Máquinas de Vectores de Soporte

Support Vector Machines (SVM)

Reconocimiento de Patrones

Departamento de Procesamiento de Señales
Instituto de Ingeniería Eléctrica
Facultad de Ingeniería, UdelaR

2018

Schölkopf & Smola, Cap. 7 – Bishop, Cap. 4 – Duda et al., Cap. 5 – Hastie et al., Cap 4

Motivación

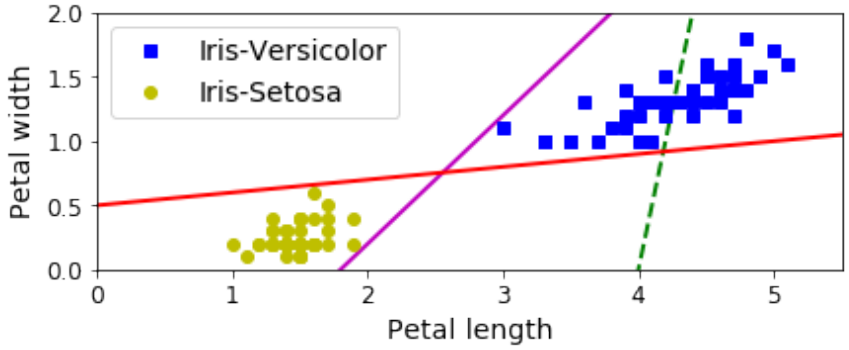
Limitantes del perceptrón:

- Sólo contempla el caso linealmente separable
- No hay unicidad de la solución

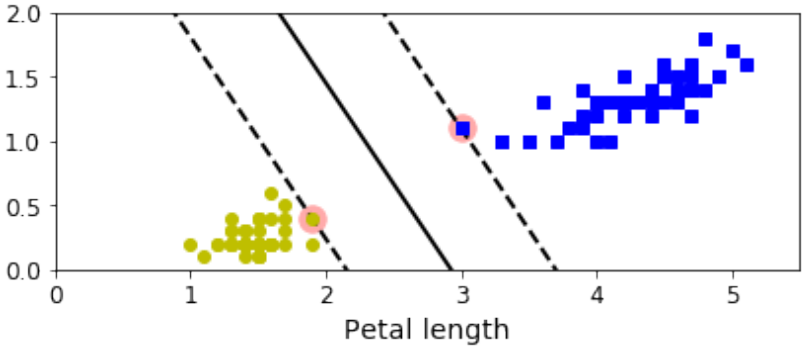
Limitantes de los métodos de Kernel:

- En training: irrealizable si se requiere evaluar $k(\mathbf{x}_n, \mathbf{x}_m)$ para todo par de muestras de entrenamiento
- En testing o clasificación de nuevo $\mathbf{x}$: costo computacional de evaluar $k(\mathbf{x}, \mathbf{x}_n)$ para todo $\mathbf{x}_n$ de entrenamiento puede ser excesivo.

Motivación
Caso linealmente separable



Tres posibles clasificadores lineales



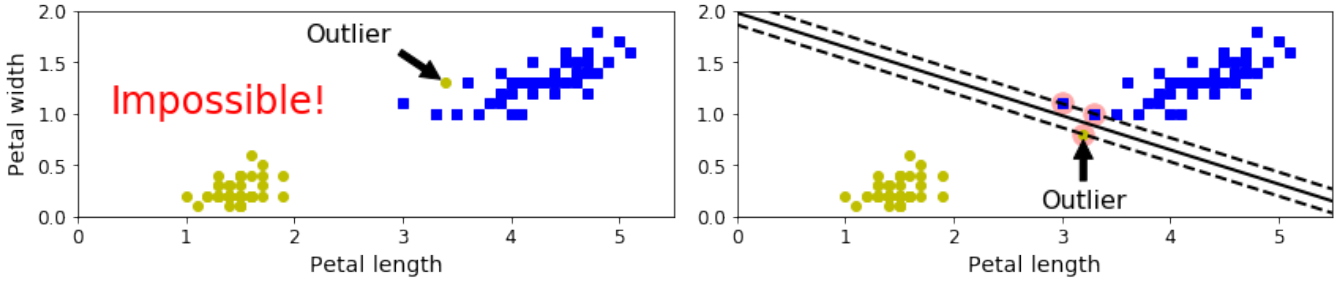
SVM lineal, con sus márgenes

Clasificador SVM: hiperplano separador más distante de las muestras de entrenamiento.

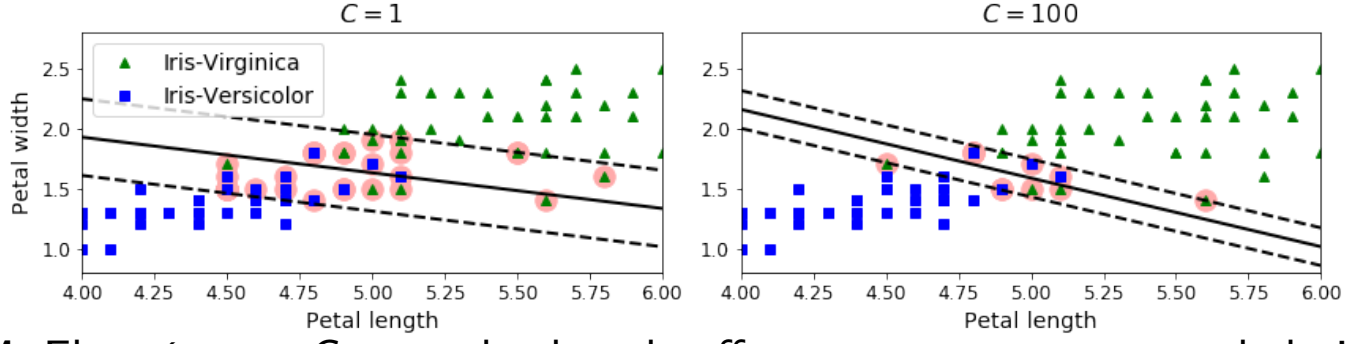
- Intuición: generaliza mejor.
- Agregar muestras fuera de los márgenes no afecta la frontera de decisión. Completamente determinada por las muestras que soportan los márgenes (**vectores de soporte**, SV).

Motivación
Caso NO linealmente separable

SVM lineal: ¿Qué ocurre en el ejemplo anterior si hay outliers?



Solución: permitirnos ampliar el margen a costa de admitir algunas muestras mal clasificadas ⇒ *soft margin* SVM.



C-SVM. El parámetro C controla el trade off margen vs. muestras mal clasificadas

Motivación

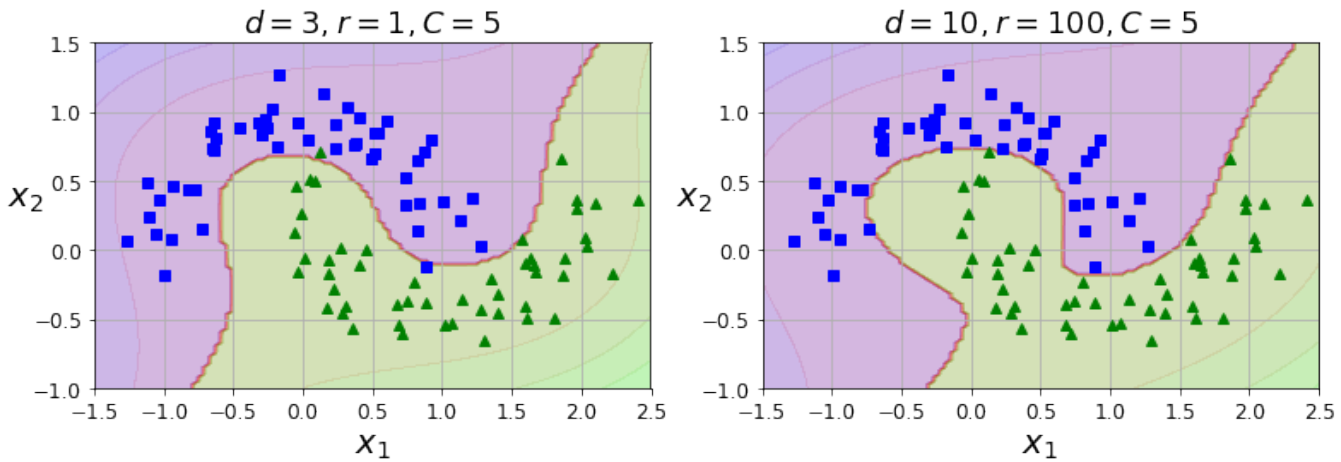
SVM no lineal, kernels, *kernel trick*

Agregando características no lineales se puede llegar a transformar un problema no linealmente separable en uno linealmente separable.

Kernel polinómico

Ejemplo: $\phi(x_1, x_2) = [1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1x_2]^T$.

Kernel Trick: $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T\phi(\mathbf{x}') = (1 + \mathbf{x}^T\mathbf{x}')^2$. Kernel polinómico de orden 2.

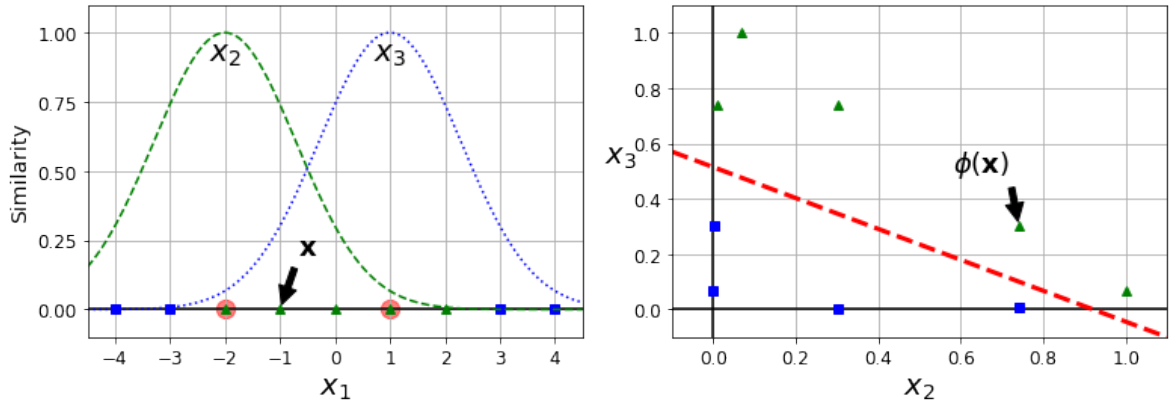


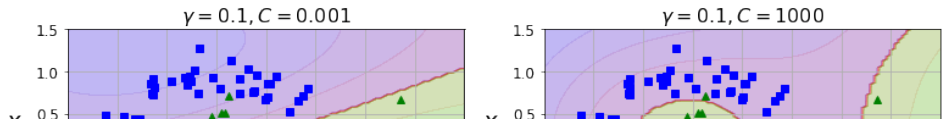
$$k(\mathbf{x}, \mathbf{x}') = (r + \mathbf{x}^T\mathbf{x}')^d$$

Kernel RBF Gaussiano

Idea: agregar características calculadas con una función de similitud que mida cuan parecida es una instancia a un landmark particular: $k(\mathbf{x}, \mathbf{x}') = \mathrm{e}^{-\gamma\|\mathbf{x}-\mathbf{x}'\|^2}$.

- Landmarks: se suelen tomar todas las muestras
- $\phi(\mathbf{x})$ mapea a un espacio de dim. infinita (Taylor de $\exp(\cdot)$).





Agenda

- Resumiendo:
  - SVM busca el hiperplano separador (en el espacio transformado) que maximiza el margen
  - SVM permite contemplar el caso no linealmente separable
  - SVM permite controlar *trade-off* entre margen y errores en training
- Veremos que:
  - **Parámetros:** se determinan resolviendo un problema convexo
  - **Clasificación de nuevas muestras $\mathbf{x}$:** depende sólo de los $k(\mathbf{x}, \mathbf{x}_n)$ con $\mathbf{x}_n$ SV (subconjunto muy reducido de muestras)
- Discutiremos qué sucede con SVM y la maldición de la dimensionalidad

Problema de dos clases

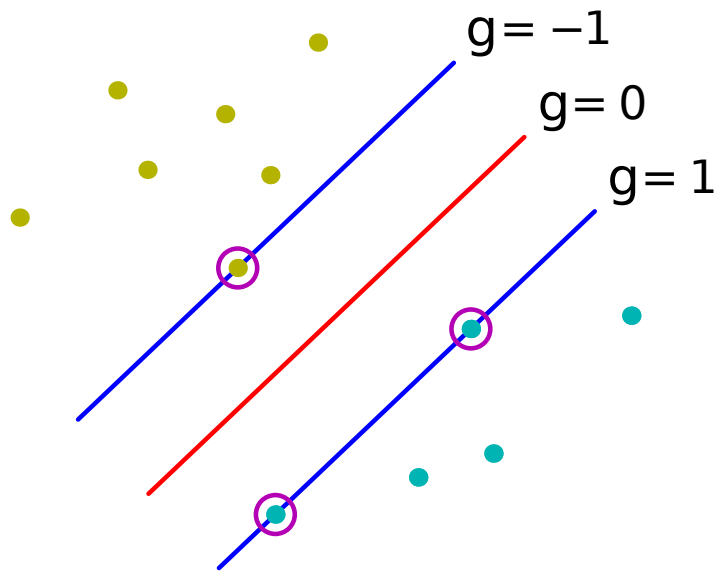
- $\{(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2), \ldots, (\mathbf{x}_N, t_N)\}$ conjunto de entrenamiento, $\mathbf{x}_n \in \mathcal{X}$ espacio de características, $t_n \in \{-1, +1\}$ etiqueta de clase.
- $\phi : \mathcal{X} \to \mathcal{Y}$ mapeo no lineal a un espacio de dimensión mayor, $\dim(\mathcal{Y}) > \dim(\mathcal{X})$.
- Superficie de decisión:
  $g(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$, $\mathcal{S} = \{\mathbf{x} : \ g(\mathbf{x}) = 0\}$.
  En el espacio $\mathcal{Y}$: hiperplano de normal $\frac{\mathbf{w}}{\|\mathbf{w}\|}$ y sesgo $\frac{-b}{\|\mathbf{w}\|}$.
- Clasificación: nueva muestra $\mathbf{x}$ se clasifica según $\text{signo}(g(\mathbf{x}))$.

Conjunto de aprendizaje linealmente separables

Problema linealmente separable $\Rightarrow$ Existe al menos un par $(\mathbf{w}, b)$ para el cual todas las muestras están bien clasificadas, i.e.

$$\forall\, n = 1, \ldots, N,\ t_n g(\mathbf{x}_n) > 0.$$

SVM: entre los $(\mathbf{w}, b)$ posibles, elegir aquél que corresponde al hiperplano separador de margen máximo.



Clasificador de máximo margen

¿Por qué? Maximizar el margen minimiza el error de generalización del clasificador (es intuitivo y se puede demostrar [Vapnik, 1996])

Tenemos:

- Distancia de un punto $\mathbf{y} = \phi(\mathbf{x})$ al plano $\{g(\mathbf{x}) = 0\}$: $\frac{|g(\mathbf{x})|}{\|w\|}$.
- Caso linealmente separable $\Rightarrow$ para cualquier hiperplano separador,

$$|g(\mathbf{x}_n)| = t_n g(\mathbf{x}_n), \quad n = 1, \ldots, N.$$

$\Rightarrow$ **Solución de máximo margen:**

$$(P) \qquad \max_{\mathbf{w},b} \min_{n \in \{1,\ldots,N\}} \left\{ \frac{t_n \left(\mathbf{w}^T \phi(\mathbf{x}_n) + b \right)}{\|\mathbf{w}\|} \right\}$$

Optimización del clasificador de máximo margen

- (P) problema de optimización complejo.
- Se puede transformar en un problema equivalente (P') sencillo de optimizar, observando que $f_n(\mathbf{w}, b) = \frac{t_n(\mathbf{w}^T \phi(\mathbf{x}_n)+b)}{\|\mathbf{w}\|}$ es invariante ante escalados:

$$\forall\, \lambda > 0,\ f_n(\lambda \mathbf{w}, \lambda b) = f_n(\mathbf{w}, b).$$

$\Rightarrow$ Podemos re-escalar para que $t_k(\mathbf{w}^T \phi(\mathbf{x}_k) + b) = 1$ para el (o los) $\mathbf{x}_k$ que realizan el margen.

$\Rightarrow$ Con esa normalización,

$$\forall\, n = 1, \ldots, N,\ t_n(\mathbf{w}^T \phi(\mathbf{x}_n) + b) \geq 1.$$

Optimización del clasificador de máximo margen

Problema equivalente para clasificador de máximo margen:

$$(P') \qquad \begin{cases} \min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{sujeto a } t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b) \geq 1, \quad n=1,\ldots,N. \end{cases}$$

- Puntos activos/inactivos:

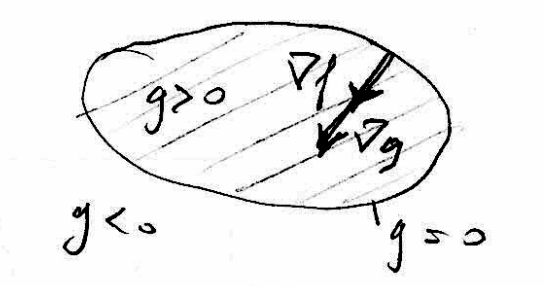
  si $t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b) \begin{cases} = 1 & \mathbf{x}_n \text{ punto activo} \\ > 1 & \mathbf{x}_n \text{ punto inactivo.} \end{cases}$

  Al maximizar el margen habrá al menos dos puntos activos.

- (P') es un problema de programación cuadrática fácil de resolver (e.g. solvers de Matlab, CVX, etc).
- (P') (problema primal) admite una formulación aún más sencilla (problema dual) *via* las condiciones de KKT.

Paréntesis: condiciones de Karush-Kuhn-Tucker

Consideramos el problema $\begin{cases} \min f(\mathbf{x}) \\ \text{s.t. } g(\mathbf{x}) \geq 0 \end{cases}$



Dos posibilidades para la solución:

(i) $\mathbf{x} \in \{g > 0\}$: entonces $\mathbf{x}$ mínimo local de $f \Rightarrow \nabla f(\mathbf{x}) = 0$.

(ii) $\mathbf{x} \in \{g = 0\}$: entonces debe ser $\nabla f(\mathbf{x}) \perp \{g = 0\}$, de lo contrario existiría $\mathbf{x}' \in \{g = 0\}$ vecino de $\mathbf{x}$ con $f(\mathbf{x}') < f(\mathbf{x})$.

- Como $\nabla g(\mathbf{x}) \perp \{g = 0\}$, $\exists \lambda \neq 0$ t.q. $\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$.
- Además debe ser $\lambda > 0$, de lo contrario existiría $\mathbf{x}' \in \{g > 0\}$ vecino de $\mathbf{x}$ con $f(\mathbf{x}') < f(\mathbf{x})$.

Paréntesis: condiciones de Karush-Kuhn-Tucker

- Definimos el Lagrangeano $\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$.
- Las dos situaciones de solución se resumen en una única (KKT):

$$\boxed{\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = 0, \quad \lambda \geq 0, \quad \lambda g(\mathbf{x}) = 0}$$

- **Obs.1:** La solución $(\mathbf{x}^*, \lambda^*)$ es un punto silla de $\mathcal{L}(\mathbf{x}, \lambda)$:

$$\forall\ (\mathbf{x}, \lambda), \quad \mathcal{L}(\mathbf{x}^*, \lambda) \leq \mathcal{L}(\mathbf{x}^*, \lambda^*) \leq \mathcal{L}(\mathbf{x}, \lambda^*).$$

Ejercicio Demostrar

- **Obs.2:** De Obs.1, la optimización correspondiente es

$$(\mathbf{x}^*, \lambda^*) = \arg \min_{\mathbf{x}} \max_{\lambda \geq 0} \mathcal{L}(\mathbf{x}, \lambda).$$

Vuelta al clasificador de máximo margen (caso separable)

Teníamos (P') $\left\{ \begin{array}{l} \min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{sujeto a } t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b) \geq 1, \quad n=1,\ldots,N. \end{array} \right.$

Llamamos $\mathbf{a} = (a_1, a_2, \ldots, a_N)$ a los multiplicadores de Lagrange:

$$\boxed{\mathcal{L}(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1}^{N} a_n \left(t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b) - 1\right).}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{w}} = 0 \Rightarrow \boxed{\mathbf{w} = \sum_{n=1}^{N} a_n t_n \phi(\mathbf{x}_n)} (*) \qquad \frac{\partial \mathcal{L}}{\partial b} = 0 \Rightarrow \boxed{\sum_{n=1}^{N} a_n t_n = 0} (**)$$

$$\overset{(*),(**)}{\Longrightarrow} \mathcal{L}(\mathbf{w}, b, \mathbf{a}) = \tilde{\mathcal{L}}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \underbrace{\phi^T(\mathbf{x}_n)\phi^(\mathbf{x}_m)}_{k(\mathbf{x}_n,\mathbf{x}_m)}$$

Optimización del clasificador de máximo margen (separable)

El nuevo problema equivalente (llamado problema dual porque involucra sólo a los multiplicadores) queda:

$$(D') \qquad \begin{cases} \max_{\mathbf{a}} \tilde{\mathcal{L}}(\mathbf{a}) \\ \text{sujeto a } \sum_{n=1}^{N} a_n t_n = 0, \quad a_n \geq 0, \;\; n = 1, \ldots, N. \end{cases}$$

Problema de programación cuadrática más sencillo (involucra únicamente a los multiplicadores)

Clasificador de máximo margen (caso separable)

- **Clasificación de nueva muestra $\mathbf{x}$:** se evalúa el signo de $g(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}_n) + b \overset{(*)}{=} \sum_{n=1}^{N} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$
  (veremos luego cómo calcular el sesgo b).

- **Vectores de soporte:** De las condiciones KKT,

$$a_n \geq 0, \quad t_n g(\mathbf{x}_n) - 1 = 0, \quad a_n(t_n g(\mathbf{x}_n) - 1) = 0.$$

  $\Rightarrow$ Dos tipos de muestra:

  (i) $a_n = 0 \rightarrow$ No influyen en la clasificación de una nueva muestra: $g(\mathbf{x}) = \sum_{a_n \neq 0} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$

  (ii) $t_n g(\mathbf{x}_n) = 1 \rightarrow$ muestras que caen sobre los hiperplanos de máximo margen. Se llaman **vectores de soporte (SV)**.

Clasificador de máximo margen (caso separable)

- **Obs. 1:** en general $\#SV \ll N \Rightarrow$ etapa de testing rápida.
- **Obs. 2:** $g(\mathbf{x}) = \sum_{\mathbf{x}_n SV} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b$ es un promedio ponderado (y con signo) de cuán similar es la muestra $\mathbf{x}$ a los SV. La decisión se toma según $\text{signo}(g(\mathbf{x}))$.
- ¿Cuánto vale el *bias* b? Ejercicio Usando la def. de SV, mostrar que

$$b = \frac{1}{\#SV} \sum_{\mathbf{x}_n SV} \left(t_n - \sum_{\mathbf{x}_m SV} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right).$$

Ejemplo XOR

Ejercicio Queremos diseñar un clasificador para el problema XOR, en donde los puntos $(-1,-1)$ y $(1,1)$ son de la clase ω_1, y $(-1,1)$ y $(1,-1)$ son de la clase ω_2.

- Considerando el mapeo $\phi : \mathbb{R}^2 \to \mathbb{R}^6$, $\mathbf{y} = \phi(x_1, x_2) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2]^T$, demostrar que el problema se vuelve separable en el espacio transformado.
- Diseñar un clasificador de máximo margen para el problem.
- Especificar los hiperplanos (separador, y de márgenes) y el margen.

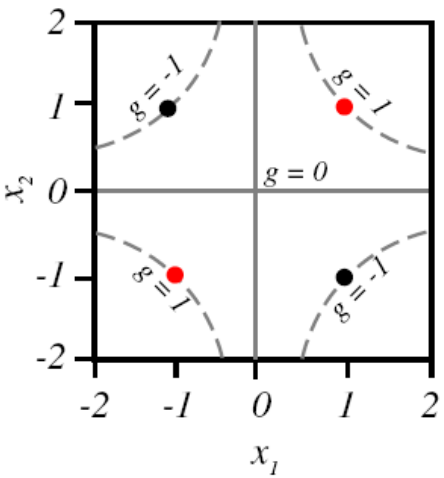
Ejemplo XOR

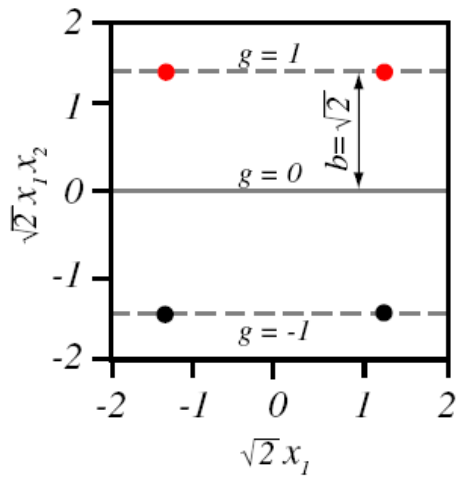
Maximizar: $\tilde{\mathcal{L}}(\mathbf{a}) = \sum_{n=1}^{4} a_k - \frac{1}{2}\sum_{n=1}^{4}\sum_{m=1}^{4} a_n a_m t_n t_m \overbrace{(\mathbf{x}_n^T \mathbf{x}_m + 1)^2}^{\phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m)}$,

Kernel pol. orden 2

sujeto a: $a_1 - a_2 + a_3 - a_4 = 0, \quad a_n \geq 0, \quad n \in \{1, 2, 3, 4\}.$

- La solución es $a_1 = a_2 = a_3 = a_4 = 1/8$ y $b = \sqrt{2}$.
- La función discriminante es $g(x) = x_1 x_2$.



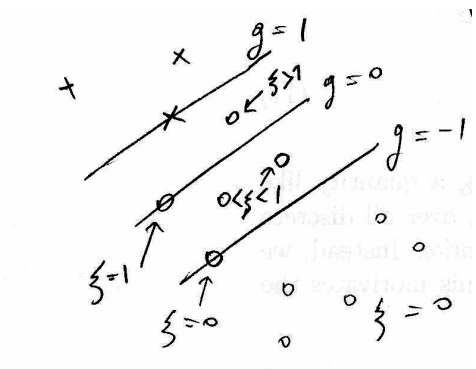


Caso no separable linealmente

Idea: modificar la formulación anterior para autorizar algunos puntos mal clasificados en entrenamiento.

Obs.: Definiendo una **penalización** para los errores, habrá un **compromiso** entre mayor margen (mejor generalización) y mayor cantidad de errores en entrenamiento.

Penalización: $\xi_n = \begin{cases} 0 & \text{si } \mathbf{x}_n \text{ bien clasificado} \\ |t_n - g(\mathbf{x}_n)| & \text{si no} \end{cases}$



- Puntos en la frontera: $g(\mathbf{x}_n) = 0 \Rightarrow \xi_n = 1$
- Puntos mal clasificados: $\xi_n > 1$

Caso no separable linealmente: C-SVM

Remplazamos la condición de clasificación perfecta (*hard margin*):

$$t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b)\geq 1, \quad n=1,\ldots,N$$

por la condición relajada (*soft margin*):

$$t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b)\geq 1-\xi_n, \quad \xi_n\geq 0, \quad n=1,\ldots,N$$

Objetivo: maximizar el margen penalizando errores:

$$\begin{cases} \min_{\mathbf{w},b}\left\{\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{n=1}^{N}\xi_n\right\}, & C>0 \\ \text{s.t.} \begin{cases} t_n(\mathbf{w}^T\phi(\mathbf{x}_n)+b)\geq 1-\xi_n, \\ \xi_n\geq 0, \qquad n=1,\ldots,N. \end{cases} \end{cases}$$

Pregunta ¿Qué representa C? $C\uparrow+\infty$?, $C\downarrow 0$?

Optimización de C-SVM

Ejercicio

- Escribir el Lagrangeano para C-SVM, en función de $\mathbf{w}, b$ y los multiplicadores $\mathbf{a}$ y $\boldsymbol{\mu}$.
- Escribir las condiciones de optimalidad y KKT.
- Demostrar que el Lagrangeano del problema dual es

$$\tilde{\mathcal{L}}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(\mathbf{x}_n, \mathbf{x}_m),$$

y que el problema dual es

$$\begin{cases} \max_{\mathbf{a}} \tilde{\mathcal{L}}(\mathbf{a}) \text{ sujeto a} \\ \sum_{n=1}^{N} a_n t_n = 0, \\ 0 \leq a_n \leq C, \quad \mu_n = C - a_n, \quad n = 1, \ldots, N. \end{cases}$$

C-SVM

- **Clasificación de una nueva muestra $\mathbf{x}$:** se evalúa el signo de

$$g(\mathbf{x}) = \sum_{n=1}^{N} a_n t_n k(\mathbf{x}_n, \mathbf{x}) + b.$$

- **Vectores de soporte:** verifican $a_n > 0, \quad t_n g(\mathbf{x}_n) = 1 - \xi_n$.

  Dos casos:

  (i) $a_n < C \Rightarrow \mu_n > 0 \overset{\mu_n \xi_n = 0 (\text{KKT})}{\Longrightarrow} \xi_n = 0$: **SV del lado correcto, fuera del margen**

  (ii) $a_n = C \overset{\mu_n = C - a_n}{\Longrightarrow} \mu_n = 0 \Rightarrow$ **si** $\xi_n \leq 1$ **bien clasificado; si** $\xi_n > 1$ **mal clasificado.**

C-SVM

- **Bias b:** SV: $0 < a_n < C \Rightarrow \xi_n = 0 \Rightarrow t_n g(\mathbf{x}_n) = 1$.
  Idem que para hard margin,

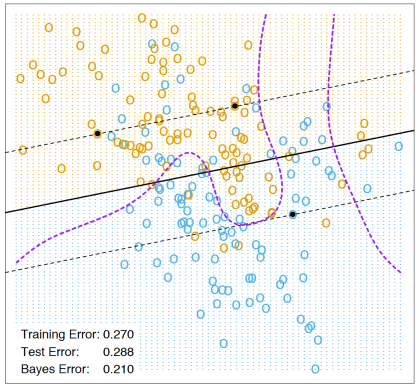
$$b = \frac{1}{\#SV} \sum_{\mathbf{x}_n SV} \left(t_n - \sum_{\mathbf{x}_m SV} a_m t_m k(\mathbf{x}_n, \mathbf{x}_m) \right),$$

  con $SV = \{\mathbf{x}_n : \ \ 0 < a_n < C, \ \ n = 1, \ldots, N\}$.

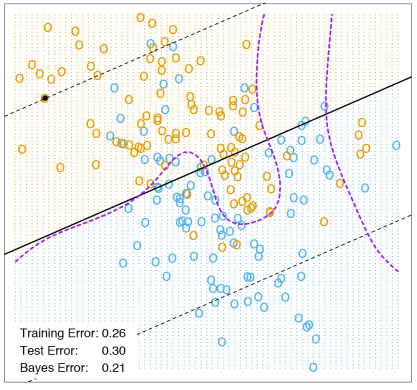
- **Elección de parámetros:** usualmente C, la elección del kernel y sus parámetros se estiman conjuntamente usando Grid Search y validación cruzada, sobre el conjunto de entrenamiento.

C-SVM

Ejemplo: Cada clase es una mezcla de 10 gaussianas de baja varianza, cuyas medias siguen una distribución gaussiana.



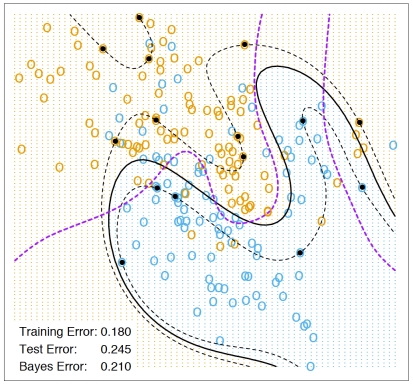
$C = 10000$



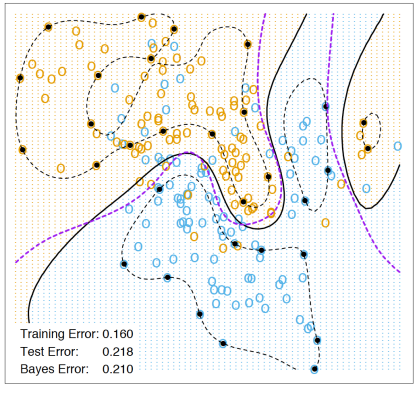
$C = 0.01$

Lineal

SVM - Degree-4 Polynomial in Feature Space



SVM - Radial Kernel in Feature Space



Polinomial (orden 4) y RBF ($\gamma = 1$). En ambos casos C ajustado por validación cruzada.

Extensión multi-clase

No hay una única forma de extender SVM.

1. **Uno contra el resto:** M clasificadores binarios, una clase y el resto para cada clase, $g^j(\mathbf{x}) = \mathbf{w}^{j\,T}\mathbf{x} + w_0^j, j = 1, ..., c$. Luego una muestra nueva $\mathbf{x}$ se asigna a la clase con mayor discriminante, $\arg\max_{j=1,...,c} g^j(\mathbf{x})$. Poco simétrico o balanceado. Regiones indefinidas.
2. **Clasificación por pares:** se consideran los $c(c-1)/2$ pares de clases, y para cada una se diseña un clasificador. Nueva muestra: se pasa por todos los clasificadores, y se asigna a la clase más votada. Hay que resolver más problemas, pero con menos muestras. Regiones indefinidas.
3. **Error Correcting Output Coding:** se generan $L = \log_2 c$ particiones binarias del conjunto de clases. A cada clase le corresponde un único código binario $\mathbf{d} = (d_1, d_2, \ldots, d_L) \in \{-1, 1\}^L$. Para cada partición se entrena un clasificador $g^1, \ldots, g^L$. Clasificación de nueva muestra $\mathbf{x}$: a la clase más cercana a $(\mathrm{sgn}(g^1(\mathbf{x})), \ldots, \mathrm{sgn}(g^L(\mathbf{x})))$ en distancia de Hamming.

SVM y la Maldición de la Dimensionalidad

Viendo que en SVM la frontera de decisión queda determinada por los SV (y que $\#SV \ll N$), tiene sentido preguntarse si este método tiene alguna ventaja al respecto.

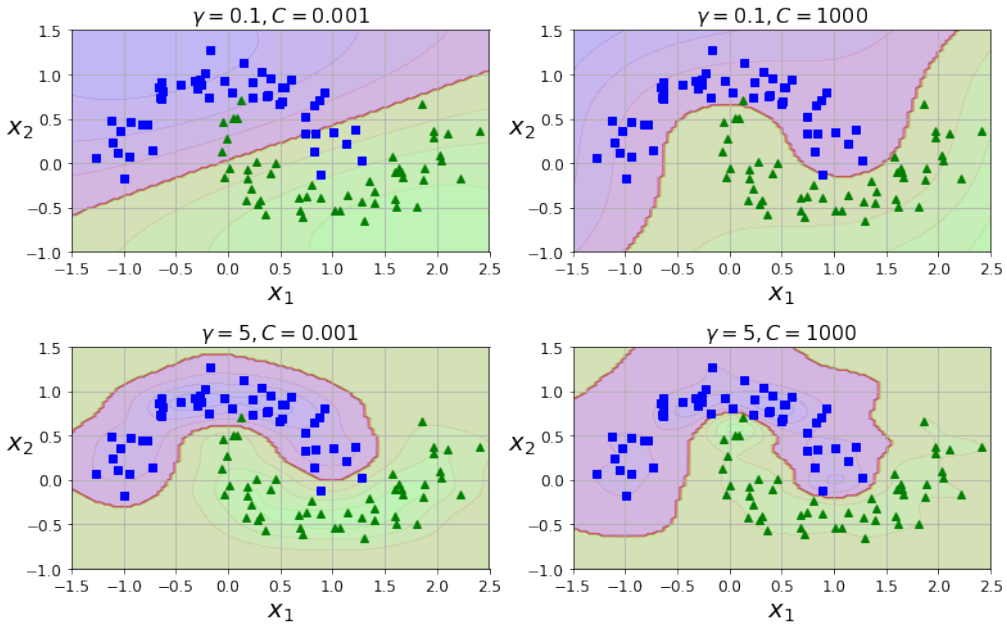
Observación 1: $k(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^T\mathbf{x}')^d$, con $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^p$

- Todo los términos tienen pesos fijos, por lo que el kernel no tiene total libertad para concentrarse en subespacios.
- Si la separación lineal se da, por ejemplo, en el subespacio lineal definido por las dos primeras coordenadas, el kernel tendrá dificultad en encontrar la estructura.
- Cuanto mayor p, más compleja la búsqueda de la estructura, y más muestras se precisan.
- Ver ejemplo numérico en Hastie et al., sección 12.3.4.

SVM y la Maldición de la Dimensionalidad

Observación 2: $k(\mathbf{x}, \mathbf{x}') = \mathrm{e}^{-\gamma\|\mathbf{x}-\mathbf{x}'\|^2}$

- Si $\gamma \to +\infty$, $\mathbf{K} = [k(\mathbf{x}_n, \mathbf{x}_m)]_{i,j=1}^{N} \to Id \Rightarrow$ Las muestras de entrenamiento sólo son consideradas similares a ellas mismas $\Rightarrow$ Overfitting, mala generalización, frontera irregular.
- Si $\gamma \to 0$, entonces $\forall \mathbf{x}, \mathbf{x}',\ k(\mathbf{x}, \mathbf{x}') \to 1 \Rightarrow$ Efecto regularizador, underfitting, frontera regular.



SVM y la Maldición de la Dimensionalidad

Observación 3: margen, generalización, rol de la dimensión

Dimensión de Vapnik-Chervonenkis (VC) (capacidad de un modelo)

- Un modelo de clasificación f con parámetro θ, se dice que separa el conjunto $\{x_1, x_2, \ldots, x_n\}$ si para todas las asignaciones de etiquetas binarias, existe un θ que clasifica sin error.
- La dimensión VC de f definido sobre el espacio X es el cardinal del mayor subconjunto de X que puede ser separado por f.
- Ejemplos: $VC(\text{rectas en } \mathbb{R}^2) = 3$;
  $VC(\text{hiperplanos en } \mathbb{R}^d) = d + 1$;
  $VC(\{\text{sgn}(\sin \omega x),\ \omega \geq 0\}) = +\infty$; $VC(\text{1-KNN}) = +\infty$

SVM y la Maldición de la Dimensionalidad

Observación 3 (cont.)

Riesgo empírico, riesgo verdadero

– $f : X \mapsto \{-1, 1\}$, $f(\mathbf{x}) = \text{sgn}(g(\mathbf{x}))$.

– $\{(\mathbf{x}_n, t_n), n = 1, \ldots, N, t_n \in \{-1, 1\}\}$ muestras etiquetadas.

– Costo 0-1: $\frac{1}{2}|f(\mathbf{x}_n) - t_n|$.

- Riesgo empírico o error promedio de training:
  $R_{emp}[f] = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2}|f(\mathbf{x}_n) - t_n|$.
- Riesgo (verdadero): $R[f] = \int \frac{1}{2}|f(\mathbf{x}) - t_n| dP(\mathbf{x}, t)$.

Cómo se vinculan $R_{emp}[f]$, $R[f]$ y la dimensión VC??

SVM y la Maldición de la Dimensionalidad

Observación 3 (cont.)

Cota de Vapnik-Chervonenkis

Con probabilidad $\geq 1 - \delta$,

$$R[f] \leq R_{emp}[f] + \underbrace{\sqrt{\frac{1}{N}\left(h\left(\ln\left(\frac{2N}{h}\right)+1\right)+\ln\left(\frac{4}{\delta}\right)\right)}}_{\text{término de capacidad}}.$$

- Cuando $N \to +\infty$, $R[f] \to R_{emp}[f]$.
- h baja $\Rightarrow$ reduce el término de capacidad.
- Pero: si h demasiado baja $\Rightarrow$ difícil reducir $R_{emp}[f]$.

SVM y la Maldición de la Dimensionalidad

Observación 3 (cont.)

- Recordemos que $VC(\text{hiperplanos en } \mathbb{R}^d) = d + 1$.
- **Un resultado fundamental:** se demuestra que, si R radio es de la esfera más pequeña que contiene los datos,
  Indep. de d, $VC(\text{hiperplanos de margen } \rho) \leq \frac{R^2}{\rho^2} + 1$.

$\Longrightarrow$ **Conclusión:** maximizar el margen es una forma de controlar la "maldición de la capacidad" al trabajar en espacios de muy alta dimensión.

SVM y la Maldición de la Dimensionalidad

Observación 4: SVM como método de regularización L^2

Consideremos el problema de regresión

$$(\mathcal{P}): \min_{\mathbf{w},b} \sum_{n=1}^{N} [1 - t_n g(\mathbf{x}_n)]_+ + \frac{\lambda}{2}\|\mathbf{w}\|^2, \quad \text{con } [x]_+ = \max\{0, x\}$$

($L(t_n, g(\mathbf{x}_n)) = [1 - t_n g(\mathbf{x}_n)]_+$ se conoce como *hinge loss function*).

Observación: la solución de $(\mathcal{P})$, con $\lambda = \frac{1}{C}$, es la solución del problema C-SVM.

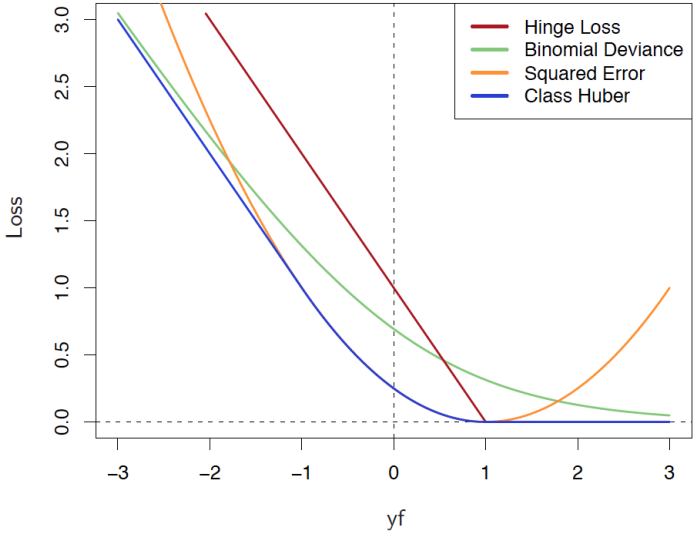
$\Longrightarrow$ **Conclusión:** C-SVM es un ajuste a datos con regularización:

- La regularización es equivalente a maximizar el margen
- Cuanto mayor λ (i.e. menor C), más regular será la solución.

Vínculo de SVM con otros métodos de clasificación

Consideremos ahora la siguiente familia de métodos de clasificación lineal ($f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$) con regularización por $\|\mathbf{w}\|_2^2$:

$$\min_{\mathbf{w},b} \sum_{n=1}^{N} L(t_n, f(\mathbf{x}_n)) + \frac{\lambda}{2}\|\mathbf{w}\|^2.$$



Loss Function	$L[y, f(x)]$	Minimizing Function
Binomial Deviance Logistic Regression	$\log[1 + e^{-yf(x)}]$	$f(x) = \log \frac{\Pr(Y = +1\|x)}{\Pr(Y = -1\|x)}$
SVM Hinge Loss SVM	$[1 - yf(x)]_+$	$f(x) = \text{sign}[\Pr(Y = +1\|x) - \frac{1}{2}]$
Squared Error LDA	$[y - f(x)]^2 = [1 - yf(x)]^2$	$f(x) = 2\Pr(Y = +1\|x) - 1$
"Huberised" Square Hinge Loss	$-4yf(x)$, $yf(x) < -1$ $[1 - yf(x)]_+^2$ otherwise	$f(x) = 2\Pr(Y = +1\|x) - 1$