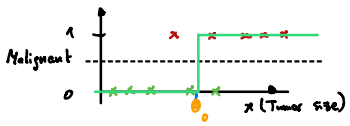
* Classification



MSE → Regression

$\min_\theta \frac{1}{2}\left(y - g(h_\theta(x))\right)^2$

$g(h_\theta(x)) = \begin{cases} 0 & h_\theta(x) < 0 \\ 1 & h_\theta(x) \geq 0 \end{cases}$ thresholding

(classification) historically

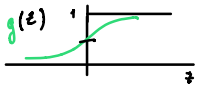
Logistic regression: Force to have $0 \leq g(h_\theta(x)) \leq 1$

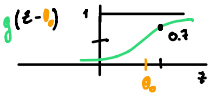
g = Heaviside function

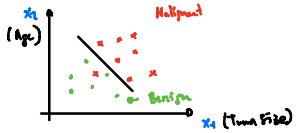
$h_\theta(x) \geq$ hypothesis function

$w_1 x_1 + w_2 x_2 + w_0 \cdot 1$ ← Representation!

* Signoid function (Logistic function): $g(z) = \frac{1}{1+e^{-z}}$;

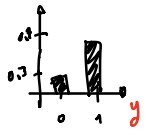






* Interpretation of $g(h_\theta(x))$: Estimated probability that $y=1$ on input x

For some x_i s.t. $g(h_\theta(x_i)) = 0.7$ ⇒ 70% chance of human being malignant

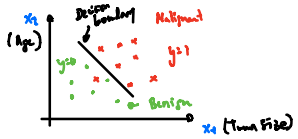


$g(h_\theta(x)) = p(y=1 \mid x=x; \theta)$

$1 - g(h_\theta(x)) = 1 - p(y=1 \mid x=x; \theta) = p(y=0 \mid x=x; \theta) = 0.3$

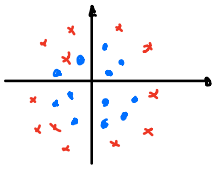
* Decision boundary

$g(\theta^T x) = \begin{cases} 1 & \theta^T x \geq 0 \\ 0 & \theta^T x < 0 \end{cases}$ linear



* Non-linear decision boundary

$g(\theta_1^T x + \theta_2^T x^2) = g\left(\begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}\begin{bmatrix}x_1 \\ x_2\end{bmatrix} + \begin{bmatrix}\theta_3 & \theta_4\end{bmatrix}\begin{bmatrix}x_1^2 \\ x_2^2\end{bmatrix}\right) = \begin{cases} 1 & \theta_1^T x + \theta_2^T x^2 \geq 0 \\ 0 & \theta_1^T x + \theta_2^T x^2 < 0 \end{cases}$



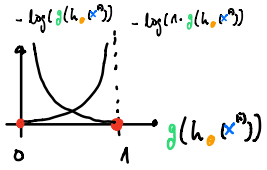
* Logistic regression model

Given $y \in \{0,1\}$, Bernulli: $Pr(y \mid X; \theta) = g(h_\theta(x))^y (1 - g(h_\theta(x)))^{1-y}$

Likelihood (iid): $L(\theta \mid y; x) = Pr(y \mid X; \theta) = \prod_i Pr(y_i \mid x_i; \theta)$

$= \prod_i^m g(h_\theta(x_i))^{y_i} (1 - g(h_\theta(x_i)))^{1-y_i}$

Taking now Log likelihood ⇒ $\min_\theta J(\theta)$ MLE!



if $y^{(i)} = 0$ ⇒ penalty $1 = g(h_\theta(x^{(i)}))$

if $y^{(i)} = 1$ ⇒ penalty $0 = g(h_\theta(x^{(i)}))$

We can take: $g(z) = \frac{1}{1+e^{-z}}$; $g'(z) = g(z)(1-g(z))$

* Cost: $J(\theta) = \frac{1}{m}\sum_{i=1}^m (1-y^i)(-\log(1-g(h_\theta(x^{(i)})))) + y^i(-\log(g(h_\theta(x^{(i)}))))$

$y_i = 0 \Rightarrow g = 0$, $y_i = 1 \Rightarrow g = 1$; $y_i = 0 \Rightarrow g = 1 \Rightarrow -\log(1-g) \to \infty$

$h = \theta x$, $\frac{\partial h}{\partial \theta}$

* Gradient

$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^m (1-y^i)\frac{-1}{(1-g(h_\theta(x^{(i)})))}\,-(g(h_\theta(x^{(i)}))(1-g(h_\theta(x^{(i)}))) x^{(i)}$

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

$= \frac{1}{m}\sum_{i=1}^m \left[g(h_\theta(x^{(i)})) - y^i\right] x^{(i)}$ ← Same as linear regression! Easy!

* Playing with { cost function definition { least squares, log } + composition of non-linear functions { linear, sigmoid, kernel }

+ **Minimization**

$\min J(\theta) \Rightarrow \begin{cases} \nabla J(\theta) = 0 \longrightarrow \text{Eq. no lineal in general} \\ \text{s.t. } \theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k) \Rightarrow J(\theta_{k+1}) < J(\theta_k) \end{cases}$

Steepest descend — step size → Line search — learning rate

+ **Ill-posedness**

Escenario 1: $\begin{bmatrix} \\ X \\ \\ \end{bmatrix}_{3} \begin{bmatrix} \\ \theta \\ \\ \end{bmatrix} \sim \begin{bmatrix} \\ y \\ \\ \end{bmatrix}$

Escenario 2: $\begin{bmatrix} & & \\ & X & \\ & & \end{bmatrix}_{3 \times 1000} \begin{bmatrix} \\ \theta \\ \\ \end{bmatrix} \sim \begin{bmatrix} \\ y \\ \\ \end{bmatrix}$

$\Rightarrow$

Es: $\min_\theta J(\theta) = \|X\theta - y\|^2 \Rightarrow \nabla J(\theta) = 0 \Rightarrow X^T X \theta = X^T y$

Regularización: $\min_\theta J(\theta) = \|X\theta - y\|^2 + \lambda^2 \theta^2 \Rightarrow \nabla J(\theta) = 0 \Rightarrow [X^T X + \lambda^2 I]\theta = X^T y$

SI: $[x^T]_{3\times 1}[x]_{3\times 1} = [x^T x] \Rightarrow \text{rank}(X^T X) = 3$

SC2: $\begin{bmatrix} x^T \end{bmatrix}_{3\times 1} \begin{bmatrix} & X & \end{bmatrix}_{3 \times 1000} = \begin{bmatrix} & X^T X & \end{bmatrix}_{1000 \times 1000} \Rightarrow \text{rank}(X^T X) = 3$

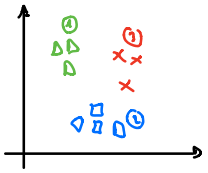
$X = [1, x_1, x_2]$ a $X = [$

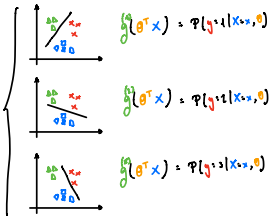
Lin. regr, Log. Reg → One vs all → NN

+ **Multi-clasification**

+ **One-vs-all:**

$y \in \{1, 2, 3\}$





¿ How is the cost?

$$J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left[\sum_{k=1}^{K} (1 - y_k^i)(-\log(1 - g(h_\theta(x^i)))) + y_k^i (-\log(g(h_\theta(x^i)))) \right] + \lambda \|\theta\|^2$$

$y_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}; \quad y_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}; \ldots$

+ **Práctica:** Extend logistic regression (clasification) to multi-clasification

- **Neural model: Logistic unit**

$h_\theta(x) = X\theta = x_1\theta_1 + x_n\theta_n + \underset{\downarrow}{1}\theta_0$

Linear case