Tema 6
Extensiones y aplicaciones
(Máquinas de vectores soporte, SVM)

José R. Berrendero

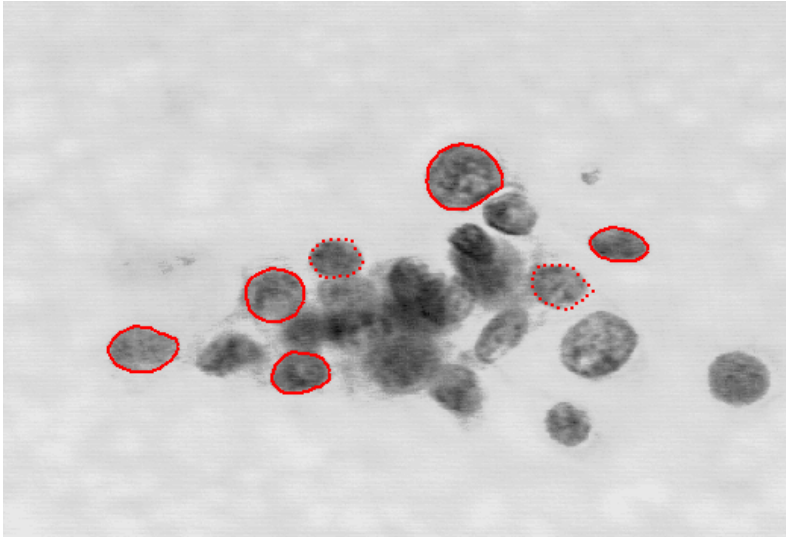
Departamento de Matemáticas
Universidad Autónoma de Madrid

Contenidos del tema 6

- El problema de clasificación supervisada: un ejemplo
- SVM para datos separables linealmente.
- SVM para datos no separables linealmente.
- Reglas de clasificación no lineales: el truco del núcleo.

Diagnóstico por imagen del cáncer de mama

- Punción con aguja fina.
- La muestra se tiñe para resaltar los núcleos de las células.
- Se determinan los límites exactos de los núcleos.
- Las variables corresponden a distintos aspectos de su forma.

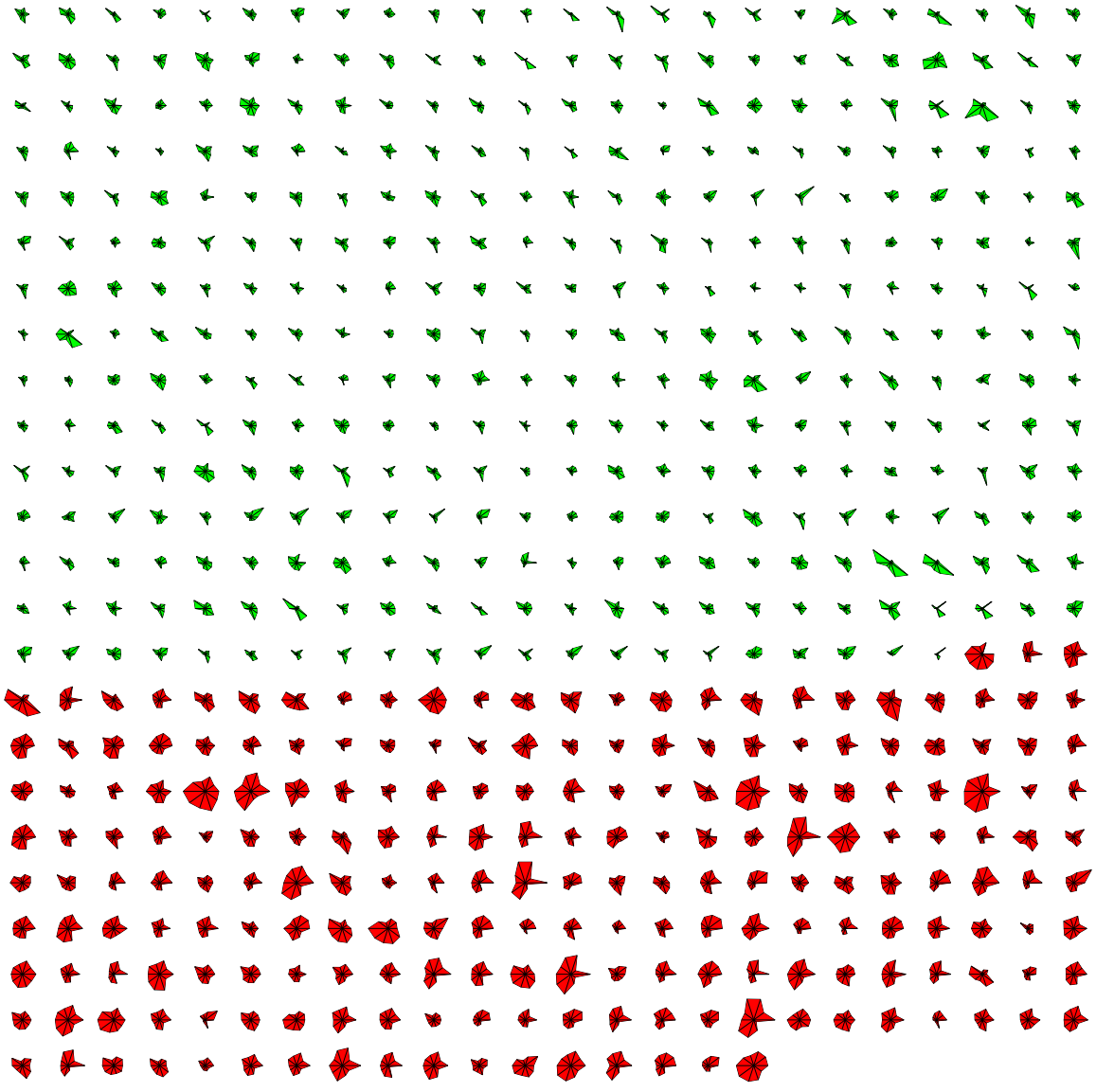


Variables

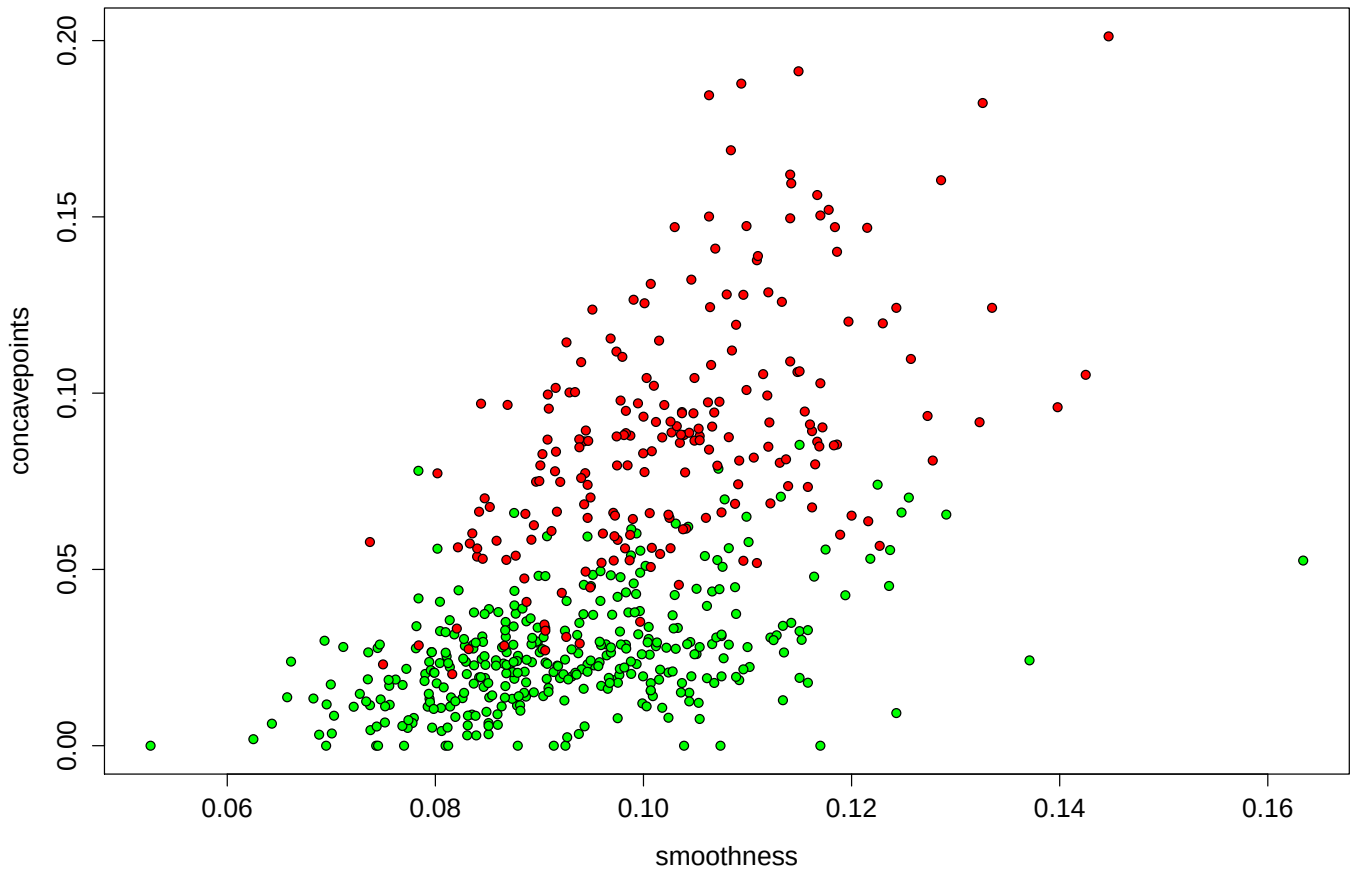
nombre	descripción
radius	radio del núcleo
texture	varianza de los niveles de gris en el interior del núcleo
perimeter	perímetro del núcleo
area	área del núcleo
smoothness	suavidad medida mediante la variación del radio
compactness	el perímetro al cuadrado dividido por el área
concavity	medida de la importancia de las concavidades
concavepoints	número de concavidades
symmetry	medida de la simetría del núcleo
fractal	dimensión fractal de la frontera

Más información sobre estos datos

Gráfico de estrellas



Variables *smoothness* y *concavepoints*



Clasificación supervisada

Disponemos de una muestra de datos bien clasificados (*training data*):

$$(x_1, y_1), \ldots, (x_n, y_n)$$

donde $x_i \in \mathbb{R}^d$ son las variables observadas e $y_i \in \{-1, 1\}$ es la etiqueta que representa la clase a la que pertenecen las observaciones.

Se observa ahora un nuevo vector x independiente de los anteriores.
El objetivo es determinar a qué clase pertenece la observación x.

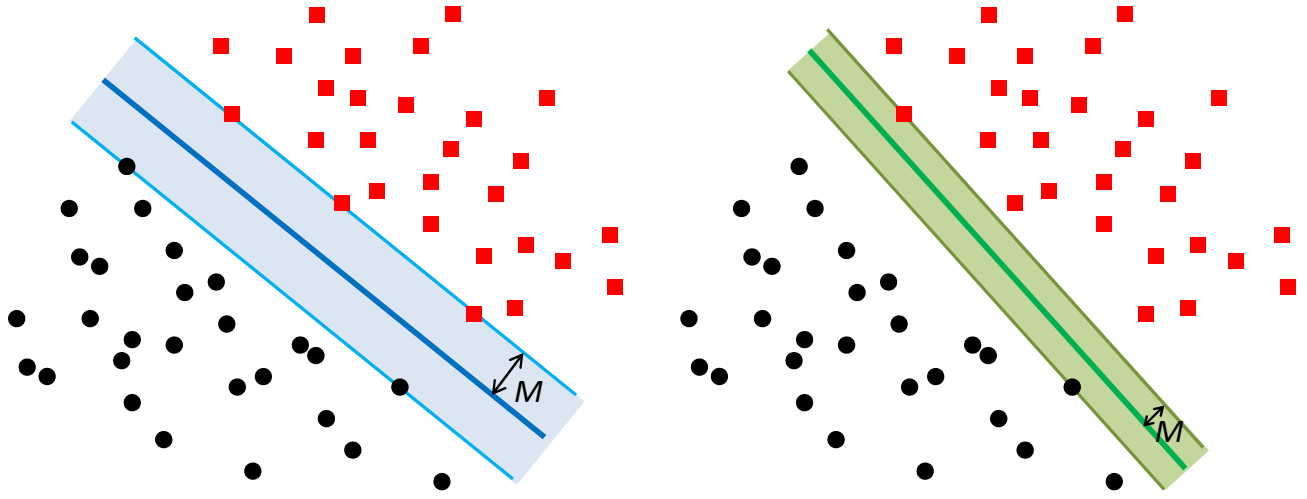
La regla óptima (regla Bayes) consiste en asignar a x el valor $y = 1$ si y solo si

$$\mathbb{P}(y = 1|x) > \mathbb{P}(y = -1|x)$$

No es aplicable en la práctica.

SVM para datos separables linealmente

- Suponemos que las muestras de ambos grupos son separables mediante un hiperplano.



- El **margen** de un hiperplano separador viene dado por la menor distancia de los puntos al hiperplano.
- El **hiperplano óptimo** es aquel que maximiza el margen.

Distancia de un punto a un hiperplano

- Distancia de un punto x al hiperplano $w'x + w_0 = 0$.

  Sea $\hat{x}$ el punto del hiperplano más cercano a x. Entonces,

$$x = \hat{x} + r\frac{w}{\|w\|} \Rightarrow w'x + w_0 = (w'\hat{x} + w_0) + r\|w\| = r\|w\|.$$

- La distancia de un punto x al hiperplano es:

$$d = |r| = \frac{|w'\hat{x} + w_0|}{\|w\|}.$$

Margen

- Disponemos de una muestra de datos clasificados (x_i, y_i), $i = 1, \ldots, n$.
- La clase es $y_i \in \{-1, 1\}$.

- Un hiperplano separador verifica $y_i(w'x_i + w_0) > 0$, para todo $i = 1, \ldots, n$.
- Siempre podemos definir w y w_0 de manera que

$$\min_i \{y_i(w'x_i + w_0)\} = 1.$$

- El margen es

$$\text{Margen} = \min_i \frac{y_i(w'x + w_0)}{\|w\|} = \frac{1}{\|w\|}.$$

Hiperplano separador óptimo

- Buscamos el hiperplano separador que maximiza el margen.
- Tenemos que resolver el problema convexo

$$\begin{array}{ll} \text{minimizar} & \|w\|^2/2 \\ \text{s.a.} & y_i(w'x_i + w_0) \geq 1, \quad i = 1, \ldots, n \end{array}$$

- La función lagrangiana de este problema es

$$L(w, w_0) = \frac{\|w\|^2}{2} - \sum_{i=1}^{n} u_i[y_i(w'x_i + w_0) - 1]$$

Condiciones KKT

Las condiciones de Karush-Kuhn-Tucker que debe satisfacer la solución de este problema son:

- El gradiente de la función lagrangiana se anula
- Se cumplen las restricciones del problema
- Los multiplicadores no son negativos.
- Se cumplen las condiciones de holgura complementaria.

Estas condiciones permiten deducir algunas propiedades importantes de la solución.

Condiciones KKT

$$\nabla L(\hat{w}, \hat{w}_0) = 0 \Rightarrow \hat{w} = \sum_{i=1}^{n} \hat{u}_i y_i x_i \quad \text{y} \quad \sum_{i=1}^{n} \hat{u}_i y_i = 0.$$

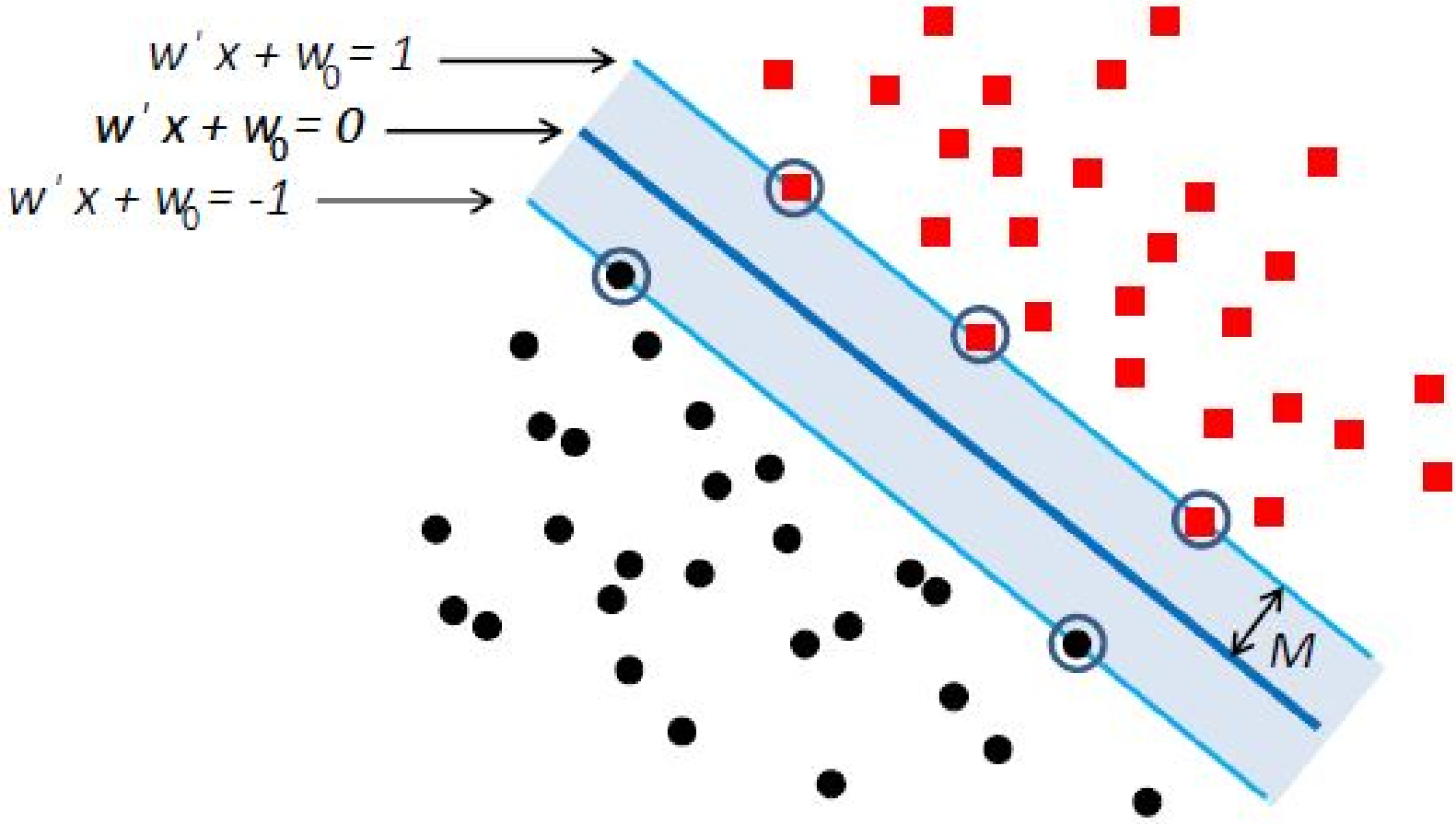
Para $i = 1, \ldots, n$,

$$y_i(\hat{w}' x_i + \hat{w}_0) \geq 1, \quad \hat{u}_i \geq 0$$

$$\hat{u}_i \big(y_i(\hat{w}' x_i + \hat{w}_0) - 1 \big) = 0$$

El hiperplano óptimo solo depende de aquellos puntos de los que está más cerca ($y_i(\hat{w}' x_i + \hat{w}_0) > 1 \Rightarrow \hat{u}_i = 0$).

Típicamente son pocos. Se llaman **vectores soporte**.


$w^T x + w_0 = 1$
$w^T x + w_0 = 0$
$w^T x + w_0 = -1$
M

Problema dual

- **Función dual** (se obtiene minimizando la función lagrangiana en w y w_0):

$$g(u) = \sum_{i=1}^{n} u_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} u_i u_j y_i y_j x_i' x_j$$

  si $\sum_{i=1}^{n} u_i y_i = 0$, y $g(u) = -\infty$ en caso contrario.

- **Problema dual**:

$$\begin{array}{ll} \text{maximizar} & g(u) \\ \text{s.a.} & \sum_{i=1}^{n} u_i y_i = 0. \\ & u_i \geq 0, \quad i = 1, \ldots, n. \end{array}$$

Problema dual

En forma matricial, si $\mathbf{1}_n = (1, \ldots, 1)' \in \mathbb{R}^n$, $y = (y_1, \ldots, y_n)'$ y H es la matriz cuyas entradas son $h_{ij} = y_i y_j x_i' x_j$,

$$\begin{array}{ll} \text{maximizar} & u'\mathbf{1}_n - \frac{1}{2} u' H u \\ \text{s.a.} & u' y = 0 \\ & u \geq 0. \end{array}$$

- Es un problema de optimización convexo (la matriz H es definida positiva).
- La solución depende de $x_1, \ldots, x_n$ únicamente a través de los productos escalares $x_i' x_j$.

Cálculo del hiperplano óptimo

- Resolvemos el problema dual mediante algún método de programación convexa estándar.

- A partir de la solución del dual, $\hat{u}$, aplicamos $\hat{w} = \sum_{i=1}^{n} \hat{u}_i y_i x_i$ para obtener $\hat{w}$.

- Sean $S = \{i : \hat{u}_i > 0\}$ los índices de los vectores soporte. Por las condiciones de holgura complementaria, para cada $i \in S$,

$$\hat{w}_0 = \frac{1 - y_i \hat{w}' x_i}{y_i} = y_i - \hat{w}' x_i.$$

- En la práctica, es numéricamente más estable usar el promedio de estos valores. Si $\#S = n_s$.

$$\hat{w}_0 = \frac{1}{n_s} \sum_{i \in S} (y_i - \hat{w}' x_i).$$

Regla de clasificación

Resulta una regla de clasificación lineal: asignamos a x el valor $y = 1$ si y solo si $\hat{w}'x + \hat{w}_0 > 0$.

$$\hat{w}_0 + \hat{w}'x > 0 \Leftrightarrow \hat{w}_0 + \left[\sum_{i \in S} y_i \hat{u}_i x_i\right]' x > 0$$

Si $\alpha_i = y_i \hat{u}_i$, también podemos escribir la regla de clasificación como:

$$y = 1 \Leftrightarrow \hat{w}_0 + \sum_{i \in S} \alpha_i (x_i' x) > 0$$

¿Cómo afecta a la clasificación una rotación de los datos?

SVM para datos no separables linealmente

En la práctica, la mayoría de los datos no son separables linealmente.

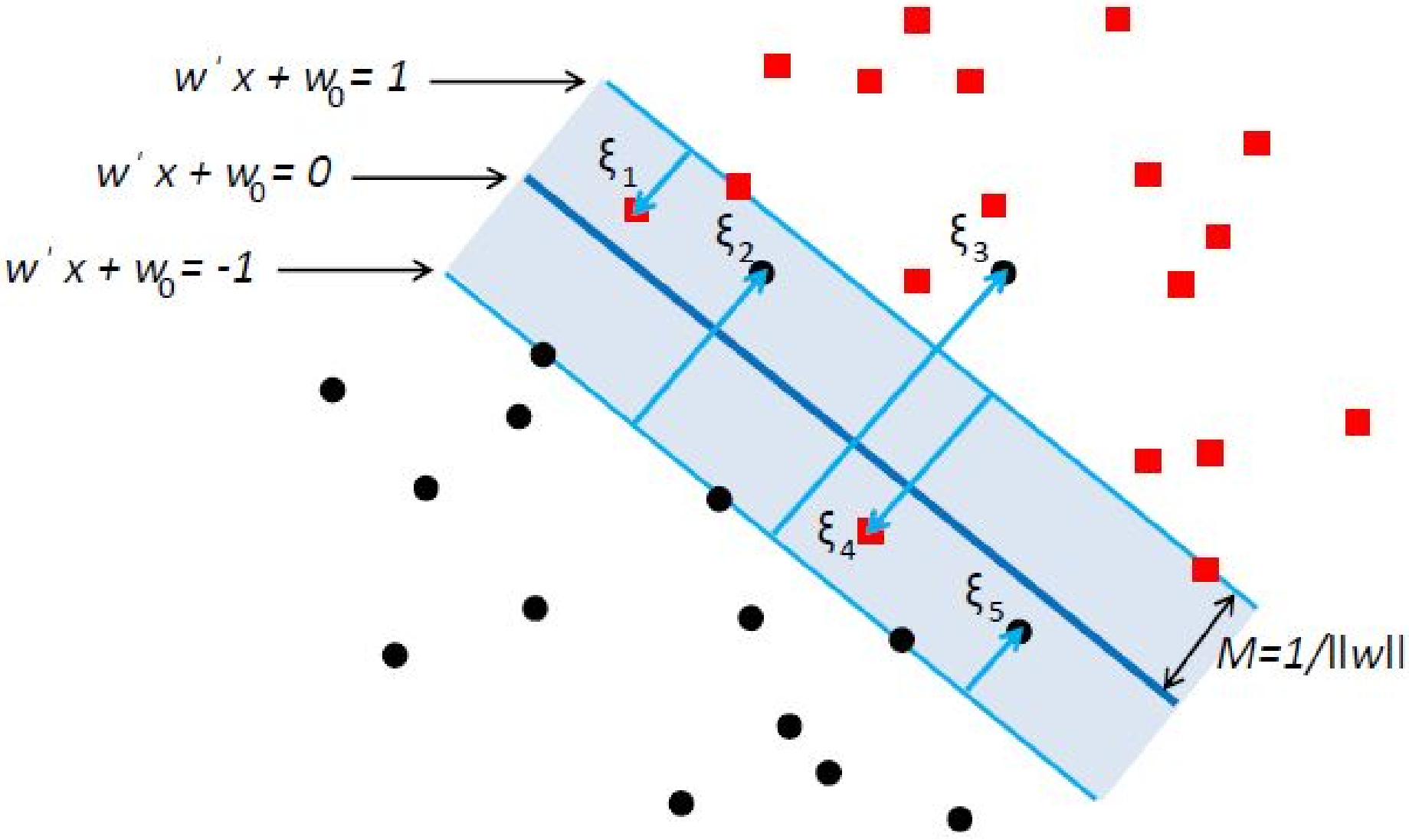
Se introducen unas variables de holgura $\xi_1, \ldots, \xi_n$ de manera que:

- se relajan las restricciones con el fin de permitir errores de clasificación,
- se cambia el objetivo para penalizar estos errores.

$$\begin{array}{lll} \text{minimizar} & \|w\|^2/2 + C\sum_{i=1}^{n} \xi_i & \\ \text{s.a.} & y_i(w'x_i + w_0) + \xi_i \geq 1, & i = 1, \ldots, n \\ & \xi_i \geq 0, & i = 1, \ldots, n \end{array}$$

La constante $C > 0$ es seleccionada por el usuario y determina si los errores se penalizan más o menos.

SVM para datos no separables linealmente



Condiciones KKT

$$L(w, w_0, u, v) = \frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n} u_i[y_i(w'x_i + w_0) + \xi_i - 1] - \sum_{i=1}^{n} v_i\xi_i$$

- Gradiente de L igual a cero:

$$w = \sum_{i=1}^{n} u_i y_i x_i; \quad \sum_{i=1}^{n} u_i y_i = 0.$$

- Factibilidad primal y dual:

$$y_i(w'x_i + w_0) + \xi_i \geq 1; \quad \xi_i \geq 0; \quad 0 \leq u_i \leq C.$$

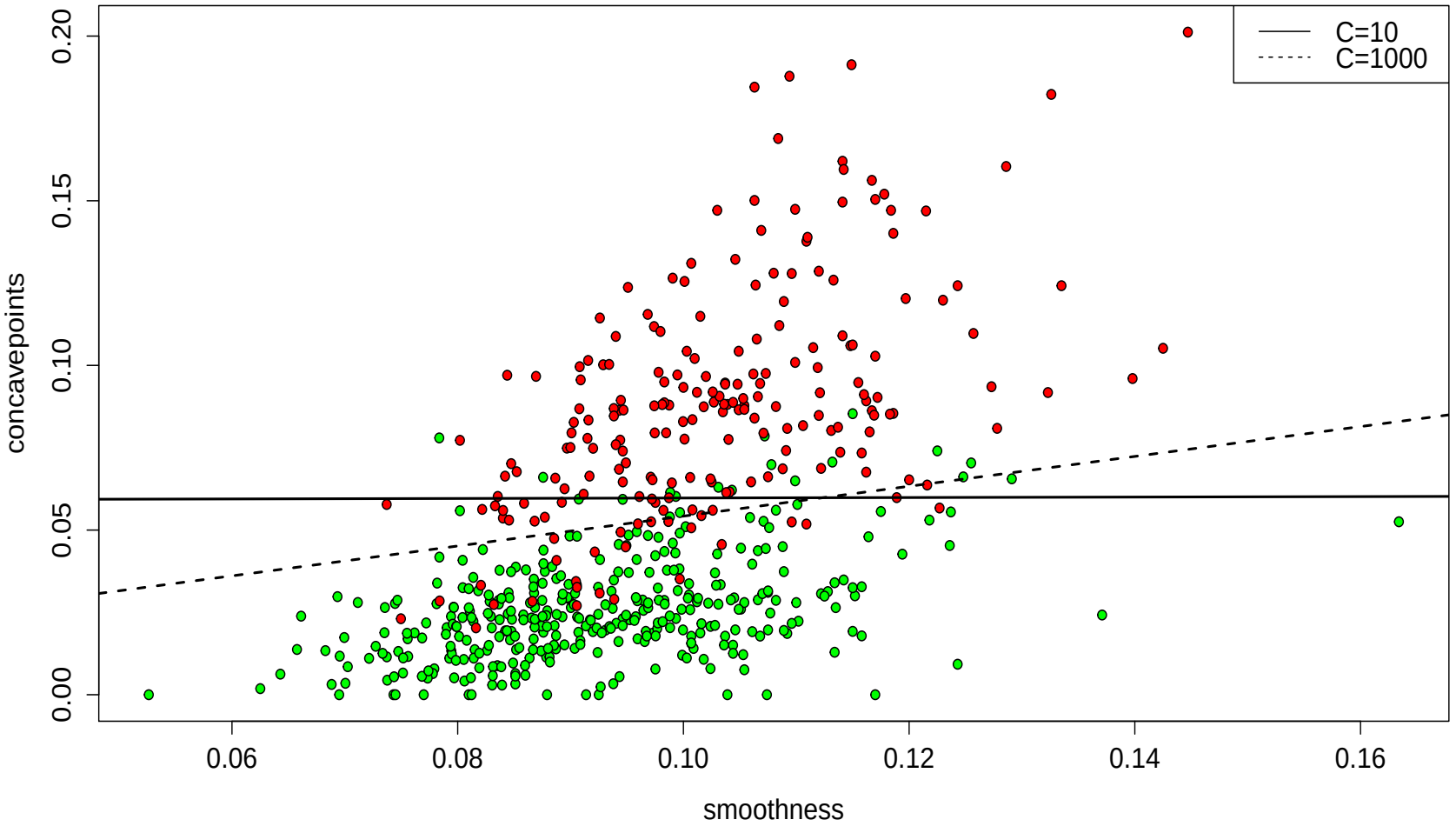
- Holgura complementaria:

$$u_i[y_i(w'x_i + w_0) + \xi_i - 1] = 0; \quad (C - u_i)\xi_i = 0.$$

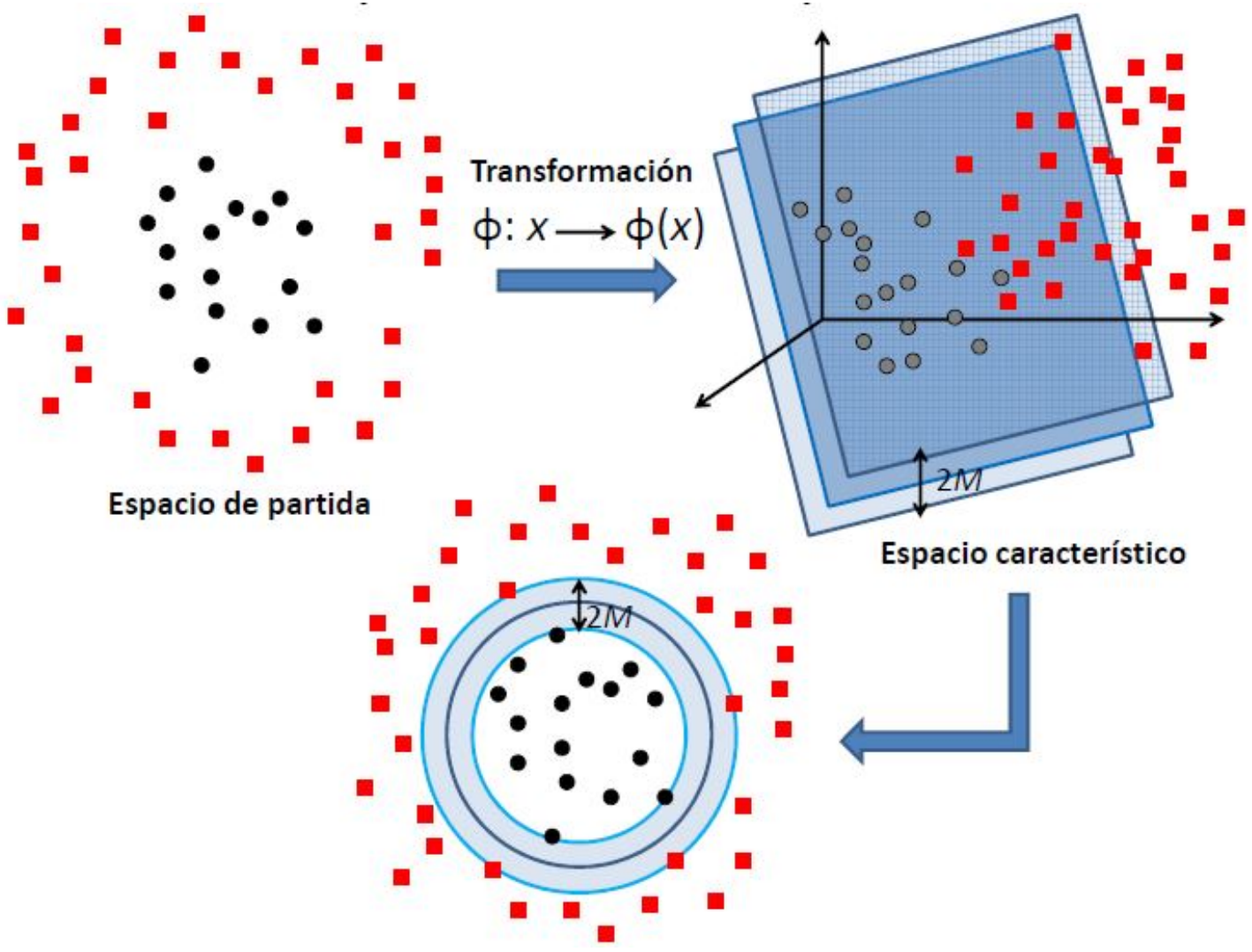
Cuestiones

- Escribe la función y el problema dual. ¿Qué diferencias se observan respecto al caso en que los datos son separables linealmente?
- ¿Qué condición deben verificar en este caso los vectores soporte?
- Si $\hat{u}_i$, $i = 1, \ldots, n$ es la solución del problema dual, ¿cómo se calculan $\hat{w}$ y $\hat{w}_0$
- Escribe la regla de clasificación.

Ejemplo. SVM para datos no separables linealmente



Extensión a reglas no lineales



Extensión a reglas no lineales

- Es posible que una regla de clasificación lineal no sea apropiada para los datos originales $x_1, \ldots, x_n$ pero sí para los datos transformados $\phi(x_1), \ldots, \phi(x_n)$, donde $\phi : \mathbb{R}^d \to \mathcal{H}$ para un espacio de Hilbert $\mathcal{H}$.

- Basta sustituir $x_i' x_j$ por $\langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$ en el problema dual.

- Típicamente $\mathcal{H} = \mathbb{R}^N$ con $N >> d$ o $\mathcal{H}$ es un espacio de funciones (dimensión infinita).

- En la práctica, puede ser difícil calcular los productos escalares $\langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$.

El truco del núcleo (*the kernel trick*)

Teorema: Una función $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ es simétrica y semidefinida positiva (SDP) si y solo si existe un espacio de Hilbert $\mathcal{H}$ y una transformación $\phi : \mathbb{R}^d \to \mathcal{H}$ tal que $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$.

- Estas funciones simétricas y SDP se llaman núcleos.
- En la práctica, en lugar de elegir $\mathcal{H}$ y ϕ, se elige un núcleo y se sustituye $x_i' x_j$ por $k(x_i, x_j)$.
- Así se obtienen reglas de clasificación de la forma:

$$y = 1 \Leftrightarrow \hat{w}_0 + \sum_{i \in S} \alpha_i k(x_i, x) > 0.$$

Algunos núcleos muy utilizados

- Polinomios de grado m:

$$k(x, y) = (x'y + c)^m.$$

- Gaussiano:

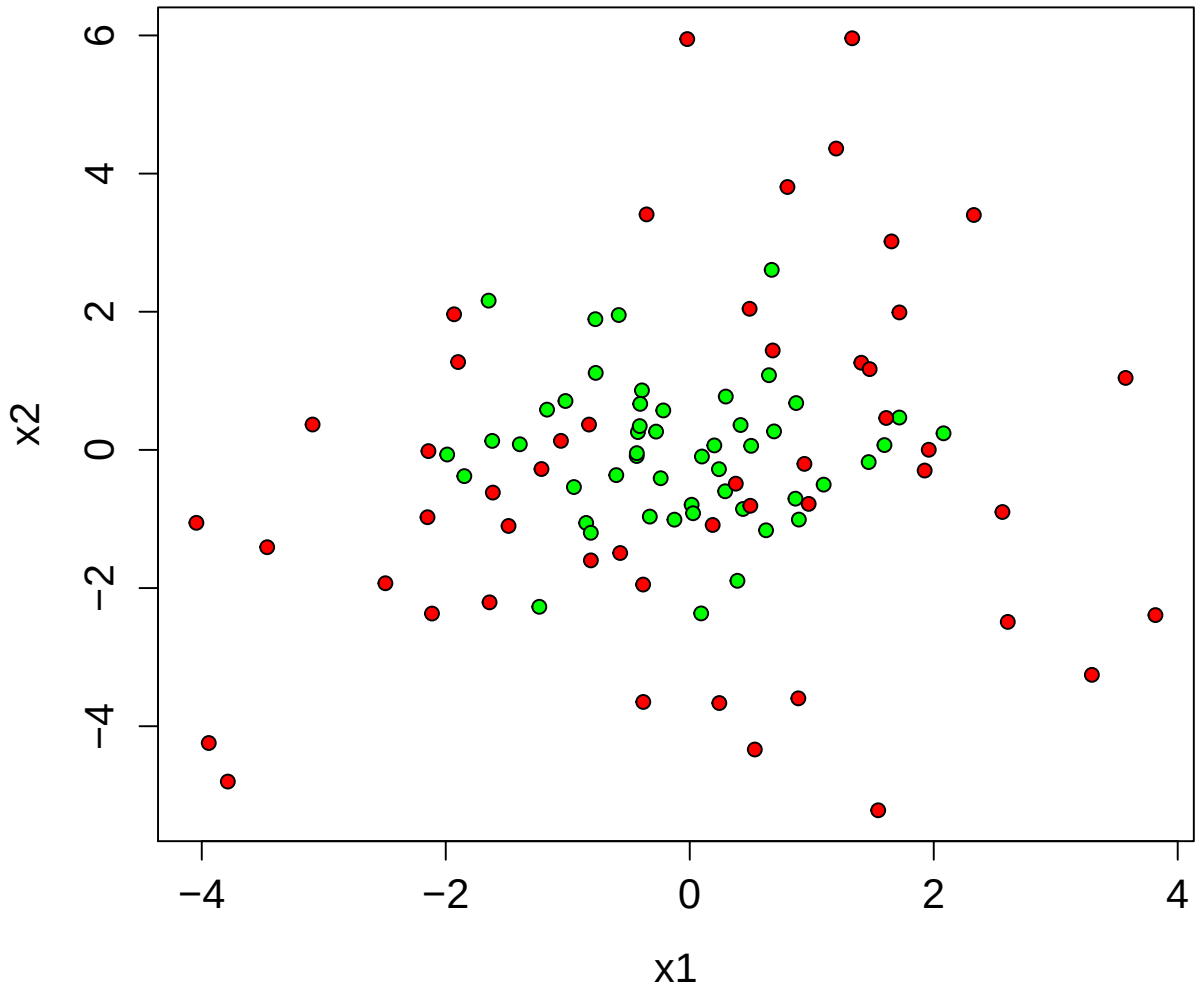
$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$$

- Laplaciano:

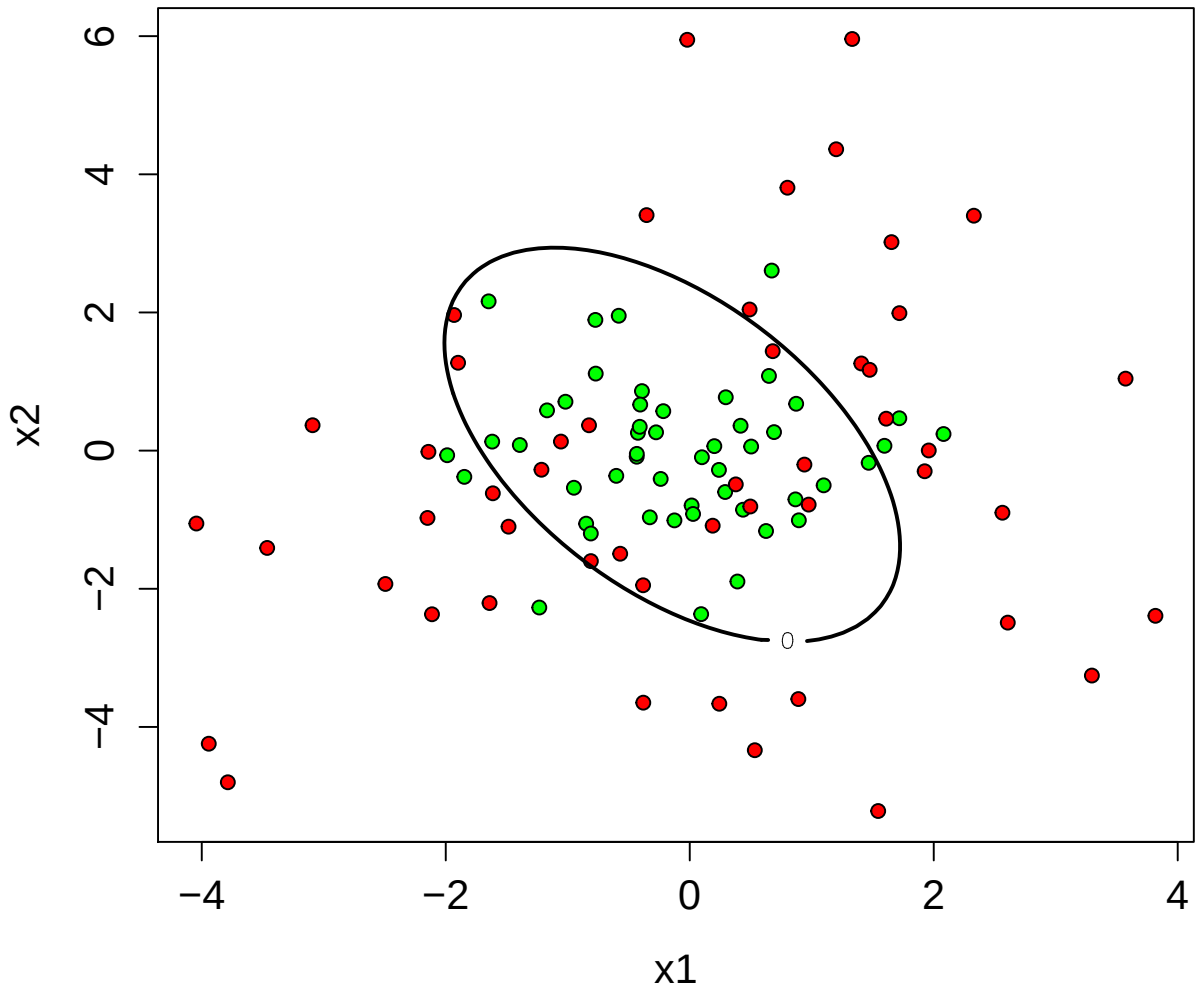
$$k(x, y) = \exp\left(-\frac{\|x - y\|}{\sigma}\right)$$

Para cada problema concreto hay que usar un núcleo apropiado.
Un polinomio de grado pequeño o el núcleo gaussiano suelen ser buenas primeras opciones.

Regla de clasificación con núcleo cuadrático

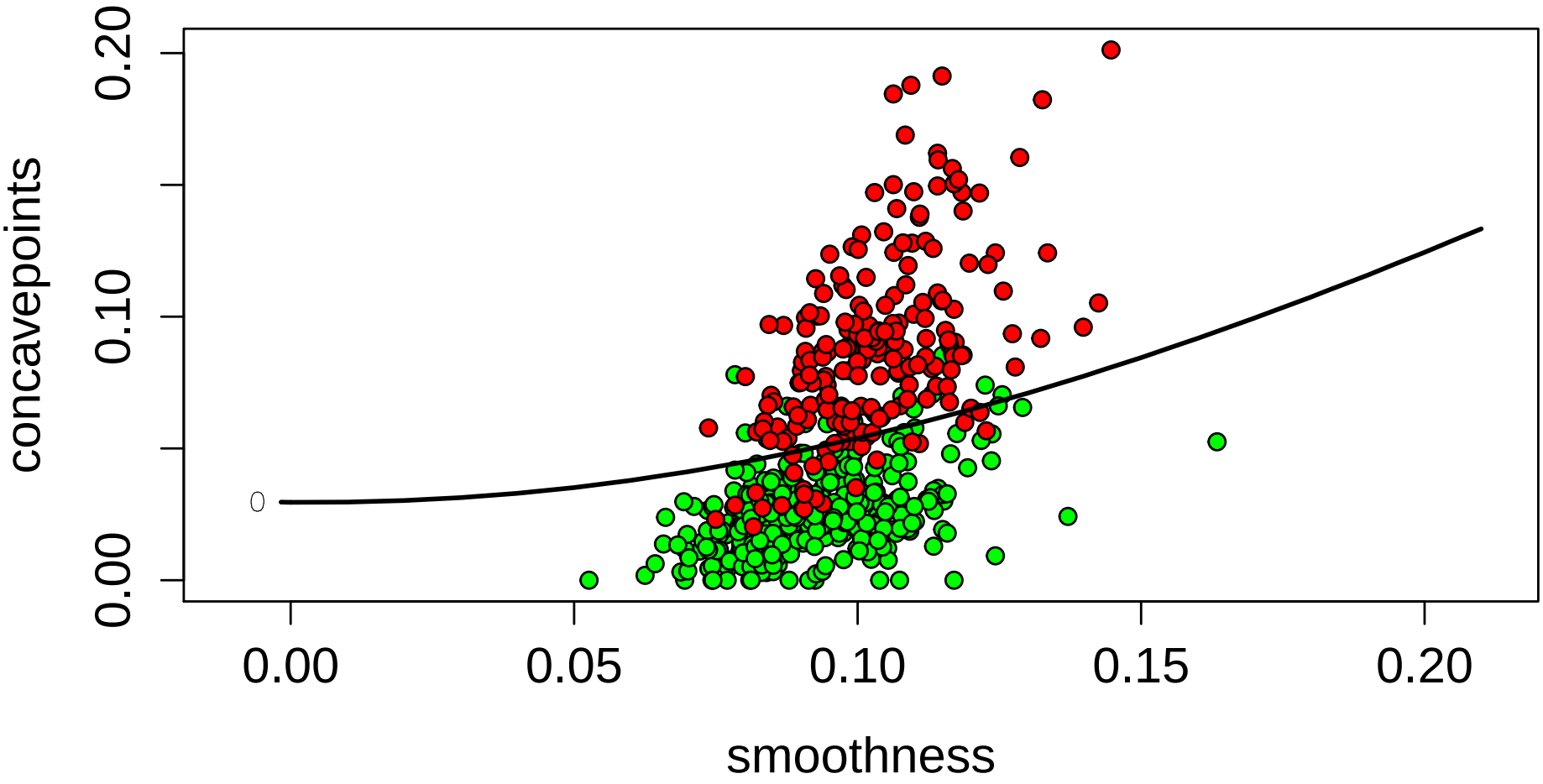


Regla de clasificación con núcleo cuadrático



Resultado para un núcleo cuadrático con $C = 100$, $m = 2$ y $c = 1$

Regla de clasificación con núcleo cuadrático



Resultado para un núcleo cuadrático con $C = 100000$, $m = 2$ y $c = 1$