Máquinas de soporte vectorial, PII.

César Gómez

28 de octubre de 2020

Que sucede cuando la frontera de decisión es no lineal?

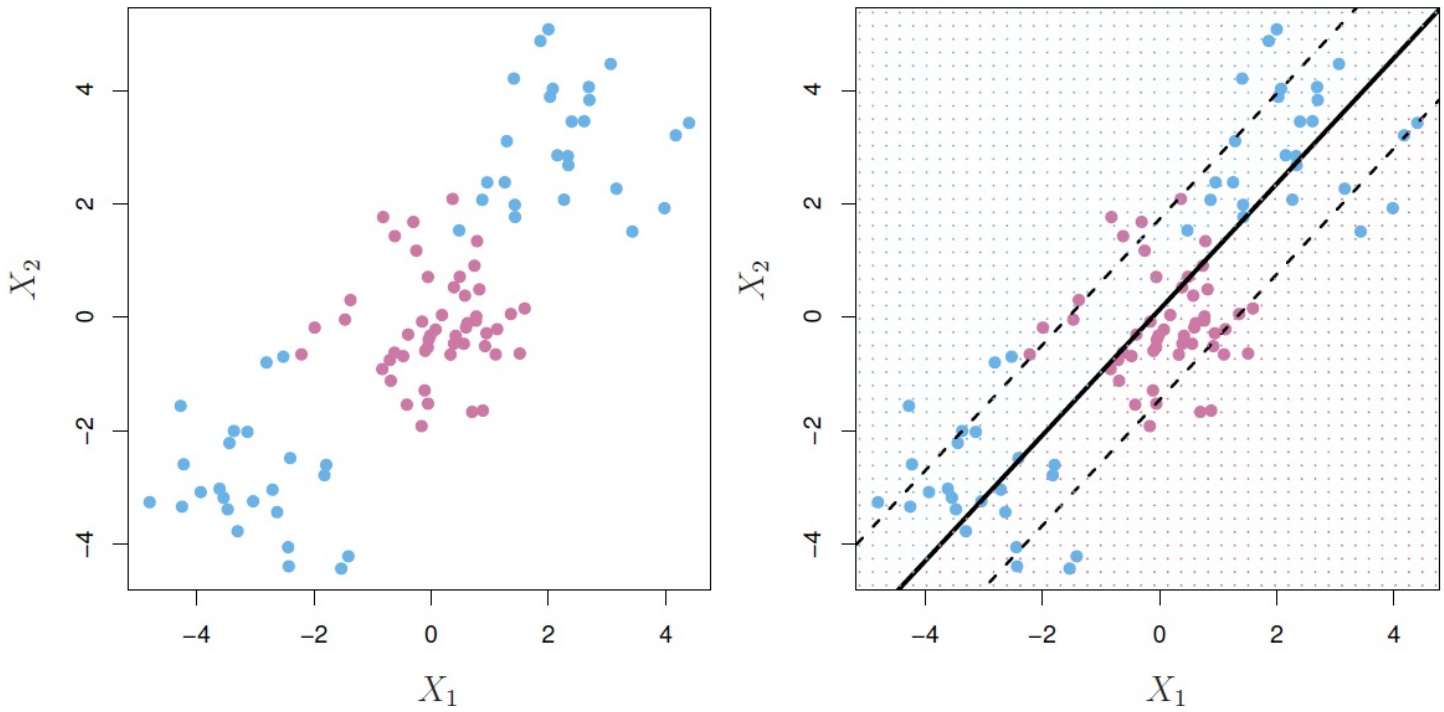

Figura 1

Clasificación con fronteras de decisión no lineales

- Una alternativa puede ser aumentando el espacio de predictores considerando funciones polinomiales de los predictores.
  Por ejemplo, en vez de ajustar un clasificador de soporte vectorial con los p predictores

$$X_1, X_2, \ldots, X_p,$$

se podía tal vez ajustar un clasificador de SV utilizando los $2p$ predictores

$$X_1, X_1^2, X_2, X_2^2, \ldots, X_p, X_p^2,$$

y consideramos el siguiente problema de optimización

$$\underset{\beta_0, \beta_1, \ldots, \beta_p, \epsilon_1, \ldots, \epsilon_n, M}{\text{Maximice}} \; M \tag{1}$$

$$\text{Sujeto a} \quad \sum_{j=1}^{p} \beta_j^2 = 1 \tag{2}$$

$$y_i \left(\beta_0 + \sum_{i=1}^{p} \beta_{1j} x_{ij} + \sum_{i=1}^{p} \beta_{2j} x_{ij}^2 \right) \geq M(1 - \epsilon_i) \quad \forall i = 1, 2, \ldots, n. \tag{3}$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C, \quad \sum_{j=1}^{p} \sum_{k=1}^{2} \beta_{kj}^2 = 1. \tag{4}$$

- Se puede pensar en agrandar el espacio de atributos con términos polinomiales de alto orden, o incluso incluyendo *términos de interacción* de la forma $X_j X_{j'}$ con $j \neq j'$.

- Se puede pensar en agrandar el espacio de atributos con términos polinomiales de alto orden, o incluso incluyendo *términos de interacción* de la forma $X_j X_{j'}$ con $j \neq j'$.
- No es difícil de ver que el espacio de predictores se puede agrandar de muchas formas y se debe ser cuidadoso por qué se puede terminar con un número muy grande de atributos y el cómputo deviene inmanejable.

- Se puede pensar en agrandar el espacio de atributos con términos polinomiales de alto orden, o incluso incluyendo *términos de interacción* de la forma $X_j X_{j'}$ con $j \neq j'$.
- No es difícil de ver que el espacio de predictores se puede agrandar de muchas formas y se debe ser cuidadoso por qué se puede terminar con un número muy grande de atributos y el cómputo deviene inmanejable.
- las **máquinas de soporte vectorial** que presentamos a continuación permiten agrandar el espacio de atributos para un clasificador de soporte vectorial de forma que los cómputos se pueden desarrollar de forma eficiente.

Máquina de Soporte Vectorial (SVM)

- Acá se representará el producto interno de 2 vectores r-dimensionales a y b por

$$\langle a, b \rangle = \sum_{i=1}^{r} a_i b_i. \tag{5}$$

Máquina de Soporte Vectorial (SVM)

- Acá se representará el producto interno de 2 vectores r-dimensionales a y b por

$$\langle a, b \rangle = \sum_{i=1}^{r} a_i b_i. \tag{5}$$

- El producto interno de 2 observaciones x_i y $x_{i'}$ viene dado por

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

- Se puede mostrar que el clasificador de soporte vectorial lineal, se puede representar por

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle. \tag{6}$$

donde hay n parámetros $\alpha_i \quad i = 1, 2, \ldots, n$ uno por cada observación.

- Se puede mostrar que el clasificador de soporte vectorial lineal, se puede representar por

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \langle x, x_i \rangle. \tag{6}$$

donde hay n parámetros $\alpha_i \quad i = 1, 2, \ldots, n$ uno por cada observación.

- También puede mostrase que los únicos parámetros α_i que son distintos de cero en (6) corresponden a los vectores de soporte. Si $\mathcal{S}$ denota el conjunto de índices correspondientes a los *vectores de soporte* entonces

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i \langle x, x_i \rangle \tag{7}$$

- Para resumir, al representar el clasificador lineal $f(x)$ todo lo que necesitamos son productos internos.

- Para resumir, al representar el clasificador lineal $f(x)$ todo lo que necesitamos son productos internos.
- Ahora se considerará una *generalización* de un producto interno de la forma

$$K(x_i, x_{i'}). \tag{8}$$

A una tal función se le denominará un **kernel**, y se trata de una función que mide la "*similitud*" entre 2 observaciones.

- Para resumir, al representar el clasificador lineal $f(x)$ todo lo que necesitamos son productos internos.
- Ahora se considerará una *generalización* de un producto interno de la forma

$$K(x_i, x_{i'}). \tag{8}$$

A una tal función se le denominará un **kernel**, y se trata de una función que mide la *"similitud"* entre 2 observaciones.

- Se tienen varios ejemplos de un *kernel*, por ejemplo, ya hemos trabajado con producto internos

$$K(x_i, x_{i'}) = \langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}, \tag{9}$$

Máquina de soporte vectorial

- También puede considerarse el **kernel polinomial** de grado d

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_i x_{i'})^d, \tag{10}$$

donde d es un entero positivo.

Máquina de soporte vectorial

- También puede considerarse el **kernel polinomial** de grado d

$$K(x_i, x_{i'}) = (1 + \sum_{j=1}^{p} x_i x_{i'})^d, \tag{10}$$

  donde d es un entero positivo.
- Cuando se utiliza un kernel polinomial con $d > 1$, entonces el algoritmo para calcular el clasificador de soporte vectorial consigue fronteras de decisión mucho más flexibles.

- Considerar un clasificador de soporte vectorial con el *kernel* en (10), esencialmente equivale a ajustar un clasificador de soporte vectorial en un espacio aumentado que contiene transformaciones polinomiales hasta e grado d de los atributos originales $X_1, \ldots, X_p$.

- Considerar un clasificador de soporte vectorial con el *kernel* en (10), esencialmente equivale a ajustar un clasificador de soporte vectorial en un espacio aumentado que contiene transformaciones polinomiales hasta e grado d de los atributos originales $X_1, \ldots, X_p$.
- Cuando el clasificador de soporte vectorial se combina con un *kernel* **no lineal** como en (10), el clasificador resultante se denomina **máquina de soporte vectorial**.

$$f(x) = \beta_0 + \sum_{i \in \mathcal{S}} \alpha_i K(x, x_i), \tag{11}$$

- Otro *kernel* no lineal alternativo, consiste en el **kernel radial**

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2). \qquad (12)$$

- Otro *kernel* no lineal alternativo, consiste en el **kernel radial**

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2). \qquad (12)$$

- En relación a este kernel, si una observación de prueba $x^* = [x_1^*, \ldots, x_p^*]$ está lejos de una observación de entrenamiento en términos de distancia euclidea, entonces $\sum_{j=1}^{p} (x_j^* - x_{ij})^2$ es **grande** por lo que $K(x^*, x_i) = \exp(-\gamma \sum_{j=1}^{p} (x_i^* - x_{ij})^2).$ es **pequeño y contribuye poco** a definir el signo de $f(x^*)$ en (11).

- Otro *kernel* no lineal alternativo, consiste en el **kernel radial**

$$K(x_i, x_{i'}) = \exp(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2). \qquad (12)$$

- En relación a este kernel, si una observación de prueba $x^* = [x_1^*, \ldots, x_p^*]$ está lejos de una observación de entrenamiento en términos de distancia euclidea, entonces $\sum_{j=1}^{p} (x_j^* - x_{ij})^2$ es **grande** por lo que $K(x^*, x_i) = \exp(-\gamma \sum_{j=1}^{p} (x_i^* - x_{ij})^2).$ es **pequeño y contribuye poco** a definir el signo de $f(x^*)$ en (11).
- Lo anterior indica que el kernel radial posee un comportamiento *local*, en el sentido de que solo observaciones de entrenamiento próximas poseen un efecto para asignar la clase a una observación de prueba.

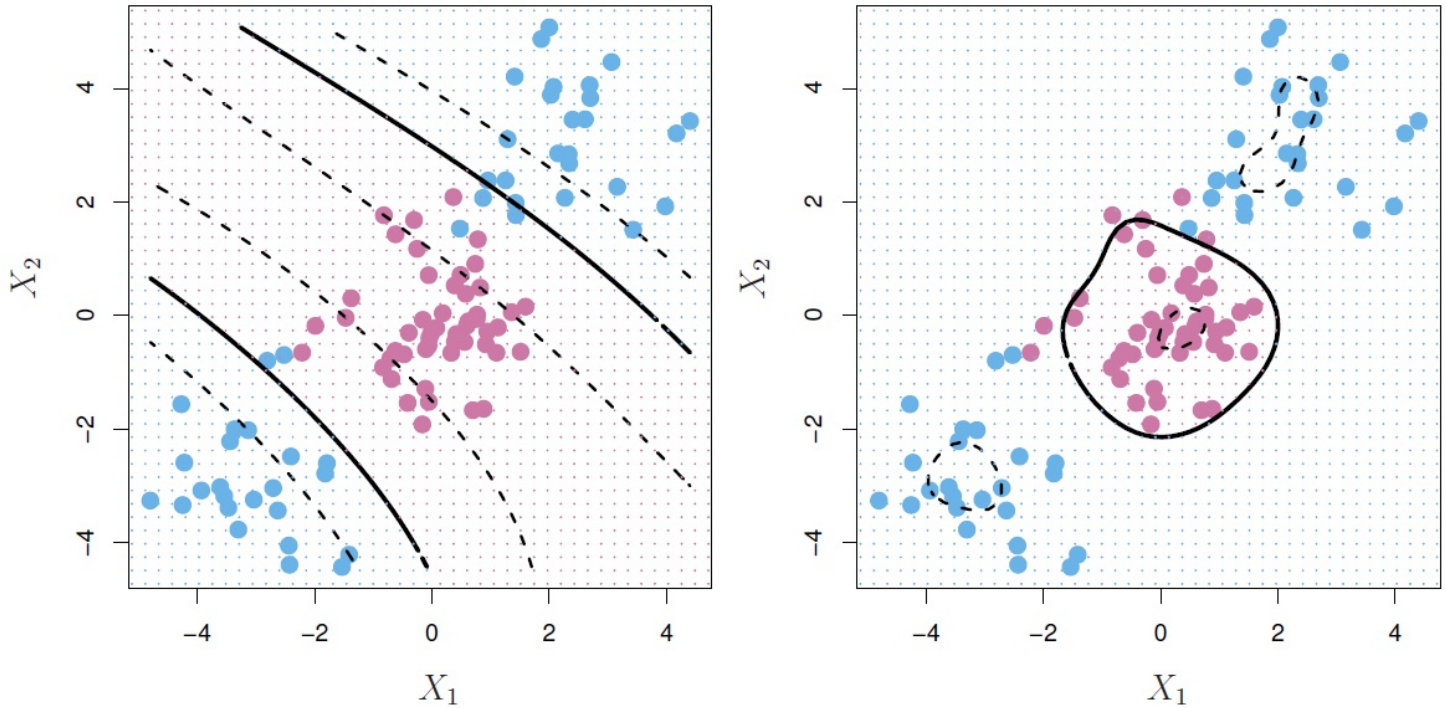

Figura 2: **Izquierda**: una MSV con kernel polinomial de grado $d = 3$ aplicada a los datos de la figura [1].
Derecha: Se ajusta una MSV con kernel radial a los mismos datos. Cada kernel es capaz de capturar la frontera de decisión.

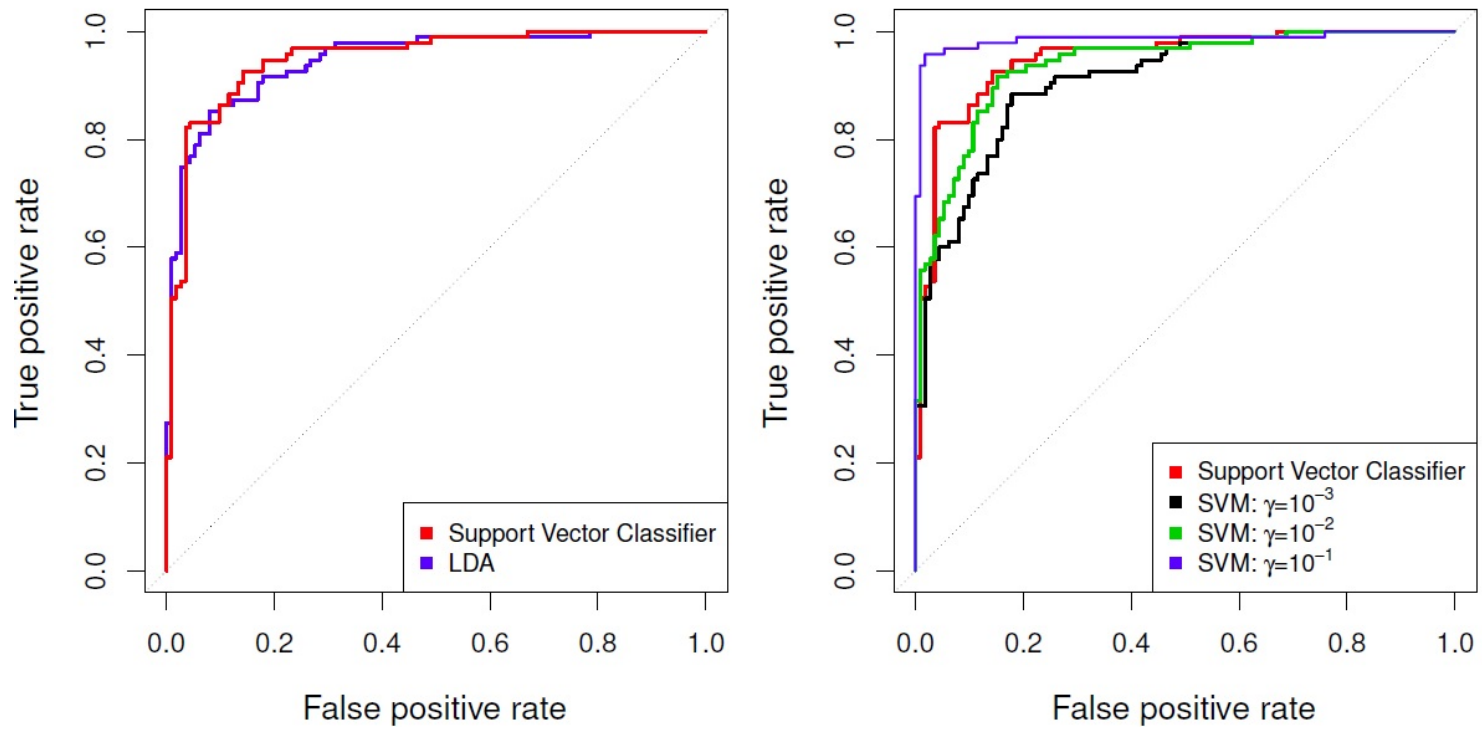

Figura 3: Curvas ROC para un conjunto de datos de entrenamiento (datos Heart). **Izquierda**: se comparan, el clasificador de soporte vectorial y LDA,**Derecha**:El clasificador de soporte vectorial comparado con una MSV que utiliza un kernel radial con $\gamma = 10^{-3},\ 10^{-2}$ y $\gamma = 10^{-1}$

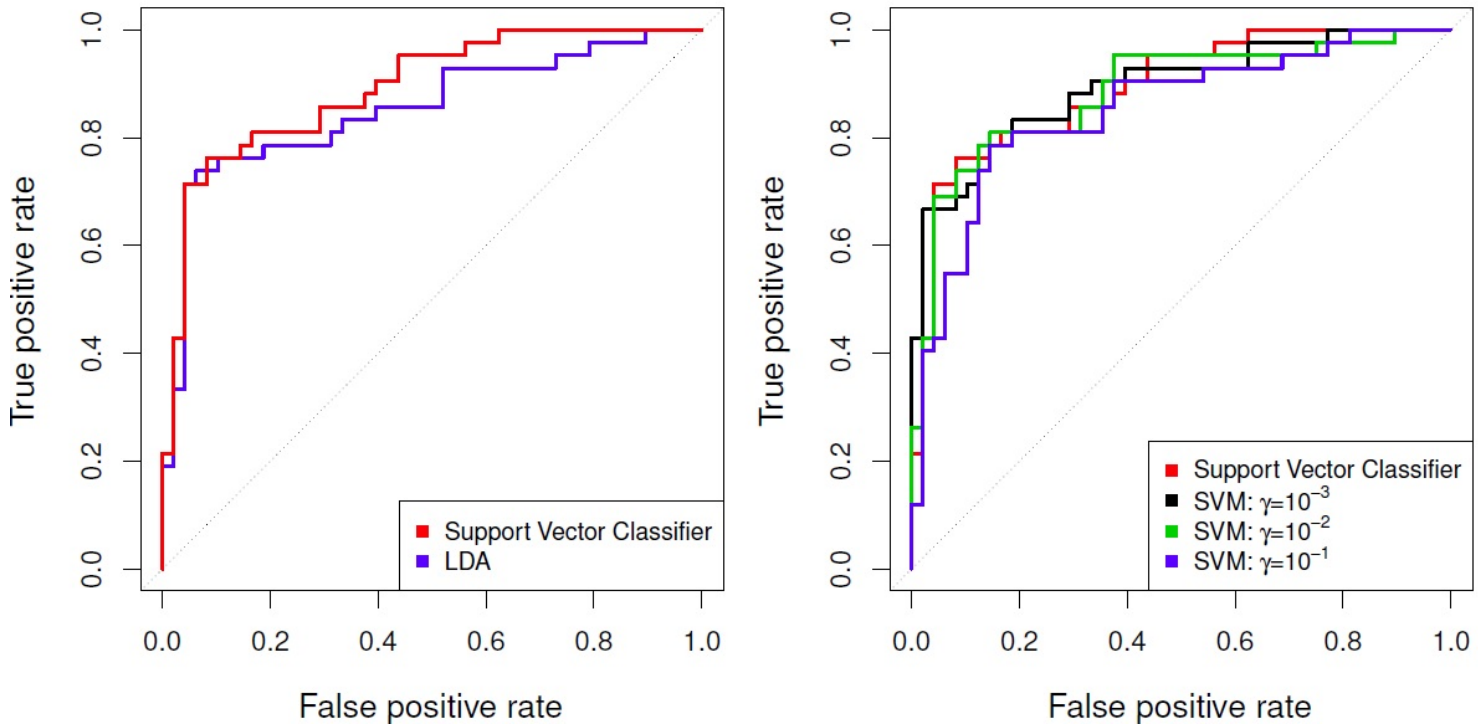

Figura 4: Curvas ROC para un conjunto de datos de prueba (datos Heart). **Izquierda**: se comparan, el clasificador de soporte vectorial y LDA,**Derecha**:El clasificador de soporte vectorial comparado con una MSV que utiliza un kernel radial con $\gamma = 10^{-3},\ 10^{-2}$ y $\gamma = 10^{-1}$

MSV con más de 2 clases

Clasificación uno vs uno (one vs one)

- Suponga que se quiere llevar a cabo una clasificación utilizando MSV, y hay $K > 2$ clases. En el procedimiento one *vs* one ó *"de todos los pares"* se ajustan $\binom{K}{2}$ MSV, cada una de las cuales compara 2 clases distintas.

MSV con más de 2 clases

Clasificación uno vs uno (one vs one)

- Suponga que se quiere llevar a cabo una clasificación utilizando MSV, y hay $K > 2$ clases. En el procedimiento one *vs* one ó *"de todos los pares"* se ajustan $\binom{K}{2}$ MSV, cada una de las cuales compara 2 clases distintas.
- Por ejemplo una de estas MSV compara la clase k codificada como $(+1)$ con la clase k' codificada como (-1).

MSV con más de 2 clases

Clasificación uno vs uno (one vs one)

- Suponga que se quiere llevar a cabo una clasificación utilizando MSV, y hay $K > 2$ clases. En el procedimiento one *vs* one ó *"de todos los pares"* se ajustan $\binom{K}{2}$ MSV, cada una de las cuales compara 2 clases distintas.
- Por ejemplo una de estas MSV compara la clase k codificada como $(+1)$ con la clase k' codificada como (-1).
- Una observación de prueba x^* es clasificada utilizando cada una de las $\binom{K}{2}$ MSV y se registra cuantas veces la observación es etiquetada en cada clase.

MSV con más de 2 clases

Clasificación uno vs uno (one vs one)

- Suponga que se quiere llevar a cabo una clasificación utilizando MSV, y hay $K > 2$ clases. En el procedimiento one *vs* one ó *"de todos los pares"* se ajustan $\binom{K}{2}$ MSV, cada una de las cuales compara 2 clases distintas.
- Por ejemplo una de estas MSV compara la clase k codificada como $(+1)$ con la clase k' codificada como (-1).
- Una observación de prueba x^* es clasificada utilizando cada una de las $\binom{K}{2}$ MSV y se registra cuantas veces la observación es etiquetada en cada clase.
- Finalmente a la observación de prueba se le asigna la clase que cuenta con más asignaciones de esta observación.

Clasificación "uno *vs* todos" (one vs all)

- Un procedimiento alternativo para clasificar entre $K > 2$ clases consiste en ajustar K MSV, por ejemplo cada una representada por

$$f_k(x_i) = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \cdots + \beta_{pk}x_{ip}, \quad k = 1, 2, \ldots, K. \tag{13}$$

donde $f_k(\cdot)$ clasifica la observación de prueba x^* en la categoría $k \in \{1, 2, \ldots, K\}$ codificada como $(+1)$, si $f(x^*) > 0$.

Clasificación "uno *vs* todos" (one vs all)

- Un procedimiento alternativo para clasificar entre $K > 2$ clases consiste en ajustar K MSV, por ejemplo cada una representada por

$$f_k(x_i) = \beta_{0k} + \beta_{1k}x_{i1} + \beta_{2k}x_{i2} + \cdots + \beta_{pk}x_{ip}, \quad k = 1, 2, \ldots, K. \tag{13}$$

  donde $f_k(\cdot)$ clasifica la observación de prueba x^* en la categoría $k \in \{1, 2, \ldots, K\}$ codificada como $(+1)$, si $f(x^*) > 0$.

- Finalmente se asigna x^* en la categoría k tal que

$$f_k(x^*) = \max_{j=1,\ldots,K} f_j(x^*), \tag{14}$$

- Por ejemplo supongamos que $K = 3$ y $p = 2$. Entonces ajustamos 3 MSV tal que cada una codifica la clase 1,2 y 3 como $(+1)$ respectivamente,

$$\begin{aligned} f_1(x_i) &= \beta_{01} + \beta_{11}x_{i1} + \beta_{21}x_{i2}, \\ f_2(x_i) &= \beta_{02} + \beta_{12}x_{i1} + \beta_{22}x_{i2}, \\ f_3(x_i) &= \beta_{03} + \beta_{13}x_{i1} + \beta_{23}x_{i2}, \end{aligned} \tag{15}$$

- Entonces a una observación de prueba x^* se le asigna la clase correspondiente al **máximo** valor de $f_1(x^*)$,$f_2(x^*)$, $f_3(x^*)$.