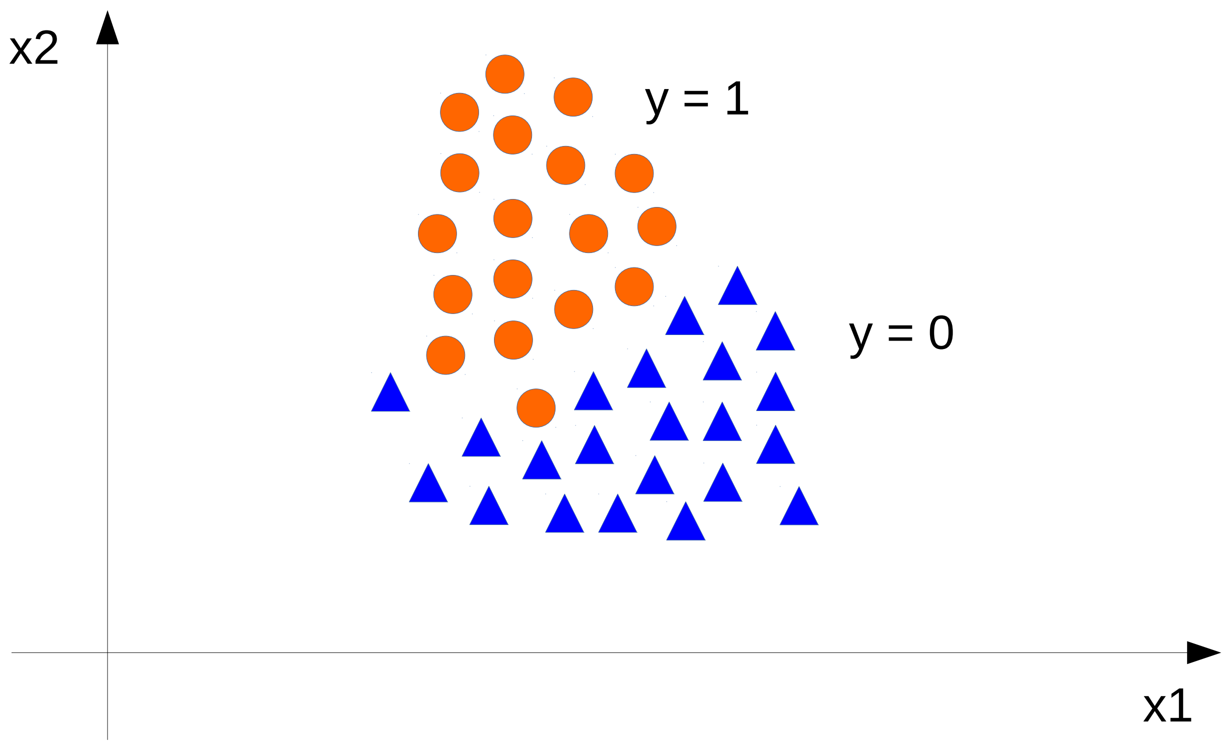
Estadística [continuación]

Santander, 2019-2020

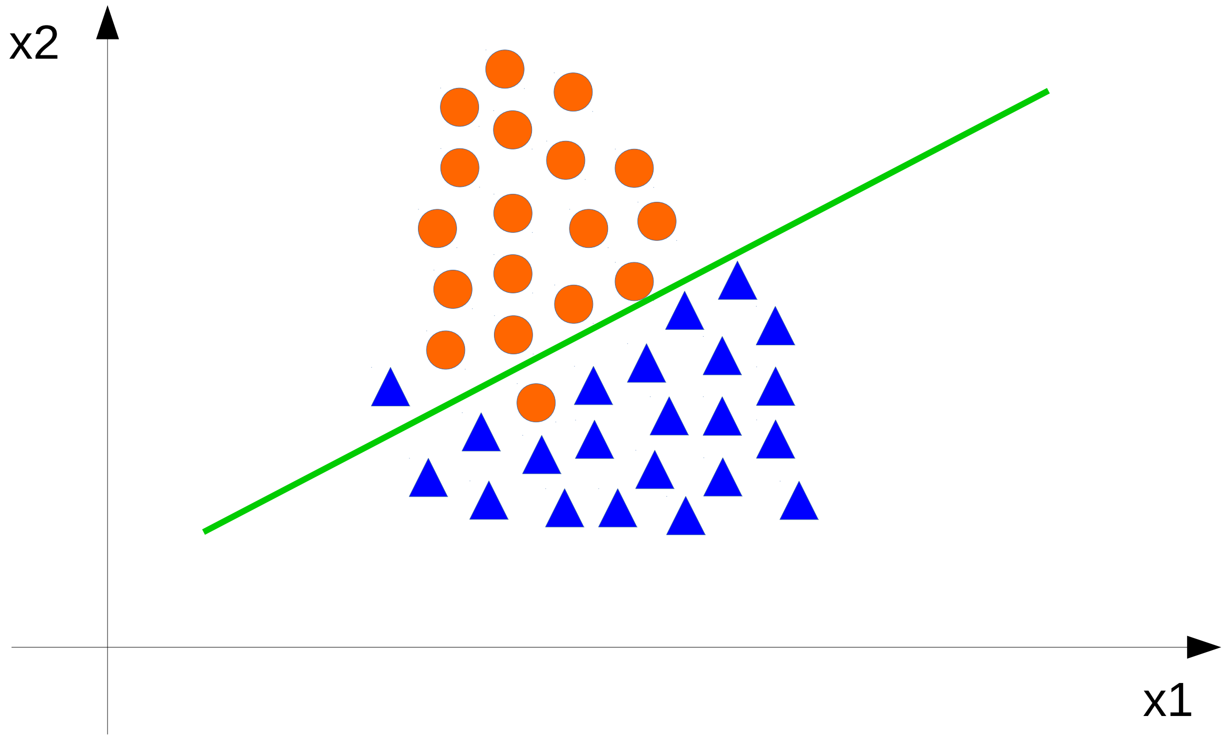
Introducción a los problemas de clasificación (I)

- A veces estamos interesados en una variable dependiente que toma unos pocos valores discretos.
- Se trata de los llamados problemas de clasificación, en los que los valores de “y” son las categorías.
- Consideremos por ejemplo un grupo de medidas con dos variables independientes y dos categorías



Introducción a los problemas de clasificación (II)

- La clasificación consiste en determinar a qué categoría pertenece una medida con valores (x1, x2)
- Puede entenderse también como la determinación de la frontera entre las dos categorías.
- Se trata de una técnica de carácter muy fundamental y que da lugar a múltiples aplicaciones.

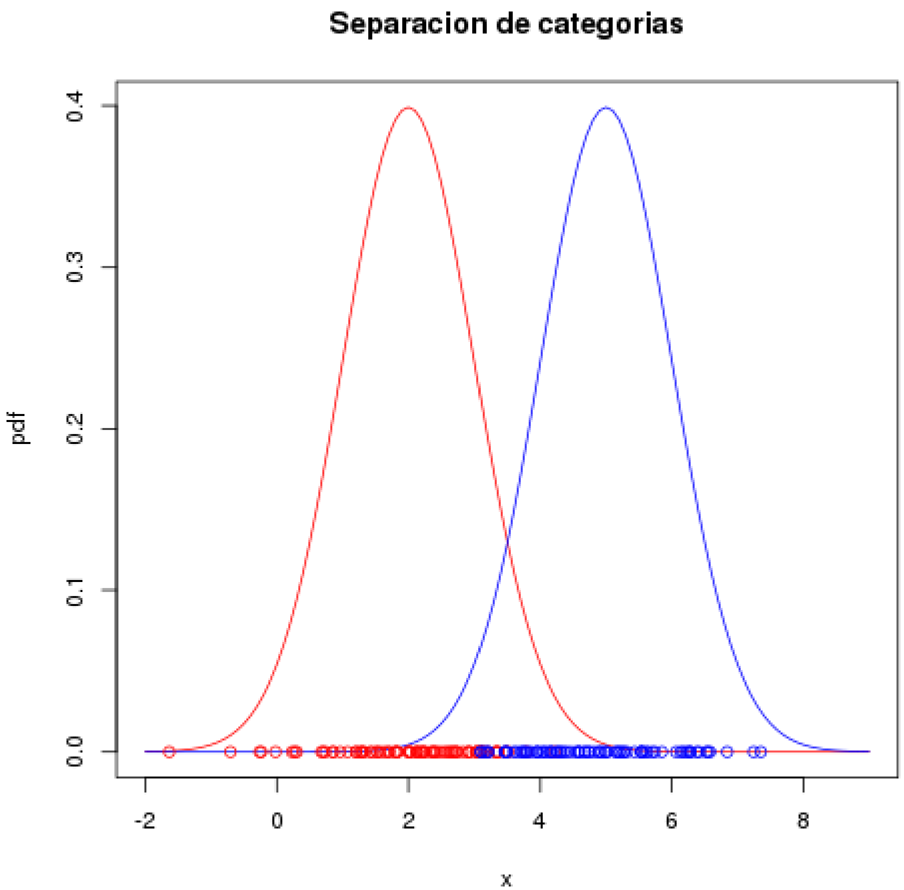


Clasificación en 1 dimensión

- Supongamos que disponemos de dos pdfs, **pdf(y=1 | x)** y **pdf(y=0 | x)**, con x un número real.
- Vamos a suponer que ambas pdfs son gaussianas con medias = 2 y 5, y sigma = 1

$$p(y=0|x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}(x-2)^2}$$

$$p(y=1|x)=\frac{1}{\sqrt{2\pi}}e^{\frac{-1}{2}(x-5)^2}$$



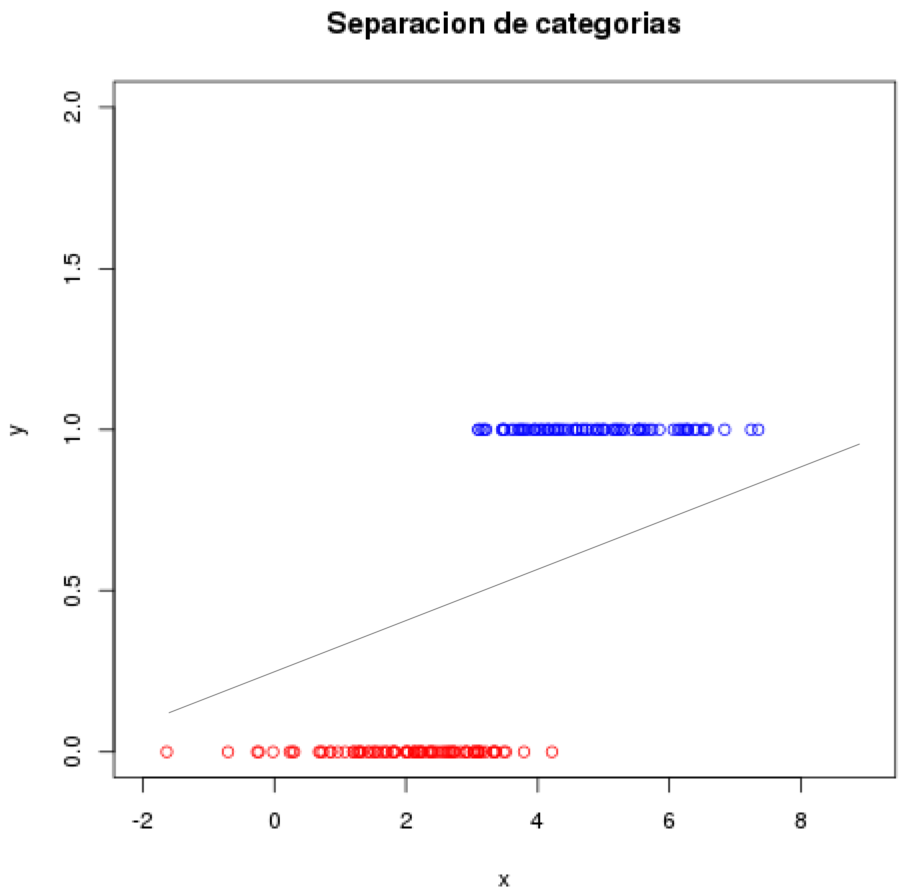
Rojo → y = 0
Azul → y = 1

Clasificación en 1 dimensión (II)

- Podemos pintar las distribuciones en función de los valores **y = 0** e **y = 1.**
- Podríamos intentar modelar estos valores de “y” con algún tipo de aproximación lineal: y = a + bx

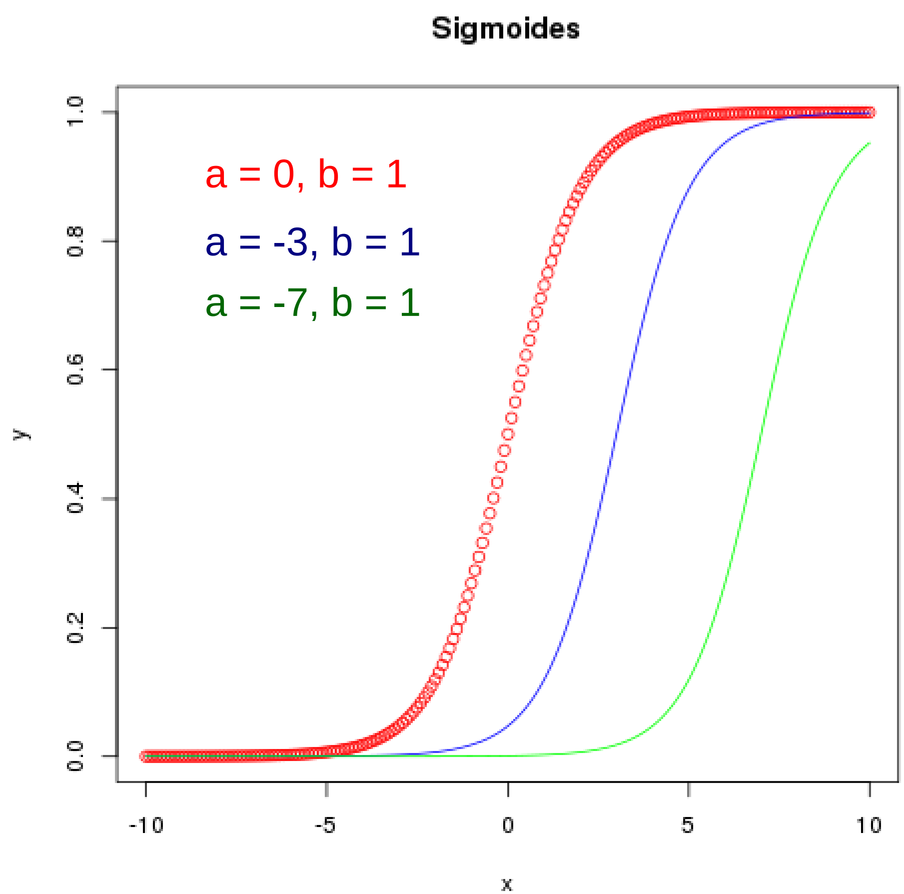
Sin embargo y = a + bx no está acotada y puede tomar valores mucho mayores que 1 o menores que 0

El ajuste tampoco es bueno en cualquier caso.



La función sigmoide

- ¿Podemos encontrar algún cambio de variable que nos resulte interesante para ajustar lo anterior?
- Estudiemos la función conocida como sigmoide



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

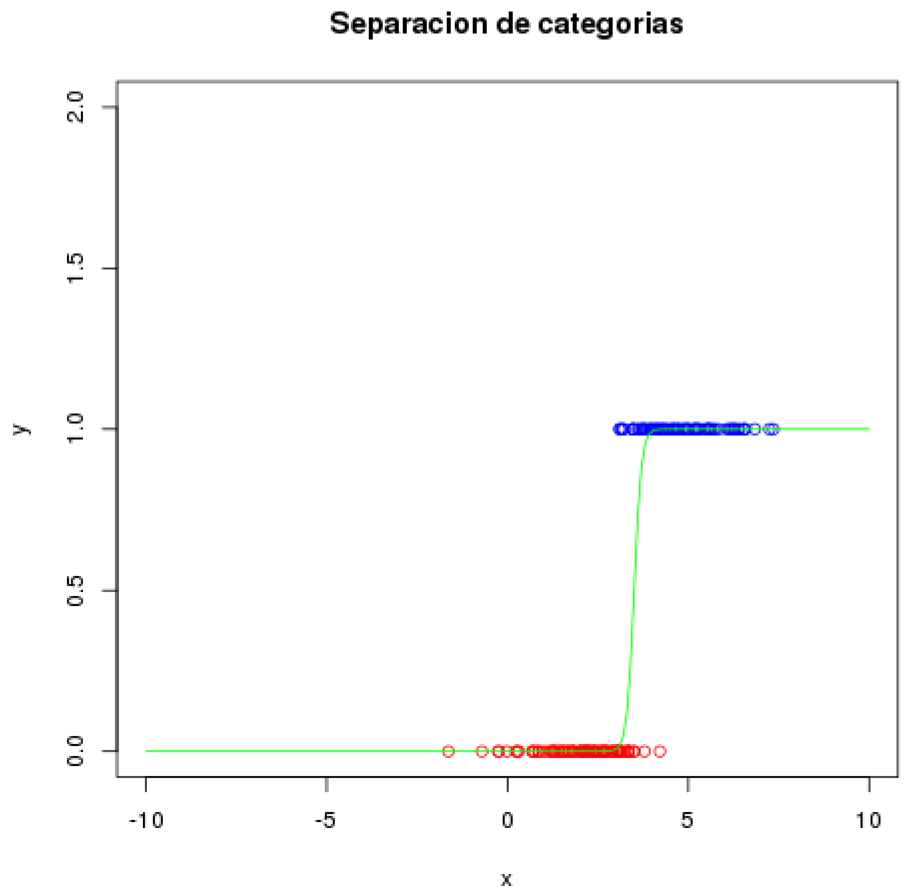
$$\lim_{x\to\infty}\sigma(x)=1$$

$$\lim_{x\to-\infty}\sigma(x)=0$$

Acotada entre 0 y 1

Ajustando a una función sigmoide

- Con el planteamiento anterior podríamos preguntarnos cuáles son las “a” y “b” que mejor “ajustan”.
- En el ejemplo de abajo he elegido a = -35 y b = 10 para obtener la sigmoide representada en verde.



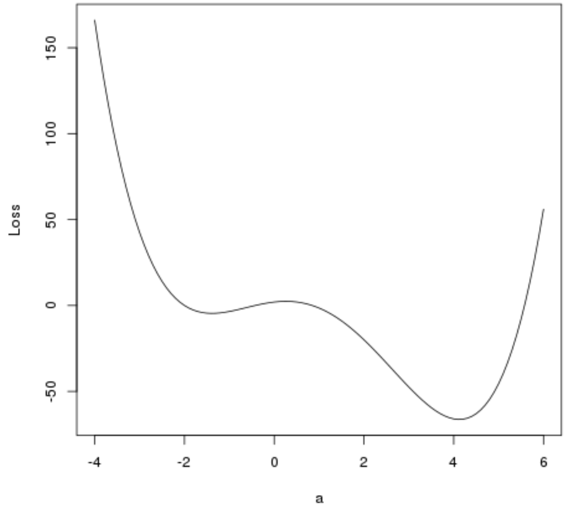
$$\sigma(x)=\frac{1}{1+e^{-(a+bx)}}$$

Ajustando a una función sigmoide

- Para ver cuales son los mejores "a" y "b" tendríamos que definir una función de coste.
- ¿Cuál es la más obvia? → Podemos probar con la misma que utilizamos para el caso lineal.

$$Loss=\sum_{i=0}^{N}\left(y^{(i)}-f\left(x^{(i)}\right)\right)^2=\sum_{j=0}^{N}\left(y^{(i)}-\frac{1}{1+e^{-\left(a+bx^{(i)}\right)}}\right)^2$$

- Sin embargo esa función de coste resulta problemática porque tiene muchos mínimos y máximos.



Función de coste para regresión logística

- Para evitar ese problema se utiliza la siguiente función de coste

$$\sigma(a+bx)=\frac{1}{1+e^{-(a+bx)}}$$

$$Loss=\frac{-1}{N}\sum_{i=0}^{N} y^{(i)}\log\left(\sigma\left(a+bx^{(i)}\right)\right)+\left(1-y^{(i)}\right)\log\left(1-\sigma\left(a+bx^{(i)}\right)\right)$$

- Analicemos cómo actua esta función de perdida para una medida “j” dentro del sumatorio:
  - Si $y^{(j)}$ es de categoría 1 entonces la segunda parte del sumatorio se anula
  - La contribución a la función de coste es por tanto -log() que sera siempre positivo porque σ C [0,1]
  - Si σ se aproxima a 1 el log va a ser casi 0 y por tanto la contribución al coste muy pequeña
  - Si σ se aproxima a 0 el log va a ser muy alto y por tanto la contribución al coste muy grande
  - Si $y^{(j)}$ es de categoría 0 entonces la primera parte del sumatorio se anula
  - La contribución a la función de coste es por tanto -log(1-σ) que sera tambien siempre positivo
  - Si σ se aproxima a 0 el log va a ser casi 0 y por tanto la contribución al coste muy pequeña
  - Si σ se aproxima a 1 el log va a ser muy alto y por tanto la contribución al coste muy grande
  - El coste crece cuando $y^{(j)}$ es 1 y los parámetros a y b hacen que sigma sea 0 y viceversa.

Minimización de la función de coste para regresión logística

- Para minimiar la función de coste necesitamos calcular el gradiente de la función de Loss
- Comencemos con algunas observaciones y derivadas previas:

$$\sigma(Z)=\frac{1}{1+e^{-Z}} \Rightarrow \frac{\partial \sigma(Z)}{\partial Z}=\sigma(Z)(1-\sigma(Z))$$

$$L=y\log(\sigma)+(1-y)\log(1-\sigma) \Rightarrow \frac{\partial L}{\partial \sigma}=\frac{y}{\sigma}-\frac{(1-y)}{(1-\sigma)}$$

$$\frac{\partial L}{\partial Z}=\frac{\partial L}{\partial \sigma}\frac{\partial \sigma}{\partial Z}=\left(\frac{y}{\sigma}-\frac{(1-y)}{(1-\sigma)}\right)\sigma(1-\sigma)=y(1-\sigma)-(1-y)\sigma=(\sigma-y)$$

- Supongamos que ahora introducimos también z = a + b x

$$\frac{\partial Loss}{\partial a}=\frac{\partial Loss}{\partial \sigma}\frac{\partial \sigma}{\partial Z}\frac{\partial Z}{\partial a}=\frac{-1}{N}\sum_{j=1}^{N}\left(\sigma(a+bx^{(j)})-y^{(j)}\right)(-1)$$

$$\frac{\partial Loss}{\partial b}=\frac{\partial Loss}{\partial \sigma}\frac{\partial \sigma}{\partial Z}\frac{\partial Z}{\partial b}=\frac{-1}{N}\sum_{j=1}^{N}\left(\sigma(a+bx^{(j)})-y^{(j)}\right)(-x^{(j)})$$

Generalización a M dimensiones

- Todo el desarrollo anterior lo hemos llevado a cabo para el caso de una sola “feature”
- Sin embargo, siempre podemos hacer que nuestra coordenada Z dependa linealmente de más de 1

$$Z^{(j)} = \alpha_0 + \alpha_1 x_1^{(j)} + \ldots + \alpha_M x_M^{(j)}$$

- Si agrupamos las alphas y las x en vectores (incluyendo un 1 en el vector) podemos escribir

$$Z^{(j)} = \alpha^T x^{(j)}$$

- Y con esta definición tenemos que podemos escribir el gradiente para muchas dimensiones como

$$\nabla Loss = \frac{1}{N} \sum_{j=1}^{N} \left(\sigma(\alpha^T x^{(j)}) - y^{(j)} \right) x^{(j)}$$

- Para encontrar el mímino vamos a utilizar un “gradient descent”.

Ejercicio 4

1) Crea una función que genere dos muestras que se distribuyen según dos gaussianas distintas. La función recibirá como valores de entrada: el número N de puntos a generar para cada categoría, y mu1, sigma1, mu2, sigma2 que son los correspondientes parámetros de las dos gaussianas. Como output devolverá un valor con longitud 2N que contenga la muestra x generada, y otro vector de longitud 2N que contenga 0 o 1 en función de la categoría asociada a ese elemento.
2) Crea una función que calcule el valor de la sigmoide para un valor de entrada Z.
3) Crea una función que calcule el valor de la función de Loss y que reciba como entrada "x" e "y" y los parámetros del modelo que vamos a asumir: $z = a + b\,x$ (es decir, a y b).
4) Crea una función que devuelva el gradiente de la función de Loss y que reciba como entrada "x" e "y" y los parámetros (a, b) del modelo que vamos a asumir.
5) Generar un par de vectores "x", "y" con N = 100, mu1 = 2, mu2 = 6, sigma1 = 1 y sigma2 = 1.
6) Calcular la función de coste y el gradiente para (a = 0, b = 0). Actualizar los valores de a y b de manera que (a, b)_nuevos = (a, b)_viejos + lambda * gradiente. Repite 3 o 4 cuatro veces y observa los valores de la función de coste. Intenta encontrar el mínimo aproximadadamente. Interpreta los resultados.