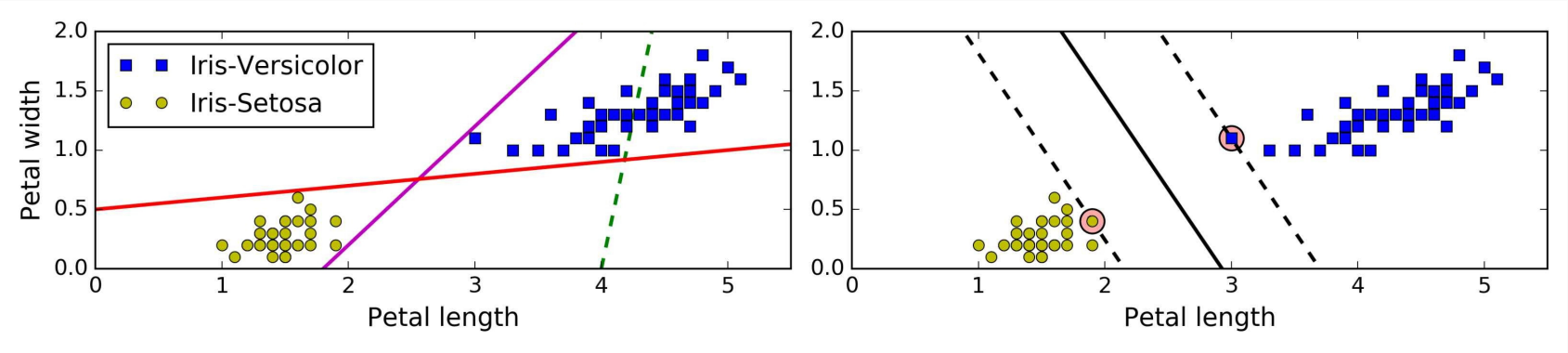
La representación dual para la regresión lineal requiere evaluar el kernel en $\boldsymbol{x}$ para **todos** los puntos en el conjunto de datos. Por lo tanto, su uso puede ser lento si el conjunto de datos es grande.

Algunos algoritmos reducen ese problema al evaluar el kernel sólo en un conjunto de entrenamiento puntos. Estos se llaman métodos de **sparse kernel**.

Clasificadores de margen máximo

Clasificadores de margen máximo

La siguiente figura muestra un conjunto de datos con fronteras de decisión lineales que producen una separación perfecta en el conjunto de datos. Intuitivamente, las nuevas instancias serán mejor clasificadas en la figura derecha ya que la frontera tiene un **margen** grande a ambas clases.



Esta es la idea que explotan las máquinas de vector soporte (SVM): construir clasificadores de margen máximo.

Clasificadores de margen máximo (II)

Supongamos que las predicciones para un clasificador binario se basan en el modelo lineal:

$$y(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b \tag{5}$$

donde b es un término de bias. Los targets son $t_n \in \{-1, 1\}$ y los puntos nuevos se clasifican según el signo de $y(\boldsymbol{x})$ (por tanto, $t_n y(\boldsymbol{x}) \geq 1$ para los puntos del conjunto de entrenamiento correctamente clasificados).

La distancia desde un punto $\boldsymbol{x}_n$ a la frontera de decisión es:

$$\frac{y(\boldsymbol{x}_n)}{\|\boldsymbol{w}\|} = \frac{\boldsymbol{w}^T \boldsymbol{x}_n + b}{\|\boldsymbol{w}\|} \tag{6}$$

Clasificadores de margen máximo (III)

De los infinitos planos posibles, nuestro objetivo es seleccionar aquel con todos los puntos de entrenamiento correctamente clasificados y con el margen máximo.

Para encontrar el hiperplano de separación óptimo se puede suponer una escala del espacio de datos tal que $\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b = 1$ para el punto más cercano (el margen mínimo). Por lo tanto, el margen será ≥ 1 para todos los puntos si están clasificados correctamente. Lograr el margen máximo requiere maximizar $1/\|\boldsymbol{w}\|$ o, a la inversa, minimizar $\|\boldsymbol{w}\|^2$.

Clasificadores de margen máximo (IV)

El problema de optimización se puede establecer en su forma *primal*:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{7}$$

sujeto a $t_n[\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b] \geq 1$ (el factor t_n fuerza las clasificaciones correctas para el conjunto de entrenamiento).

Estas restricciones se pueden agregar a la función de optimización mediante multiplicadores de Lagrange $\{a_n\}$,

$$L(\boldsymbol{w}, b, \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} a_n[t_n(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + b) - 1] \tag{8}$$

Clasificadores de margen máximo (V)

Estableciento los gradientes con respecto a $\boldsymbol{w}$, b, y $\boldsymbol{a}$ a 0, se obtiene:

$$\boldsymbol{w} = \sum_{n=1}^{N} a_n t_n \boldsymbol{\phi}(\boldsymbol{x}_n) \tag{9}$$

$$0 = \sum_{n=1}^{N} a_n t_n \tag{10}$$

Al sustituir ambos en la ec. (8), se obtiene la representación dual:

$$\max_{\boldsymbol{a}} \widetilde{L}(\boldsymbol{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\boldsymbol{x}_n, \boldsymbol{x}_m) \tag{11}$$

sujeto a $a_n \geq 0$, para $n = 1, \ldots, N$, y $\sum_{n=1}^{N} a_n t_n = 0$.

Clasificadores de margen máximo (VI)

Tanto la forma primal como la dual son problemas de programación cuadrática, y por lo tanto, **convexos**, ya que la región de las restricciones es convexa, y hay un solo óptimo.

Resolver problemas cuadráticos tiene un coste computacional cúbico en el número de variables. Por lo tanto, el primario es $O(M^3)$, y el dual es $O(N^3)$. Es decir, resolver el dual es ventajoso para espacios de características de alta dimensión. Además, el truco del kernel permite el cálculo directo de los kernels en el espacio de entrada.

Clasificadores de margen máximo (VII)

Un óptimo en un problema cuadrático debe cumplir las condiciones Karush-Kuhn-Tucker.

En la forma dual se traduce a:

$$a_n \geq 0$$
$$t_n y(\boldsymbol{x}_n) - 1 \geq 0$$
$$a_n(t_n y(\boldsymbol{x}_n) - 1) = 0$$

En palabras: un dato del conjunto de entrenamiento se encuentra en los límites establecido por el margen ($t_n y(\boldsymbol{x}_n) - 1) = 0$), o más allá y su multiplicador de Lagrange está desactivado ($a_n = 0$).

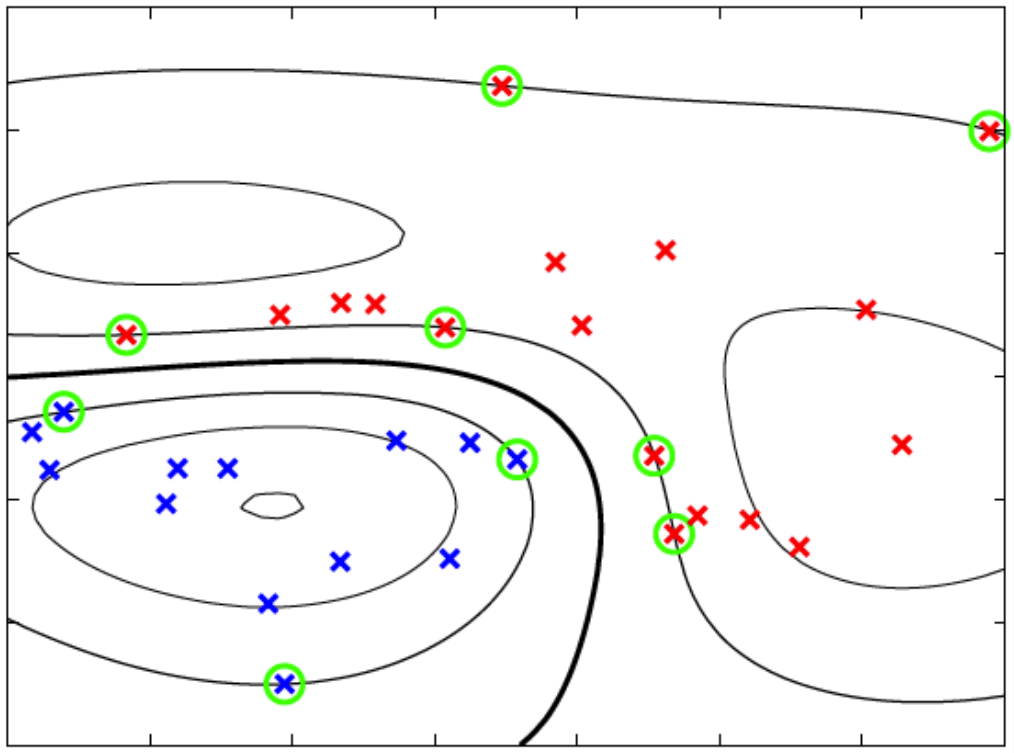
Clasificadores de margen máximo (VII)

Los puntos del margen se denominan **vectores soporte**. En el resto de los puntos, como $a_n = 0$, el cálculo del kernel en la ecuación (11) no se necesita, lo que conduce a un modelo *sparse*, donde sólo se usan los vectores soporte.

Esta observación también permite heurísticos eficientes para acelerar la resolución del programa dual. Su solución proporciona $\boldsymbol{a}$, que a su vez permite calcular $\boldsymbol{w}$ con la eq. (9). Entonces, el sesgo b se calcula seleccionando algún vector soporte, que debe cumplir con $\boldsymbol{w}^T\boldsymbol{x} + b = 1$. Las predicciones para nuevas instancias de datos se toman evaluando el signo de:

$$y(\boldsymbol{x}) = \sum_{x_n \text{ is SV}} a_n t_n k(\boldsymbol{x}, \boldsymbol{x_n}) + b$$

Clasificadores de margen máximo (VIII)

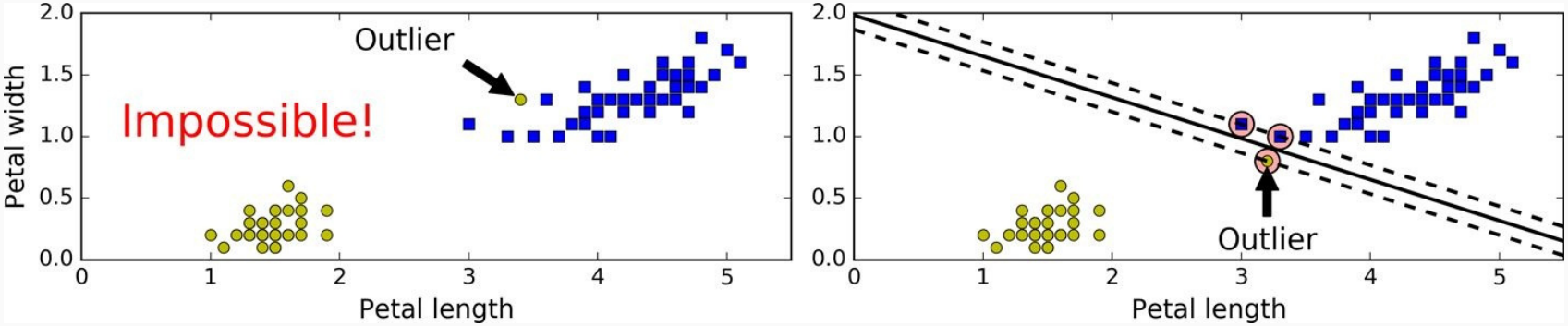


Frontera de la SVM calculada utilizando el kernel gaussiano,
$k(\boldsymbol{x}, \boldsymbol{x'}) = \exp\left(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\sigma^2\right.$

Clasificadores de margen suave

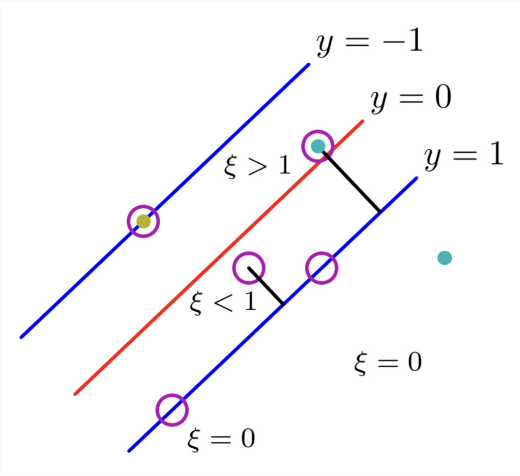
Clasificadores de margen suave

LA SVM funciona muy bien si el conjunto de entrenamiento es linealmente separable. Pero, en caso de que no lo sea, o en presencia de *outliers*, su rendimiento se degrada rápidamente, como se muestra en la figura a continuación.

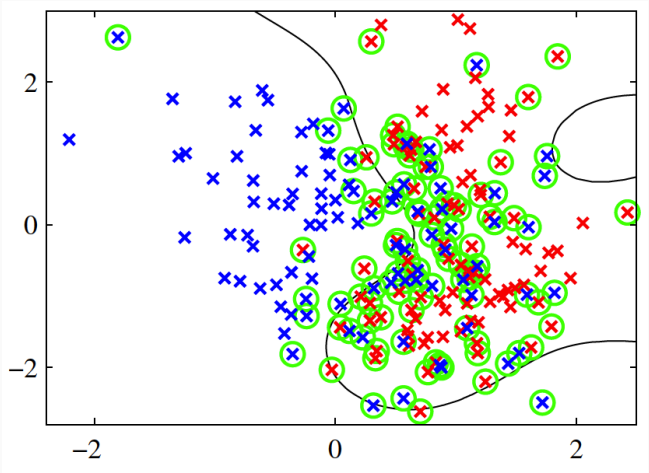


Clasificadores de margen suave (II)

Para corregir este problema se usan **funciones de penalización**, que permiten a algunos puntos estar más allá del margin, o incluso en el lado equivocado de la frontera. Una elección típica para esta penalización se muestra a continuación (izquierda). Como resultado, se obtienen versiones ligeramente modificadas de la forma primal, que se puede resolver aplicando ideas similares.



Las penalizaciones se introducen en la forma primal como variables *slack* ε_n.



Frontera de decisón de la SVM para un conjunto de entrenamiento que no es linealmente separable con el kernel gaussiano.