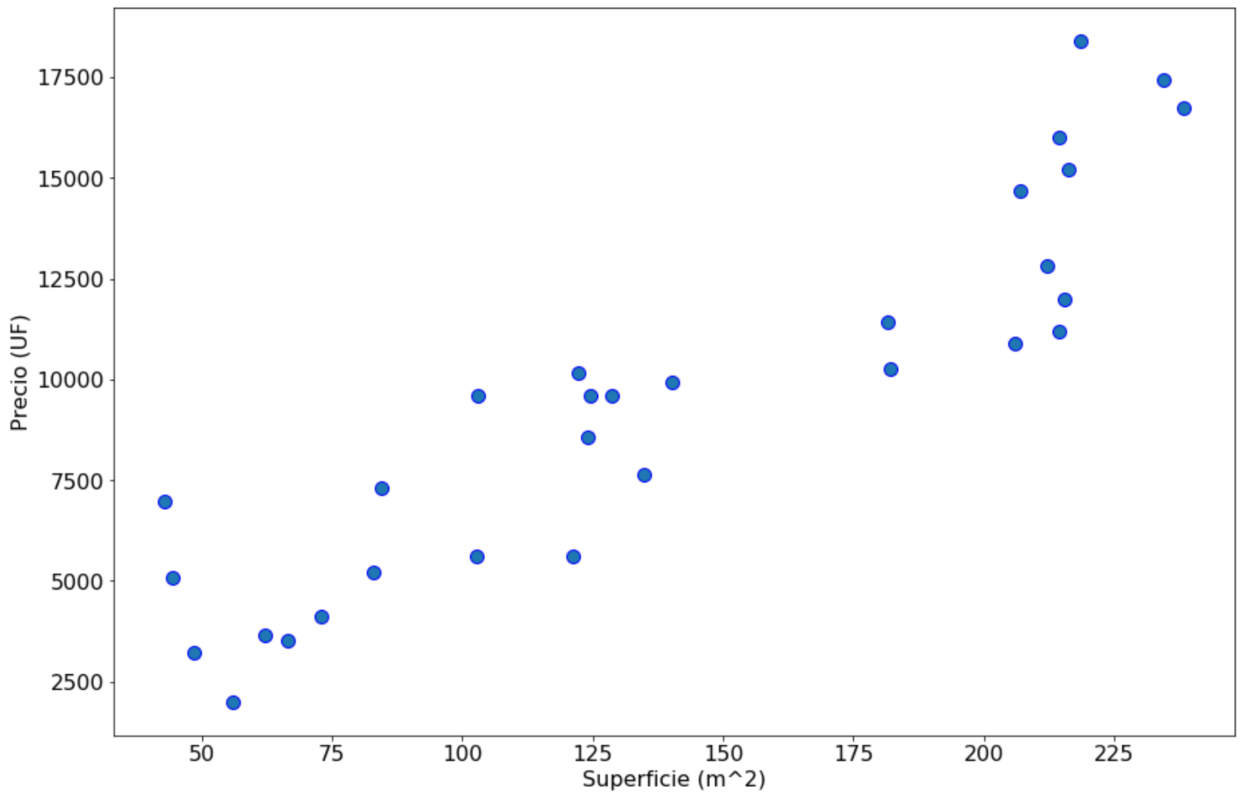
Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2613 – Inteligencia Artificial

Análisis de regresión

Profesor: Hans Löbel

Consideremos el siguiente conjunto de puntos, que presentan el precio de un departamento en función de su superficie

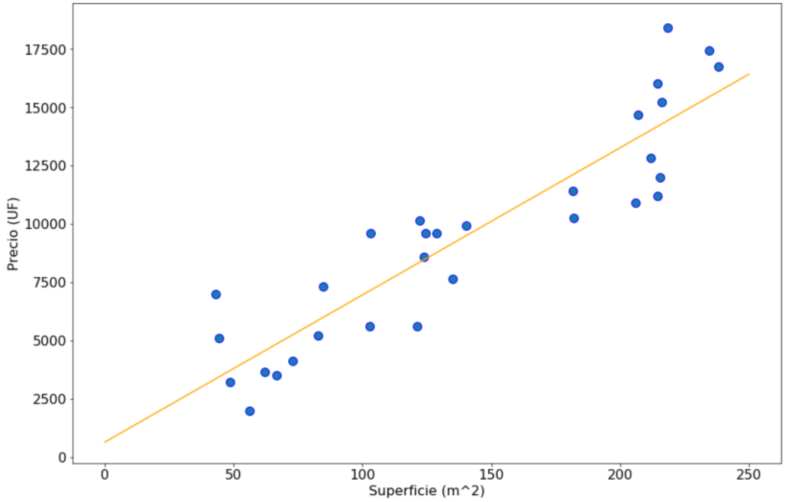


¿Cómo podemos estimar el precio de un departamento de $150m^2$?

¿Y el de uno de $30m^2$?

Análisis de regresión es una técnica simple y poderosa

- Técnica clásica de estimación de funciones.
- Altamente flexible e interpretable.
- Permite realizar regresión y clasificación.
- En la actualidad, su rendimiento puede ser fácilmente superado por técnicas más modernas.
- Nos centraremos en dos tipos de regresión: lineal y logística.



Regresión lineal permite estimar funciones continuas de manera supervisada

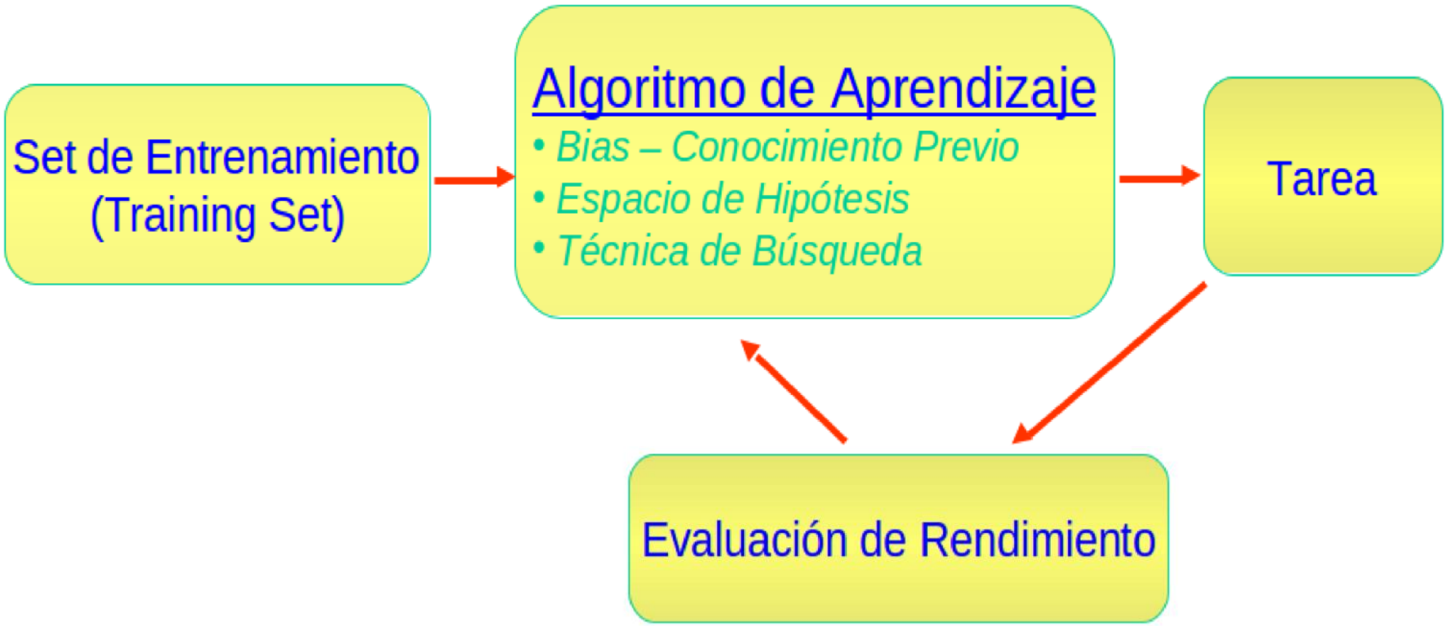
- El espacio de hipótesis de una regresión lineal es el conjunto de las funciones lineales que mapean los vectores $x^{(i)}$ del dominio, a escalares $y^{(i)}$.

- Al parametrizar este espacio de hipótesis a través del vector de parámetros (pesos) θ, obtenemos la siguiente expresión:

$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x$$

¿Cómo podemos obtener los valores de los pesos θ ?

Recordemos que Machine Learning se centra en algoritmos que mejoran su rendimiento en una tarea, a través de la experiencia



Buscamos la solución más adecuada en el espacio de hipótesis, usando conocimiento previo y datos de entrenamiento para guiar la búsqueda.

Regresión lineal permite estimar funciones continuas de manera supervisada

- Si tenemos suficientes datos, pares $(x^{(i)}, y^{(i)})$, es posible construir una función que nos indique cuán buena es en promedio la estimación.

- Definimos entonces la función de pérdida (o costo, o error) de la regresión de la siguiente manera:

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

- El valor para los pesos que minimice esta pérdida, entregará la mejor estimación de la función original (¿por qué?, ¿seguro?).

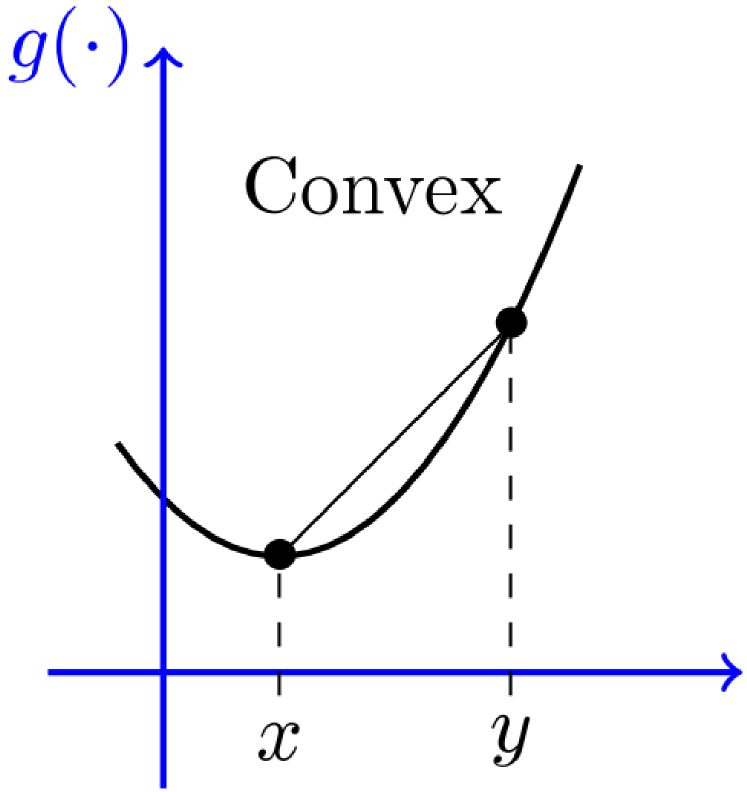
$$Err(x) = E\left[(Y - \hat{f}(x))^2\right]$$

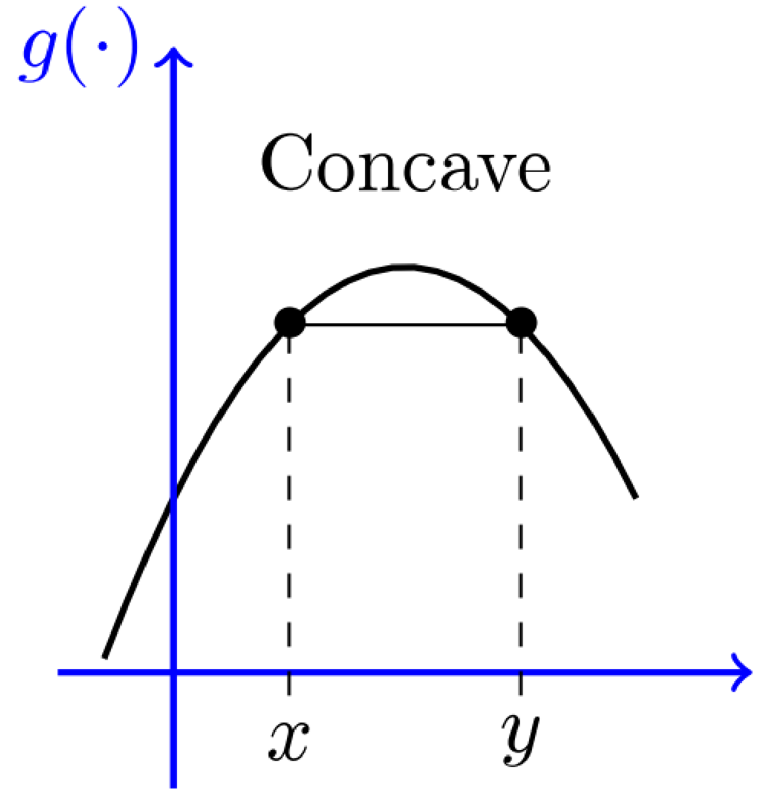
$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma_e^2$$

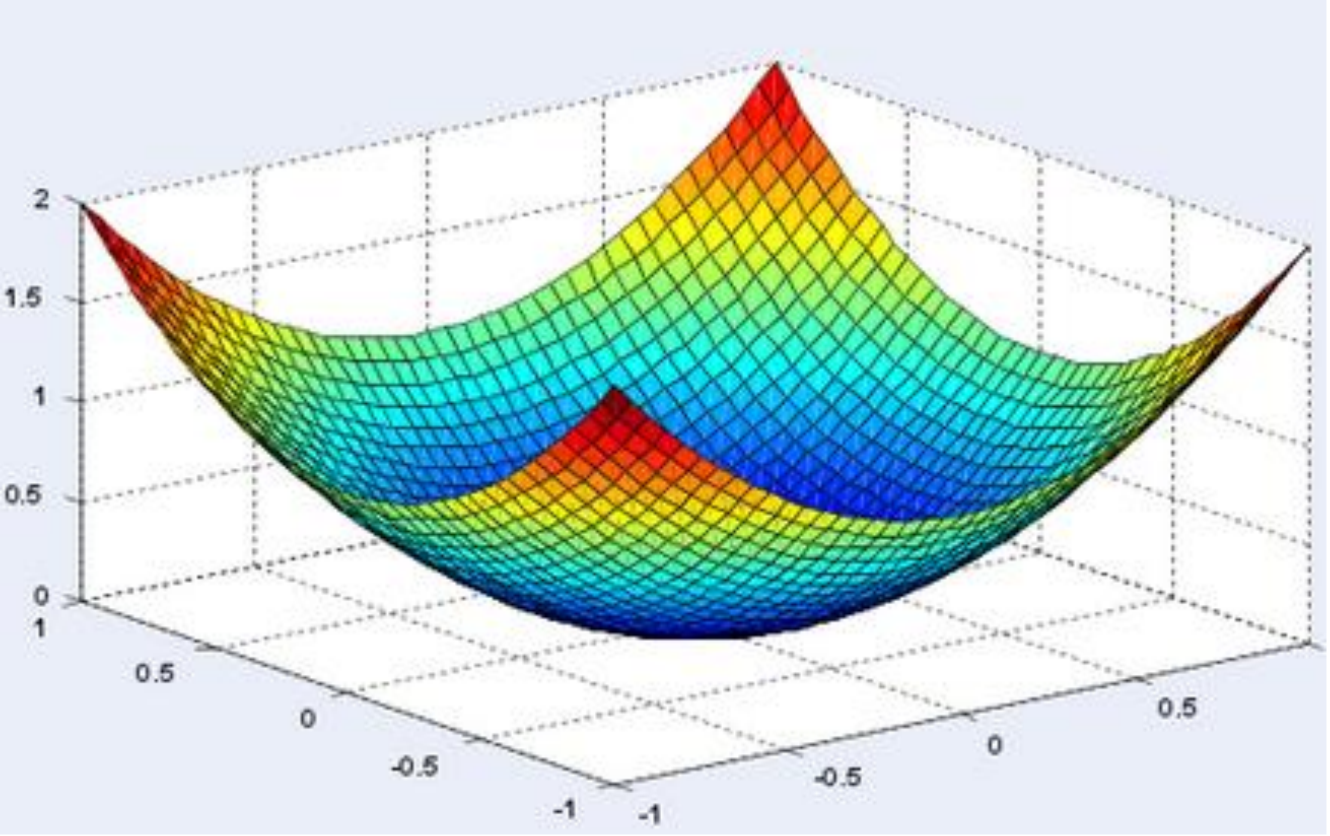
$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$

Usemos un poco de optimización convexa clásica para solucionar el problema

- Dado que la pérdida es convexa (¿por qué?), podemos usar un algoritmo de descenso desde cualquier punto de inicio para encontrar el óptimo (¿ah?, ¿qué?).


$g(\cdot)$
Convex
x
y


$g(\cdot)$
Concave
x
y


2
1.5
1
0.5
0
1
0.5
0
-0.5
-1
-1
-0.5
0
0.5

Usemos un poco de optimización convexa clásica para solucionar el problema

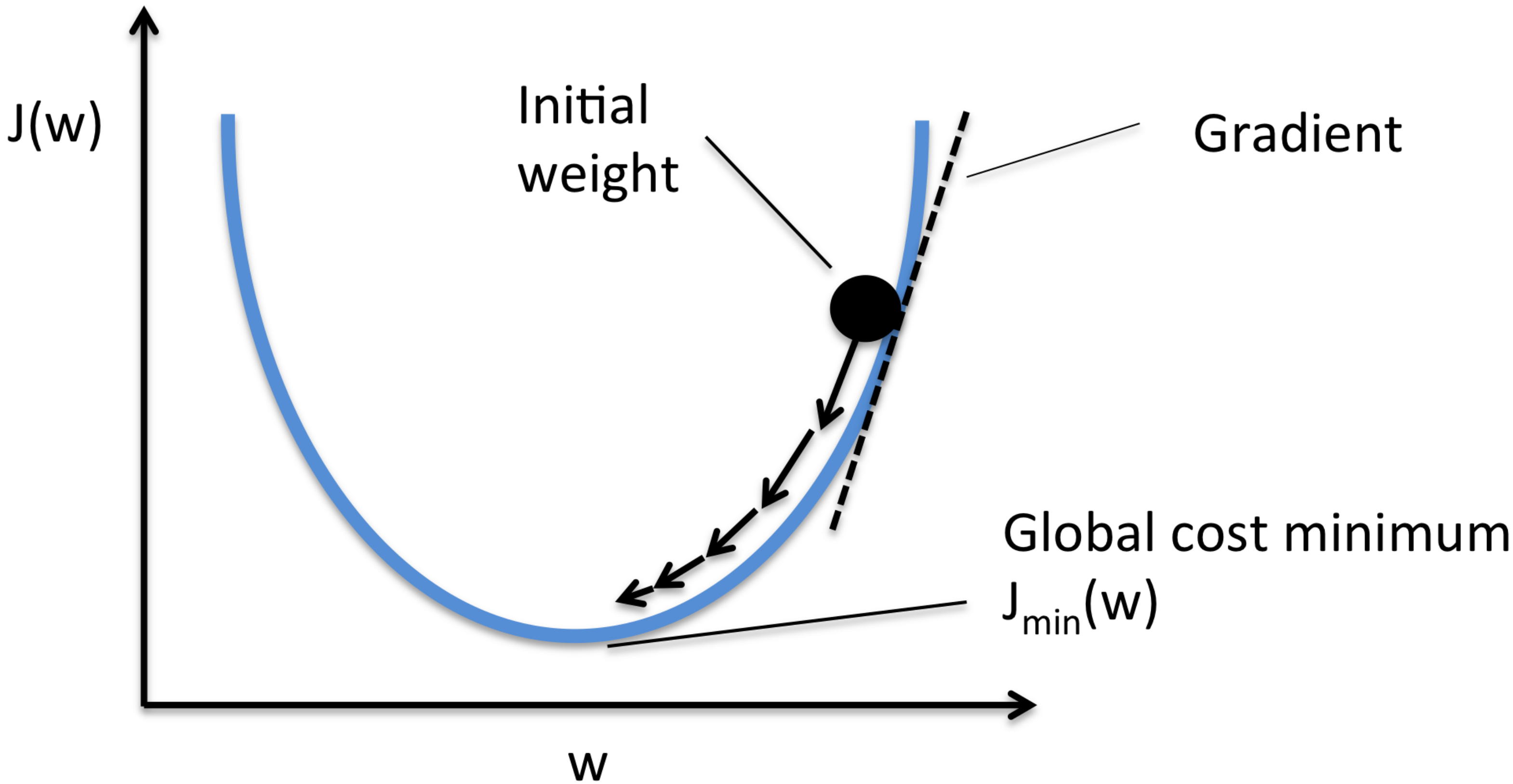
- Dado que la pérdida es convexa (sin duda), podemos usar un algoritmo de descenso desde cualquier punto de inicio para encontrar el óptimo (obvio).

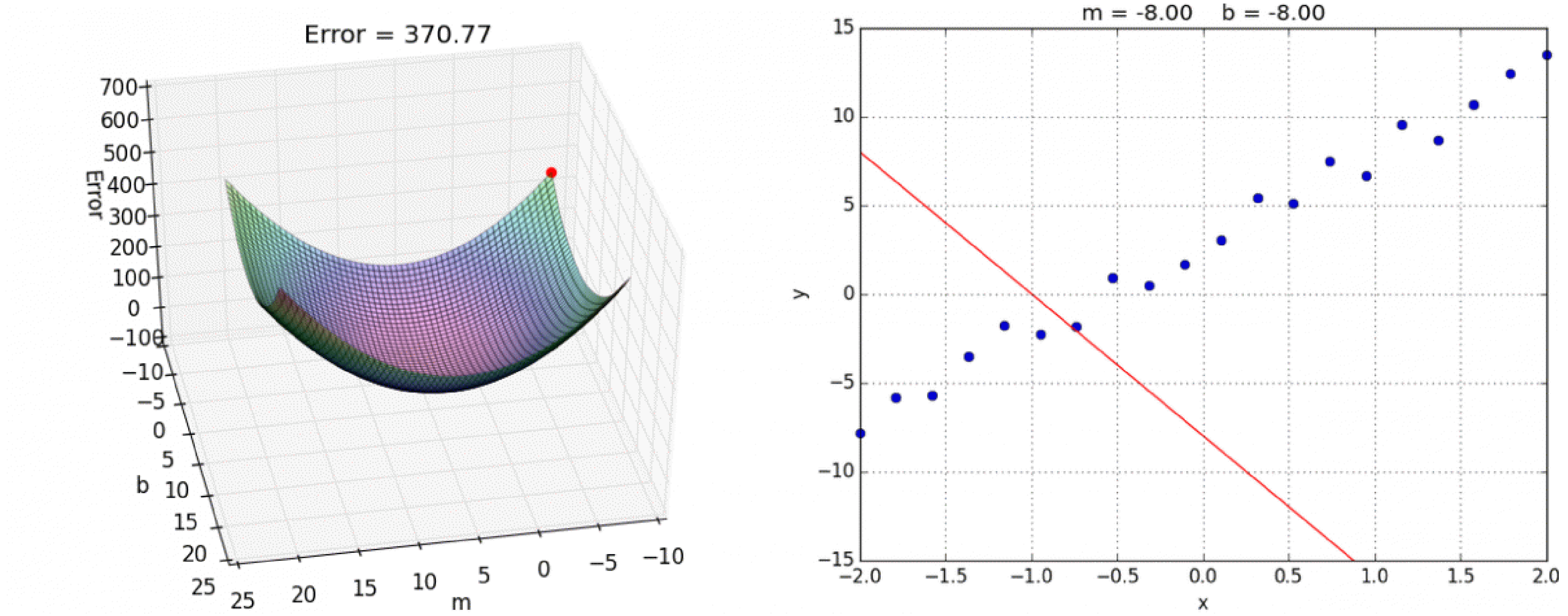
- En particular, podemos utilizar la dirección opuesta a la entregada por el gradiente de la pérdida (máximo decrecimiento de la función).

- Si avanzamos en esta dirección usando pequeños pasos, tenemos un algoritmo iterativo para encontrar el óptimo:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

- Donde α es conocido como tasa de aprendizaje o *learning rate*.


J(w)
Initial
weight
Gradient
Global cost minimum
$J_{min}(w)$
w


Error = 370.77
Error
m
b
m = -8.00   b = -8.00
x
y

Falta todavía la parte más entretenida, calcular la derivada 😒

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \left(h_\theta(x) - y\right)^2 \\
&= 2 \cdot \frac{1}{2} \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y\right) \\
&= \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{n} \theta_i x_i - y \right) \\
&= \left(h_\theta(x) - y\right) x_j
\end{aligned}
$$

Algoritmo final es simple e intuitivo

- Podemos finalmente darle una forma a nuestro algoritmo de descenso basado en mínimos cuadrados:

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

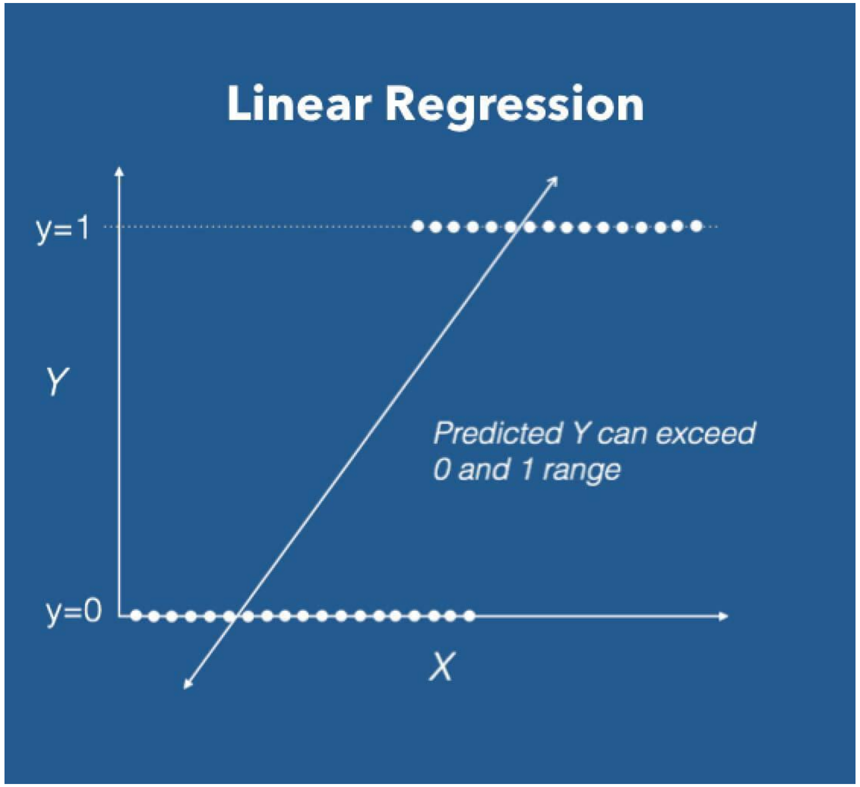
- Una propiedad interesante de este algoritmo, es que la actualización de los parámetros es proporcional al error (mientras más cerca del óptimo, menos se corrige).

- Otra propiedad derivada de esto, es que los ejemplos con que ya están bien estimados, no colaboran en la pérdida.

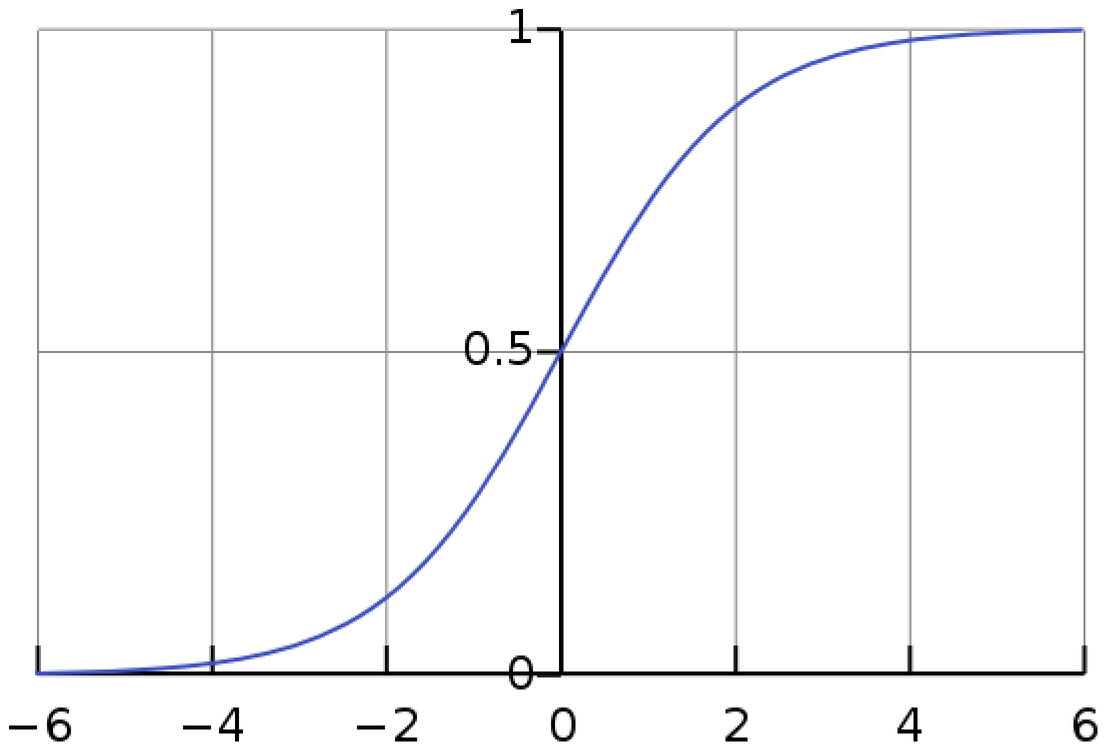
- ¿Qué podría pasar con este algoritmo si la pérdida no es convexa?

Veamos ahora como resolver un problema de clasificación con análisis de regresión

- Un problema de clasificación difiere de una regresión en que la salida de este puede tomar ahora sólo un pequeña cantidad de valores discretos.
- En particular, un problema de clasificación de alto interés es la clasificación binaria (1 = clase positiva, 0 = clase negativa).
- ¿Cómo funcionaría una regresión lineal para resolver un problema de clasificación?


Linear Regression
y=1
Y
Predicted Y can exceed
0 and 1 range
y=0
X

La función logística permite modelar el cambio en la probabilidad de ocurrencia de un evento



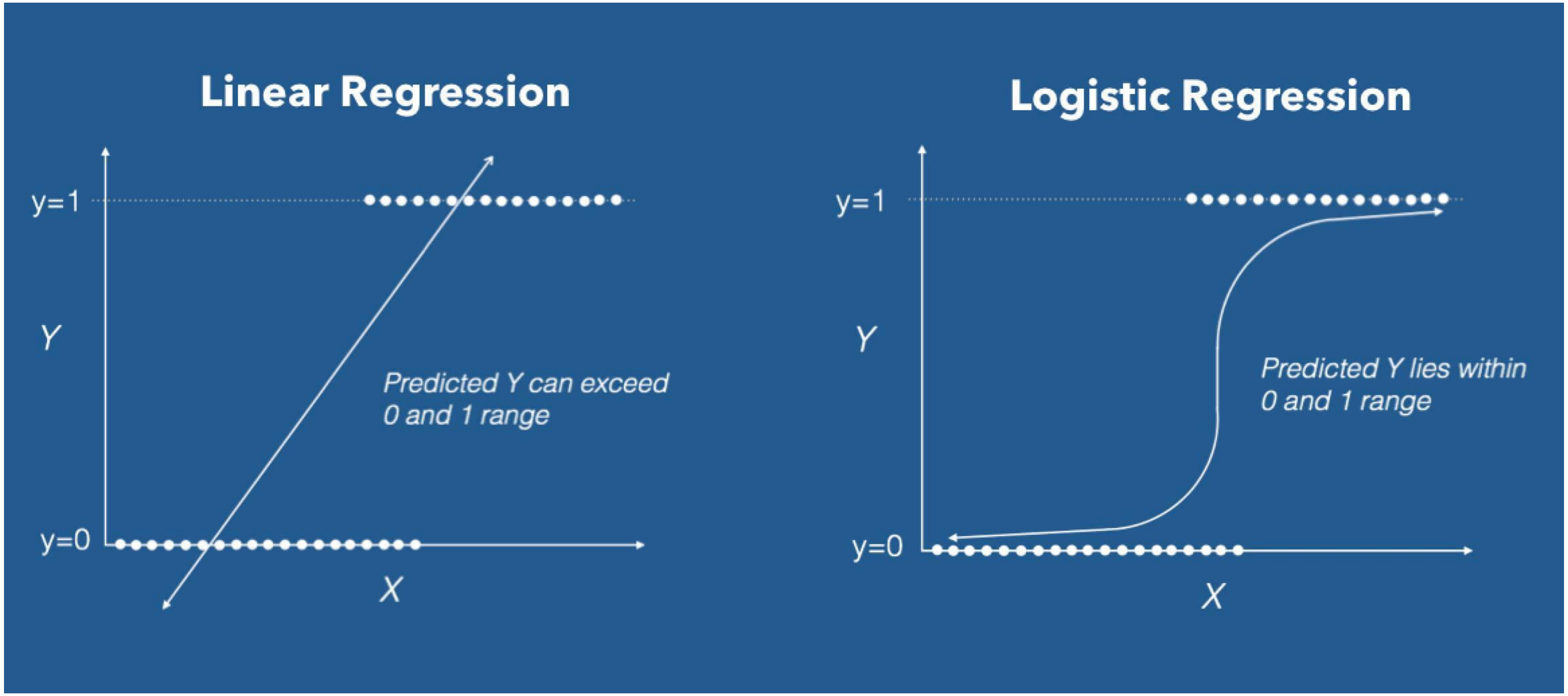
$$g(z) = \frac{1}{1 + e^{-z}}$$

La regresión logística permite resolver los problemas de la regresión lineal al clasificar

- La principal diferencia entre ambas regresiones, es que la logística utiliza un espacio de hipótesis distinto. Más específicamente, parametrizamos el espacio de hipótesis usando la función lógística de la siguiente manera:

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

¿A qué corresponde ahora el espacio de hipótesis?



Linear Regression
y=1
Y
Predicted Y can exceed
0 and 1 range
y=0
X
Logistic Regression
y=1
Y
Predicted Y lies within
0 and 1 range
y=0
X

La regresión logística permite resolver los problemas de la regresión lineal al clasificar

- Una propiedad interesante de la función logística, es que su derivada puede escribirse como una función de ella misma:

$$
\begin{aligned}
g'(z) &= \frac{d}{dz}\frac{1}{1+e^{-z}} \\
&= \frac{1}{(1+e^{-z})^2}\left(e^{-z}\right) \\
&= \frac{1}{(1+e^{-z})}\cdot\left(1-\frac{1}{(1+e^{-z})}\right) \\
&= g(z)(1-g(z)).
\end{aligned}
$$

Utilicemos ahora un enfoque probabilístico para obtener los pesos óptimos

- Definamos inicialmente las probabilidades de cada una de las dos clases:

$$
\begin{aligned}
P(y=1 \mid x;\theta) &= h_\theta(x) \\
P(y=0 \mid x;\theta) &= 1-h_\theta(x)
\end{aligned}
$$

- Esto puede escribirse de manera más conveniente como:

$$
p(y \mid x;\theta) = (h_\theta(x))^y\,(1-h_\theta(x))^{1-y}
$$

Utilicemos ahora un enfoque probabilístico para obtener los pesos óptimos

- Si asumimos que los puntos de la muestra fueron generados de manera independiente, podemos calcular la verosimilitud (*likelihood*) de los parámetros:

$$\begin{aligned} L(\theta) &= p(\vec{y} \mid X; \theta) \\ &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) \\ &= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}} \end{aligned}$$

- Y como siempre, es más sencillo si tomamos la log-verosimilitud (*log likelihood*):

$$\begin{aligned} \ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \end{aligned}$$

Utilicemos ahora un enfoque probabilístico para obtener los pesos óptimos

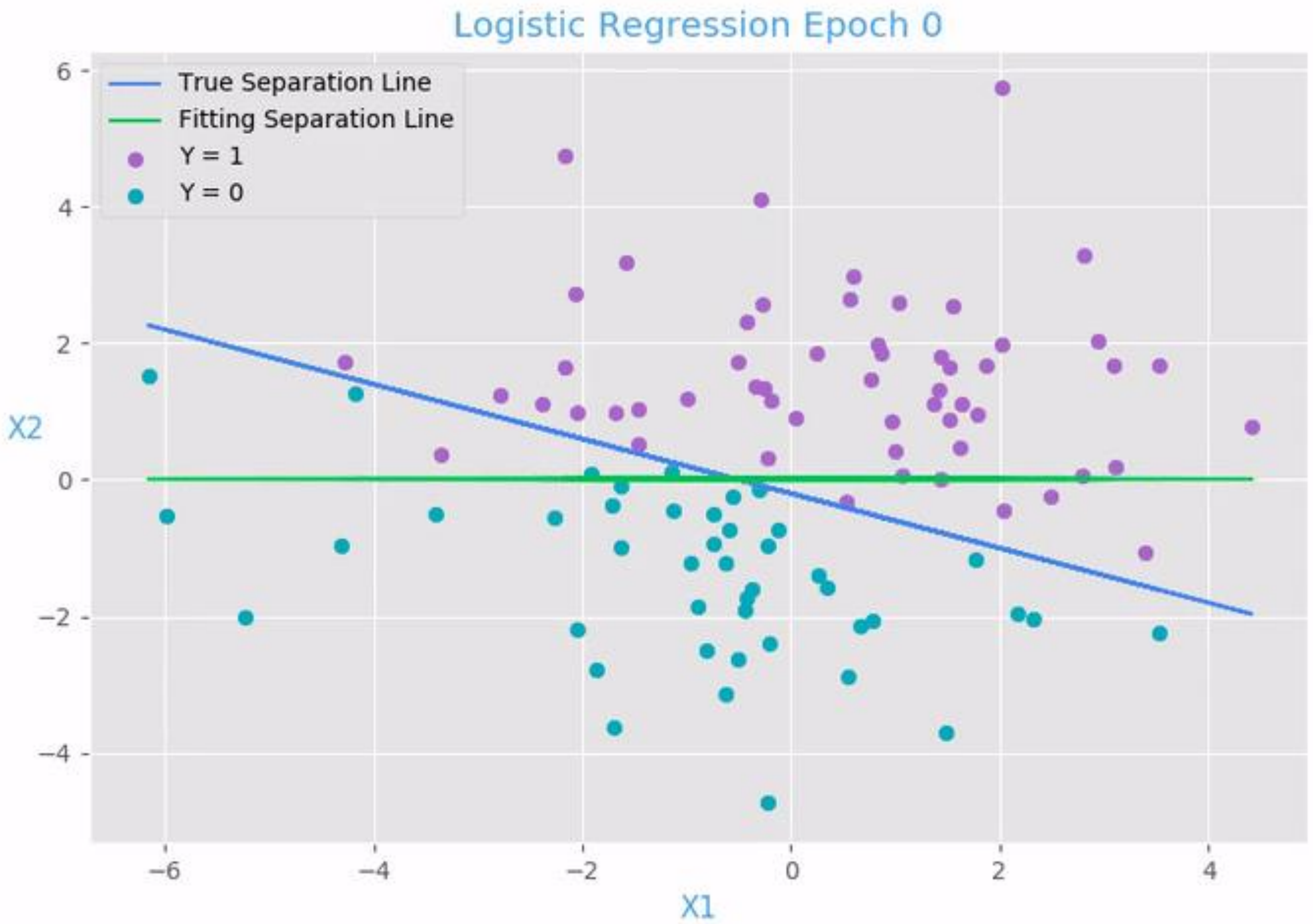
- Finalmente, para obtener los valores óptimos, podemos nuevamente utilizar un algoritmo basado en el gradiente:

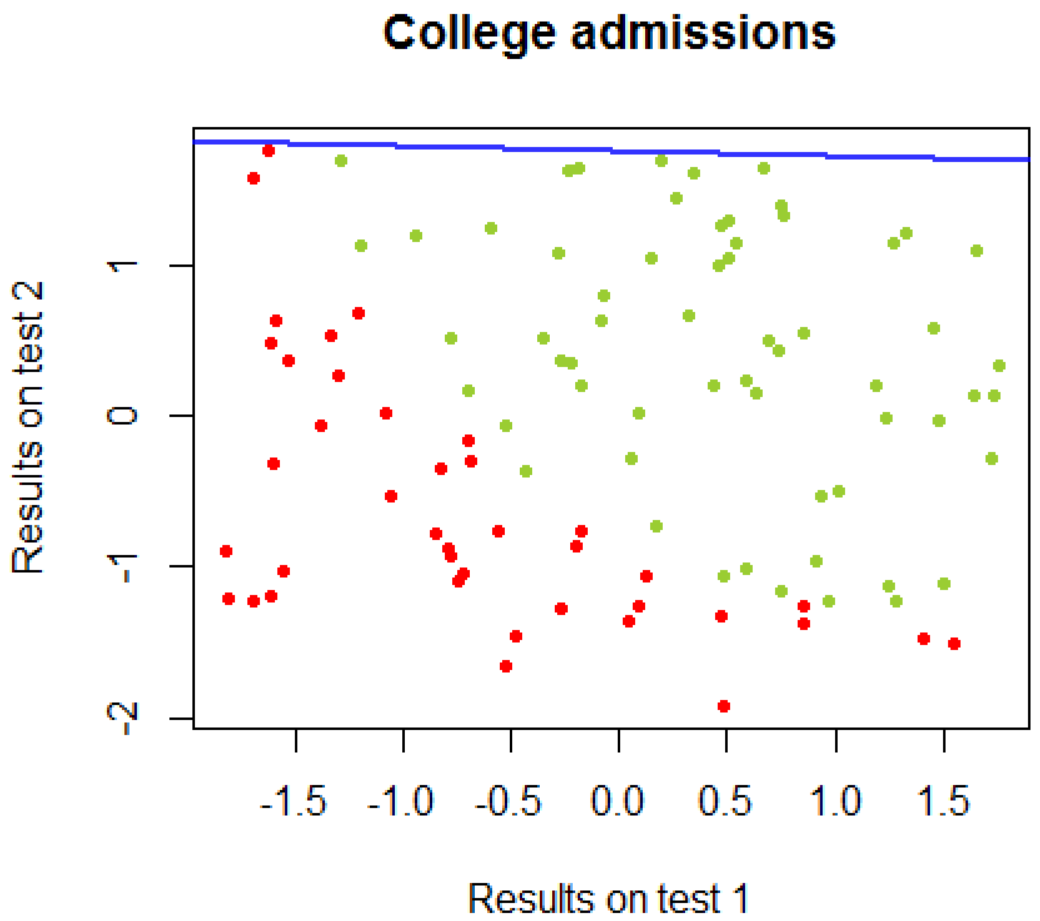
$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \ell(\theta) &= \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\
&= \left(y \frac{1}{g(\theta^T x)} - (1-y) \frac{1}{1 - g(\theta^T x)} \right) g(\theta^T x)(1 - g(\theta^T x) \frac{\partial}{\partial \theta_j} \theta^T x \\
&= \left(y(1 - g(\theta^T x)) - (1-y) g(\theta^T x) \right) x_j \\
&= (y - h_\theta(x))\, x_j
\end{aligned}
$$

- Con lo que nuestra regla de actualización queda de la siguiente manera:

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

- Esta expresión la hemos visto en algún lado, ¿o no?


Logistic Regression Epoch 0
True Separation Line
Fitting Separation Line
Y = 1
Y = 0
X2
X1
6
4
2
0
−2
−4
−6
−4
−2
0
2
4


College admissions
Results on test 2
Results on test 1
-2
-1
0
1
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

Pontificia Universidad Católica de Chile
Escuela de Ingeniería
Departamento de Ciencia de la Computación

IIC2613 – Inteligencia Artificial

Análisis de regresión

Profesor: Hans Löbel