Regresión vs clasificación

DRA. CONSUELO VARINIA GARCÍA MENDOZA

Regresión vs clasificación

En los problemas vistos previamente se crean modelos capaces de predecir valores continuos

- Precios de casas
- Clima
- Valores de acciones

Para este tipo de problemas los algoritmos de regresión lineal obtienen buenos resultados

Pero existen otros problemas donde los valores que se desean predecir no son continuos si no categóricos

Clasificación

Algunos problemas con valores categóricos a predecir son:

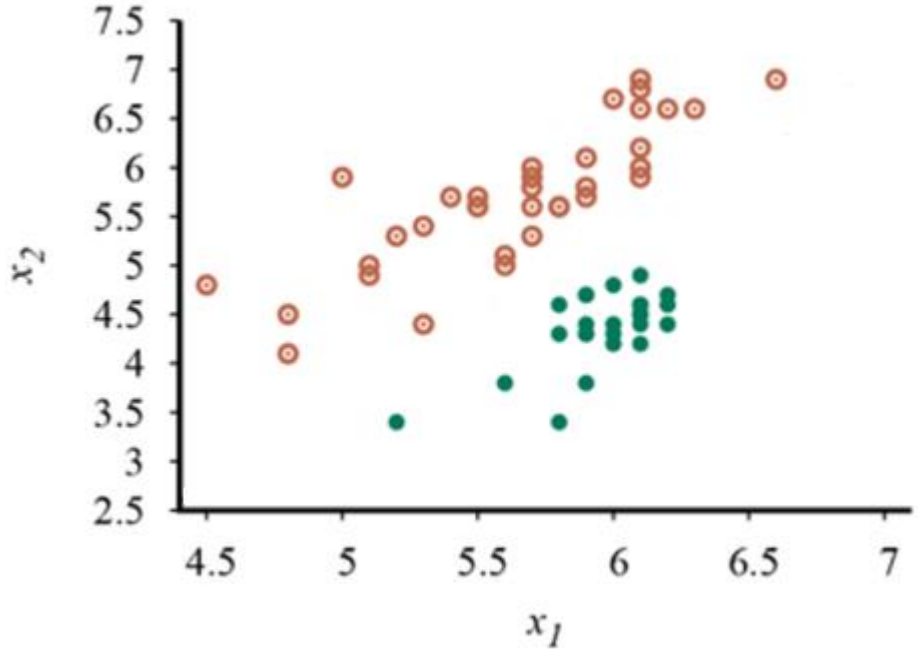
- Identificar si un correo es spam o no
- Identificar si un paciente tiene cáncer o no
- Identificar la polaridad de una opinión (positiva, negativa, neutra)

A este tipo de problemas se les conoce como problemas de clasificación

Los dos primeros problemas son de clasificación binaria, mientras el último es un problema de clasificación multiclase

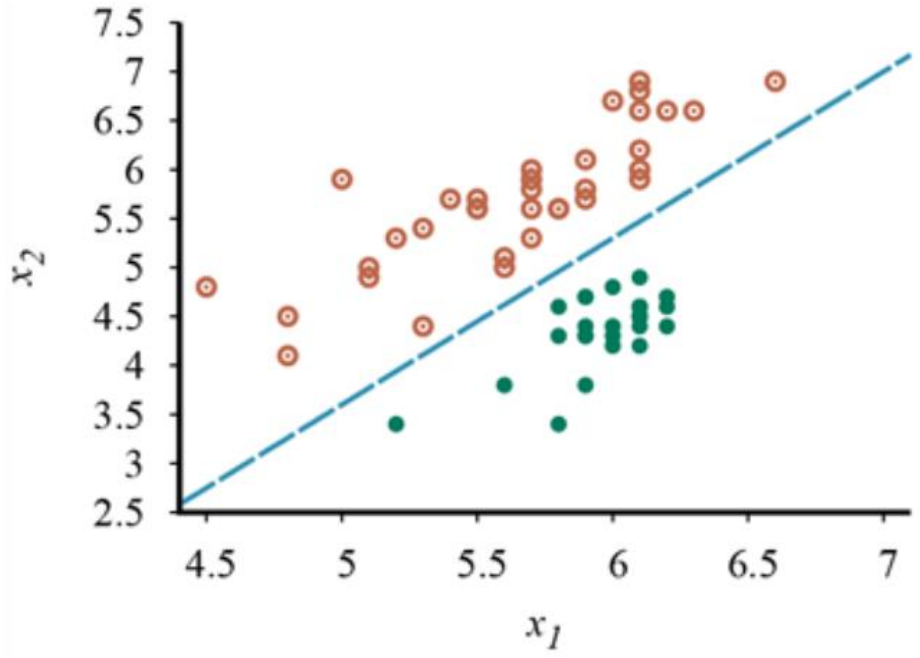
¿Se puede utilizar regresión lineal para problemas de clasificación?

¿Se puede utilizar regresión lineal para problemas de clasificación?



Límite de decisión

Una forma de aplicar la regresión lineal a los problemas de clasificación es estableciendo un límite de decisión (decision boundary)



Separador lineal

$x_2 = 1.7x_1 - 4.9 \quad \text{or} \quad -4.9 + 1.7x_1 - x_2 = 0$

Explosiones

$-4.9 + 1.7x_1 - x_2 > 0$

Terremotos

$-4.9 + 1.7x_1 - x_2 < 0$

Simplificando la ecuación

$-4.9x_0 + 1.7x_1 - x_2 = 0$

Representación como un vector de pesos

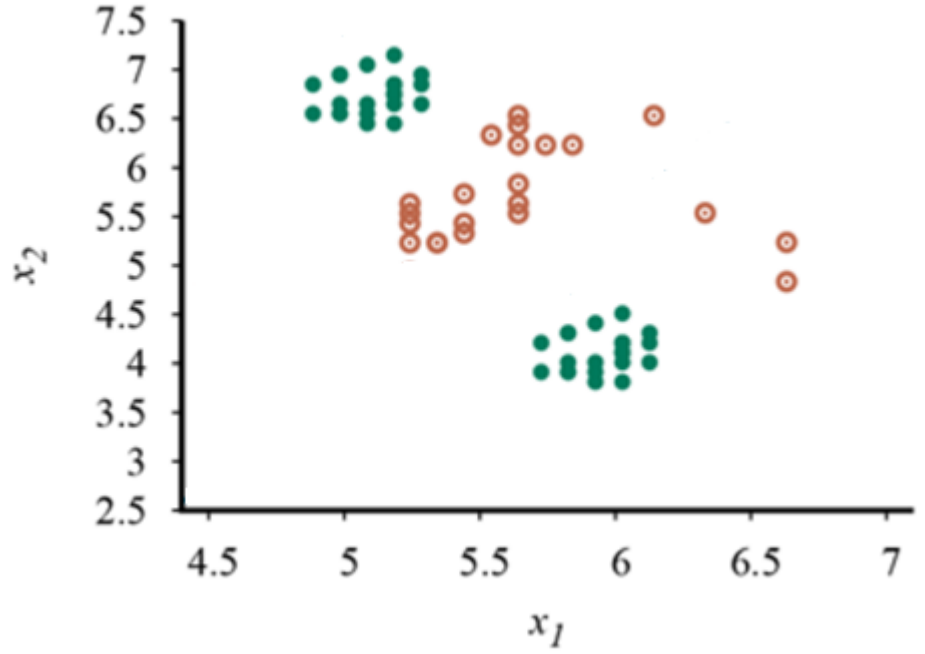
$\mathbf{w} = \langle -4.9, 1.7, -1 \rangle$

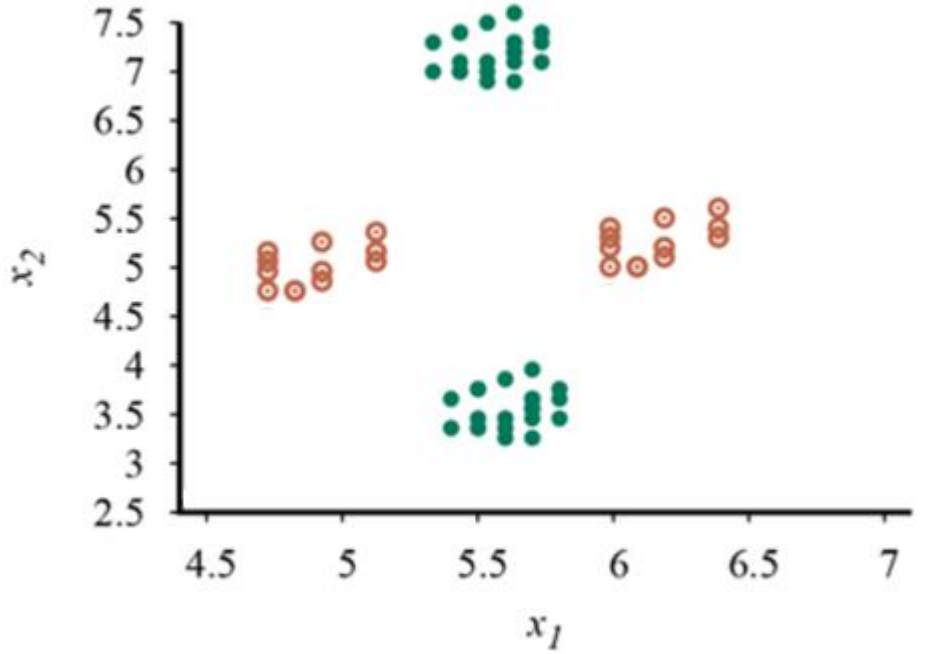
Hipótesis de clasificación

$h_{\mathbf{w}}(\mathbf{x}) = 1 \text{ if } \mathbf{w} \cdot \mathbf{x} \geq 0 \text{ and } 0 \text{ otherwise}$

Limitaciones del uso de la regresión lineal para la clasificación

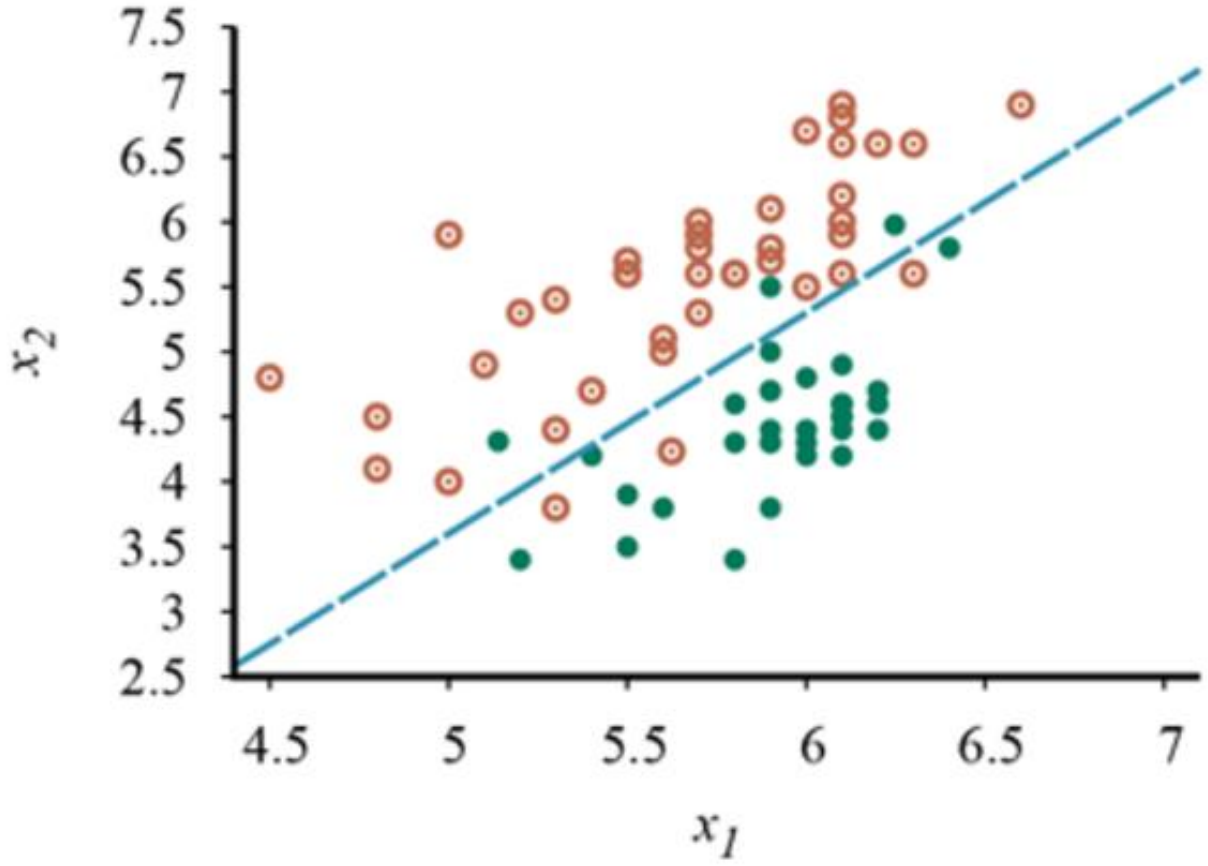
La regresión lineal asume que los datos son linealmente separables





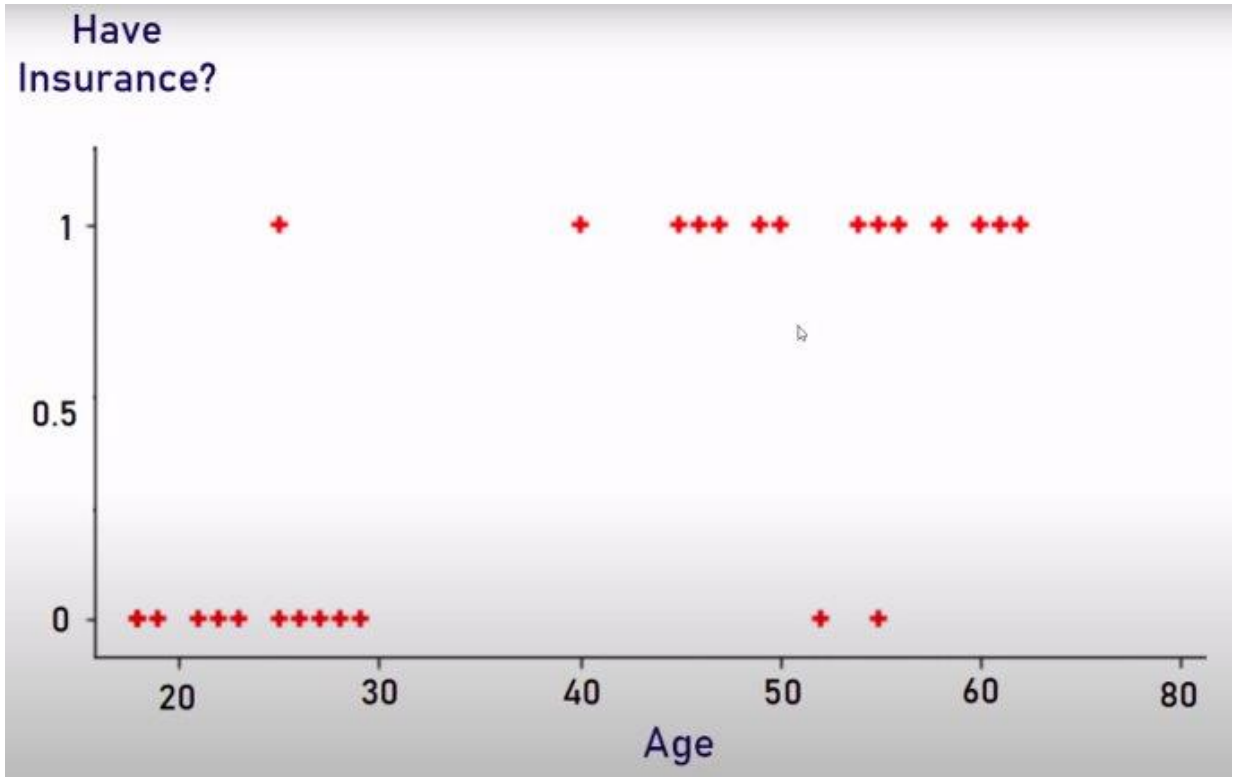
Limitaciones del uso de la regresión lineal para la clasificación

Además, el límite de decisión que se establece se ve muy afectado cuando los datos son dispersos

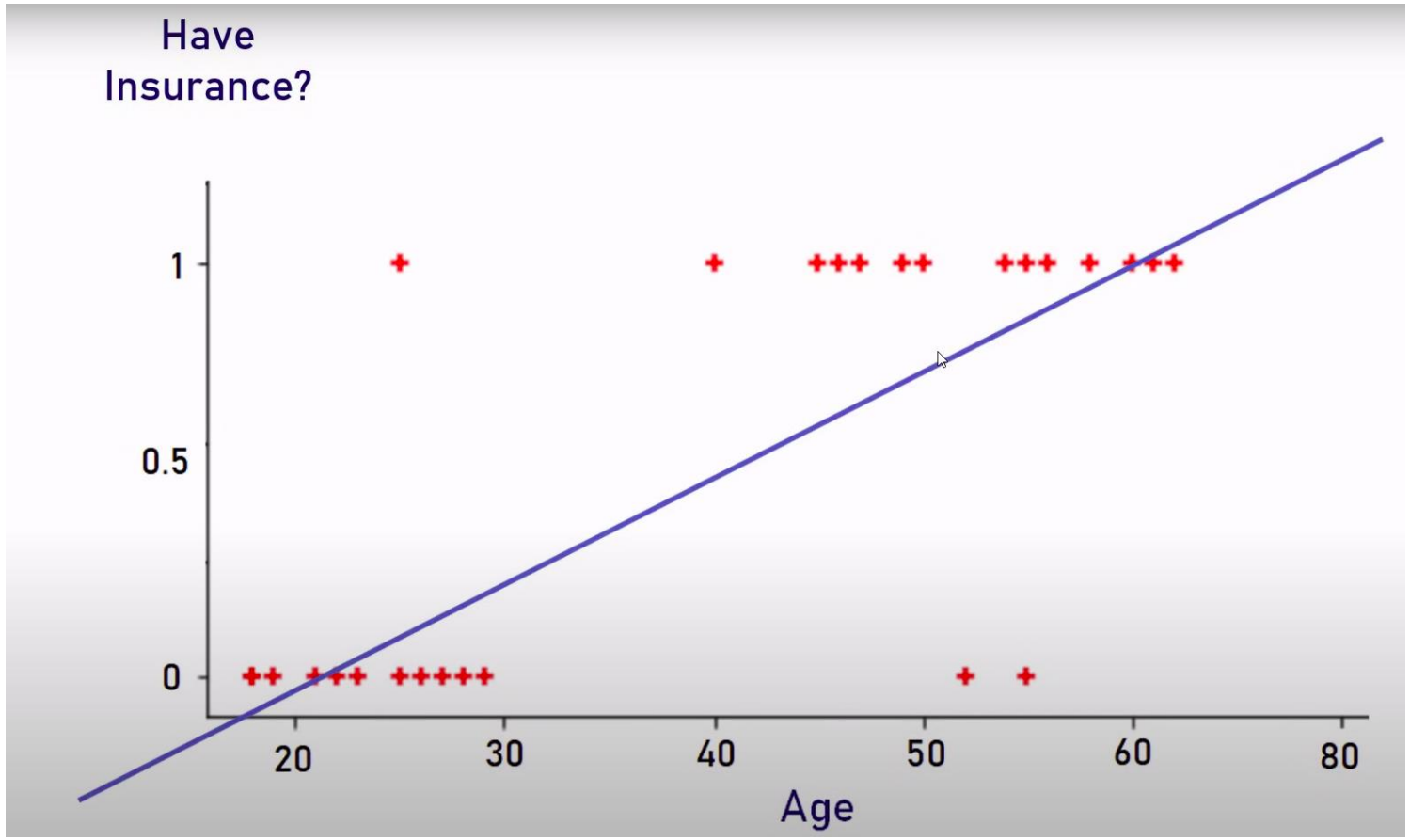


Representación de un problema de clasificación binaria

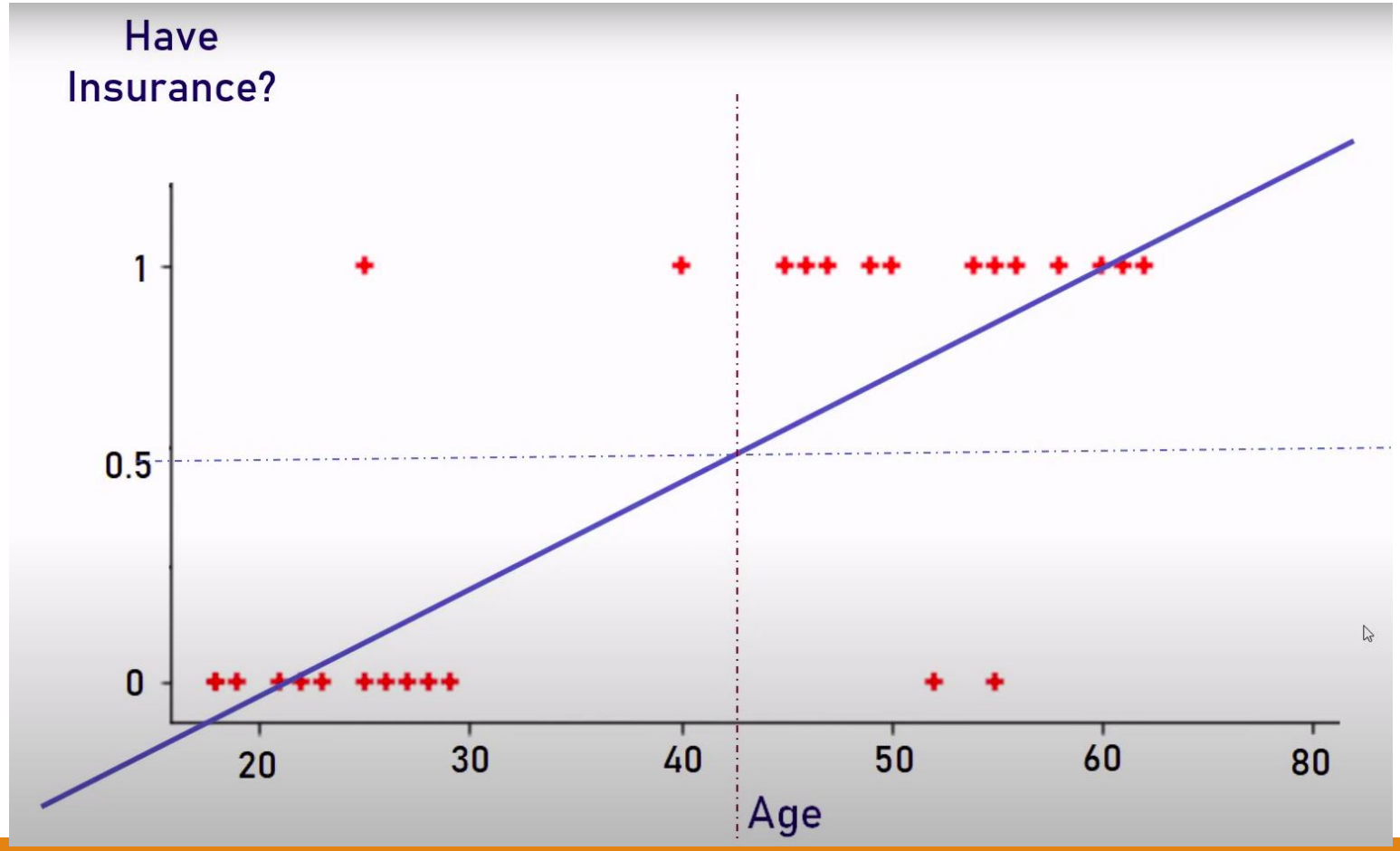
age	have_insurance
22	0
25	0
47	1
52	0
46	1
56	1
55	0
60	1
62	1
61	1
18	0
28	0
27	0
29	0
49	1



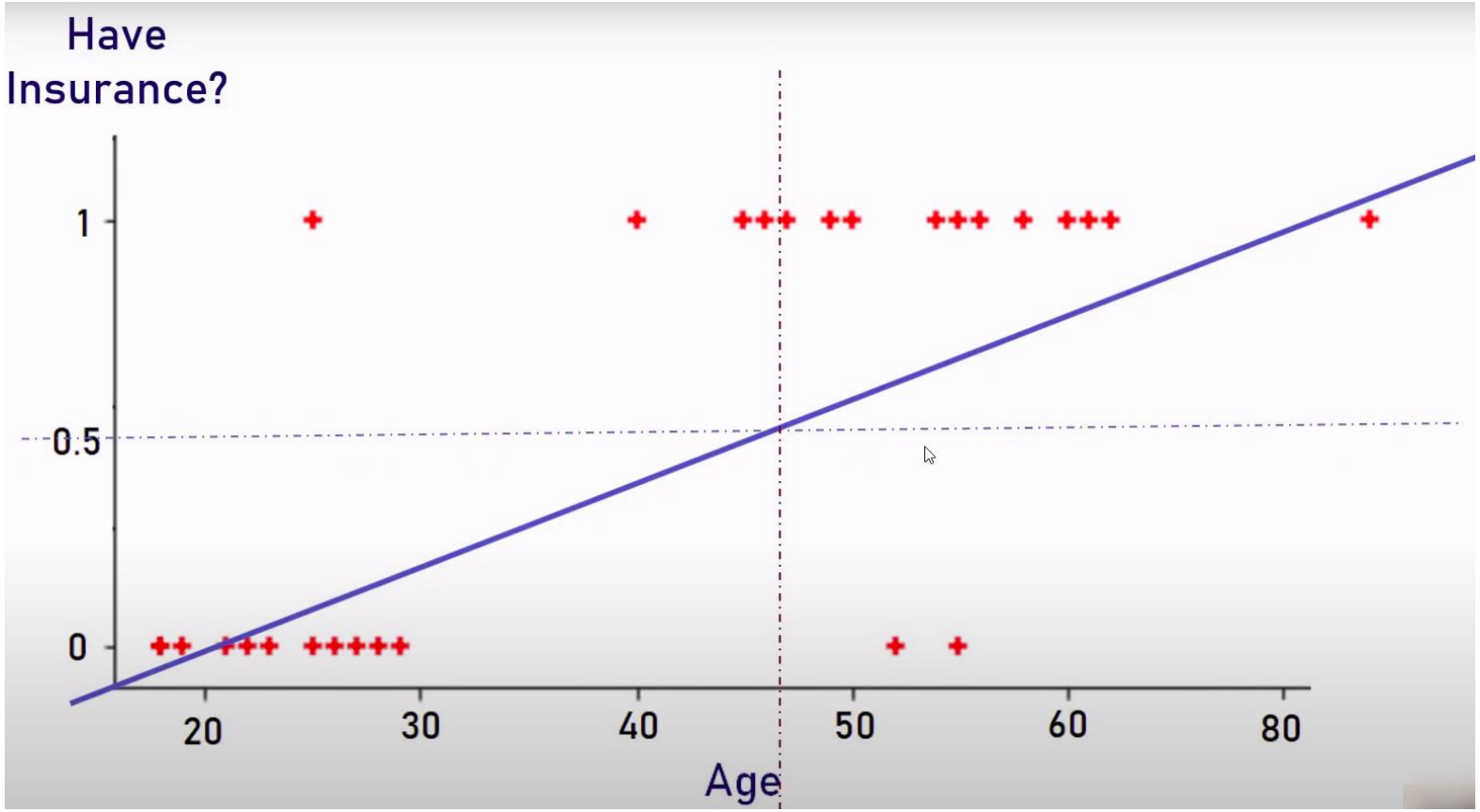
Representación de un problema de clasificación binaria



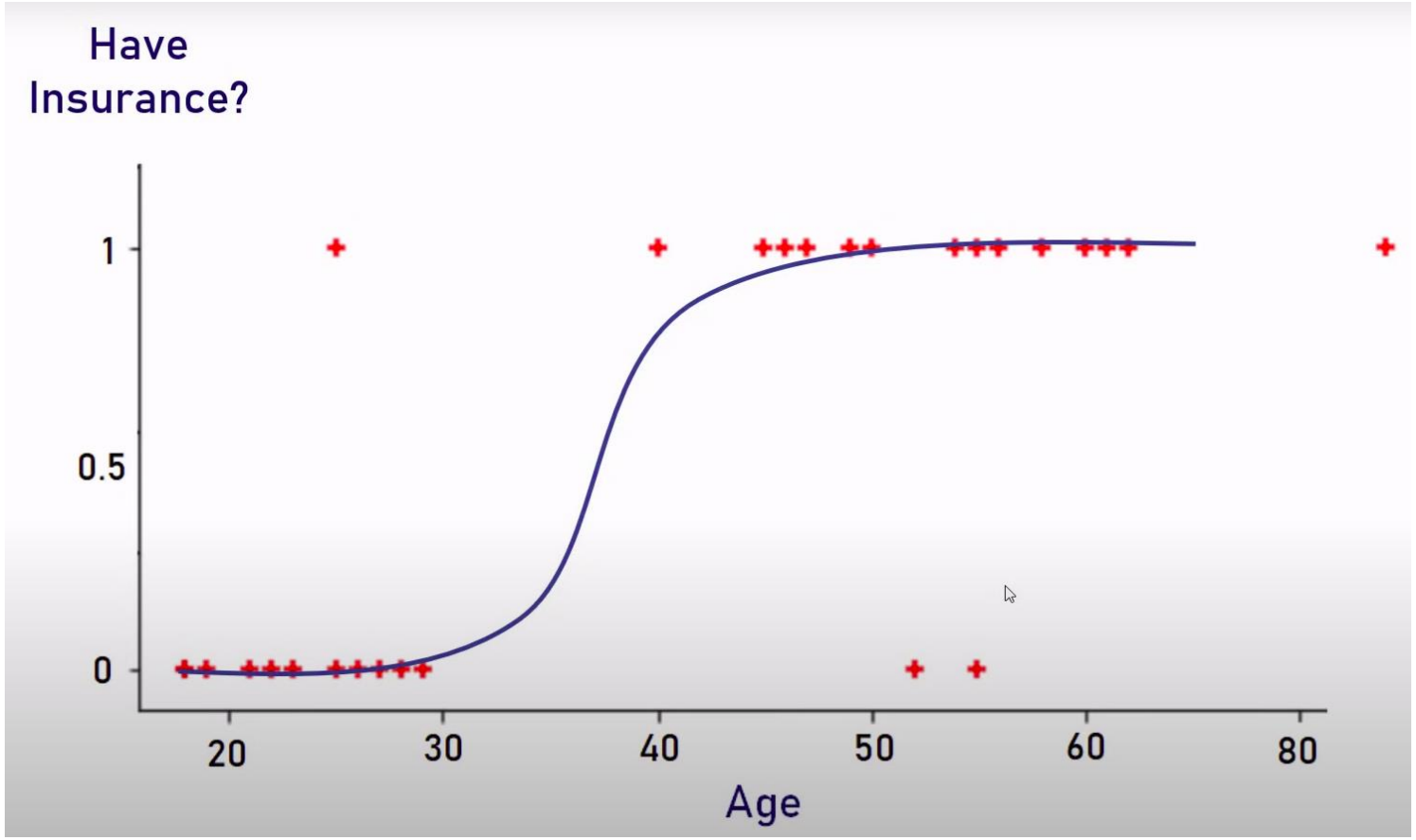
Límite de decisión



Efecto de datos dispersos



Función logística o sigmoidal



Características de la función sigmoide

Tienen una progresión temporal en niveles bajos al inicio

Después de un tiempo se acerca a un climax y se produce una transición acelerada mostrando valores altos

Después de esta aceleración los valores se mantienen en este nivel alto

La gráfica muestra una forma típica en forma de "S"

La ecuación de la función sigmoide se define de la siguiente manera:

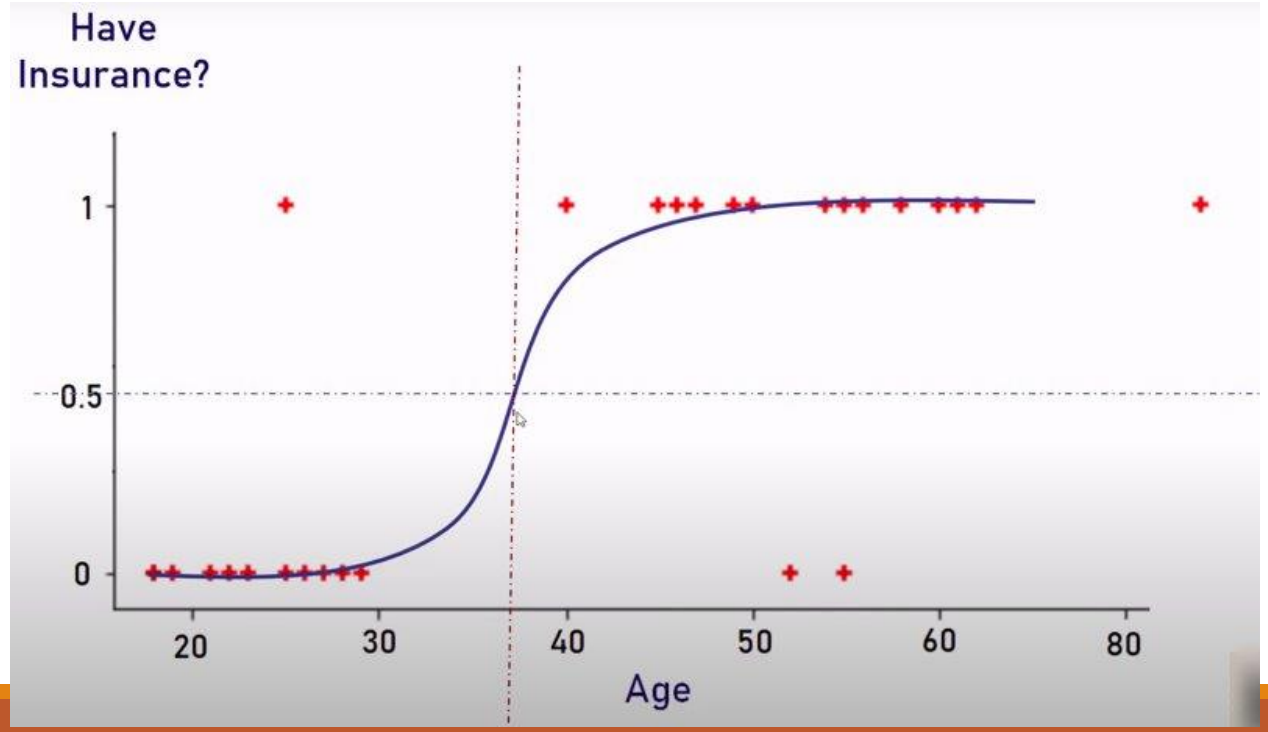
$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

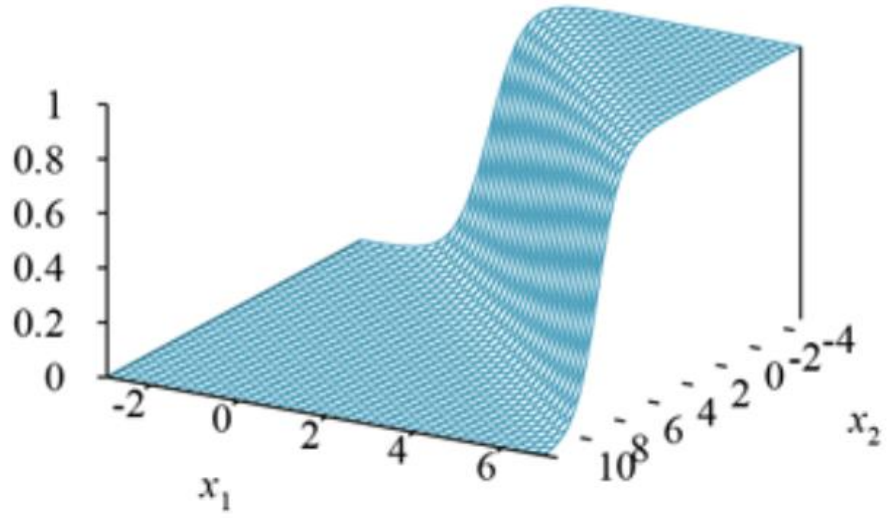
e = número de Euler ~ 2.71828

Características de la función sigmoide

La función sigmoide genera valores entre 0 y 1

Con estos valores se simplifica el establecer un límite de decisión





Hipótesis de la regresión logística

$$h_w(x) = w_1 x + w_0$$

$$sigmoid(z) = \frac{1}{1 + e^{-z}}$$

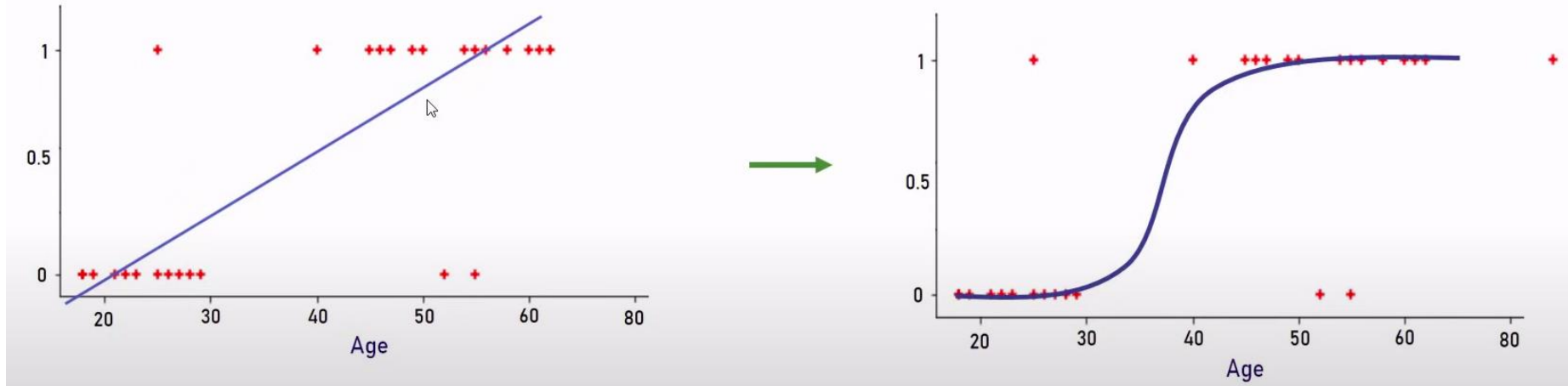
$$h_w(x) = g(w_1 x + x_0) = \left(\frac{1}{1 + e^{-(w_1 x + x_0)}} \right)$$

Hipótesis de la regresión lineal

Hipótesis de la regresión logística

$$y = m * x + b$$

$$y = \frac{1}{1 + e^{-(m*x+b)}}$$



Distribución de la probabilidad

Los valores que resultan de la regresión logística pueden ser interpretados como valores de probabilidad y presentan la siguiente distribución

- $P(y=1|x;w)=h_w(x)$
- $P(y=0|x;w)=1-h_w(x)$

Estas distribuciones se pueden integrar en una sola ecuación

- $P(y|x;w)=h_w(x)^y(1-h_w(x))^{1-y}$
  - Si y=1 -> $h_w(x)^1(1-h_w(x))^0$ -> $h_w(x)^1(1)$ -> $h_w(x)$
  - *Si y=0* -> $h_w(x)^0(1-h_w(x))^1$ -> $(1)(1-h_w(x))$ -> $1-h_w(x)$

Función de pérdida en la regresión logística

Para ajustar los valores de los pesos en la regresión lineal usamos la siguiente ecuación

$$w_i = w_i - \alpha \frac{\partial}{\partial w_i} f(w)$$

- Aquí la función *f(w)* es una función de pérdida que se determina mediante el error cuadrado

$$f(w) = \sum_{i=1}^{m} (w \cdot x - y)^2$$

- En la regresión logística se utiliza como función de pérdida la máxima verosimilitud (maximum likelihood)

Máxima verosimilitud

La estimación de la máxima verosimilitud es un método para determinar los valores de los parámetros de un modelo

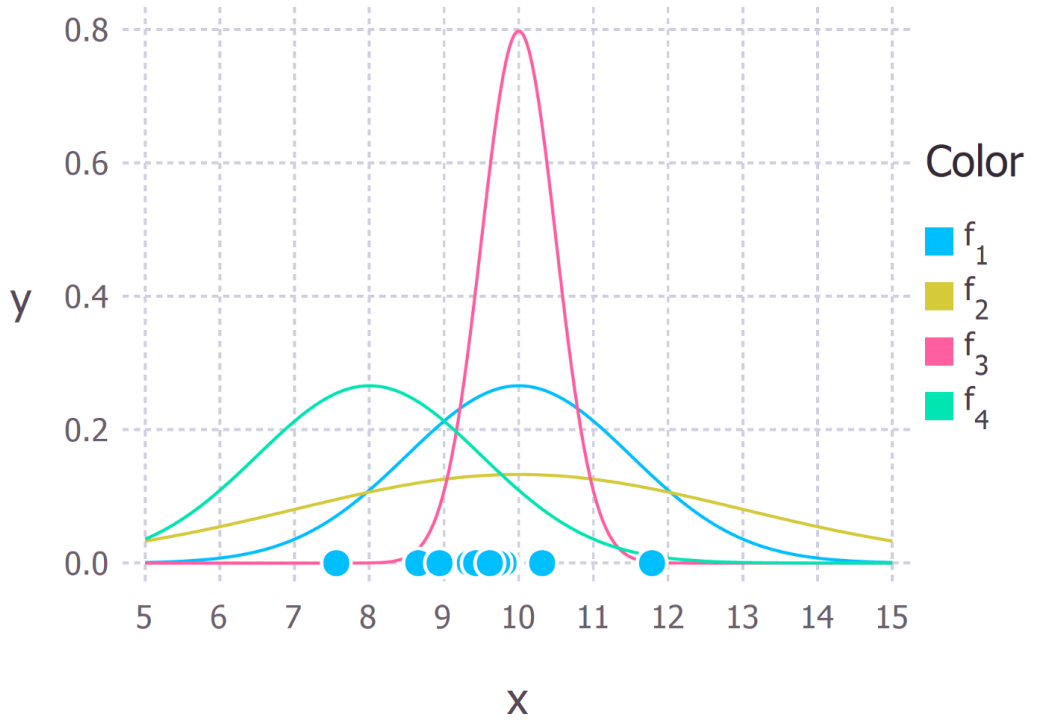
Los valores de los parámetros que se quieren encontrar son aquellos que maximizan la verosimilitud de tal forma que el proceso descrito por el modelo produce los datos que fueron observados

Cálculo de la máxima verosimilitud

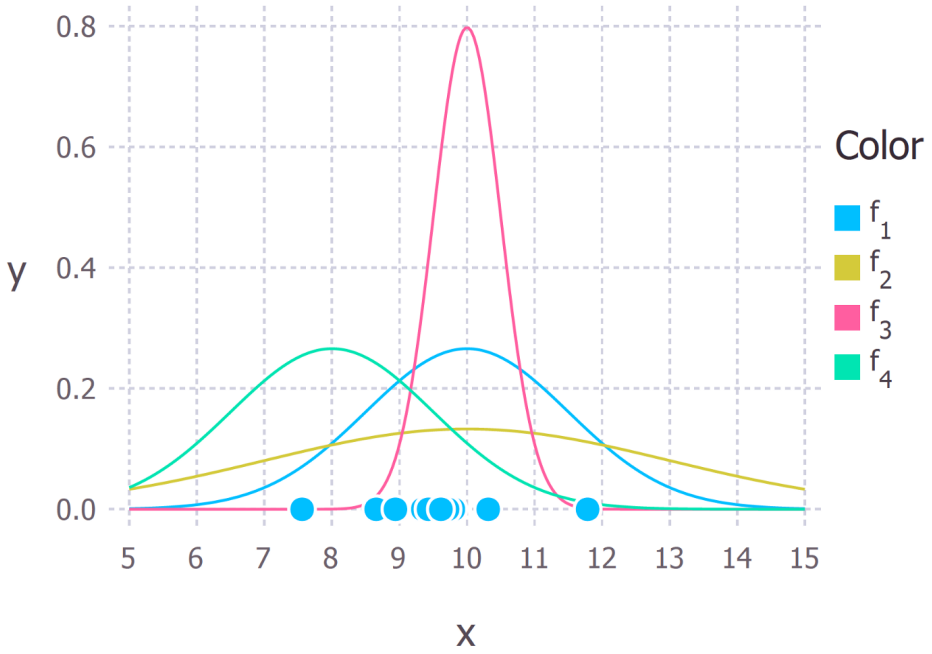
Suponga que se tienen una serie de datos que muestran una distribución normal

La distribución normal (Gaussiana) tiene dos parámetros, la media y la desviación estándar. Diferentes valores de estos parámetros generan diferentes curvas

Lo que se quiere es saber qué curva es la más probable para crear los puntos de datos que se observan



Cálculo de la máxima verosimilitud



Distribución normal a través de la densidad de probabilidad

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

La máxima verosimilitud se calcularía encontrando la función de densidad conjunta de todas las observaciones de la siguiente forma

$$\prod_{i=1}^{n} P(x_i;\mu,\sigma)$$

Ejemplo 1. Cálculo de la máxima verosimilitud

Si se tienen los datos 9, 9.5 y 11, la máxima verosimilitud se calcularía encontrando la función de densidad conjunta de todas las observaciones de la siguiente forma

$$\prod_{i=1}^{n} P(x_i; \mu, \sigma)$$

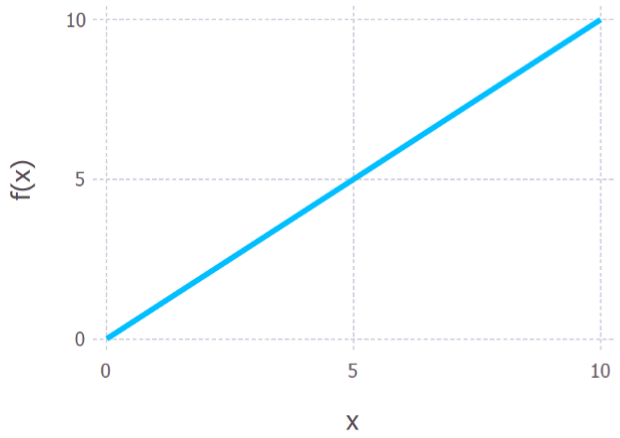
$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

Logaritmo de la máxima verosimilitud

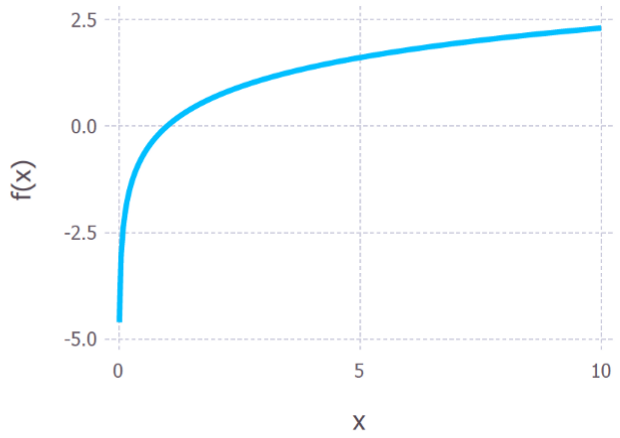
El cálculo anterior requiere encontrar los valores que maximicen la verosimilitud y para esto tenemos que aplicar la derivada

Una forma de simplificar el cálculo de la derivada es tomando el logaritmo natural de la expresión

Esto se puede hacer ya que el logaritmo natural es una función monótona creciente



(a) $f(x) = x$



(b) $f(x) = \ln(x)$

Ejemplo 1. Logaritmo de la máxima verosimilitud

$$P(9, 9.5, 11; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(9.5-\mu)^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(11-\mu)^2}{2\sigma^2}\right)$$

$$\ln(P(x; \mu, \sigma)) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(9.5-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(11-\mu)^2}{2\sigma^2}$$

- Usando las leyes de los logaritmos se obtiene

$$\ln(P(x; \mu, \sigma)) = -3\ln(\sigma) - \frac{3}{2}\ln(2\pi) - \frac{1}{2\sigma^2}\left[(9-\mu)^2 + (9.5-\mu)^2 + (11-\mu)^2\right]$$

- Aplicando la derivada

$$\frac{\partial \ln(P(x; \mu, \sigma))}{\partial \mu} = \frac{1}{\sigma^2}\left[9 + 9.5 + 11 - 3\mu\right].$$

$$\mu = \frac{9 + 9.5 + 11}{3} = 9.833$$

Regresión logística
Cálculo de la máxima verosimilitud

Distribución normal a través de la densidad de probabilidad

$$P(y|x;w) = h_w(x)^y\big(1 - h_w(x)\big)^{1-y}$$

Máxima verosimilitud

$$\prod_{i=1}^{m} P(y_i|x_i;w) = \prod_{i=i}^{n} h_w(x_i)^{y_i}\big(1 - h_w(x_i)\big)^{1-y_i}$$

Aplicando el logaritmo

$$\sum_{i=i}^{m} y_i \log\big(h_w(x_i)\big) + (1 - y_i)\log\big(1 - h_w(x_i)\big)$$

Gradiente ascendiente

Al igual que se hizo con la regresión lineal, en la regresión logística se requiere un algoritmo que permita realizar el ajuste de los pesos

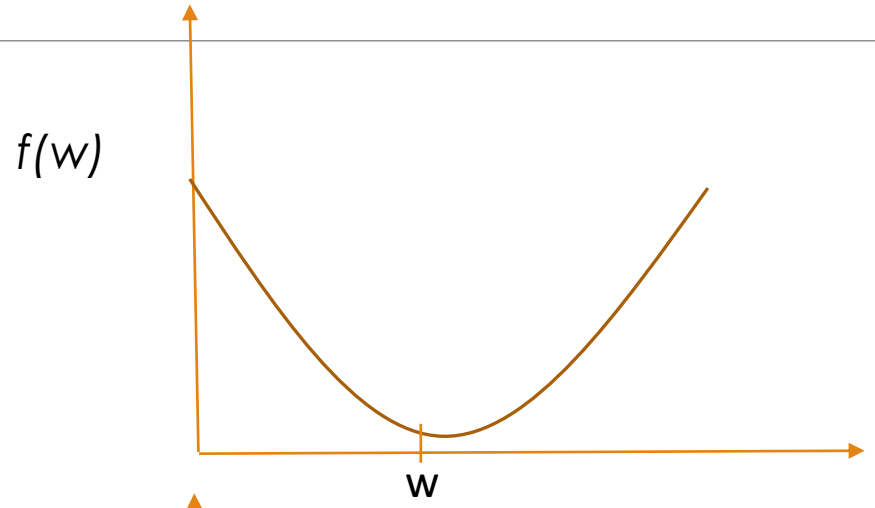
En la regresión lineal se ajustan los pesos mediante el algoritmo de gradiente descendiente **minimizando** el error cuadrado medio (función de pérdida)

En la regresión logística la función de pérdida es la máxima verosimilitud, y aquí lo que se desea es **maximizar** este valor

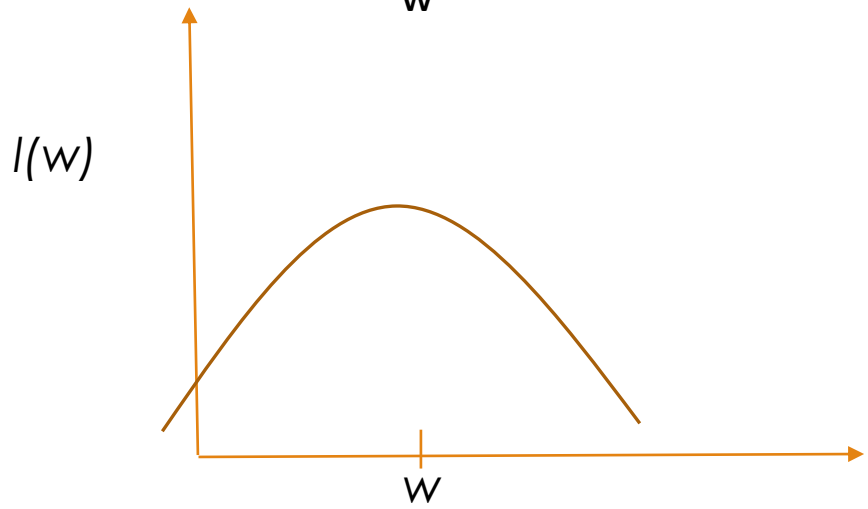
Por lo anterior se puede usar una nueva versión del algoritmo de gradiente que en vez de descender vaya ascendiendo

A este algoritmo se le conoce como gradiente ascendiente

Gradiente ascendiente



$$w \coloneqq w - \alpha \frac{\partial}{\partial f(w)} f(w)$$



? $$w \coloneqq w + \alpha \frac{\partial}{\partial l(w)} l(w)$$

Gradiente ascendiente

$$w \coloneqq w + \alpha \frac{\partial}{\partial l(w)} l(w)$$

$$l(w) = \sum_{i=1}^{m} y_i \log(h_w(x_i)) + (1 - y_i) \log(1 - h_w(x_i))$$

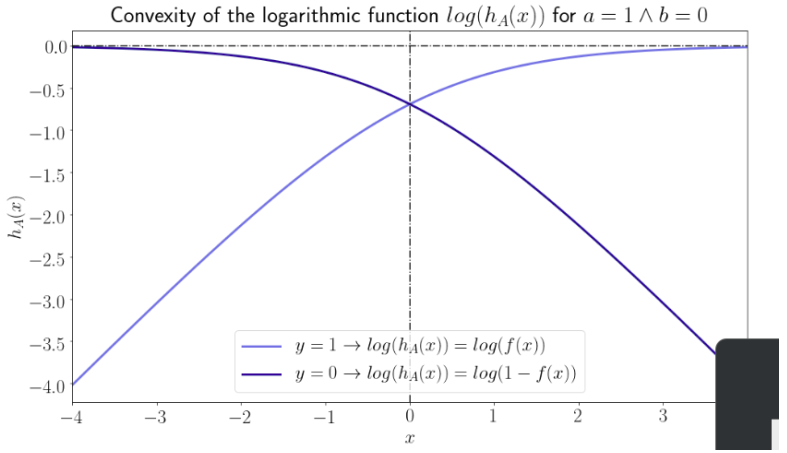
$$w = w + \alpha \frac{\partial}{\partial l(w)} \sum_{i=1}^{m} y_i \log(h_w(x_i)) + (1 - y_i) \log(1 - h_w(x_i))$$

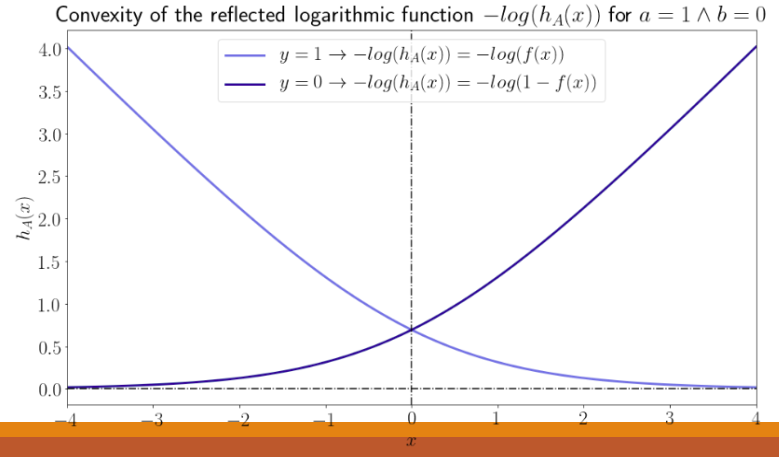
$$w \coloneqq w + \alpha \sum_{i=i}^{m} (y_i - h_w(x_i)) \cdot x_i$$

$$\sum_{i=1}^{m} 2(w \cdot x - y) \cdot x$$

Regresión logística con gradiente descendiente

- Se puede hacer una modificación para utilizar gradiente descendiente
- Cuando se aplica la función logaritmo se obtiene una forma cóncava
- Dado que gradiente descendiente busca minimizar el error, la función debe tener una forma convexa
- Para esto se cambia el signo de la función logaritmo





Regresión logística con gradiente descendiente

Considerando lo anterior se obtendría la siguiente ecuación para la actualización de los pesos

$$w = w - \alpha \sum_{i=1}^{m} (h_w(x_i) - y_i) \cdot x_i$$

- Esta ecuación es idéntica a la que utilizamos en la regresión lineal, pero se debe recordar que la función h_w ahora genera valores entre 0 y 1 ya que se pasa a una forma sigmoide

$$h_w(x) = g(w_1 x + x_0) = \left(\frac{1}{1 + e^{-(w_1 x + x_0)}} \right)$$

Clasificación multiclase

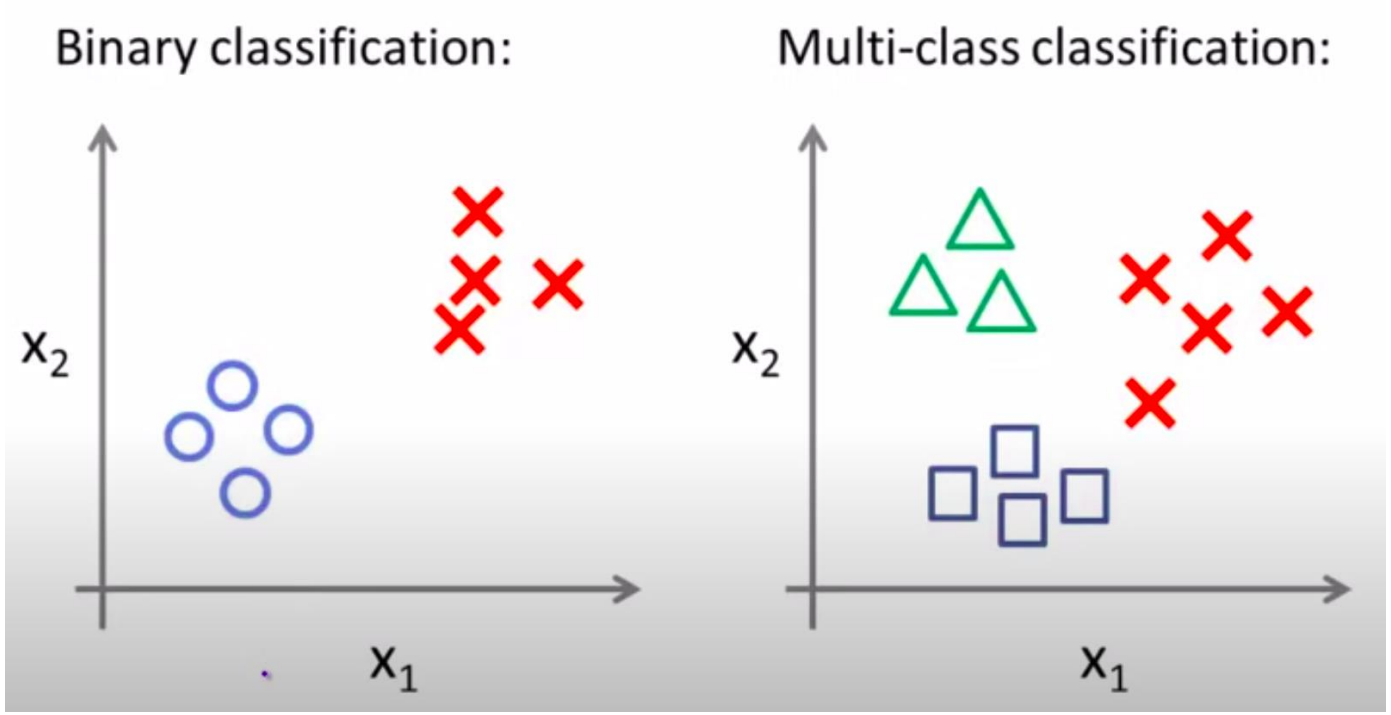
En el problema de clasificación multiclase los valores a predecir pueden tomar más de 2 valores distintos

Ejemplos:

- Polaridad de opinión: [positivo, negativo, neutro]
- Calificación de un producto (estrellas): [1,2,3,4,5]
- Clima: [soleado, lluvioso, nublado, nevado]

Es importante aclarar que la clase a predecir sólo puede tomar uno de estos valores

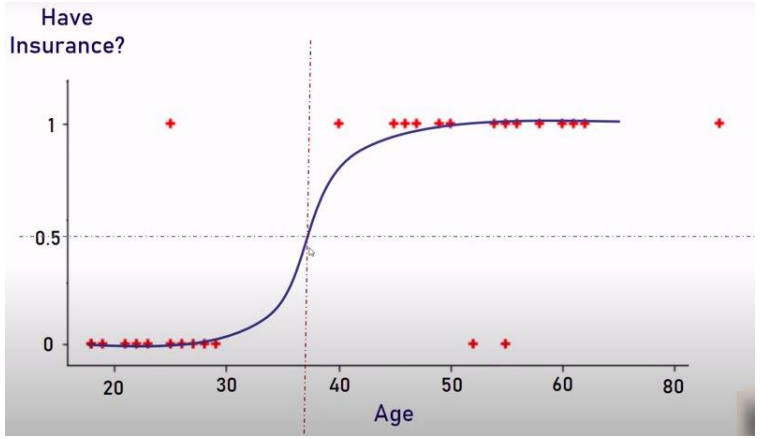
Clasificación multiclase



Regresión logística para clasificación multiclase

Como se explicó anteriormente la regresión logística permite resolver problemas de clasificación binaria

Por su funcionamiento este algoritmo no resuelve problemas multiclase de forma directa



¿Cómo se puede resolver el problema de clasificación multiclase mediante regresión logística?

Estrategias para clasificación multiclase

Existen algoritmos de aprendizaje automático que de forma nativa resuelven problemas de clasificación binaria

- Naïve Bayes
- Árboles de decisión
- Redes neuronales

Para aquellos algoritmos que no manejan la clasificación multiclase de forma nativa existen estrategias que extienden su funcionamiento para poder manejar estos problemas

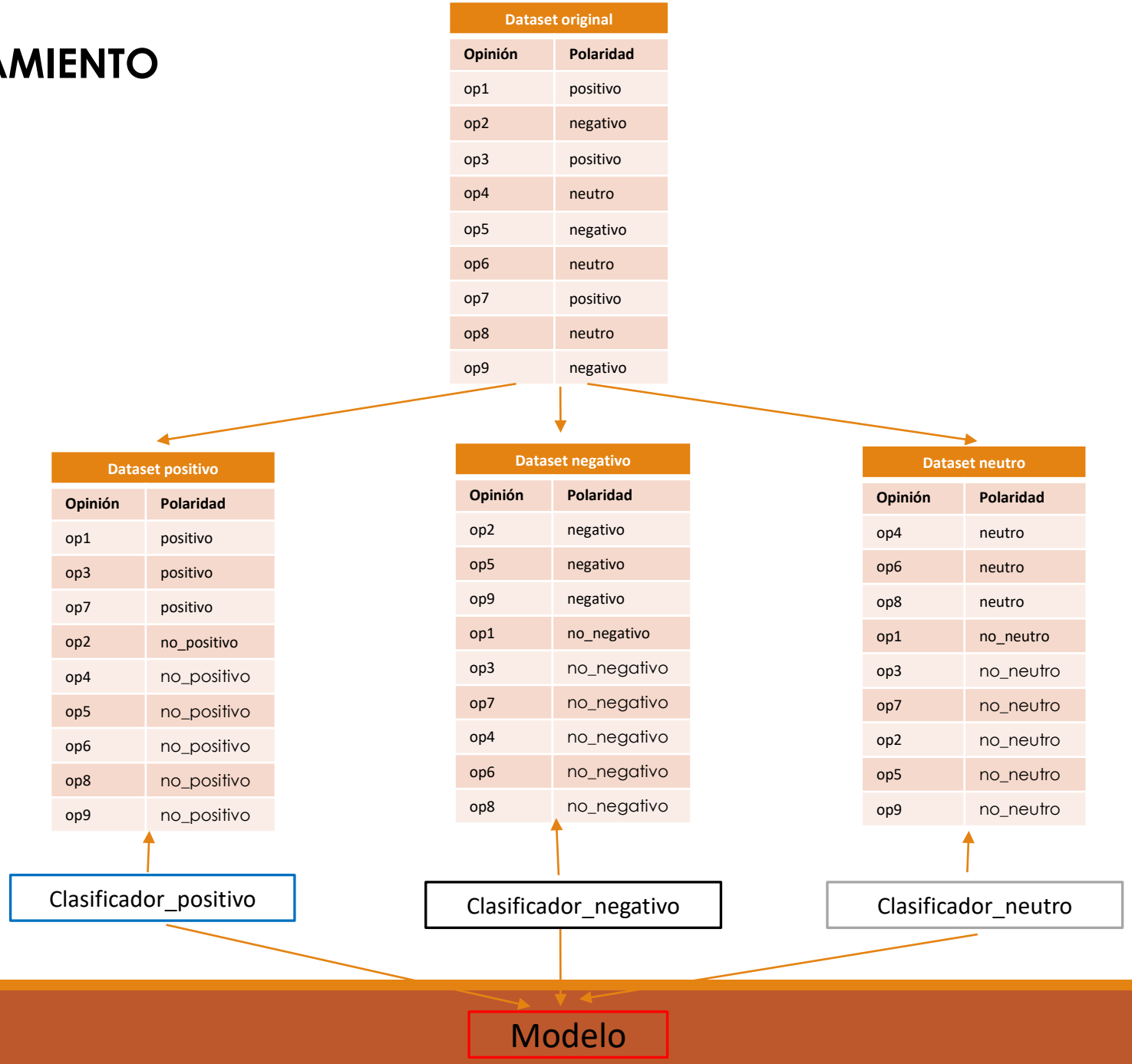
- *One vs All* (también llamado *One vs Rest*)
- *One vs One*

One vs All

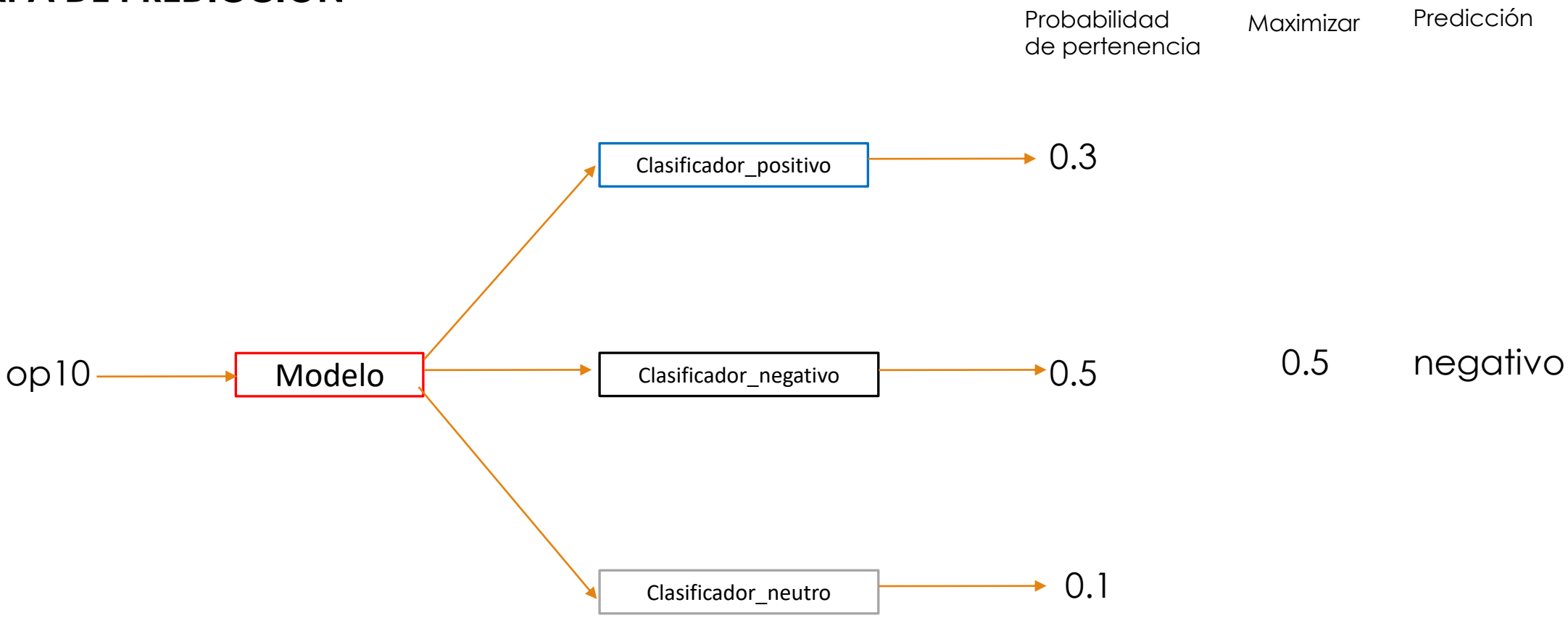
La estrategia parte el problema de clasificación multiclase en múltiples conjuntos de datos de clasificación binaria

Posteriormente entrena un modelo de clasificación binaria en cada conjunto de datos

ETAPA DE ENTRENAMIENTO



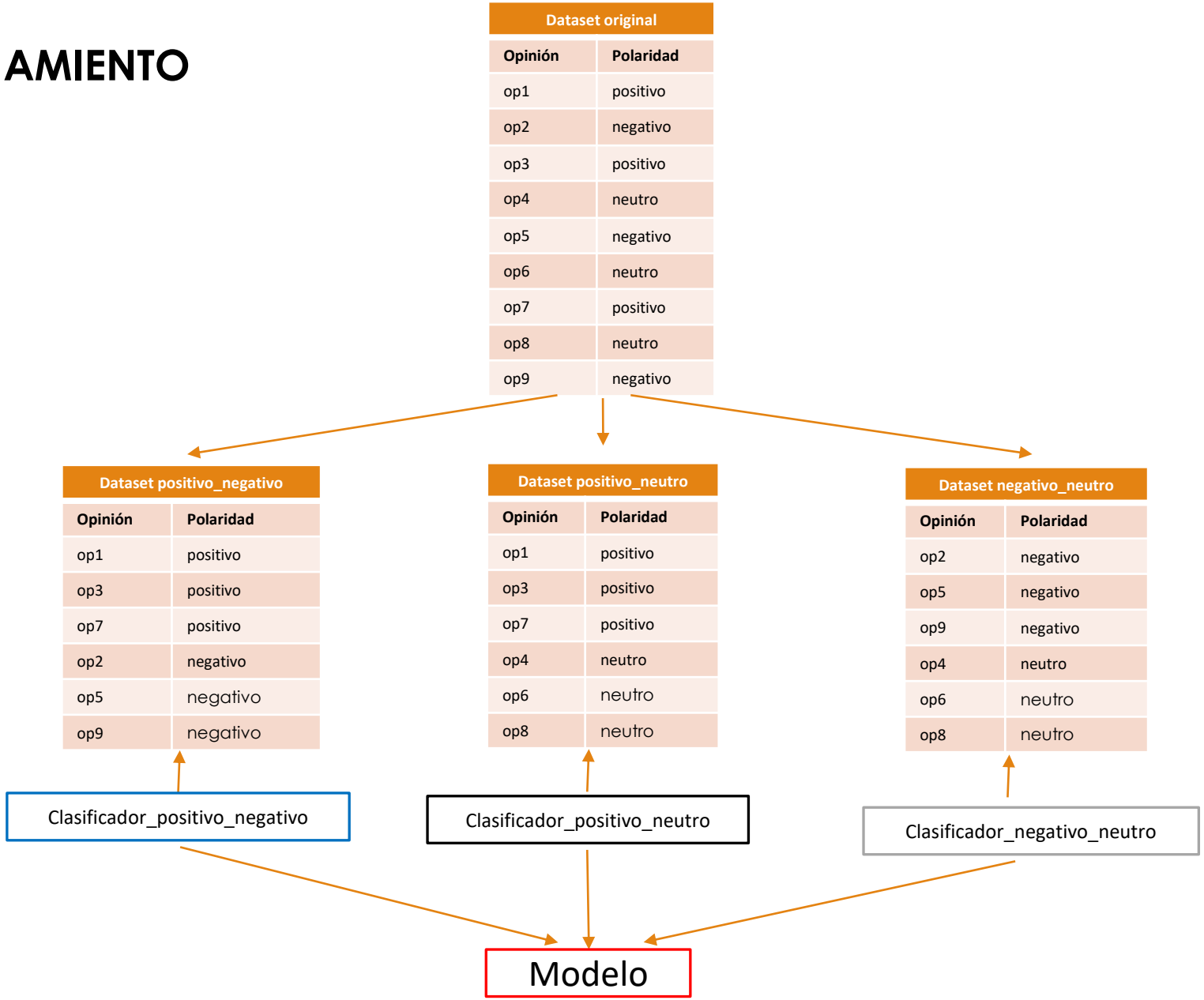
ETAPA DE PREDICCIÓN



One vs One

Es una estrategia muy similar a la anterior, pero ahora se crean nuevos datasets formados por todos los pares de clases disponibles

ETAPA DE ENTRENAMIENTO



ETAPA DE PREDICCIÓN