ESERCIZI

DEMO SESSION 2022/2023

5) – $h_\theta(x) = \theta_0 + \theta_1 x$ $\alpha = 0{,}01$ APPLICAZIONE SGD

$(\theta_0, \theta_1) = (0,0)$

$(x^*, y^*) = (5,2)$

– $y_{PRED} = h_\theta(x^*) = 0 + 0 \cdot 5 = 0$

$e = y_{PRED} - y = 0 - 2 = -2$

$$\frac{d}{d\theta_0} J(\theta) = \frac{d}{d\theta_0}\left[\frac{1}{2}(h_\theta(x^*) - y)^2\right] = \frac{1}{\cancel{2}}\cancel{2}\left[h_\theta(x^*) - y\right]\frac{d}{d\theta_0} h_\theta(x^*) = (0-2)\cdot 1 = -2$$

$$\frac{d}{d\theta_1} J(\theta) = (0-2)\cdot x^* = -10$$

$$\theta_0 := \theta_0 - \alpha \frac{d}{d\theta_0} J(\theta) = 0 - 0.01 \cdot (-2) = 0{,}02$$

$$\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta) = 0 - 0.01 \cdot (-10) = 0{,}1$$

6) ESERCIZIO PCA

– DATASET $X = \begin{pmatrix} 1 & 5 & 3 \\ 3 & 4 & 8 \\ 5 & 1 & 6 \end{pmatrix}$, MATRICE DELLA COVARIANZA $\Sigma = \begin{pmatrix} 4 & 1 & -4 \\ 1 & 7 & 3.5 \\ -4 & 3.5 & 7 \end{pmatrix}$

MATRICE AUTOVETTORI $U = \begin{pmatrix} -0{,}34517975 & 0{,}56157965 & 0{,}76509614 \\ -0{,}62904883 & -0{,}73898939 & 0{,}24123107 \\ -0{,}69652603 & 0{,}39801488 & 0{,}59702231 \end{pmatrix}$

MATRICE AUTOVALORI $S = (\underset{\lambda_1}{11{,}2912878} \quad \underset{\lambda_2}{6{,}70871215} \quad \underset{\lambda_3}{3{,}91717725 10 \cdot 10^{-16}})$

– VARIANZA TOTALE $= \lambda_1 + \lambda_2 + \lambda_3 \simeq 18$

– CALCOLI RAPPORTI DI VARIANZA, MANTENENDO IL 60% DEL DATASET ORIGINALE

Devo scegliere un valore di K t.c. si abbia il 60 % del dataset originale:

Primo $\underline{K=1}$ $\frac{\lambda_1}{\text{VAR.TOT.}} \simeq 0.63 \geq 0.6$

Primo $K=2$ $\frac{\lambda_1 + \lambda_2}{\text{VAR.TOT.}} \simeq 1$

} scelgo K=1 perché mi basta per mantenere il 60% del dataset, avessi ottenuto un val ≤ 0.6 avrei dovuto considerare K=2.

– SCELTA DEL NUM. FINITO DI COMPONENTI NECESSARIE

Ho scelto k=1 perché mi permette di mantenere la % del dataset, quindi la prima colonna

$$U_{REDUCE} = \begin{pmatrix} -0{,}34517975 \\ -0{,}62904883 \\ -0{,}69652603 \end{pmatrix}$$

$$* \left[(v_1, v_2, v_3) \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = (v_1 a_{11} + v_2 a_{21} + v_3 a_{31},\ v_1 a_{12} + v_2 a_{22} + v_3 a_{32},\ v_1 a_{13} + v_2 a_{23} + v_3 a_{33}) \right]$$

– PROIEZIONE DEL DATASET ORIGINALE SU U_{REDUCE}

$$z = U_{REDUCE}^T \cdot X = (-0.34517975, -0.62904883, -0.69652603) \begin{pmatrix} 1 & 5 & 3 \\ 3 & 4 & 8 \\ 5 & 1 & 6 \end{pmatrix}^* = (-5.71, -4.94, -10{,}25)$$

$$* \; [u_1, u_2, u_3] \cdot \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & u_1v_3 \\ u_2v_1 & u_2v_2 & u_2v_3 \\ u_3v_1 & u_3v_2 & u_3v_3 \end{bmatrix}$$

· RICOSTRUZIONE DATI ORIGINALI

$$\tilde{X} = U_{REDUCE} \cdot z = \begin{pmatrix} -0,34517975 \\ -0,62904883 \\ -0,69652603 \end{pmatrix} (-5.71, -4.94, -10.25) \overset{*}{=} \begin{pmatrix} 1.97 & 1,71 & 3.54 \\ 3.59 & 3,10 & 6.45 \\ 3.59 & 3.44 & 7.14 \end{pmatrix} \simeq \begin{pmatrix} 2 & 2 & 4 \\ 4 & 3 & 6 \\ 4 & 3 & 7 \end{pmatrix}$$

DA CAPIRE SE HA SENSO

ESERCIZIO ROC (ATSOC)

1) Avendo dei dati come questi:

Esempio	Etichetta Reale	Probabilità Predetta
1	Positivo	0.95
2	Negativo	0.85
3	Positivo	0.80
4	Negativo	0.70
5	Positivo	0.65
6	Negativo	0.60
7	Positivo	0.40
8	Negativo	0.35
9	Positivo	0.20
10	Negativo	0.10

ricaviamo TPR (recall) e FPR (1 - specificity), impostando NOI le soglie:

$$TPR = \frac{TP}{TP + FN} \qquad e \qquad FPR = \frac{FP}{TN + FP}$$

Threshold = 0.45 ⇒ il modello considera positivi tutti i valori sopra la (o uguale) soglia di 0.45:

Ricaviamo da questo TP = 3 (Es. 1, 3 e 5) FP = 3 (Es. 2, 4, 6) TN = 2 (Es. 8 e 10) FN = 2 (7 e 9)

⇒ TPR = 0.6 FPR = 0.6

Threshold = 0.85 ⇒ TP = 1 FP = 1 TN = 4 FN = 4 ⇒ TPR = 0.25 FPR = 0.25

ES. 1 STOCHASTIC

$$h_\theta(x) = \theta_0 + \theta_1 x \qquad (x_1, y_1) = (5,2) \quad (x_2, y_2) = (6,3) \quad \alpha = 0{,}01$$

$$(\theta_0, \theta_1) = (0,0)$$

$$h_\theta(x_1) = 0 \qquad e_1 = 0 - 2 = -2$$

$$\frac{d}{d\theta_0} J(\theta) = \frac{1 \cancel{2}}{\cancel{2}} \left(h_\theta(x_1) - y_1 \right) \frac{d}{d\theta_0} h_\theta(x_1) = (-2) \cdot 1 = -2$$

$$\frac{d}{d\theta_1} J(\theta) = (-2) \cdot 5 = -10 \qquad \theta_0 := \theta_0 - \alpha(-2) = -0{,}02 \qquad \theta_1 := \theta_1 - \alpha(-10) = -0.1$$

$$h_\theta(x_2) = \theta_0 + \theta_1 x_2 = -0.02 + 0.01 \cdot 6 = 0.04 \qquad e_2 = 0.04 - 3 = -2.96$$

$$\frac{d}{d\theta_0} J(\theta) = (-2.96) \cdot 1 = -2.96 \qquad \frac{d}{d\theta_1} J(\theta) = (-2.96) \cdot 6 = -17.76$$

$$\theta_0 := \theta_0 - \alpha(-2.96) = -0.02 + 0.0296 = 9.6 \cdot 10^{-3} \qquad \theta_1 := \theta_1 - \alpha(-17.76) = -0.1 + 0.1776 = 0.0776$$

BATCH CON STESSA FUNZIONE E DATASET

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x_i) - y_i \right)^2$$

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^i) - y^i\right) x_0^i$$

$e_1 = -2 \qquad e_2 = -3$

$$\theta_0 := \theta_0 - \frac{\alpha}{m}(-2-3) = 0 - \frac{0.01}{2} \cdot -5 = 0.025$$

$$\theta_1 := \theta_1 - \frac{\alpha}{m}(-2\ 5 - 3\cdot 6) = 0 - \frac{0.01}{2} \cdot (-21) = 0.14$$

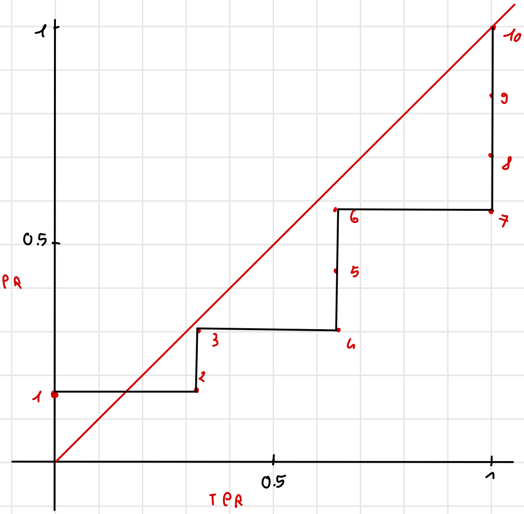
ESERCIZIO ROC

0 = POS 1 = NEG

Campione	Classe Reale	Probabilità Predetta
1	0	0.90
2	1	0.85
3	0	0.75
4	1	0.70
5	0	0.60
6	0	0.55
7	1	0.50
8	0	0.45
9	0	0.30
10	0	0.10

$$TPR = \frac{TP}{TP + FN = 7} \qquad FPR = \frac{FP}{FP + TN = 3}$$

TRESHOLD	FPR	TPR	N°
0.90	0.14	0	1
0.85	0.14	0.33	2
0.75	0.29	0.33	3
0.70	0.29	0.66	4
0.60	0.42	0.66	5
0.55	0.57	0.66	6
0.50	0.57	1	7
0.45	0.71	1	8
0.30	0.85	1	9
0.10	1	1	10



ESERCIZIO QUARTILI ASTOC

2, 4, 5, | 7, 8, 9, | 10, 12, 14, | 15, 30, 35

Q_1 Q_2 Q_3

sono già ordinati, se non lo sono ordinarli. Bisogna trovare la mediana

$$Q_1 = \frac{5+7}{2} = 6 \qquad Q_2 = \frac{9+10}{2} = 9.5 \qquad Q_3 = \frac{14+15}{2} = 14.5$$

$IQR = Q_3 - Q_1 = 8.5$

$Q_1 - IQR \cdot 1.5 = 6 - 12.75 = -6.75$

$Q_3 + IQR \cdot 1.5 = 14.5 + 8.5 \cdot 1.5 = 27.25$

Scarto 30 e 35.

$\Rightarrow 2, 4, 5, 7, 8, | 9, 10, 12, 14, 15$

Q_1 Q_2 Q_3

$Q_2 = \frac{8+9}{2} = 8.5$

$IQR = Q_3 - Q_1 = 7$

$Q_1 - IQR \cdot 1.5 = -5.5$

$Q_3 + IQR \cdot 1.5 = 22.5$ fine $\Rightarrow$ NO OUTLIERS

ESERCIZIO CLASSIFICATION TREE DEMO EXAM

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
~~D1~~	~~Sunny~~	~~Hot~~	~~High~~	~~Weak~~	~~No~~
~~D2~~	~~Sunny~~	~~Hot~~	~~High~~	~~Strong~~	~~No~~
~~D3~~	~~Overcast~~	~~Hot~~	~~High~~	~~Weak~~	~~Yes~~
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
~~D7~~	~~Overcast~~	~~Cool~~	~~Normal~~	~~Strong~~	~~Yes~~
~~D8~~	~~Sunny~~	~~Mild~~	~~High~~	~~Weak~~	~~No~~
~~D9~~	~~Sunny~~	~~Cool~~	~~Normal~~	~~Weak~~	~~Yes~~
D10	Rain	Mild	Normal	Weak	Yes
~~D11~~	~~Sunny~~	~~Mild~~	~~Normal~~	~~Strong~~	~~Yes~~
~~D12~~	~~Overcast~~	~~Mild~~	~~High~~	~~Strong~~	~~Yes~~
~~D13~~	~~Overcast~~	~~Hot~~	~~Normal~~	~~Weak~~	~~Yes~~
D14	Rain	Mild	High	Strong	No

Costruisco l'albero di classificazione da questo. Partiamo dalla formula generale dell'information gain per capire quale feature ha più impatto.

ENTROPIA: $H\left(\frac{T}{T+F}, \frac{F}{T+F}\right) = -\frac{T}{T+F}\log_2\frac{T}{T+F} - \frac{F}{T+F}\log_2\frac{F}{T+F}$

dove con T ed F si intendono le righe il cui valore target finale è TRUE e FALSE rispettivamente (utilizzabile sia per l'entropia totale, sia per quella di una singola feature, in quel caso si considerano solo le righe che hanno un determinato valore per quella feature)

RESIDUO: $R(\text{feature}) = \sum_{i=1}^{v} \frac{T_i + F_i}{T+F} H\left(\frac{T_i}{T_i+F_i}, \frac{F_i}{T_i+F_i}\right)$

dove con v si intende il numero di valori che assume una feature, T_i ed F_i: valori veri e falsi solo delle righe in cui la feature assume un determinato valore

GUADAGNO INFORMATIVO $IG(\text{feature}) = H\left(\frac{T}{T+F}, \frac{F}{T+F}\right) - R(\text{feature})$

Per scegliere subito bene il primo attributo vediamo quale permette una buona distinzione fra True e False, cioè se per esempio l'attributo ha 3 valori possibili, ad uno è associato sempre TRUE, ad un altro sempre FALSE, è una buona possibilità che sia l'attributo migliore.

$H = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.41 + 0.53 = 0.94$

$IG(OUTLOOK) = 0.94 - \left[\frac{5}{14} H\left(\frac{2}{5}, \frac{3}{5}\right) + \frac{4}{14} H(1, 0) + \frac{5}{14} H\left(\frac{3}{5}, \frac{2}{5}\right)\right] = 1 - \left[\frac{5}{14}\left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{4}{14}\left(-1\log_2 1\right) + \frac{5}{14}\left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)\right]$

$= 0.94 - \left[\frac{5}{14}(0.53 + 0.44) + \frac{5}{14}(0.44 + 0.53)\right] = 0.94 - 0.69 = 0.25$

$IG(TEMP) = 0.94 - \left[\frac{4}{14}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{6}{14}\left(-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right) + \frac{4}{14}\left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right)\right] = 0.94 - \left[\frac{4}{14} + \frac{6}{14}(0.39 + 0.53) + \frac{4}{14}(0.31 + 0.5)\right] =$

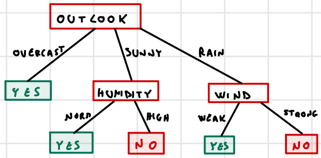
$= 0.94 - (0.29 + 0.39 + 0.23) = 0.94 - 0.91 = 0.03$

PROVA DI ALBERO A SENTIMENTO

1) 1/3 - 4 0/3 0/3 0/2

2) a. 2/3 - 3 2/2 0/2

b. 0/2 0/2 2/2



DOMANDE DI TEORIA

1. SPIEGARE HOLD-OUT, K-FOLD CROSS VALIDATION, RANDOM SAMPLING
2. VANTAGGI E SVANTAGGI DELLA REGOLARIZZAZIONE CON L1 E L2.
3. SPIEGARE L'USO DEI QUARTILI E ESERCIZIO SEMPLICE.
4. ESERCIZIO DI AGGIORNAMENTO DEI THETA NEL CASO FULL BATCH E STOCHASTIC.
5. ESERCIZIO SUL ROC.
6. DATI 3 GRAFICI DISEGNARE E SPIEGARE LE LEARNING CURVE.
7. SPIEGARE COSA SONO LE LEARNING CURVE E I LORO COLLEGAMENTI CON BIAS E VARIANZA.
8. PSEUDO CODICE BATCH, STOCHASTIC E MINIBATCH.
9. SPIEGARE SUPERVISED E UNSUPERVISED CON ESEMPI.
10. MECCANISMO ONE VS ALL E ALL VS ALL.

BATCH GRADIENT DISCENT

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

PSEUDO CODICE

```
theta = rand
while (not convergence) {
    for (j from 1 to n) {
        update = 0
        for (i from 1 to m) {
            update = update + (h(x[i]) - y[i]) * x[i]
        }
        theta_new[j] = theta[j] - α * update/m
    }
    theta = theta_new
}
```

STOCHASTIC GRADIENT DISCENT

$$\theta_j = \theta_j - \alpha (h(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

PSEUDO CODICE

```
theta = rand()
while (not convergence) {
    for (i from 1 to m) {
        for (j from 1 to n) {
            theta_new[j] = theta[j] - α * (h(x[i]) - y[i]) * 1/m * x[i]
        }
    theta = theta_new
}
```

MINI-BATCH GRADIENT DISCENT

PSEUDO CODICE

```
b = 50
while (not convergence) {
    while (i < m) {
        for (j from 1 to n) {
            update = 0
            for (z from i to i+b) {
                update = update + (h(x[z]) - y[z]) * x[z]
            }
            theta[j] = theta[j] - α * update/b
        }
        i = i + b;
    }
}
```

NORMAL EQUATIONS DERIVATIVES

TESI: $\nabla J(\theta) = X^T X \theta - X^T y = 0 \Rightarrow \hat{\theta} = (X^T X)^{-1} \cdot X^T y$

DIM: $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2$ vett: $J(\theta) = \frac{1}{2} (X\theta - y)^T (X\theta - y)$

$$(X\theta - y) = \begin{pmatrix} h(x^{(1)}) - y^{(1)} \\ \dots \\ h(x^{(m)}) - y^{(m)} \end{pmatrix}$$

sappiamo che $(X\theta - y)^T (X\theta - y) = \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})^2$

scal. $\frac{d}{d\theta_k} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)}) x_k^{(i)}$ vett: $\nabla J(\theta) = (h(x^{(1)}) - y^{(1)}, \dots, h(x^{(m)}) - y^{(m)}) \begin{pmatrix} x_k^{(1)} \\ \vdots \\ x_k^{(m)} \end{pmatrix}$

$$\Rightarrow \nabla J(\theta) = \begin{pmatrix} \sum_{i=1}^{m} h(x^{(i)}) - y^{(i)} \\ \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})\, x_1^{(i)} \\ \vdots \\ \sum_{i=1}^{m} (h(x^{(i)}) - y^{(i)})\, x_n^{(i)} \end{pmatrix} \quad \Rightarrow \quad X^T X \theta - X^T y = 0 \quad \Rightarrow \quad \hat{\theta} = (X^T X)^{-1} + X^T y$$

APPROCCIO PROBABILISTICO DELLA REGRESSIONE LINEARE

TESI: $y^{(i)} = h(x^{(i)}) + e^{(i)}$, $\max L(\theta) = \min J(\theta)$

DIM: $p(e^{(i)}) = N(0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(e^{(i)})^2}{2\sigma^2}}$, dato $e^{(i)} = y^{(i)} - h(x^{(i)})$

$$\Rightarrow p(y^{(i)} \mid x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^{(i)} - h(x^{(i)}))^2}{2\sigma^2}}$$

$$L(\theta) = L(\theta; X; y) = p(y \mid X; \theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^{(i)} - h(x^{(i)}))^2}{2\sigma^2}}$$

$$\log L(\theta) = \ell(\theta)$$

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \log L(\theta) = \arg\max_{\theta} \ell(\theta) = \arg\max_{\theta} \log \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^{(i)} - h(x^{(i)}))^2}{2\sigma^2}} =$$

$$= \max \sum_{i=1}^{m} \log \left[\frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(y^{(i)} - h(x^{(i)}))^2}{2\sigma^2}} \right] = \max \frac{1}{m} \log \frac{1}{\sqrt{2\pi\sigma}} - \quad \text{ecc...}$$

AGGIORNAMENTO PESI BACKPROPAGATION

calcolo gradiente stato di output: $\delta_k^{(L)} = \frac{dJ(\theta)}{da_k^{(L)}} = (a_k^L - y)\, g'(z_k^{(L)})$

calcolo il grad per un neurone in uno stato precedente come somma degli errori degli stati successivi:

$$\delta_j^{(l)} = \sum_k (\theta_{kj}^{(l)} \delta_k^{(l+1)})\, g'(z_j^{(l)})$$

Da questo approccio con l'uso di ciascun neurone influenza sul precedente a ottenere.

$$\frac{dJ(\theta)}{d\theta_{ji}} = \delta_i^{(l+1)} a_j^{(l)}$$

$$\delta_k^{(L)} = \frac{dJ(\theta)}{da_k^L} = (a_k^L - y)\, g'(z_k^{(L)})$$

$$\delta_j^{(l)} = \sum_k (\theta_{kj}^{(l)} \delta_k^{(l+1)})\, g'(z_j^{(l)})$$

$$\frac{dJ(\theta)}{\theta_{kj}} = \delta_k^{(l+1)} a_j^{(l)}$$

$$\theta_{kj} := \theta_{kj} - \alpha \frac{dJ(\theta)}{\theta_{kj}}$$