EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Máquinas de soporte vectorial.
Objetivo: Que veremos.
Lectura recomendada: Referencia.

1. INTRODUCCIÓN

Las máquinas de soporte vectorial (`SVM`) son modelos no lineales que buscan resolver problemas predictivos. Por simplicidad, estudiaremos `SVM` en el contexto de clasificación.

La forma en que operan las `SVM` es a través de buscar clasificar por un `separación con un hiperplano` en el espacio de atributos.

Si los datos no nos permiten realizar dicha separación entonces nos volcamos a relajar los supuestos del modelo:

1. el concepto de `separación`.
2. el espacio de atributos.

2. SEPARACIÓN CON UN HIPERPLANO

- Un hiperplano de dimensión p es un subespacio de dimensiones $p-1$.
- La ecuación que define a un hiperplano tiene la forma

$$0 = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{1}$$

- En dos dimensiones es una línea.
- Si $\beta_0 = 0$ entonces el hiperplano atraviesa el origen.
- El vector de coeficientes $(\beta_1, \ldots, \beta_p)$ se llama el vector `normal` del hiperplano.

2.1. Para problemas de clasificación

- Nota que si definimos $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, entonces $f(X) > 0$ para X de un lado del hiperplano. Si $f(X) < 0$ entonces X se encuentra del otro lado.
- Para los datos en Fig. 1, si los puntos azules son aquellos donde $Y_i = 1$ y los morados son aquellos que tienen $Y_i = -1$, entonces

$$Y_i \cdot f(X_i) > 0\,, \qquad \forall i\,. \tag{2}$$

- El caso $f(X) = 0$ define al `hiperplano separador`.

2.2. Máxima separación

En Fig. 1 vemos que existe una infinidad de planos que pueden separar nuestro conjunto de datos. Nuestra noción del `mejor` separador es aquel que `maximice el margen` de separación.

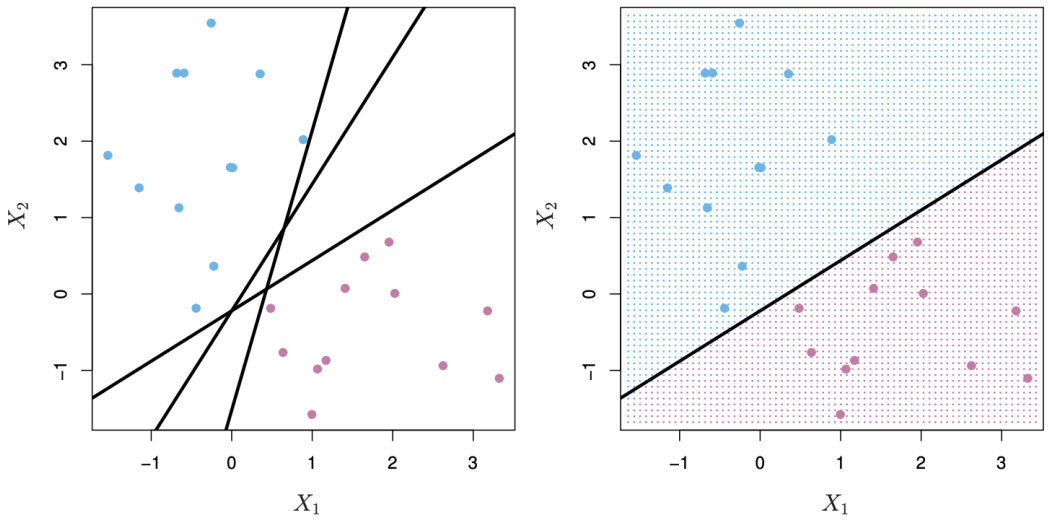
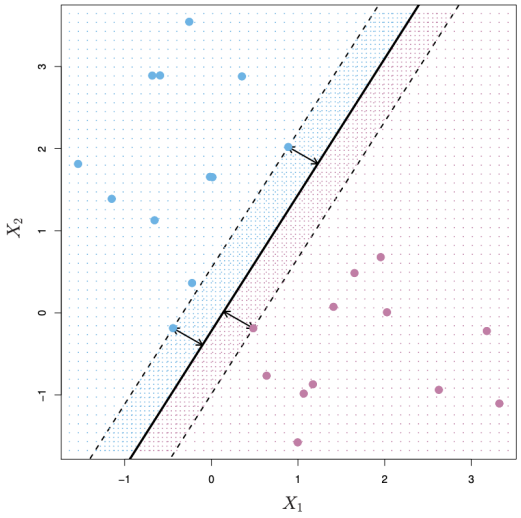


FIGURA 1. *Imagen tomada de [1].*



Esto lo escribimos como la solución al problema de optimización con restricciones

$$\max_{\beta_0,\beta_1,\ldots,\beta_p} M$$

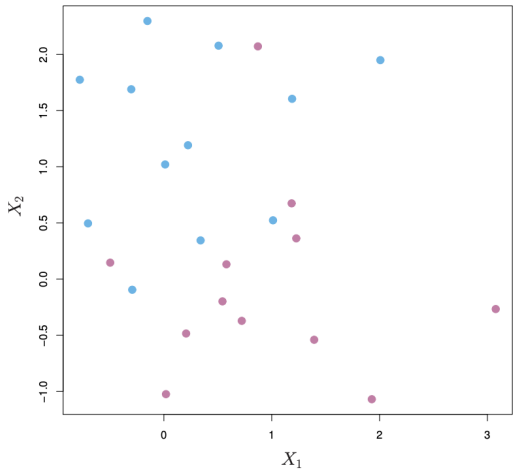
$$\text{sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1\,,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M\,, \quad \forall i\,.$$

El problema de optimización descrito se denomina `formulación primal` y se puede rescribir en términos de un problema de optimización cuadrático por medio de la `formulación dual`. Esto permite la resolución del problema de optimización por medio de herramientas de optimización convexa para problemas cuadráticos.

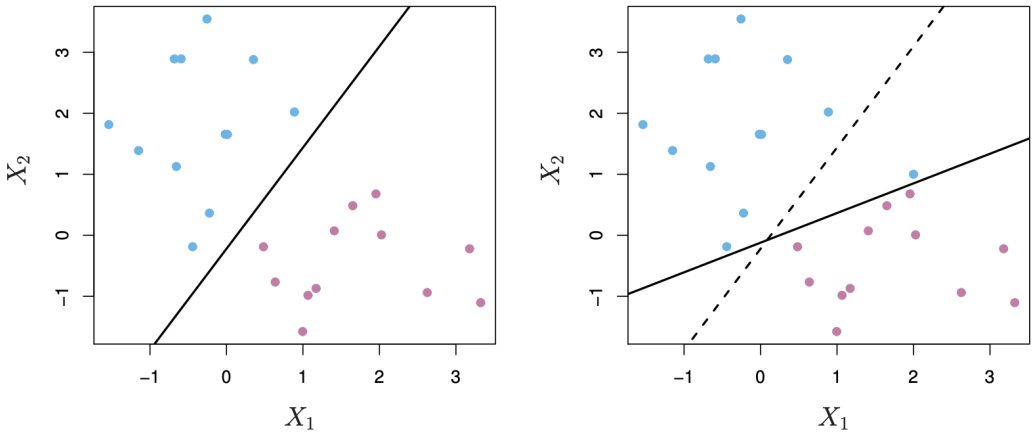
2.3. Datos no separables

- Usualmente sucede esto en la práctica (a menos que $n < p$).



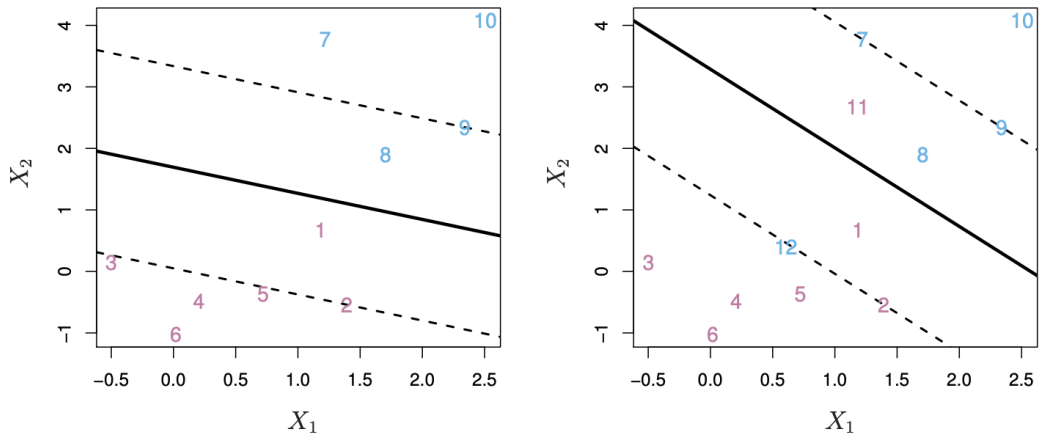
2.4. Ruido en las mediciones

Los datos a veces son separables, pero el ruido en las observaciones puede hacer que un clasificador por margen máximo tenga generalización deficiente.



3. CLASIFICADOR BASADO EN VECTORES DE SOPORTE

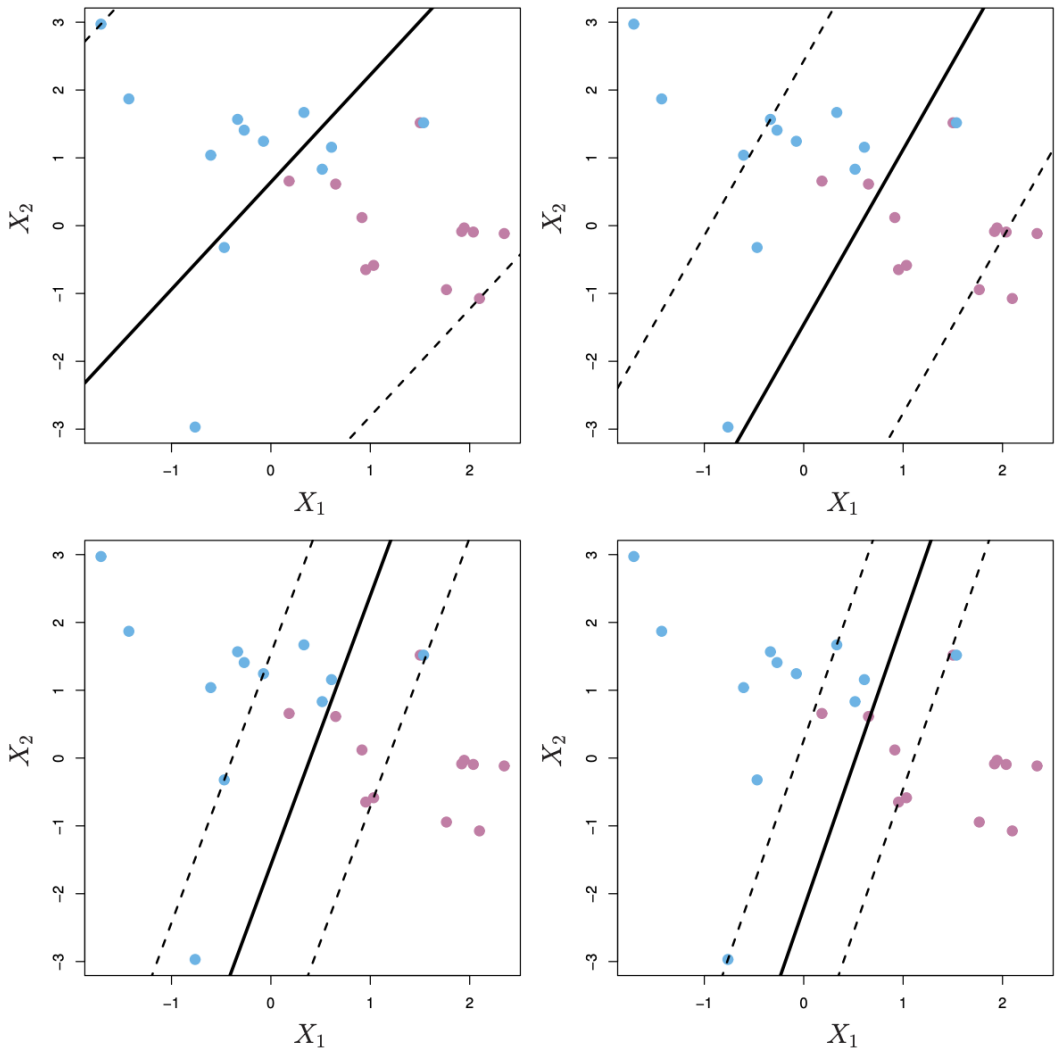
Se pueden relajar las restricciones del clasificador para incorporar un `margen suave`.



3.1. Formulación

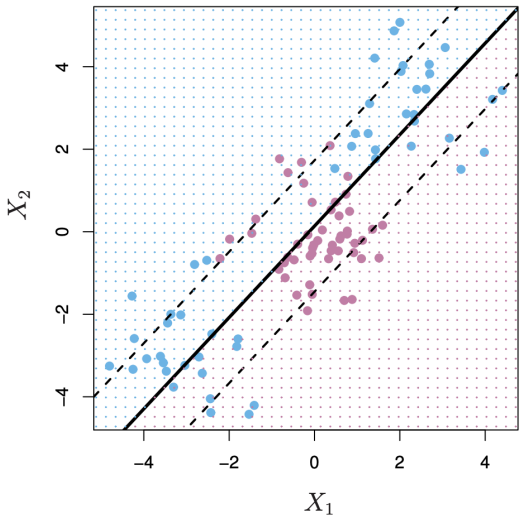
$$\max_{\beta_0,\beta_1,\ldots,\beta_p} M$$
$$\text{sujeto a } \sum_{j=1}^{p} \beta_j^2 = 1\,,$$
$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M(1-\epsilon_i)\,, \quad \forall i\,,$$
$$\epsilon_i \geq 0\,, \qquad \sum_{i=1}^{n} \epsilon_i \leq C\,.$$

El término C es dicta cuántos datos mal clasificados estamos dispuestos a cometer.



4. SEPARACIÓN LINEAL

En algunas situaciones la separación lineal no será suficiente.



4.1. Ingeniería de características

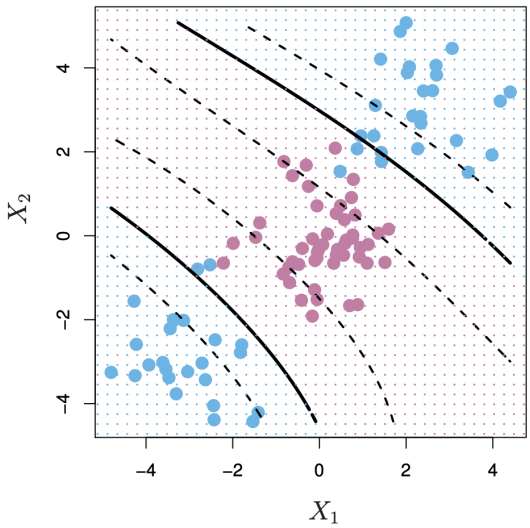
- Buscamos una expansión de los atributos por medio de transformaciones

$$X_1^2, X_2^2, X_1X_2, \ldots \tag{3}$$

  lo cual permite expandir el espacio $\mathbb{R}^p \to \mathbb{R}^M$ con $M > p$.
- Ajustamos un clasificador por vectores de soporte en el espacio expandido.
- Esto nos ayuda a incorporar decisión de separación no lineales en el espacio original de atributos.

4.2. Solución con polinomios cúbicos

- Si utilizamos una expansión con polinomios cúbicos, pasamos de 2 dimensiones a 9.
- La separación lineal la logramos en este nuevo espacio.



5. SEPARACIONES NO LINEALES Y KERNELS

- Los polinomios (especialmente en varias dimensiones) pueden tener un comportamiento oscilatorio muy fuerte y sobre-ajustar sin cuidado.
- Hay un mecanismo matemáticamente mas elegante para introducir no-linealidades.
- Especialmente útil en problemas donde podamos usar `productos interiores`.

5.1. Productos interiores y vectores de soporte

- Recordemos que

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j} \,. \tag{4}$$

- El clasificador se puede expresar como

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle \tag{5}$$

- Necesitamos los $\binom{n}{2}$ productos interiores para poder estimar los parámetros.
- Pero, la mayoría de las α_i son cero,

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle \,. \tag{6}$$

5.2. Kernels y máquinas de soporte vectorial

- Podríamos calcular productos interiores y construir un clasificador por medio de vectores de soporte.
- Hay algunos *kernels* que calculan lo que necesitamos. Por ejemplo,

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \,, \tag{7}$$

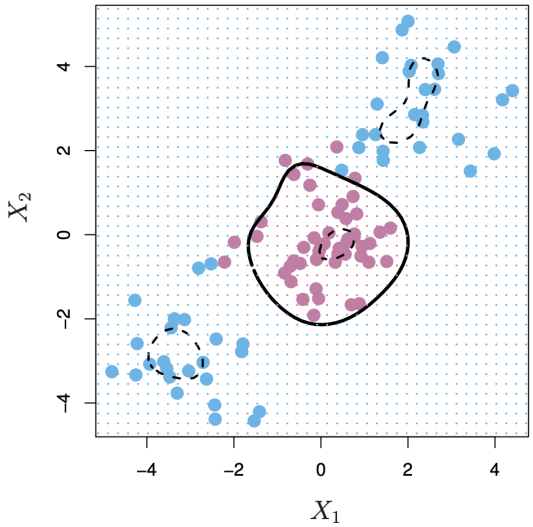
  calcula los productos internos de la expansión en polinomios.
- La solución del problema, entonces, tiene la forma

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i) \,. \tag{8}$$

5.2.1. Kernel *radial:* Consideremos el *kernel*

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_i - x_{i'})^2\right) \,, \tag{9}$$

lo cual nos permite ajustar superficies de decisión como la que muestra



6. CLASIFICACIÓN MULTICLASE

Una pregunta natural es cómo extender la formulación de una `SVM` para el problema de clasificación multiclase.

Tenemos dos estrategias:

1. `Una clase contra todas las anteriores`. Clasifica la clase con mayor $\hat{f}_k(x)$.
2. `Todas las posibles tareas de dos clases`. Clasifica la clase que gana la mayor de las veces.

7. COMPARACIÓN CON OTROS MODELOS

El problema de optimización se puede rescribir como

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \left\{ \sum_{i=1}^{n} \max[0, 1 - y_i f(x_i)] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} . \tag{10}$$

La función de pérdida se conoce como la función `hinge`. Es muy similar a la pérdida por entropía cruzada (regresión logística).

7.1. Regresión logística ó SVM

- Cuando las clases son casi separables, preferimos `SVM`.
- Cuando no, regresión logística + Ridge es muy parecido a `SVM`.
- Si nos interesa calcular probabilidades, usamos regresión logística.
- Se pueden utilizar *kernels* con otros modelos (regresión logística o LDA) pero es mas costoso.

8. CONCLUSIONES

- Las `SVM` son modelos que pueden acomodar no linealidades para hacer modelos predictivos.
- En la práctica aún siguen siendo utilizadas pues es de las pocas alternativas para ajustar no linealidades.
- Siguen siendo muy útiles en aplicaciones donde se tienen identificadas las características mas importantes para una tarea predictiva.

REFERENCIAS

[1] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. 2