Clasificación

Clasificación Binaria

Ejemplos

- Es spam o no es spam
- Transacciones fraudulentas
- Tumor benigno o maligno
- La persona pagará su préstamo de vuelta o no
- Los usuarios darán click en la publicidad
- Nombre de la persona en tu foto en Facebook

Clasificación

- Predice un label Y discreto.

- Se asignan nuevas observaciones en una clase a la cual probablemente corresponden.
  - Nos basamos en un modelo construido durante entrenamiento

Supervised Learning: Classification

	Observation #	Input image (X)	Label (Y)
training set	1		“dog”
	2		“cat”
	3		“dog”
	…	…	…
	N		“dog”
test set	1		???
	2		???

Clasificación binaria vs multiclasificación

- Clasificación binaria
  - Tenemos clase negativa y clase positiva. El término es ambiguo

$$y \in \{0, 1\}$$

- Multiclasificación
  - Muchas clases

$$y \in \{0, 1, 3, 4, 5\}$$

¿Cómo podemos resolver este problema con las herramientas que tenemos?

¿Cómo podemos resolver este problema con las herramientas que tenemos?

- Usamos regresión lineal como antes
- Ponemos un límite en el clasificador en 0.5
  - Si la hipótesis es mayor a 0.5, predecimos que y=1
  - Si la hipótesis es menor a 0.5, predecimos que y=0

$$h_\theta = \theta^T x$$

¿Cómo podemos resolver este problema con las herramientas que tenemos?

- Usamos regresión lineal como antes
- Ponemos un límite en el clasificador en 0.5
  - Si la hipótesis es mayor a 0.5, predecimos que y=1
  - Si la hipótesis es menor a 0.5, predecimos que y=0

$$h_\theta = \theta^T x$$

- ¿Qué pasaría si tenemos un tumor grande que no es maligno?

¿Cómo podemos resolver este problema con las herramientas que tenemos?

- Usamos regresión lineal como antes
- Ponemos un límite en el clasificador en 0.5
  - Si la hipótesis es mayor a 0.5, predecimos que y=1
  - Si la hipótesis es menor a 0.5, predecimos que y=0

$$h_\theta = \theta^T x$$

- ¿Qué pasaría si tenemos un tumor grande que no es maligno?
  - Se mueve toda la línea.

Regresión Logística

- Prestamos
- Si somos un banco, ¿damos un préstamo o no?

- Regresión logística calcula la probabilidad de que la aplicación sea fraude.

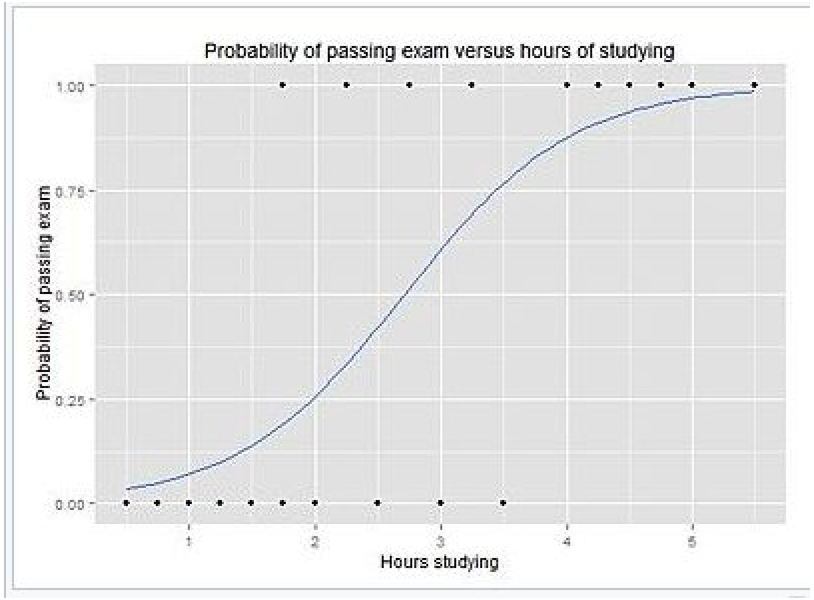
Regresión Logística

$$0 \leq h_\theta(x) \leq 1$$

$$h_\theta(x) = g(\Theta^T x)$$

Función sigmoidal o logística

$$S(x) = \frac{1}{1+e^{-x}}.$$



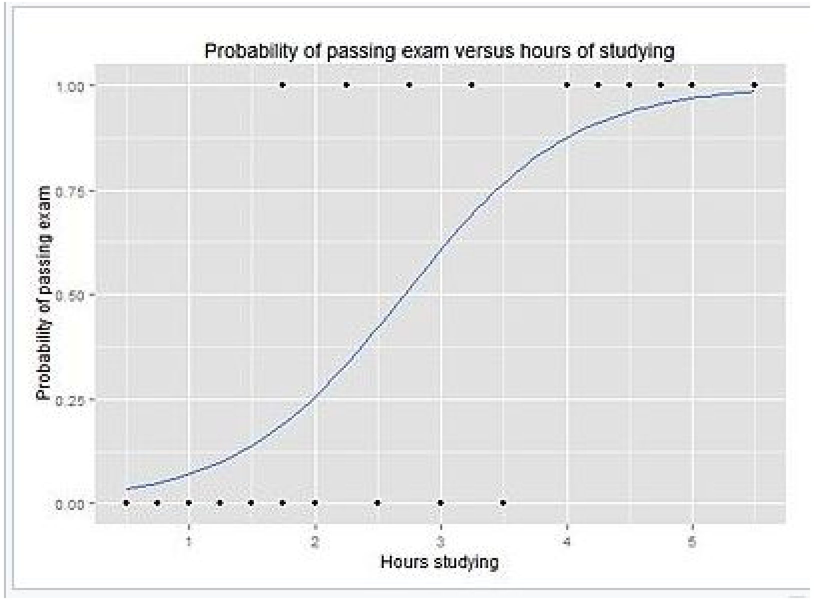
Regresión Logística

$$0 \leq h_\theta(x) \leq 1$$

$$h_\theta(x) = g(\Theta^T x)$$

$$S(x) = \frac{1}{1 + e^{-x}}.$$

$$h_\theta(x) = \frac{1}{1 + e^{-\Theta^T x}}$$



Regresión Logística

- Objetivo: Encontrar los mejores parámetros Θ
- $h_\theta(x)$ = probabilidad estimada que y = 1 dado el input x
  - $P(y = 1 \mid x; \theta)$
  - $P(y = 0 \mid x; \theta) = 1 - P(y = 1 \mid x; \theta)$

$$h_\theta(x) = g(\Theta^T x) = P(y = 1|X; \Theta)$$

Regresión Logística $h_\theta(x) = g(\Theta^T x) = P(y = 1|X; \Theta)$

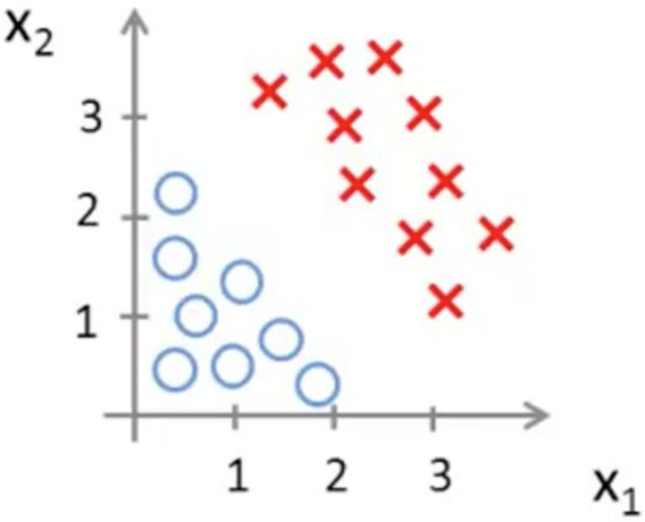
- ¿En qué momento la función g(z) es igual o mayor a 0.5?

Regresión Logística $h_\theta(x) = g(\Theta^T x) = P(y=1|X;\Theta)$

- ¿En qué momento la función g(z) es igual o mayor a 0.5?
- Cuando z es mayor o igual a 0.
- O: $\Theta^T x \geq 0$
- Cuando esto pasa predecimos que y = 1

Ejemplo

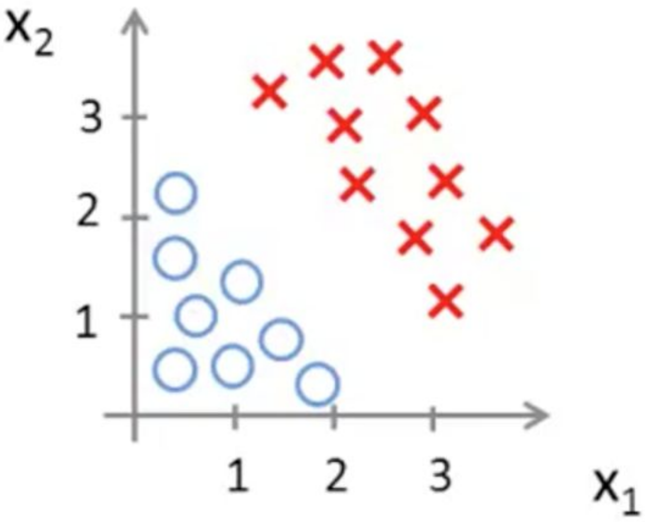
- $h_\theta(x) = g(-3 + x_1 + x_2)$
- ¿Cuándo predecimos que y = 1?



Ejemplo - Límite de decisión

- $h_\theta(x) = g(-3 + x_1 + x_2)$
- ¿Cuándo predecimos que y = 1?

- Esto se llama límite de decisión
  - Permite construir un límite sin tener dat



Límites no lineares

- Clasificación con círculo
- Como en regresión polinomial, podemos agregar términos de mayor grado.

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

Límites no lineares

- Como en regresión polinomial, podemos agregar términos de mayor grado.

$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_3 x_1^2 + \theta_4 x_2^2)$$

- Asumimos el siguiente valor para los parámetros

$$\Theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

$$-1 + (x_1)^2 + (x_2)^2 \geq 0$$

$$(x_1)^2 + (x_2)^2 \geq 1$$

Modelo de Regresión Logística

Regresión Logística

- Tenemos un training set con m ejemplos

$$x_0 = 1$$

$$x = \begin{bmatrix} x_0 \\ x_1 \\ \ldots \\ x_n \end{bmatrix}$$

$$h_\theta(x) = \frac{1}{1 + e^{-\Theta^T x}}$$

Función de pérdida

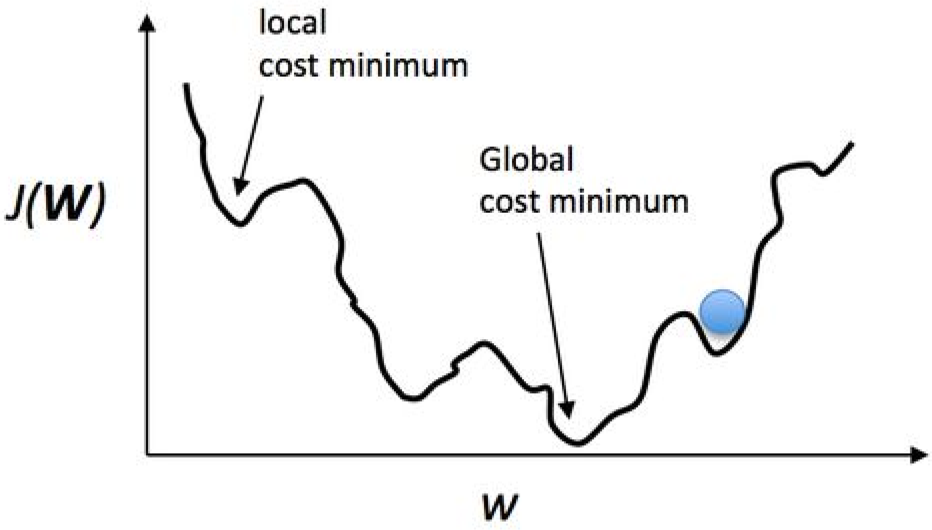
$$J(\theta_0, \theta_1, \ldots, \theta_n) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J(\Theta) = (\frac{1}{m})\sum_{i=1}^{m}\frac{1}{2}(h_\theta(x^i) - y^{(i)})^2 =$$

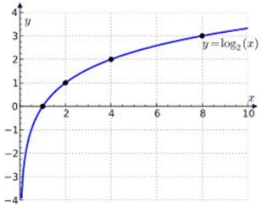
$$(\frac{1}{m})\sum_{i=1}^{m} loss(h_\theta(x^i), y^{(i)})$$

Regresión Logística

- La anterior función no era convexa por la hipótesis (no linear)
  - La función sigmoide afecta



Loss Function



Gráfica de logaritmo

- Como la anterior no funcionaba, podemos proponer una nueva que si sea convexa

$$\mathrm{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

- ¿Cómo se comporta cuando y=1?
- ¿Cómo se comporta cuando y=0?

Loss Function

$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

Cost(h_θ(x), y) = -ylog(h_θ(x)) - (1-y) log(1 - h_θ(x))

$$J(\theta) = -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))]$$

Esta fórmula se puede derivar de un concepto llamado maximum likehood estimator. Es convexa para este problema

Gradient Descent

- Ajustamos los parámetros para minimizar la pérdida utilizando Gradient Descent

Repeat $\{$

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

(simultaneously update θ_j for $j = 0, \ldots, n$)

$\}$

- Lo único que cambia es la hipótesis

Algoritmos de Optimización

- No sólo está Gradient Descent

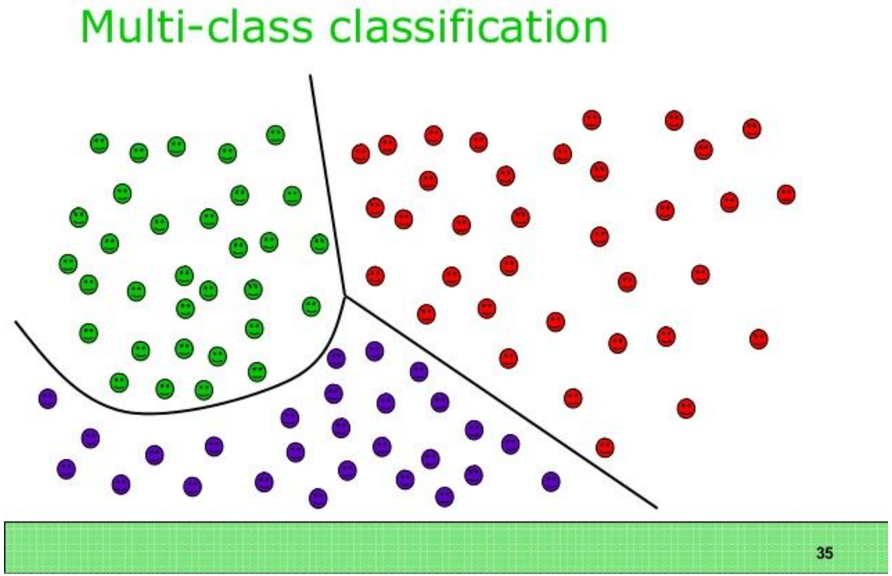
- Otros son
  - Conjugate Gradient
  - BFGG
  - L-BFGS

- Ventajas: No necesitas elegir learning rate y pueden ser más rápidos
- Desventajas: Más complejos - expertos de análisis numérico

Multiclasificación

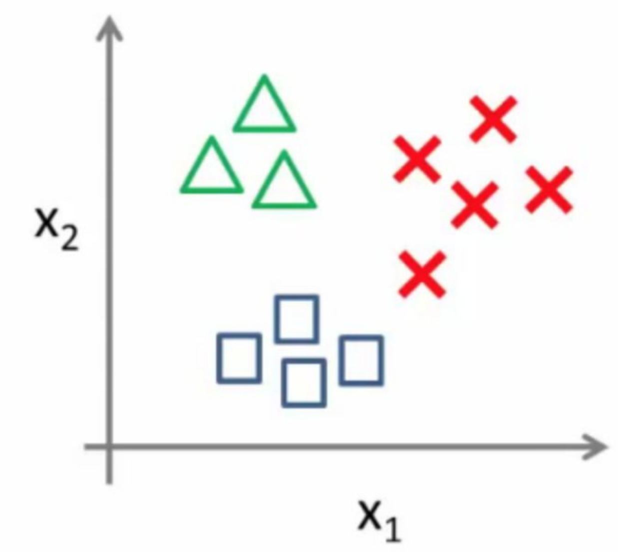
Ejemplos

- Etiquetado en correos
- Diagramas médicos
- Clasificación de clima



Clasificación uno vs todos

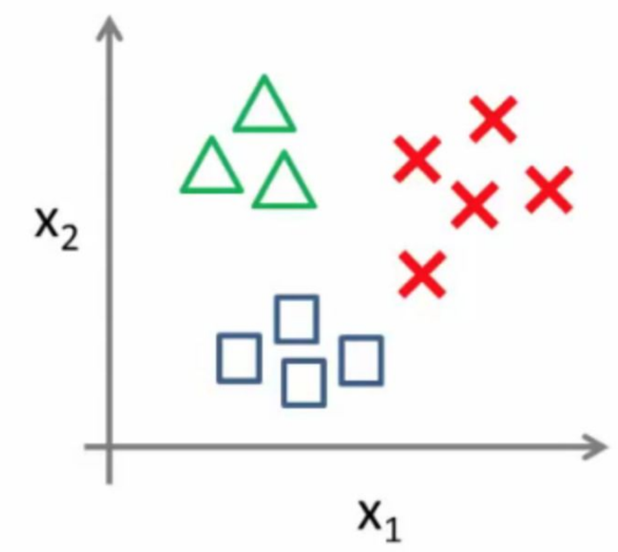
- Tenemos 3 clases



- ¿Ideas?

Clasificación uno vs todos

- Tenemos 3 clases



- ¿Ideas?
  - Podemos crear 3 clasificadores

Multiclasificación

- Entrenamos un clasificador de regresión logística para cada clase i.

- Dada una nueva entrada, elegimos la clase i que maximice la hipótesis.

Otras técnicas

- Redes neuronales
- K-nearest neighbors
- Naive bayes
- Árboles de decisiones
- SVMs