Lezione 2

mercoledì 14 ottobre 2020 13:34

MINIMI QUADRATI - NOTAZIONE MATRICIALE

$$X_{N\times d} = \begin{bmatrix} 1 & x_1(1) & \cdots & x_{d-1}(1) \\ 1 & x_1(2) & & \\ \vdots & \vdots & & \\ 1 & x_1(N) & \cdots & x_{d-1}(N) \end{bmatrix} \quad \varphi(1)^T \ (1\times d)$$

$$\vartheta_{d\times 1} = \begin{bmatrix} \vartheta_0 \\ \vartheta_1 \\ \vdots \\ \vartheta_{d-1} \end{bmatrix} \qquad Y_{N\times 1} = \begin{bmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{bmatrix} \qquad E_{N\times 1} = \begin{bmatrix} e(1) \\ e(2) \\ \vdots \\ e(N) \end{bmatrix}$$

$$= \begin{bmatrix} \varphi(1)^T \\ \varphi(2)^T \\ \vdots \\ \varphi(N)^T \end{bmatrix}$$

$$\boxed{Y_{N\times 1} = X_{N\times d}\cdot \vartheta_{d\times 1} + E_{N\times 1}}$$

$$(X\vartheta)^T = \vartheta^T \cdot X^T$$

$$J(\vartheta) = \frac{1}{N}\|Y - X\vartheta\|_2^2 = \frac{1}{N}\underbrace{(Y - X\vartheta)^T}_{1\times N}\underbrace{(Y - X\vartheta)}_{N\times 1}$$

$$\|x\|_2^2 = \sqrt{(x_1)^2 + \cdots + (x_d)^2} \quad \text{(NORMA)} \qquad x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

$$\sum_{i=1}^{N} x_i^2 \qquad x_i = e(i)$$

$$= \frac{1}{N}\left(\underbrace{Y^TY}_{1\times N\ N\times 1} - \underbrace{Y^TX\vartheta}_{1\times N\ N\times d\ d\times 1} - \underbrace{\vartheta^TX^TY}_{1\times d\ d\times N\ N\times 1} + \underbrace{\vartheta^TX^TX\vartheta}_{1\times d\ d\times N\ N\times d\ d\times 1}\right)$$

$$= \frac{1}{N}\left(Y^TY - 2\vartheta^TX^TY + \vartheta^TX^TX\vartheta\right)$$

- $\nabla_\vartheta\left(\underbrace{\vartheta^T}_{1\times d}\underbrace{A}_{d\times d}\underbrace{\vartheta}_{d\times 1}\right) = (A + A^T)\vartheta$

$\alpha\cdot x^2 \qquad \frac{d(\alpha\cdot x^2)}{dx} = 2\alpha x = (\alpha+\alpha)\cdot x$

- $\nabla_\vartheta\left(\underbrace{\vartheta^T}_{1\times d}\underbrace{b}_{d\times 1}\right) = \underbrace{b}_{d\times 1}$

$$\nabla_\vartheta J(\vartheta) = \underset{d\times 1}{0} \Rightarrow \frac{1}{N}\left(-2X^TY + 2X^TX\vartheta\right) = 0$$

$$\Rightarrow \boxed{\hat{\vartheta} = (X^TX)^{-1}X^TY}$$

$$\nabla_\vartheta\left(\vartheta^T\underbrace{X^TX}_{A}\vartheta\right) = (A + A^T)\vartheta = \left(X^TX + [X^TX]^T\right)\vartheta = \left(X^TX + X^TX\right)\vartheta$$

SISTEMA "VERO": $y(i) = \varphi(i)^T\vartheta^0 + v(i)$ — ϑ^0: VALORE VERO PARAMETRI

- Supponendo che $v(i)$ sia $\mathbb{E}[v(i)] = 0$, $v(i) \perp \varphi(i)$

$$\mathbb{E}[\hat{\vartheta}] = \vartheta^0 \quad \text{CORRETTO}$$

- Supponendo inoltre che i numeri $v(i)$ siano indipendenti $\mathbb{E}[v(i)\cdot v(j)] = 0$

- Supponendo inoltre che i numeri $v(i)$ siano indipendenti $\mathbb{E}[v(i)\cdot v(j)] = 0 \quad \forall i \neq j$

$$\mathrm{Var}[\hat{\vartheta}] = \lambda^2 \cdot \left[\sum_{i=1}^{N} \underset{d\times 1}{\varphi(i)} \cdot \underset{1\times d}{\varphi(i)^T}\right]^{-1}$$

$$\Downarrow$$

$$\mathrm{Var}[v(i)] = \lambda^2 \quad \forall i$$

$\Rightarrow$ CONSISTENTE

STIMA A MASSIMA VEROSIMIGLIANZA 4

STIMA PARAMETRICA

- $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} y(i)$ MEDIA CAMPIONARIA $\quad \vartheta = \mu \in \mathbb{R}$
- $S^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}\left(y(i) - \hat{\mu}\right)^2$ VAR. CAMPIONARIA $\quad \vartheta = \sigma^2 \in \mathbb{R}$
- STIMA A MINIMI QUADRATI

$$y(i) = \vartheta_0 + \vartheta_1 x_1(i) + \dots + \vartheta_{d-1} x_{d-1}(i) + e(i)$$

$$\vartheta = [\vartheta_0 \ \vartheta_1 \ \dots \ \vartheta_{d-1}]^T \in \mathbb{R}^{d\times 1}$$

$$e(i) \ \text{i.i.d.} \ (0, \lambda^2)$$

Il metodo della MASSIMA VEROSIMIGLIANZA è una procedura di stima che, dato un modello probabilistico, stima i suoi parametri in modo che essa sia la più consistente possibile con quanto osservato

$Y_{N\times 1} = [y(1) \ y(2) \ \dots \ y(N)]^T$: N osservazioni della variabile y

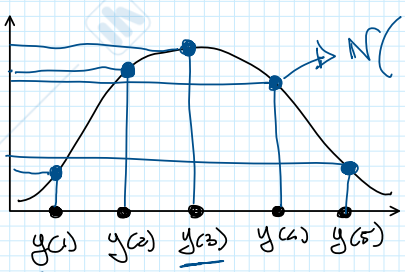
- $y(i) \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d

La pdf congiunta di Y è:

$$f(\underbrace{y(1), y(2), \dots, y(N)}_{Y_{N\times 1}} \mid \mu, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y(i) \mid \mu, \sigma^2)$$

("DATO")

$$= f(Y \mid \mu, \sigma^2)$$



$\mathcal{N}(y(i) \mid \mu, \sigma^2)$ "PRODOTTO DEI PALLINI BLU"

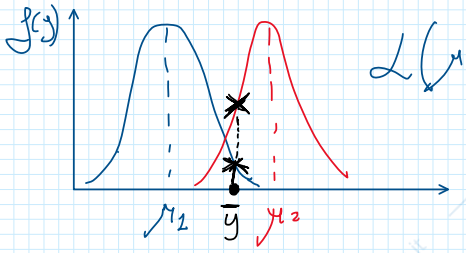
- Se $f(Y \mid \mu, \sigma^2)$ (VARIABILE, NOTO, NOTO) è funzione di Y $\Rightarrow$ è una pdf N-dimensionale
- Se $f(Y \mid \mu, \sigma^2)$ (NOTI, VARIABILI) è vista in funzione di μ e $\sigma^2 \Rightarrow$ LIKELIHOOD

- Se $f(Y \mid \mu, \sigma^2)$ è vista in funzione di μ e $\sigma^2 \Rightarrow$ LIKELIHOOD

(f: NOTI; Y, μ, σ^2: VARIABILI)

$$\rightarrow \mathcal{L}(\mu, \sigma^2 \mid Y)$$

$$\mathcal{L}(\mu, \sigma^2 \mid y) = \mathcal{N}(\mu, \sigma^2 \mid y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$$



σ^2 NOTO

$$\mathcal{L}(\mu = \mu_1 \mid \bar{y}) < \mathcal{L}(\mu = \mu_2 \mid \bar{y})$$

$$\hat{\mu} = \bar{y}$$

$$\hat{\vartheta}_{ML} = \begin{bmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{bmatrix} = \arg\max_{\vartheta} \mathcal{L}(\vartheta \mid Y) = \arg\max_{\vartheta} \prod_{i=1}^{N} \mathcal{N}(y_{(i)} \mid \vartheta)$$

$\vartheta = [\mu, \sigma^2]$

$\forall m$ con f generica pdf $f(Y \mid \vartheta)$

$$\hat{\vartheta}_{ML} = \arg\max_{\vartheta} \prod_{i=1}^{N} f(y_{(i)} \mid \vartheta)$$

$$\hat{\vartheta}_{ML} = \arg\max_{\vartheta} \ln\left[\mathcal{L}(\vartheta \mid Y)\right]$$

PROPRIETÀ STIMA A MASSIMA VEROSIMIGLIANZA

1) Asintoticamente corretta $\lim_{N \to +\infty} \mathbb{E}\left[\hat{\vartheta}_{ML}\right] = \vartheta^0$
2) Consistente
3) Asintoticamente efficiente $\lim_{N \to +\infty} \mathrm{Var}\left[\hat{\vartheta}_{ML}\right] = M^{-1}$
4) Asintoticamente normale $\hat{\vartheta}_{ML} \sim \mathcal{N}(\vartheta^0, M^{-1})$ $N \to +\infty$

Esempio

$y_{(i)} \sim \mathcal{N}(\mu, \sigma^2 = 1)$, $i = 1, 2$, c.i.d. Stimare $\hat{\mu}_{ML}$

$y_{(1)} = 4$ $\quad y_{(2)} = 6$

$y^{(1)} = 4 \qquad y^{(2)} = 6$

$$f\left(y^{(1)}=4 \mid \mu, \sigma^2=1\right) = \frac{1}{\sqrt{2\pi\cdot 1}} e^{-\frac{1}{2}\left(\frac{4-\mu}{1}\right)^2}$$

$$f\left(y^{(2)}=6 \mid \mu, \sigma^2=1\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2}$$

$$f\left(y^{(1)}=4, y^{(2)}=6 \mid \mu, \sigma^2=1\right) = \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2}\right)\cdot\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2}\right)$$

⇓ È FUNZIONE SOLO DI μ

$$= \mathcal{L}\left(\underset{\text{inc}}{\mu} \mid \underset{\theta}{Y}\right)$$

$$\hat{\mu}_{ML} = \underset{\mu}{\operatorname{argmax}}\, \mathcal{L}\left(\mu \mid y^{(1)}=4,\ y^{(2)}=6\right)$$

Calc le log-likelihood

$$\ln\left[\mathcal{L}(\mu \mid \cdot)\right] = \ln\left[\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(4-\mu)^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(6-\mu)^2}\right]$$

$$= \ln\left[\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(4-\mu)^2}\right] + \ln\left[\frac{1}{\sqrt{2\pi}}\cdot e^{-\frac{1}{2}(6-\mu)^2}\right]$$

$$= \ln\frac{1}{\sqrt{2\pi}} + \ln\left[e^{-\frac{1}{2}(4-\mu)^2}\right] + \ln\frac{1}{\sqrt{2\pi}} + \ln\left[e^{-\frac{1}{2}(6-\mu)^2}\right]$$

$$= 2\ln\frac{1}{\sqrt{2\pi}} - \frac{1}{2}(4-\mu)^2 \ln e - \frac{1}{2}(6-\mu)^2 \ln e$$

$$= 2\ln\frac{1}{\sqrt{2\pi}} - \frac{1}{2}(4-\mu)^2 - \frac{1}{2}(6-\mu)^2$$

Trov il massimo

$$\frac{d\ln[\mathcal{L}]}{d\mu} = 0 \Rightarrow 0 + \frac{2}{2}(4-\mu) + \frac{2}{2}(6-\mu) = 0$$

$$\Rightarrow 4+6 = 2\mu \Rightarrow \boxed{\hat{\mu}_{ML} = \frac{6+4}{2} = 5}$$

Osservazioni

$$\hat{\vartheta}_{ML} = \arg\max_{\vartheta} \ln\left[\mathcal{L}(\vartheta \mid Y)\right]$$

$$= \arg\min_{\vartheta} -\ln\left[\mathcal{L}(\vartheta \mid Y)\right]$$

MINIMI QUADRATI

$$J(\vartheta) = \frac{1}{N}\sum_{i=1}^{N}\left(y(i) - \varphi(i)^T\vartheta\right)^2$$

$$\hat{\vartheta} = \arg\min_{\vartheta} J(\vartheta)$$

STIMA A MASSIMA VEROSIMIGLIANZA DI MODELLI LINEARI

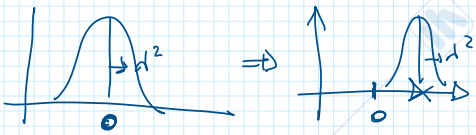
$$y(i) = \vartheta_0 + \vartheta_1 x_1(i) + \vartheta_2 x_2(i) + \dots + \vartheta_{d-1} x_{d-1}(i) + e(i)$$

$$= \underbrace{\varphi(i)^T}_{1\times d}\underbrace{\vartheta}_{d\times 1} + e(i)$$

$$\varphi(i) = \begin{bmatrix} 1 \\ x_1(i) \\ x_2(i) \\ \vdots \\ x_{d-1}(i) \end{bmatrix} \qquad \vartheta = \begin{bmatrix} \vartheta_0 \\ \vartheta_1 \\ \vdots \\ \vartheta_{d-1} \end{bmatrix}$$

$$e(i) \sim \mathcal{N}(0, \lambda^2) \quad \text{i.i.d.}$$

$$e(i) \perp \varphi(i)$$



$$\Downarrow$$

$$y(i) \sim \mathcal{N}\left(\underbrace{\varphi(i)^T\vartheta}_{\mu(i)}, \lambda^2\right)$$

La probabilità di osservare i dati mutuamente è data da:

$$f\left(y(1), y(2), \dots, y(N) \mid X, \vartheta, \lambda^2\right) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{N} f\left(y(i) \mid \varphi(i), \vartheta, \lambda^2\right)$$

$$X = \begin{bmatrix} \varphi(1)^T \\ \vdots \\ \varphi(N)^T \end{bmatrix}$$

$$= \prod_{i=1}^{N} \mathcal{N}\left(\varphi(i)^T\vartheta, \lambda^2\right)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{1}{2}\left(\frac{y(i) - \varphi(i)^T\vartheta}{\lambda}\right)^2} = \mathcal{L}\left(\vartheta, \lambda^2 \mid Y, X\right)$$

↳ NOTO (PER IPOTESI)

1) Calcolo la log-verosimiglianza

$$\ln\left[\mathcal{L}(\vartheta \mid X, Y)\right] = \ln\left[\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda^2}} e^{-\frac{1}{2}\left(\frac{y(i) - \varphi(i)^T\vartheta}{\lambda}\right)^2}\right]$$

$$= \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\lambda^2}} \cdot e^{-\frac{1}{2}\left(\frac{y(i) - \varphi(i)^T\vartheta}{\lambda^2}\right)^2}\right]$$

$$= \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\lambda^2}}\right] + \ln\left[e^{-\frac{1}{2}\left(\frac{y(i) - \varphi(i)^T\vartheta}{\lambda^2}\right)^2}\right]$$

$$= \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\lambda^2}}\right] - \frac{1}{2}\left(\frac{y(i) - \varphi(i)^T\vartheta}{\lambda^2}\right)^2$$

$$= \sum_{i=1}^{N} \ln\left[\frac{1}{\sqrt{2\pi\lambda^2}}\right] - \frac{1}{2}\sum_{i=1}^{N}\left(\frac{y(i) - \varphi(i)^T\vartheta}{\lambda^2}\right)^2$$

$$= \sum_{i=1}^{N} \ln \frac{1}{\sqrt{2\pi\lambda^2}} + \sum_{i=1}^{N} \ln \left[e^{-\frac{1}{2}\left(\frac{y_{(i)} - \varphi_{(i)}^T \vartheta}{\lambda} \right)^2} \right]$$

$$= N \cdot \ln \left(2\pi\lambda^2 \right)^{-\frac{1}{2}} + \sum_{i=1}^{N} -\frac{1}{2} \left(\frac{y_{(i)} - \varphi_{(i)}^T \vartheta}{\lambda} \right)^2 \ln e^1$$

$$= -\frac{1}{2} N \ln \left[2\pi\lambda^2 \right] - \frac{1}{2\lambda^2} \sum_{i=1}^{N} \left(y_{(i)} - \varphi_{(i)}^T \vartheta \right)^2$$

2) Calcolo del massimo = minimizzare lo $-\ln \left[\mathcal{L}(\vartheta | X, Y) \right]$

$$-\ln \left[\mathcal{L}(\vartheta | X, Y) \right] = +\frac{1}{2} N \ln \left[2\pi\lambda^2 \right] + \frac{1}{2\lambda^2} \sum_{i=1}^{N} \left(y_{(i)} - \varphi_{(i)}^T \vartheta \right)^2$$

$$\Rightarrow \hat{\vartheta}_{ML} = \arg\min_{\vartheta} \frac{1}{2\lambda^2} \sum_{i=1}^{N} \left(y_{(i)} - \varphi_{(i)}^T \vartheta \right)^2$$

$\frac{1}{N}$

Lo stimo ML del modello $y_{(i)} = \varphi_{(i)}^T \vartheta + e_{(i)}$, $i = 1 \dots N$, dove $e_{(i)} \sim N(0, \lambda^2)$ i.i.d, è equivalente alla stima LS

⇓

REGRESSIONE LINEARE

REGRESSIONE LOGISTICA 5

REGRESSIONE LINEARE: modello dei metrici → $y \in \mathbb{R}$

REGRESSIONE LOGISTICA: modello dei binari/dicotomici (categorici)

↳ $y \in \{0, 1\}$

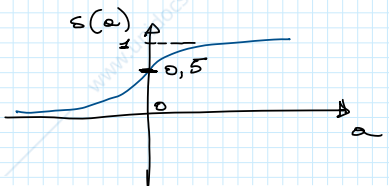
In questo caso, non ha senso utilizzare un modello lineare del tipo $y_{(i)} = \varphi_{(i)}^T \vartheta + e_{(i)}$

- non ha senso sommare un errore continuo ($\in \mathbb{R}$) ad una variabile y che può assumere solo 2 valori
- il modello potrebbe assumere valori < 0 e > 1. Non c'è niente che ci "limita" l'uscita tra 0 e 1

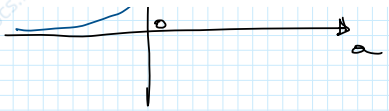
⇓

FUNZIONE LOGISTICA (SIGMOIDE)

$$s(a) = \frac{1}{1 + e^{-a}} = \frac{e^a}{1 + e^a}$$



- se $a \gg 0 \Rightarrow s(a) \to 1$



- se $a \gg 0 \Rightarrow s(a) \to 1$
- se $a \ll 0 \Rightarrow s(a) \to 0$

OBIETTIVO. modellare la probabilità che $y = 1$ tramite un modello lineare

$$P(y=1 \mid \underset{d\times 1}{\varphi}) = s(\varphi^T \theta) = \frac{1}{1 + e^{-(\varphi^T \theta)}}$$

- se $\varphi^T \theta \gg 0 \Rightarrow P(y=1 \mid \varphi) \to 1$
- se $\varphi^T \theta \ll 0 \Rightarrow P(y=1 \mid \varphi) \to 0$

$$\Rightarrow \begin{cases} y=1 & P(y=1) = \pi \\ y=0 & P(y=0) = 1-\pi \end{cases} \Rightarrow \pi^y \cdot (1-\pi)^{(1-y)}$$

REGRESSIONE LINEARE

$y \sim N(\mu, \lambda^2)$

$\mu = \underset{1\times d}{\varphi^T} \cdot \underset{d\times 1}{\theta} = \theta_0 + \theta_1 x_1 + \dots + \theta_{d-1} x_{d-1}$

REGRESSIONE LOGISTICA

$y \sim \text{Bernoulli}(\pi)$

$\pi = s(\underset{1\times d}{\varphi^T} \underset{d\times 1}{\theta})$

STIMA ML DI UN MODELLO DI REGRESSIONE LOGISTICA

Sia dato un dataset $D = \{(\varphi(1), y(1)), (\varphi(2), y(2)), \dots, (\varphi(N), y(N))\}$

$\varphi(i) \in \mathbb{R}^{d\times 1}$, $y(i) \in \{0, 1\}$, $i = 1 \dots N$, (i.i.d)

Stimare un modello di regressione logistica

$$P(y(i)=1 \mid \varphi(i)) = \frac{1}{1 + e^{-\varphi(i)^T \theta}} \equiv \pi(i)$$

1) Calcolo la verosimiglianza dei dati

$$\mathcal{L}(\pi \mid Y) = \prod_{i=1}^{N} \pi(i)^{y(i)} \cdot (1-\pi(i))^{1-y(i)} \Rightarrow \text{Calcolo la } \underline{\text{neg}}\text{-log-likelihood}$$

$$-\ln[\mathcal{L}(\pi \mid Y)] = -\ln\left[\prod_{i=1}^{N} \pi(i)^{y(i)} \cdot (1-\pi(i))^{1-y(i)}\right]$$

$$= -\sum_{i=1}^{N} \ln\left[\pi(i)^{y(i)} \cdot (1-\pi(i))^{1-y(i)}\right]$$

$$= -\sum_{i=1}^{N} \left(\ln[\pi(i)^{y(i)}] + \ln[1-\pi(i)]^{1-y(i)}\right)$$

$$= -\sum_{i=1}^{N} \left(y(i) \ln \pi(i) + (1-y(i)) \ln[1-\pi(i)]\right) = J(\theta)$$

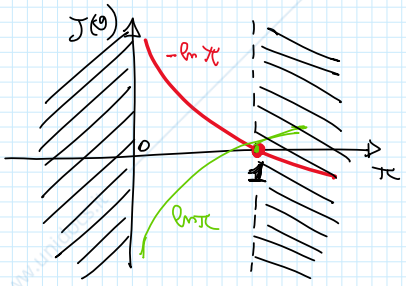
2) Interpretazione della funzione di costo

Supponiamo di avere un set di dati $D = \{(\varphi, y)\}$

$$J(\vartheta) = \begin{cases} -\ln \pi & \text{se } y = 1 \\ -\ln[1-\pi] & \text{se } y = 0 \end{cases}$$

CASO $y = 1$

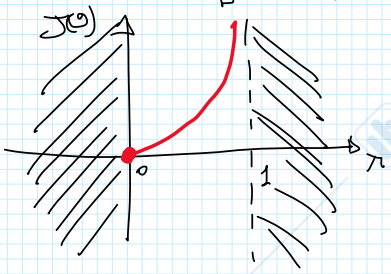
$J(\vartheta) = -\ln \pi \Rightarrow J(\vartheta) \to 0$ SE $y = 1$ && $\pi \to 1$



$J(\vartheta) \to +\infty$ SE $y = 1$ && $\pi \to 0$

CASO $y = 0$

$J(\vartheta) = -\ln[1-\pi]$



$J(\vartheta) \to 0$ SE $y = 0$ && $\pi \to 0$

$J(\vartheta) \to +\infty$ SE $y = 0$ && $\pi \to 1$

CALCOLO DEL MINIMO

$\nabla_\vartheta J(\vartheta) = 0$

$$\nabla_\vartheta J(\vartheta) = \sum_{i=1}^{N} \varphi(i) \cdot \left(\pi(i) - y(i) \right)$$

$\pi(i) = \frac{1}{1 + e^{-\varphi(i)^T \vartheta}}$

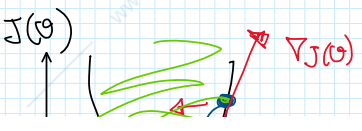
$$\frac{d\, s(a)}{da} = s(a) \cdot [1 - s(a)]$$

$$\frac{d\, s(\varphi^T \vartheta)}{d\vartheta} = \nabla_\vartheta\, s(\varphi^T \vartheta) = \varphi \cdot \pi \cdot (1 - \pi)$$

OTTIMIZZAZIONE TRAMITE GRADIENT DESCENT

$$\hat{\vartheta}(j+1) = \hat{\vartheta}(j) - \alpha \cdot \nabla_\vartheta J(\vartheta) \Big|_{\vartheta = \hat{\vartheta}(j)}$$

- α: LEARNING RATE
- $\hat{\vartheta}(0)$ è inizializzato RANDOM



- se $\nabla J(\vartheta)\Big|_{\vartheta=\hat{\vartheta}(j)} > 0 \Rightarrow \hat{\vartheta}(j+1) < \hat{\vartheta}(j)$

- se $\nabla J(\vartheta)\Big|_{\vartheta=\hat{\vartheta}(j)} < 0 \Rightarrow \hat{\vartheta}(j+1) > \hat{\vartheta}(j)$