

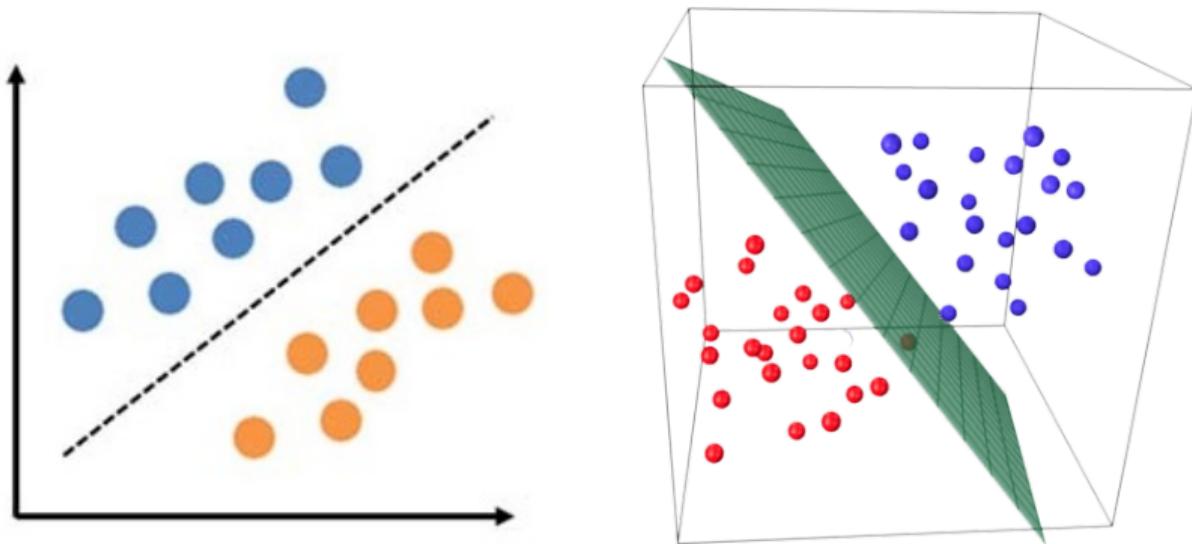
REGRESIÓN LOGÍSTICA

ALAN REYES-FIGUEROA
APRENDIZAJE ESTADÍSTICO

(AULA 23) 08.MAYO.2024

Clasificadores Lineales

Vimos algunos clasificadores: K -nn, bayesiano, análisis discriminante. Queremos estudiar otra familia de clasificadores sencillos: aquellos que dependen de una ecuación lineal (2 clases).



Clasificadores lineales

En este caso, buscamos una frontera de clasificación en la forma de un hiperplano en \mathbb{R}^d , dada por

$$w_0 + \mathbf{w}^T \mathbf{x} = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d = 0, \quad (1)$$

donde $\mathbf{w} = (w_1, \dots, w_d) \in \mathbb{R}^d$ y $w_0 \in \mathbb{R}$.

Por simplicidad, haremos una identificación del conjunto de datos \mathbb{X} en \mathbb{R}^{d+1} mediante el mapa biyectivo $i: \mathbb{R}^d \rightarrow \mathbb{R}^{d+1}$ dado por

$$i(\mathbf{x}) = (1, \mathbf{x}).$$

Similarmente, denotaremos al vector $(w_0, w_1, \dots, w_d) \in \mathbb{R}^{d+1}$ simplemente por \mathbf{w} . Así, la ecuación lineal (1) se escribe como $\mathbf{w}^T \mathbf{x} = 0$:

$$\mathbf{w}^T \mathbf{x} = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_d x_d = 0. \quad (2)$$

En general, separar un conjunto de datos (consistente de dos clases) mediante un hiperplano no siempre es posible. Distinguimos dos casos de conjuntos:

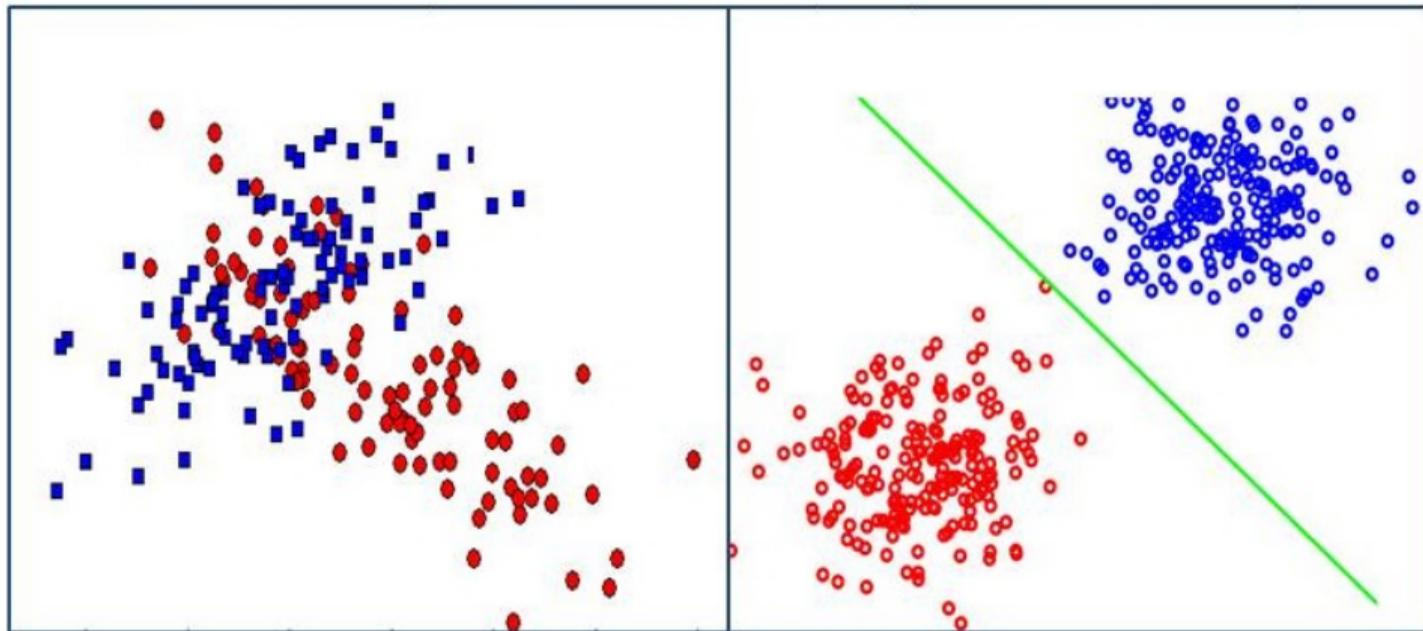
Definición

Un conjunto de datos $\mathbb{X} \in \mathbb{R}^{n \times (d+1)}$ que consiste de dos clases $y_i \in \{0, 1\}$ se llama **linealmente separable**, si existe un vector $\mathbf{w} = (w_0, w_1, \dots, w_d) \in \mathbb{R}^{d+1}$ tal que la ecuación lineal $\mathbf{w}^T \mathbf{x} = 0$ es una frontera de clasificación del conjunto \mathbb{X} . Esto es

$$y_i = \mathbf{1}(\mathbf{w}^T \mathbf{x}_i > 0), \quad \text{para todo } i = 1, 2, \dots, n.$$

Caso contrario, diremos que \mathbb{X} no es linealmente separable.

Clasificadores Lineales



Separabilidad lineal: (a) un conjunto no linealmente separable; (b) un conjunto linealmente separable.

Clasificadores Lineales

Típicamente los clasificadores lineales se trabajan de dos formas

- Etiquetas 0 y 1:

En este caso, la clasificación se obtiene mediante el criterio

$$y(\mathbf{x}) = \mathbf{1}(\mathbf{w}^T \mathbf{x} > 0).$$

- Etiquetas -1 y 1 :

En este caso, la clasificación se obtiene mediante el criterio

$$y(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x}).$$

En ambos casos, si queremos hallar el hiperplano separante óptimo $\mathbf{w} \in \mathbb{R}^{d+1}$, ambos criterios usan una función no-diferenciable.

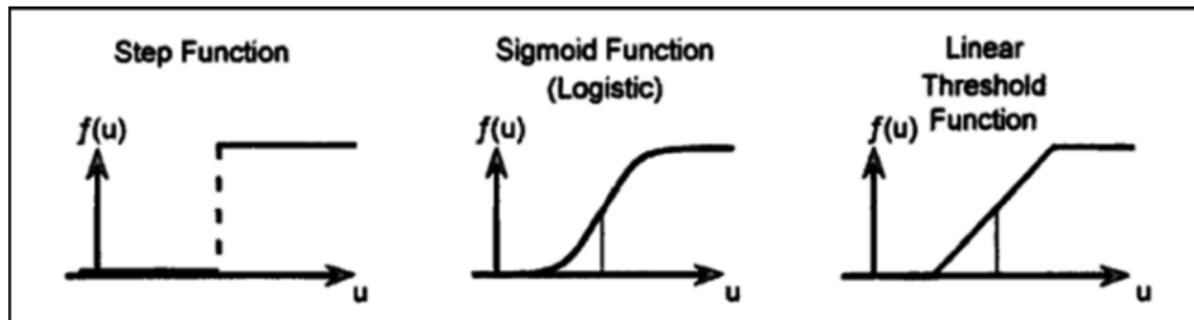
Regresión Logística

Consideramos el caso del clasificador logístico Aquí consideramos etiquetas $\{0, 1\}$, $\mathbf{x} \in \mathbb{R}$ y usamos el criterio de clasificación $\mathbf{1}(\mathbf{x} > 0)$.

El clasificador logístico utiliza la **función sigmoide estándar**

$$\sigma(\mathbf{x}) = \frac{1}{1 + e^{-x}}$$

como una aproximación suave de la función $\mathbf{1}(\mathbf{x} > 0)$.



Regresión Logística

Observaciones:

- $\sigma : \mathbb{R} \rightarrow (0, 1)$ es una función de clase C^∞ que transforma números reales en valores que pueden interpretarse como probabilidades.
- En general, $\sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}$ es una aproximación suave de $\mathbf{1}(\mathbf{w}^T \mathbf{x} > 0)$.
- σ tiene la siguiente propiedad: $\frac{d}{d\mathbf{x}}\sigma(\mathbf{x}) = \sigma(\mathbf{x})(1 - \sigma(\mathbf{x}))$.

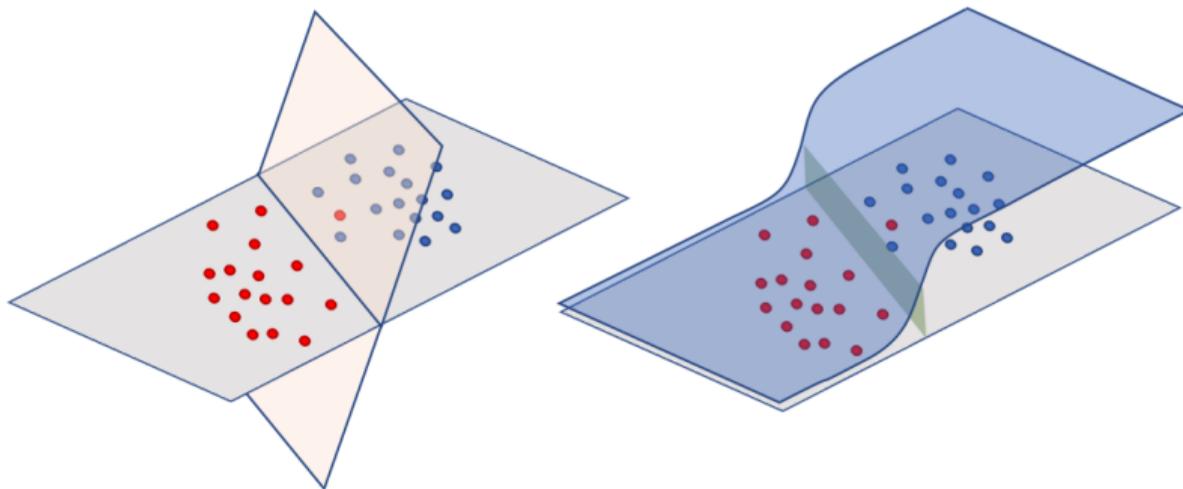
Prueba:

$$\begin{aligned}\frac{d}{d\mathbf{x}}\sigma(\mathbf{x}) &= \frac{e^{-\mathbf{x}}}{(1 + e^{-\mathbf{x}})^2} = \left(\frac{1}{1 + e^{-\mathbf{x}}}\right) \left(\frac{e^{-\mathbf{x}}}{1 + e^{-\mathbf{x}}}\right) \\ &= \sigma(\mathbf{x})(1 - \sigma(\mathbf{x})).\end{aligned}$$

Regresión Logística

Dado un conjunto de datos $\mathbb{X} \in \mathbb{R}^{n \times (d+1)}$ con etiquetas binarias, nuestro interés es hallar el vector óptimo de separación $\mathbf{w} \in \mathbb{R}^{d+1}$ tal que $y(\mathbf{x})$ sea lo más próximo al clasificador logístico

$$\hat{y}(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}.$$



Regresión Logística

Recordatorio: Regresión lineal.

Recordemos la función de pérdida en el caso de regresión. Tenemos

$$L = \mathbb{E} L(y_i, \hat{y}_i) = \frac{1}{n} \|\mathbf{y} - \mathbb{X}\mathbf{w}\|^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \mathbf{w}^T \mathbf{x}_i)^2.$$

En este caso podemos resolver de forma directa los coeficientes óptimos \mathbf{w} . Para ello, basta diferenciar con respecto de \mathbf{w} :

$$\begin{aligned} \nabla_{\mathbf{w}} L &= \nabla_{\mathbf{w}} \frac{1}{n} \|\mathbf{y} - \mathbb{X}\mathbf{w}\|^2 = \nabla_{\mathbf{w}} \frac{1}{n} \langle \mathbf{y} - \mathbb{X}\mathbf{w}, \mathbf{y} - \mathbb{X}\mathbf{w} \rangle \\ &= -\frac{2}{n} \langle \mathbb{X}, \mathbf{y} - \mathbb{X}\mathbf{w} \rangle = -\frac{2}{n} (\mathbb{X}^T \mathbf{y} - \mathbb{X}^T \mathbb{X} \mathbf{w}) = \mathbf{0}. \end{aligned}$$

$\Rightarrow \mathbb{X}^T \mathbb{X} \mathbf{w} = \mathbb{X}^T \mathbf{y}$, lo que conduce a la solución óptima $\mathbf{w}^* = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbf{y}$.

Regresión Logística

En el caso de la clasificación logística, tenemos la función de pérdida

$$L = \mathbb{E} L(y_i, \hat{y}_i) = \frac{1}{n} \|\mathbf{y} - \sigma(\mathbb{X}\mathbf{w})\|^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \sigma(\mathbf{w}^T \mathbf{x}_i))^2.$$

Al replicar la estrategia anterior, resulta:

$$\nabla_{\mathbf{w}} L = \nabla_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^n (y_i - \sigma(\mathbf{w}^T \mathbf{x}_i))^2 = -\frac{2}{n} \sum_{i=1}^n (y_i - \sigma(\mathbf{w}^T \mathbf{x}_i)) \sigma(\mathbf{w}^T \mathbf{x}_i) (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) \mathbf{x}_i.$$

Esta ecuación ya no produce una solución directa para \mathbf{w} . Sin embargo, es posible utilizar métodos iterativos para hallar el óptimo. Por ejemplo, podemos usar métodos de *descenso gradiente*

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \alpha \nabla_{\mathbf{w}} L(\mathbf{w}^{(k)}).$$

Enfoque probabilístico

Sea Z una v.a. con distribución $Ber(p)$, $0 < p < 1$. Tenemos las probabilidades condicionales

$$\begin{aligned}\mathbb{P}(Z = 1; p) &= \mathbb{P}(Z = 1; \mu = p) = p, \\ \mathbb{P}(Z = 0; p) &= \mathbb{P}(Z = 0; \mu = p) = 1 - p.\end{aligned}$$

Dado un conjunto de datos (\mathbf{x}_i, y_i) donde los $y_i \in \{0, 1\}$, podemos modelar el comportamiento de las y_i como una v.a. $Y \sim Ber(p)$, donde $\hat{p} = \mathbb{E}(y_i = 1) = \frac{m}{n}$, con m el número de datos en la clase $y = 1$.

Tenemos

$$\begin{aligned}\mathbb{P}(y_i = 1 \mid \mathbf{x}_i; p) &= p, \\ \mathbb{P}(y_i = 0 \mid \mathbf{x}_i; p) &= 1 - p.\end{aligned}$$

Enfoque probabilístico

Sin embargo, queremos que nuestro modelo represente p en términos del parámetro lineal $\mathbf{w} \in \mathbb{R}^{d+1}$. Hacemos

$$\begin{aligned}\mathbb{P}(y_i = 1 \mid \mathbf{x}_i; \mathbf{w}) &= \sigma(\mathbf{w}^T \mathbf{x}_i), \\ \mathbb{P}(y_i = 0 \mid \mathbf{x}_i; \mathbf{w}) &= 1 - \sigma(\mathbf{w}^T \mathbf{x}_i).\end{aligned}$$

Como Y es Bernoulli, podemos escribir la distribución condicional por

$$\mathbb{P}(y_i \mid \mathbf{x}_i; \mathbf{w}) = \sigma(\mathbf{w}^T \mathbf{x}_i)^{y_i} (1 - \sigma(\mathbf{w}^T \mathbf{x}_i))^{1-y_i}.$$

Asumiendo independencia de las y_i , la verosimilitud de \mathbf{w} dados los datos es

$$\mathcal{L}(\mathbf{w}) = \mathbb{P}(\mathbf{y} \mid \mathbb{X}; \mathbf{w}) = \prod_{i=1}^n \mathbb{P}(y_i \mid \mathbf{x}_i; \mathbf{w}) = \prod_{i=1}^n \sigma(\mathbf{w}^T \mathbf{x}_i)^{y_i} (1 - \sigma(\mathbf{w}^T \mathbf{x}_i))^{1-y_i}.$$

Enfoque probabilístico

Para hallar el \mathbf{w} óptimo, maximizamos la log-verosimilitud

$$\begin{aligned}\ell(\mathbf{w}) &= \log \mathcal{L}(\mathbf{w}) = \log \prod_{i=1}^n \sigma(\mathbf{w}^T \mathbf{x}_i)^{y_i} (1 - \sigma(\mathbf{w}^T \mathbf{x}_i))^{1-y_i} \\ &= \sum_{i=1}^n \left[y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) \right].\end{aligned}$$

Diferenciando en \mathbf{w} , resulta

$$\begin{aligned}\nabla_{\mathbf{w}} \ell(\mathbf{w}) &= \nabla_{\mathbf{w}} \sum_{i=1}^n \left[y_i \log \sigma(\mathbf{w}^T \mathbf{x}_i) + (1 - y_i) \log (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) \right] \\ &= \sum_{i=1}^n \left[\frac{y_i}{\sigma(\mathbf{w}^T \mathbf{x}_i)} \sigma(\mathbf{w}^T \mathbf{x}_i) (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) \mathbf{x}_i - \frac{1 - y_i}{1 - \sigma(\mathbf{w}^T \mathbf{x}_i)} \sigma(\mathbf{w}^T \mathbf{x}_i) (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) \mathbf{x}_i \right] \\ &= \sum_{i=1}^n \left(y_i (1 - \sigma(\mathbf{w}^T \mathbf{x}_i)) - (1 - y_i) \sigma(\mathbf{w}^T \mathbf{x}_i) \right) \mathbf{x}_i.\end{aligned}$$

Enfoque probabilístico

Así,

$$\nabla_{\mathbf{w}} \ell(\mathbf{w}) = \sum_{i=1}^n (y_i - \sigma(\mathbf{w}^T \mathbf{x}_i)) \mathbf{x}_i. \quad (3)$$

Finalmente, usamos (3) en el método de descenso gradiente

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \alpha \nabla_{\mathbf{w}} \ell(\mathbf{w}^{(k)}).$$

Tenemos el siguiente

Algoritmo:

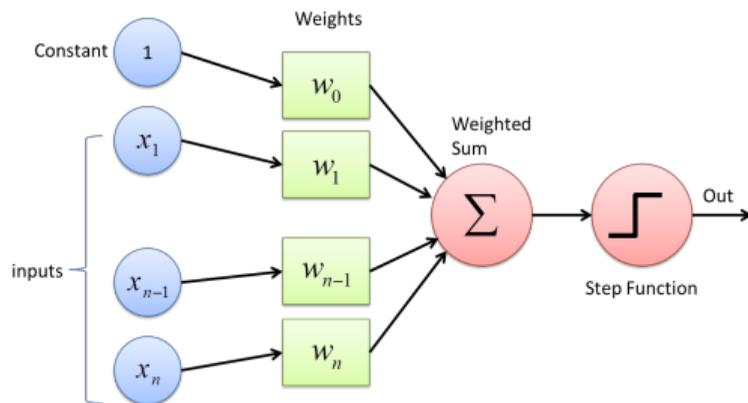
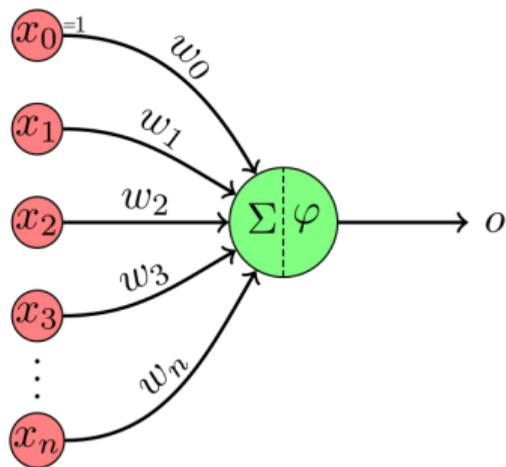
- 1.) Inicio: Elegir $\alpha > 0$, $\mathbf{w}^{(0)} \in \mathbb{R}^{d+1}$ arbitrario.
- 2.) Repetir para $k = 0, 1, 2, \dots$ (hasta cierto criterio de paro):
 - Calcular $\nabla_{\mathbf{w}} \ell(\mathbf{w}^{(k)})$ como en (3).
 - Recalcular $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \alpha \nabla_{\mathbf{w}} \ell(\mathbf{w}^{(k)})$.

Enfoque probabilístico

Observaciones:

- El método de descenso “mueve” $\mathbf{w}^{(k)}$ según la contribución de los datos mal clasificados (proporcional a la diferencia $y_i - \sigma(\mathbf{w}^T \mathbf{x}_i)$).
- La convergencia de este método depende del conjunto de datos:
 - El método de descenso gradiente siempre converge (a un mínimo local) para el caso de un conjunto linealmente separable.
 - La convergencia puede verse afectada en el caso no separable. Esto puede resolverse modificando o usando un método de descenso más elaborado (curso de Optimización).
- Las ideas aquí descritas dan origen a modelos lineales de transferencia de información (modelos neuronales). Por ejemplo, **el perceptrón**.

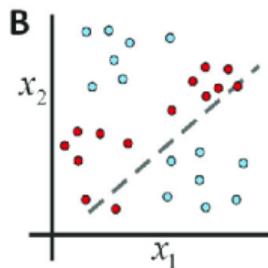
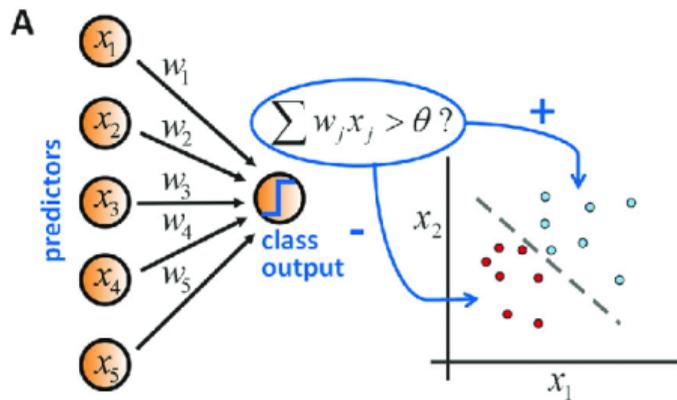
El perceptrón



El modelo perceptrón.

La salida es de la forma $y = \varphi(\mathbf{w}^T \mathbf{x})$, donde $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ es una *función de transferencia* o *función de activación*.

El perceptrón



x_1	x_2	y
0	0	0
1	0	1
0	1	1
1	1	0

