La Regressione Logistica

- E' un problema di classificazione
- L'output è composto da un numero finito di elementi

- Classificazione binaria:
  - l'output può essere solo 0 o 1
- Classificazione su classi multiple:
  - l'output è composto da più di due valori categorici

La Regressione Logistica

Usualmente si mappa su 0 un esito negativo e su 1 un esito positivo della osservazione

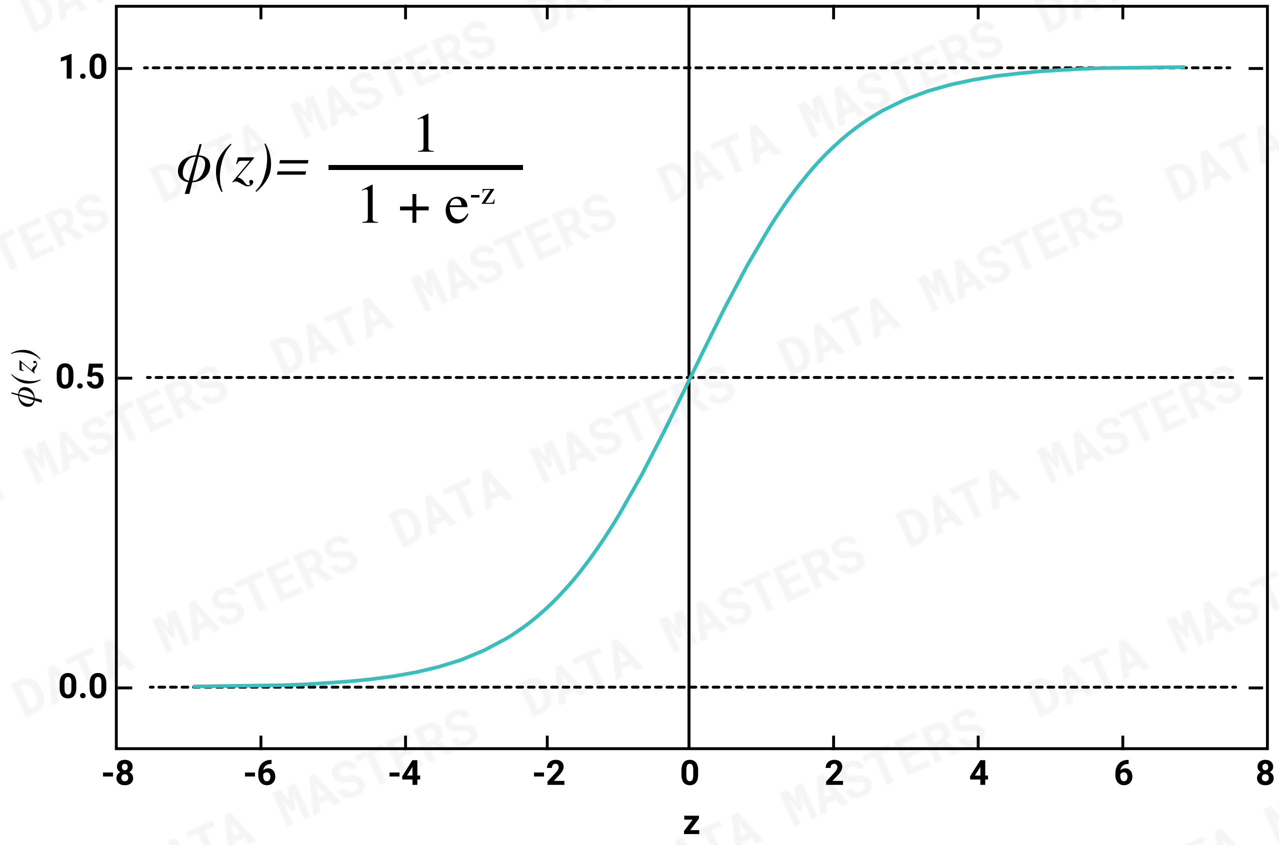
	Classificazione su due classi	
y = [0, 1]	1 – Classe OK	0 – Classe KO
Messaggio e-mail	Spam	Non Spam
Tumore	Maligno	Benigno
Transazione	Fraudolenta	Non fraudolenta

Classificazione binaria

La funzione ipotesi ha come risultato un numero tra 0 o 1

Non è possibile utilizzare una funzione lineare perché, anche impostando un confine decisionale, i valori inferiori a zero o superiori ad 1 non avrebbero senso

Funzione Sigmoide



Un esempio di funzione logistica è **la funzione sigmoide**, che restituisce un valore compreso tra 0 ed 1 per ogni valore reale passato in input

Classificazione binaria

Nel caso della regressione lineare si è utilizzata la formula:

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x$$

Per poterla sfruttare nel caso di una regressione logistica, la si può comporre con la funzione sigmoide:

$$\hat{y} = h_\theta(x) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 X)}}$$

Classificazione binaria

$h_\theta(x)$ fornisce la probabilità che l'output sia pari ad 1

Ad esempio se risulta pari a 0.7 indica il **70% di probabilità** che, per quel dato input, l'output sia 1

$$h_\theta(x) = P(y=1|x;\theta) = 1 - P(y=0|x;\theta)$$

$$P(y=0|x;\theta) + P(y=1|x;\theta) = 1$$

Classificazione binaria

Altra rappresentazione equivalente: $h_\theta(x) = g(\theta^T x)$

Vettore parametri θ

sigmoide

$z = \theta^T x$

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

Confine decisionale

Per ottenere una classificazione discreta tra 0 e 1, si deve definirne il confine decisionale, che rappresenta la linea che separa l'area dove y=0 da quella dove y=1

ad es. :

$$h_\theta(x) \geq 0.5 \rightarrow y = 1$$
$$h_\theta(x) < 0.5 \rightarrow y = 0$$

$$g(z) \geq 0.5$$
$$when\ z \geq 0$$

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

$$z = 0, e^0 = 1 \Rightarrow g(z) = 1/2$$
$$z \rightarrow \infty, e^{-\infty} \rightarrow 0 \Rightarrow g(z) = 1$$
$$z \rightarrow -\infty, e^{\infty} \rightarrow \infty \Rightarrow g(z) = 0$$

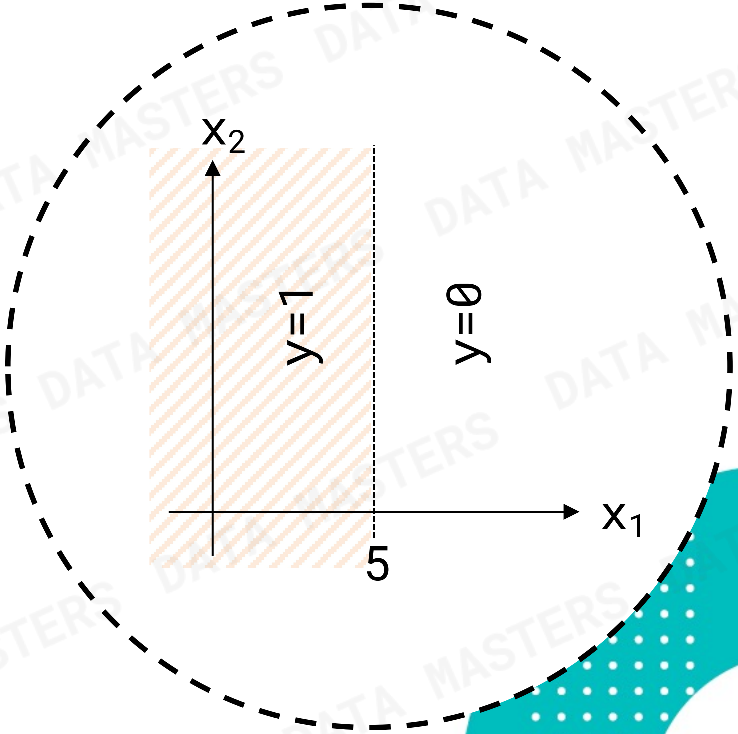
Confine decisionale

Altro esempio:

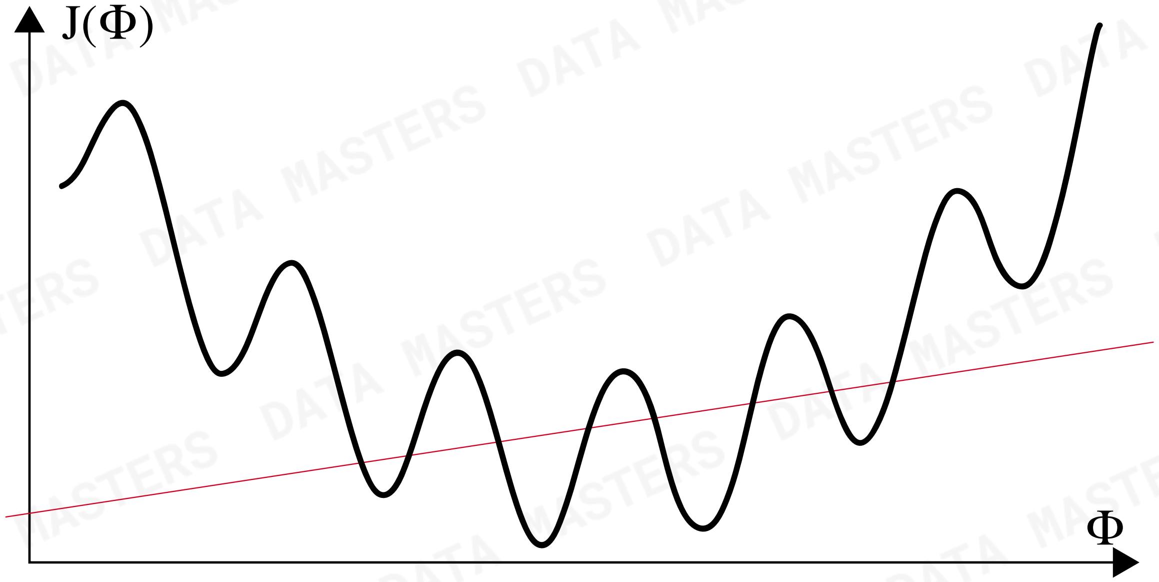
$$\theta = \begin{bmatrix} 5 \\ -1 \\ 0 \end{bmatrix} \qquad y = 1 \; if \; 5 + (-1)x_1 + 0x_2 \geq 0$$

$$5 - x_1 \geq 0$$

$$x_1 \leq 5$$



Funzione di costo per la classif. binaria



Come nel caso della regressione lineare, rappresenta la funzione obiettivo dell'ottimizzazione del modello.

L'utilizzo della **funzione di costo** utilizzata nella regressione lineare risulta in una funzione non convessa, dove è più difficile trovare il minimo assoluto tra diversi minimi relativi

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2 = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x_i) - y_i)^2$$

Funzione di costo per la classif. binaria

Definizione di una funzione di costo ***«comoda»*** per la classificazione binaria, basandoci sui seguenti vincoli:

$$\mathrm{Cost}(h_\theta(x), y) = 0 \text{ if } h_\theta(x) = y$$
$$\mathrm{Cost}(h_\theta(x), y) \rightarrow \infty \text{ if } y = 0 \text{ and } h_\theta(x) \rightarrow 1$$
$$\mathrm{Cost}(h_\theta(x), y) \rightarrow \infty \text{ if } y = 1 \text{ and } h_\theta(x) \rightarrow 0$$

$$\mathrm{Cost}(h_\theta(x), y) = -y \log(h_\theta(x)) - (1-y)\log(1 - h_\theta(x))$$

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

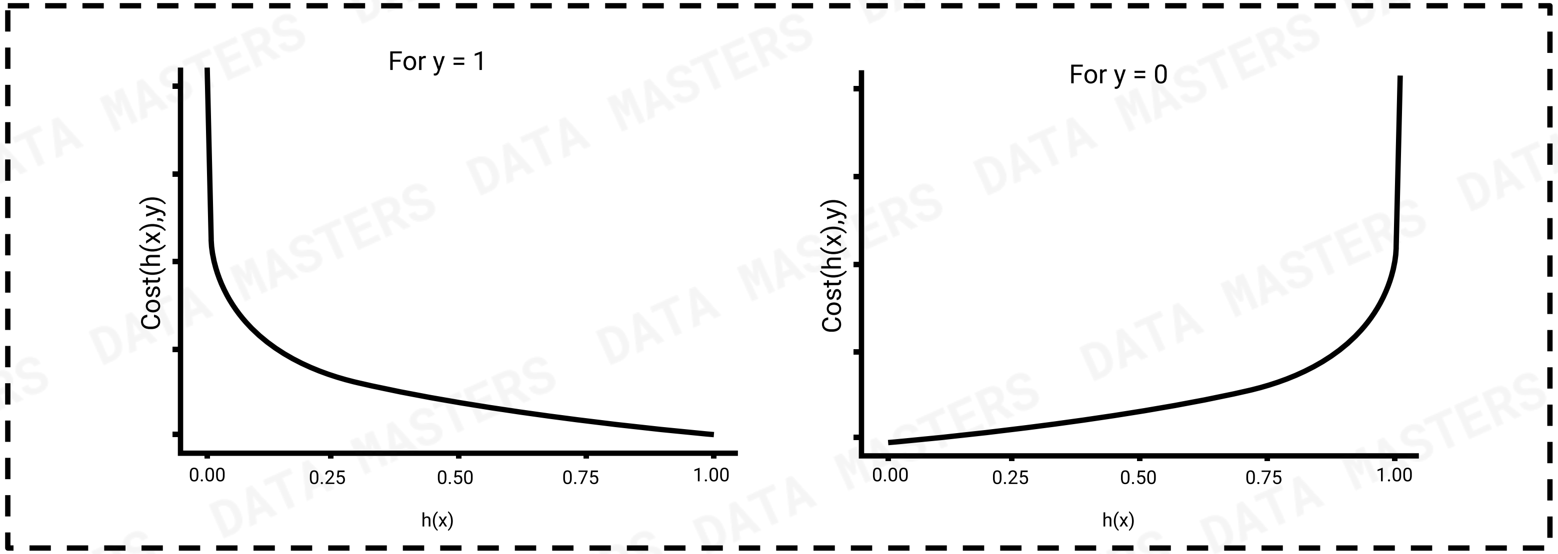
N.B.

***log** è il logaritmo naturale*

Funzione di costo per la classif. binaria

Analisi grafica della funzione di costo per la classificazione binaria:

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



Funzione di costo per la classif. binaria

Altra forma della funzione di costo:

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

Riportata in forma vettoriale:

$$h = g(X\theta)$$

$$J(\theta) = \frac{1}{m}\cdot\left(-y^T\log(h) - (1-y)^T\log(1-h)\right)$$

Discesa gradiente per la regressione logistica

Forma generale della discesa del gradiente:

repeat until convergence: $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$

Ricordiamo la definizione della funzione di costo per la Classificazione Binaria:

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

Derivata prima della funzione sigmoide

$$\boxed{\sigma(x)' = \sigma(x)(1 - \sigma(x))}$$

$$\sigma(x)' = \left(\frac{1}{1+e^{-x}}\right)' = \frac{-(1+e^{-x})'}{(1+e^{-x})^2} = \frac{-1' - (e^{-x})'}{(1+e^{-x})^2} = \frac{0 - (-x)'(e^{-x})}{(1+e^{-x})^2} = \frac{-(-1)(e^{-x})}{(1+e^{-x})^2} = \frac{e^{-x}}{(1+e^{-x})^2}$$

$$= \left(\frac{1}{1+e^{-x}}\right)\left(\frac{e^{-x}}{1+e^{-x}}\right) = \sigma(x)\left(\frac{+1-1+e^{-x}}{1+e^{-x}}\right) = \sigma(x)\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right) = \sigma(x)(1-\sigma(x))$$

Derivate prime parziali della funzione di costo

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{-1}{m} \sum_{i=1}^{m} \left[y^{(i)} log(h_\theta(x^{(i)})) + (1-y^{(i)}) log(1-h_\theta(x^{(i)})) \right] \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} \frac{\partial}{\partial \theta_j} log(h_\theta(x^{(i)})) + (1-y^{(i)}) \frac{\partial}{\partial \theta_j} log(1-h_\theta(x^{(i)})) \right] \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} \frac{\partial}{\partial \theta_j} h_\theta(x^{(i)})}{h_\theta(x^{(i)})} + \frac{(1-y^{(i)}) \frac{\partial}{\partial \theta_j} (1-h_\theta(x^{(i)}))}{1-h_\theta(x^{(i)})} \right] \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} \frac{\partial}{\partial \theta_j} \sigma(\theta^T x^{(i)})}{h_\theta(x^{(i)})} + \frac{(1-y^{(i)}) \frac{\partial}{\partial \theta_j} (1-\sigma(\theta^T x^{(i)}))}{1-h_\theta(x^{(i)})} \right] \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} \sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)})) \frac{\partial}{\partial \theta_j} \theta^T x^{(i)}}{h_\theta(x^{(i)})} + \frac{-(1-y^{(i)}) \sigma(\theta^T x^{(i)})(1-\sigma(\theta^T x^{(i)})) \frac{\partial}{\partial \theta_j} \theta^T x^{(i)}}{1-h_\theta(x^{(i)})} \right] \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[\frac{y^{(i)} h_\theta(x^{(i)})(1-h_\theta(x^{(i)})) \frac{\partial}{\partial \theta_j} \theta^T x^{(i)}}{h_\theta(x^{(i)})} - \frac{(1-y^{(i)}) h_\theta(x^{(i)})(1-h_\theta(x^{(i)})) \frac{\partial}{\partial \theta_j} \theta^T x^{(i)}}{1-h_\theta(x^{(i)})} \right] \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} (1-h_\theta(x^{(i)})) x_j^{(i)} - (1-y^{(i)}) h_\theta(x^{(i)}) x_j^{(i)} \right] \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} (1-h_\theta(x^{(i)})) - (1-y^{(i)}) h_\theta(x^{(i)}) \right] x_j^{(i)} \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} - y^{(i)} h_\theta(x^{(i)}) - h_\theta(x^{(i)}) + y^{(i)} h_\theta(x^{(i)}) \right] x_j^{(i)} \\
&= -\frac{1}{m} \sum_{i=1}^{m} \left[y^{(i)} - h_\theta(x^{(i)}) \right] x_j^{(i)} \\
&= \frac{1}{m} \sum_{i=1}^{m} \left[h_\theta(x^{(i)}) - y^{(i)} \right] x_j^{(i)}
\end{aligned}
$$

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

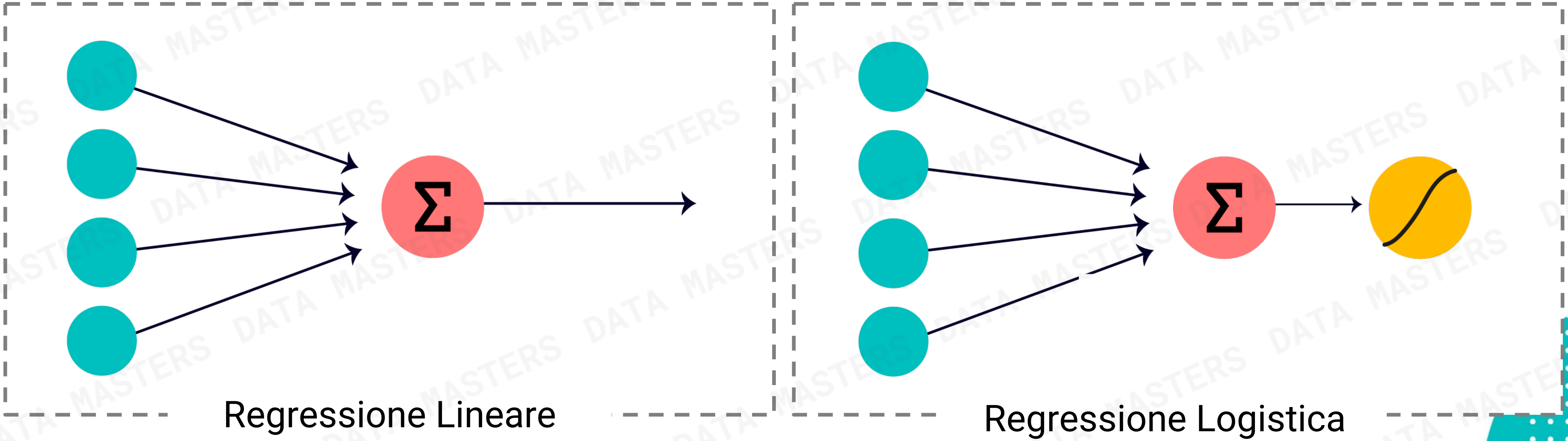
Forma vettoriale:

$$
\boxed{\nabla J(\theta) = \frac{1}{m} \cdot X^T \cdot \left(g(X \cdot \theta) - \vec{y} \right)}
$$

Esercitazione

Es01_regressione_logistica_discesa_del_gradiente.ipynb

Regressione lineare vs regressione logistica



Discesa gradiente per la regressione logistica

$$y = \{0, 1 \ldots n\}$$

In questo caso si divide il problema in n+1 (+1 perché l'indice inizia da 0) problemi di classificazione binaria; in ciascuno, si prevede la probabilità che 'y' sia un membro della relativa classe.

$$h_\theta^{(0)}(x) = P(y = 0|x;\theta)$$
$$h_\theta^{(1)}(x) = P(y = 1|x;\theta)$$
$$\cdots$$
$$h_\theta^{(n)}(x) = P(y = n|x;\theta)$$

$$\mathrm{prediction} = \max_i(h_\theta^{(i)}(x))$$

Esercitazione

Es02_regressione_logistica_su_più_classi.ipynb