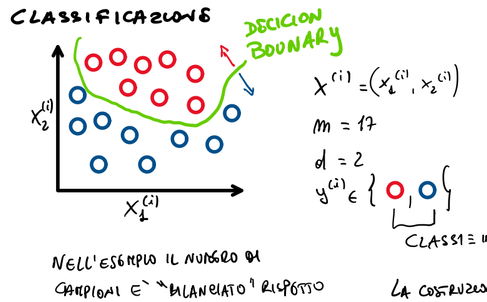
CLASSIFICAZIONE



$x^{(i)} = (x_1^{(i)}, x_2^{(i)})$

$m = 17$

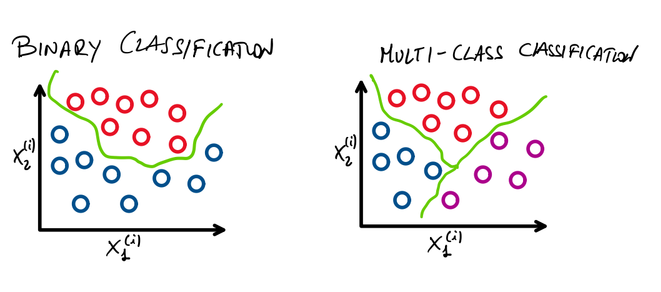
$d = 2$

$y^{(i)} \in \{ ○, ○ \}$ CLASSI ≡ INSIEME NUMERABILE

NELL'ESEMPIO IL NUMERO DI CAMPIONI È "BILANCIATO" RISPETTO ALLE CLASSI #○ ≃ #○

LA COSTRUZIONE DEL TRAINING SET DEVE RISPECCHIARE LA POPOLAZIONE DELLE CLASSI

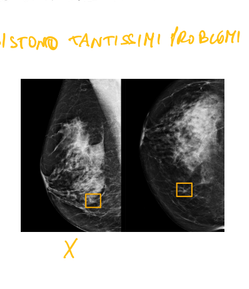
GOAL: DEFINIRE UN MODELLO PARAMETRIZZATO, UNA FUNZIONE COSTO E UN ALGORITMO DI LEARNING UTILI A TROVARE IN AUTOMATICO IL BOUNDARY DI DECISIONE (≡ CLASSIFICATORE) CONSIDERANDO IL TRAINING SET (≡ PUNTI IN UNO SPAZIO DELLE FEATURES)



UN MODO PER TRATTARE PROBLEMI MULTI-CLASSE È QUELLO DI SUDDIVIDERE IL PROBLEMA MULTI-CLASSE IN TANTI PROBLEMI BINARI. SI PROCEDERÀ RISOLVENDO I DIVERSI PROBLEMI BINARI E POI SI METTERANNO INSIEME LE SOLUZIONI OTTENUTE DAI VARI CLASSIFICATORI BINARI AL FINE DI OTTENERE UNA SOLUZIONE AL PROBLEMA MULTI CLASSE ⇒ ONE VS ALL / ONE VS ONE] APPROCCI

⇒ È SICURAMENTE UTILE CAPIRE COME COSTRUIRE UN CLASSIFICATORE BINARIO VISTO CHE PUÒ ESSERE UNA BASE PER PROBLEMI PIÙ COMPLESSI

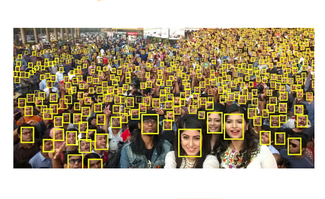
ESISTONO TANTISSIMI PROBLEMI DI CLASSIFICAZIONE BINARIA INTERESSANTI (E "DIFFICILI") ⇒ NON SOTTOVALUTARE L'UTILITÀ DEI CLASSIFICATORI BINARI



BREAST CANCER

y □ → 1 TUMOR (POSITIVE CLASS) / 0 NOT TUMOR (NEGATIVE CLASS)

IMAGE PATCH



FACE DETECTION

IS THERE A FACE IN THE PATCH?

SCALE

FACE < 1 / 0

EMAIL < SPAM (1) / NOT SPAM (0)

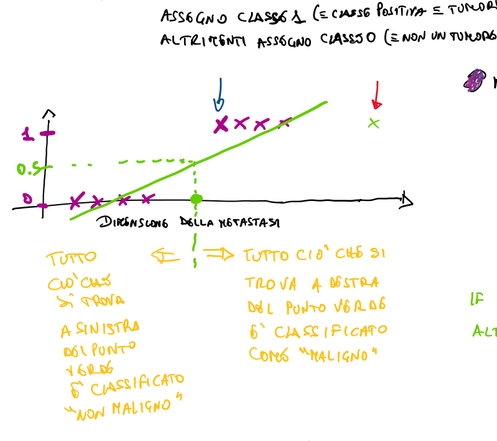
IS THIS DIGITAL CONTENT TRUE OR FAKE?

1 FAKE — 0 TRUE

- DOCUMENT
- IMG
- TEXT
- NEWS

POSSIAMO USARE IL MODELLO DI REGRESSIONE PER LA CLASSIFICAZIONE BINARIA?

IDEA ERRATA: SE $h_\vartheta(x) > 0.5$ ALLORA ASSEGNO CLASSE 1 (≡ CLASSE POSITIVA ≡ TUMORE MALIGNO) ALTRIMENTI ASSEGNO CLASSE 0 (≡ NON UN TUMORE MALIGNO)



METASTASI
- DIMENSIONE
- ETÀ PAZIENTE
- COLORE
- FORMA/REGOLARITÀ

ESPERTI DEL DOMINIO SONO FONDAMENTALI PER DEFINIRE LE CARATTERISTICHE

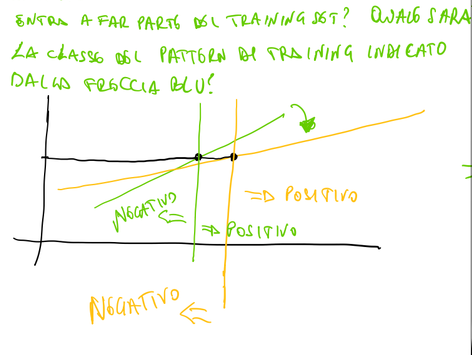
IF $h_\vartheta(x) \geq 0.5 \Rightarrow$ PREDICI 1

ALTRIMENTI (OVVERO $h_\vartheta(x) < 0.5$) PREDICI 0

A QUALE CLASSE VIENE ASSEGNATO IL PUNTO INDICATO DALLA FRECCIA ROSSA NEL GRAFICO UTILIZZANDO QUESTA "STRATEGIA" ERRATA PER LA CLASSIFICAZIONE

A QUALE CLASSE È ASSEGNATO IL PUNTO INDICATO DALLA FRECCIA BL

COSA ACCADE SE IL CAMPIONE ETICHETTATO IN ROSSO ENTRA A FAR PARTE DEL TRAINING SET? QUALE SARÀ LA CLASSE DEL PATTERN DI TRAINING INDICATO DALLA FRECCIA BLU?



IL PROBLEMA PRINCIPALE STA NEL FATTO CHE LE NOSTRE $y \in \{0,1\}$ MENTRE IL MODELLO $h_\vartheta(x)$ VISTO PER LA REGRESSIONE PUÒ ASSUMERE VALORI > 1 E < 0

⇓

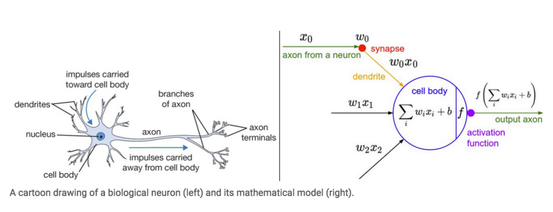
LOGISTIC REGRESSION

$0 \leq h_\vartheta(x) \leq 1$

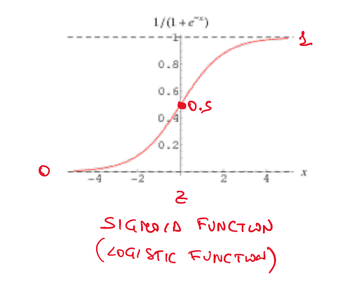
GOAL → MODELLO CON ASINTOTO SUPERIORE 1 E INFERIORE 0

⇒ UN METODO DI REGRESS
NON DEVE ESSERE USAT
UN PROBLEMA DI CL

⇒ UN METODO DI RE
ESSERE OPPORTUNAMENTE ADATTATO AF
UTILE A RISOLVERE UN PROB



A cartoon drawing of a biological neuron (left) and its mathematical model (right).



SIGMOID FUNCTION (LOGISTIC FUNCTION)

$h_\vartheta(x) = \sigma(\vartheta^T x)$

$\sigma(z) = \frac{1}{1+e^{-z}}$, $z = \vartheta^T x$

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

$\sigma(z)$ È UNA FUNZIONE NON LINEARE A CUI POSSIAMO DARE UNA INTERPRETAZIONE PROBABILISTICA.

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

IL MODELLO LOGISTICO $h_\vartheta(x)$ PUÒ ESSERE INTERPRETATO COME LA PROBABILITÀ DI ESSERE DI CLASSE 1 (CLASSE POSITIVA) AVENDO OSSERVATO x C INPUT E ϑ COME PARAMETRI

$h_\vartheta(x) = P(y=1 \mid x, \vartheta)$

$P(y=1 \mid x, \vartheta) + P(y=0 \mid x, \vartheta) = 1$

$\Rightarrow P(y=0 \mid x, \vartheta) = 1 - P(y=1 \mid x$

IL NOSTRO GOAL È QUINDI QUELLO DI TROVARE
DI QUESTO NUOVO MODELLO (CHE È UNA COM

$h_\vartheta(x) = \sigma(\vartheta^T x) = \frac{1}{1+e^{-\vartheta^T x}} = \frac{1}{1+e^{-(\vartheta_0 x_0 + \vartheta_1 x_1 + \ldots + \vartheta_d x_d)}}$

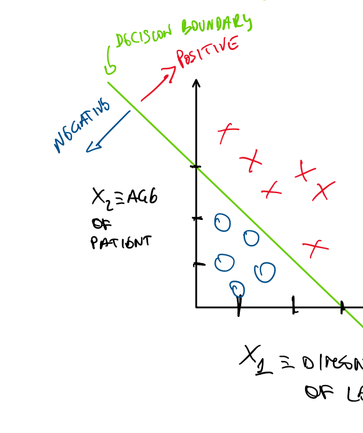
$h_\vartheta(x) \geq 0.5 \Leftrightarrow \vartheta^T x \geq 0$

$h_\vartheta(x) < 0.5 \Leftrightarrow \vartheta^T x < 0$

SI AVRÀ UNA PREDIZIONE $\hat{y} = 1$
DELLA FUNZIONE $\sigma(z)$ SARÀ MAG

$\vartheta^T x$ DEFINISCE IL BOUNDARY
$\vartheta_0 x_0 + \vartheta_1 x_1 + \ldots + \vartheta_d x_d$

QUESTO BOUNDARY DI DECISIONE È
DI PROBLEMA DI CLASSIFICAZIO
PRESENTI 2 FEATURE $(x_1, x_2$
NEL CASO IN CUI IL NOSTR
TRE FEATURES ...



SUPPONIAMO $\vartheta =$

$\hat{y} = 1 \Leftrightarrow$

$\hat{y} = 0 \Leftrightarrow x_1 + x_2$

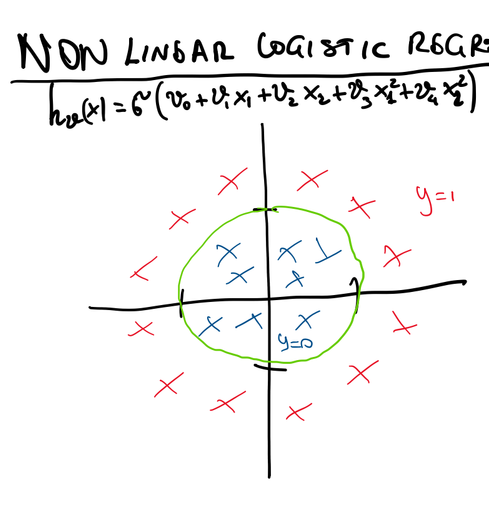
IL BOUNDARY DI DECISIONE
E DAI PARAMETRI. CAM
SI OTTERRANNO DECISION

CON IL MODELLO CHE ST
TROVARE LA MIGLIORE L
OVVERO I MIGLIORI PARAMET
CLASSE 1 STANNO
MENTRE QUELLI DI C

⇒ VOGLIAMO COMME
⇒ SERVE UNA

NON LINEAR LOGISTIC REGRESSION

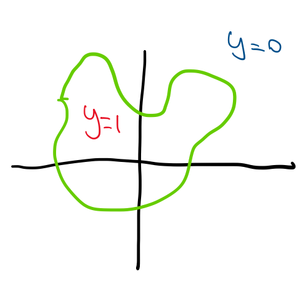
$h_\vartheta(x) = \sigma(\vartheta_0 + \vartheta_1 x_1 + \vartheta_2 x_2 + \vartheta_3 x_1^2 + \vartheta_4 x_2^2)$



$\vartheta =$

$\hat{y} =$

x_1^2



TRASF
CON
(ES,
RISO
CLA
NON
SE
NE

COME FACCIAMO A S
- BISOGNA DEF
DI LOSS (CO

$J(\vartheta)$

DEFINIZIONE
LOS

CASO
y=1

CASO y=0
↑
ETICHETTA TRAINING

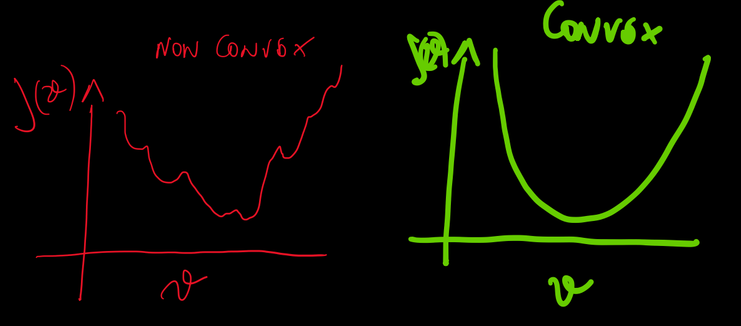
QUELLI TROVATI MINIMIZZANDO LA FUNZIONE COSTO GLOBALE

INTERPRETAZIONE

NON POSSIAMO USARE LA FUNZIONE COSTO VISTA PER LA REGRESSIONE

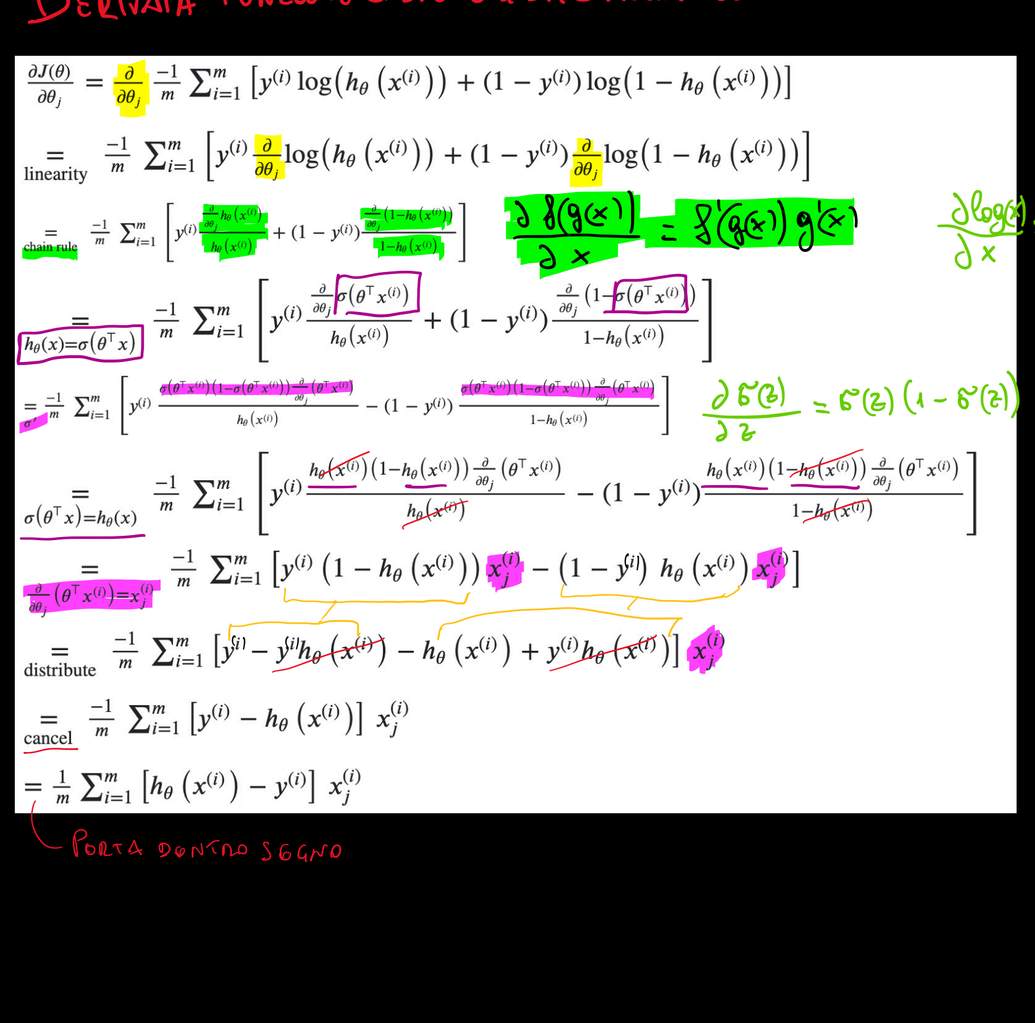
$Cost(h_\vartheta(x), y) = \frac{1}{2}(h_\vartheta(x) - y)^2$

INFATTI $h_\vartheta(x)$ NASCONDE UNA FUNZIONE NON LINEARE $\sigma(z) = \frac{1}{1+e^{-z}}$ CHE RENDEREBBE $J(\vartheta)$ NON CONVESSA QUINDI LA DISCESA DEL GRADIENTE NON GARANTIREBBE L'OTTIMO



LA FUNZIONE DI LOSS CHE USIAMO È LA CROSS-ENTROPY (CHE È CONVESSA

DERIVATA FUNZIONE COSTO LOGISTIC REGRESSION



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

$$\underset{\text{linearity}}{=} \frac{-1}{m} \sum_{i=1}^m \left[y^{(i)} \frac{\partial}{\partial \theta_j} \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \frac{\partial}{\partial \theta_j} \log(1 - h_\theta(x^{(i)})) \right]$$

$$\underset{\text{chain rule}}{=} \frac{-1}{m} \sum_{i=1}^m \left[y^{(i)} \frac{\frac{\partial}{\partial \theta_j} h_\theta(x^{(i)})}{h_\theta(x^{(i)})} + (1-y^{(i)}) \frac{\frac{\partial}{\partial \theta_j}(1 - h_\theta(x^{(i)}))}{1 - h_\theta(x^{(i)})} \right]$$

$\frac{\partial f(g(x))}{\partial x} = f'(g(x))\, g'(x)$ $\qquad \frac{\partial \log(x)}{\partial x} = \frac{1}{x}$

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

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

$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1 - \sigma(z))$

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

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

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

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

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

PORTA DENTRO SEGNO