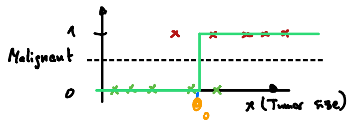
* Classification




MSE → Regr.

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

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

thresholding; $x \cdot \theta_0$

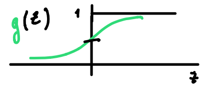
(classification) historically

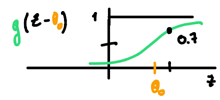
Logistic regression: Force to have $0 < g(h_\theta(x)) < 1$

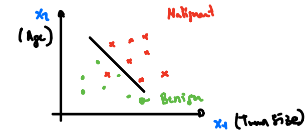
$g$ ≠ Heaviside function; $h_\theta(x)$ = hypothesis function

$w_1 x_1 + w_2 x_2 + \underbrace{w_0 \cdot 1}_{b}$ ← Representation!

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

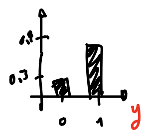



* Interpretation of $g(h_\theta(x))$: Estimated probability that $y=1$ on input $x$

For some $x_1$ ⇒ $g(h_\theta(x_1)) = 0.7$ ⇒ 70% chance of tumor being malignant

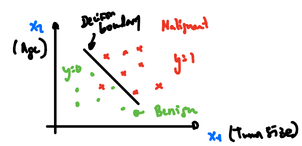



$$g(h_\theta(x)) = p(y=1 \mid X=x; \theta)$$

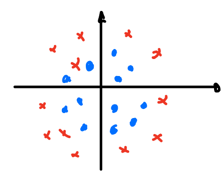
$$1 - g(h_\theta(x)) = 1 - p(y=1 \mid X=x; \theta) = p(y=0 \mid X=x; \theta) = 0.3$$

* Decision boundary

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




* Non-linear decision boundary



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

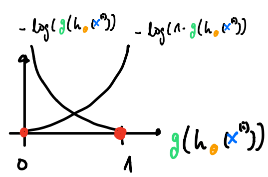
* Logistic regression model

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




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

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

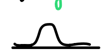
* Cost: $J(\theta) = \frac{1}{m} \sum_{i=1}^m (1-y^i)(-\log(1 - g(h_\theta(x^i)))) + y^i(-\log(g(h_\theta(x^i))))$

$y_i = 0 \Leftrightarrow g = 0$

$y_i = 1 \Leftrightarrow g = 1$

$y_i = 0$ ⇒ $g = 1$ ⇒ $-\log(1-g) \to \infty$

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



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

* Gradient

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

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

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

← Same as linear regression! Easy!

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

$$\min J(\theta) \Rightarrow \begin{cases} \nabla J(\theta^*) = 0 \longrightarrow \text{Eq. Lineal} \\ \theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k) \Rightarrow J(\theta_{k+1}) < J(\theta_k) \end{cases}$$

s.t.

Steepest descend

step size → Line search

Learning rate