Algoritmi Avanzati, A.A. 2016–2017

Seconda prova scritta — Traccia della soluzione

Martedì 7 febbraio 2017

Esercizio 1

1.1) Descrivere brevemente il metodo della K-fold Cross Validation. Quali sono i vantaggi rispetto ad un partizionamento con un solo training set ed un solo validation set?
1.2) Definire i concetti di feature numeriche (continue e discrete) e feature categoriche. In che modo è possibile passare da feature categoriche a feature numeriche? Quando questa operazione si rende necessaria?

Soluzione 1

1.1) Vedere la sezione 2.2.2 delle dispense. Il principale vantaggio del metodo è la riduzione della varianza dello score di validazione, continuando allo stesso tempo ad usare quasi l'intero dataset per il training ad ogni iterazione.
1.2) Vedere la sezione 3.1.1 delle dispense. Le feature numeriche sono caratterizzate da valori continui o discreti in $\mathbb{R}^n$ (o in un qualunque spazio dotato dei concetti di ordine e distanza) mentre le feature categoriche sono rappresentate da etichette alfanumeriche per le quali non esiste il concetto di ordine. Nel caso di algoritmi che fanno uso di distanze definite su $\mathbb{R}^n$ (ad esempio K-Nearest Neighbors) ed in presenza di feature eterogenee, un possibile approccio è la sostituzione di una feature categorica con tante feature numeriche quante sono le categorie, utilizzando una rappresentazione unaria per deciderne i valori.

Esercizio 2

Sia dato il seguente dataset di $m = 6$ campioni, con 1 feature numerica x e 1 output numerico y:

i	1	2	3	4	5	6
x	1	2	-2	0	-1	0
y	3	8	0	15	-1	0

2.1) Si richiede di apprendere il modello $y \sim \beta\phi(x)$ attraverso il metodo della regressione lineare ai minimi quadrati, applicato al dataset risultante dalla trasformazione non lineare $\phi(x) = x^2$.
2.2) Ripetere con il modello $y \sim \beta_1\phi_1(x) + \beta_2\phi_2(x)$ e le trasformazioni $\phi_1(x) = x^2$ e $\phi_2(x) = x$. Per quale dei due modelli si ottiene un valore di RMSE più basso?
Suggerimento: per rispondere all'ultima domanda non occorre calcolare esplicitamente l'RMSE con radice quadrata e divisione.

Soluzione 2

Vedere la sezione 3.2.1 delle dispense.

2.1) Costruiamo il vettore z in modo che $z_i = x_i^2$:

$$z = \begin{pmatrix} 1 \\ 4 \\ 4 \\ 0 \\ 1 \\ 0 \end{pmatrix}.$$

Considerando z e y come vettori colonna, il parametro β si calcola come segue:

$$\beta = (z^T z)^{-1} z^T y = 1.$$

La soluzione corrisponde alla formula specifica per il caso unidimensionale:

$$\beta = \frac{\sum_{i=1}^{6} y_i z_i}{\sum_{i=1}^{6} z_i^2} = \frac{34}{34} = 1.$$

Alternativa: applicazione diretta del metodo dei minimi quadrati — Consideriamo la somma dei quadrati degli scarti in funzione di β:

$$f(\beta) = (\beta - 3)^2 + (4\beta - 8)^2 + (4\beta)^2 + (-15)^2 + (\beta + 1)^2,$$

e cerchiamone il valore stazionario azzerandone la derivata prima rispetto a β:

$$0 = \frac{\mathrm{d}f(\beta)}{\mathrm{d}\beta} = 2(\beta - 3) + 2 \cdot 4(4\beta - 8) + 2 \cdot 4(4\beta) + 2(\beta + 1) = 68\beta - 68,$$

la cui soluzione è nuovamente $\beta = 1$.
2.2) In questo caso costruiamo la matrice Z in modo che $z_{i1} = x_i^2$ e $z_{i2} = x_i$:

$$Z = \begin{pmatrix} 1 & 1 \\ 4 & 2 \\ 4 & -2 \\ 0 & 0 \\ 1 & -1 \\ 0 & 0 \end{pmatrix}.$$

I parametri β_1 e β_2 si calcolano come segue:

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = (Z^T Z)^{-1} Z^T y = \begin{pmatrix} 1 \\ 2 \end{pmatrix}.$$

Alternativa: applicazione diretta del metodo dei minimi quadrati — Consideriamo la somma dei quadrati degli scarti in funzione di β_1 e β_2:

$$f(\beta_1, \beta_2) = (\beta_1 + \beta_2 - 3)^2 + (4\beta_1 + 2\beta_2 - 8)^2 + (4\beta_1 - 2\beta_2)^2 + (-15)^2 + (\beta_1 - \beta_2 + 1)^2,$$

e cerchiamone il valore stazionario azzerandone le derivate parziali rispetto ai coefficienti:

$$0 = \frac{\partial f(\beta_1, \beta_2)}{\partial \beta_1} = 2(\beta_1 + \beta_2 - 3) + 2 \cdot 4(4\beta_1 + 2\beta_2 - 8) + 2 \cdot 4(4\beta_1 - 2\beta_2) + 2(\beta_1 - \beta_2 + 1) = 68\beta_1 - 68,$$

la cui soluzione è nuovamente $\beta_1 = 1$ (si noti che β_2 si semplifica), e

$$0 = \frac{\partial f(\beta_1, \beta_2)}{\partial \beta_2} = 2(\beta_1 + \beta_2 - 3) + 2 \cdot 2(4\beta_1 + 2\beta_2 - 8) - 2 \cdot 2(4\beta_1 - 2\beta_2) - 2(\beta_1 - \beta_2 + 1) = 20\beta_2 - 40,$$

per cui $\beta_2 = 2$.
L'RMSE per il primo modello è calcolato come segue:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{m} (\beta x_i^2 - y_i)^2}{m}} = \sqrt{\frac{265}{6}},$$

mentre per il secondo modello:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{m} (\beta_1 x_i^2 + \beta_2 x_i - y_i)^2}{m}} = \sqrt{\frac{225}{6}}.$$

È evidente che nel secondo caso l'errore è minore, quindi il secondo modello descrive meglio la relazione che intercorre tra i punti del dataset.

Esercizio 3

Sia dato il seguente dataset di $m = 4$ campioni, con 3 feature discrete x_1, x_2, x_3 e 1 variabile di classe y a valori binari:

i	1	2	3	4
x_1	0	1	2	3
x_2	0	1	0	1
x_3	0	0	0	1
y	0	0	1	1

3.1) Definire il concetto di entropia di una variabile aleatoria discreta. Calcolare il valore in bit di $H(y)$, usando le frequenze relative per stimare la distribuzione di probabilità discreta.
3.2) Definire il concetto di informazione mutua tra due variabili aleatorie discrete. Indicare esplicitamente anche la relazione che intercorre tra informazione mutua, entropia ed entropia condizionata.
3.3) Calcolare $I(x_1; y)$, $I(x_2; y)$ e $I(x_3; y)$, usando le frequenze relative per stimare le probabilità degli eventi congiunti e marginali. Semplificare le espressioni ottenute senza calcolare esplicitamente i logaritmi.

Soluzione 3

3.1) L'entropia è definita dalla formula (3.8) delle dispense. Per calcolare $H(y)$, notiamo che y assume i valori 0 e 1 con $p_0 = p_1 = \frac{1}{2}$. Ne segue:

$$H(y) = -p_0 \log_2(p_0) - p_1 \log_2(p_1) = \frac{1}{2}\log_2(2) + \frac{1}{2}\log_2(2) = \log_2(2) = 1 \text{ bit.}$$

3.2) La definizione di informazione mutua corrisponde alla formula (4.5) delle dispense. La relazione tra informazione mutua, entropia ed entropia condizionata corrisponde invece alla formula (4.4).
3.3) Indichiamo con $p_{i,j}$ la probabilità che la feature in esame assuma il valore i e che la classe sia uguale a j. $p_{i,-}$ e $p_{-,j}$ sono le probabilità marginali rispettivamente della feature e della classe. Otteniamo quindi le seguenti probabilità marginali:

x_1	$p_{0,-} = 1/4$	$p_{1,-} = 1/4$	$p_{2,-} = 1/4$	$p_{3,-} = 1/4$
x_2	$p_{0,-} = 1/2$	$p_{1,-} = 1/2$		
x_3	$p_{0,-} = 3/4$	$p_{1,-} = 1/4$		
y	$p_{-,0} = 1/2$	$p_{-,1} = 1/2$		

e le seguenti probabilità congiunte:

x_1	$p_{0,0} = 1/4$	$p_{1,0} = 1/4$	$p_{2,0} = 0$	$p_{3,0} = 0$
	$p_{0,1} = 0$	$p_{1,1} = 0$	$p_{2,1} = 1/4$	$p_{3,1} = 1/4$
x_2	$p_{0,0} = 1/4$	$p_{1,0} = 1/4$		
	$p_{0,1} = 1/4$	$p_{1,1} = 1/4$		
x_3	$p_{0,0} = 1/2$	$p_{1,0} = 0$		
	$p_{0,1} = 1/4$	$p_{1,1} = 1/4$		

Usando la formula (4.5) delle dispense, otteniamo:

$$I(x_1; y) = p_{0,0} \log_2 \frac{p_{0,0}}{p_{0,-} \cdot p_{-,0}} + p_{1,0} \log_2 \frac{p_{1,0}}{p_{1,-} \cdot p_{-,0}} + p_{2,1} \log_2 \frac{p_{2,1}}{p_{2,-} \cdot p_{-,1}} + p_{3,1} \log_2 \frac{p_{3,1}}{p_{3,-} \cdot p_{-,1}} = \log_2(2) = 1.$$

In alternativa, senza usare formule, possiamo osservare dalla tabella che la conoscenza di x_1 comporta la conoscenza di y, quindi l'informazione mutua è massima (pari all'entropia di y).
Procedendo in modo simile otteniamo anche:

$$I(x_2; y) = \log_2(1) = 0.$$

Oppure possiamo osservare che, qualunque sia il valore di x_2, y resta uniformemente distribuita, quindi la sua entropia non cambia. Infine:

$$I(x_3; y) = \frac{1}{2}\log_2\left(\frac{4}{3}\right) + \frac{1}{4}\log_2(2) + \frac{1}{4}\log_2\left(\frac{2}{3}\right) = \frac{3}{2} - \frac{3}{4}\log_2(3).$$

Esercizio 4

Sia dato il seguente dataset di $m = 10$ campioni, con $n = 2$ feature numeriche X_1, X_2 e un output Y categorico (binario):

i	1	2	3	4	5	6	7	8	9	10
X_1	1	2	3	4	5	6	7	8	9	10
X_2	1	4	4	3	2	7	6	1	3	9
Y	Vero	Falso	Falso	Vero	Falso	Falso	Vero	Falso	Vero	Vero

4.1) Stimare l'impurità di Gini della variabile Y (output) del dataset.
4.2) Scegliere la feature da utilizzare come discriminante alla radice di un albero di decisione costruito in modo greedy sulla base dell'impurità di Gini attesa nei figli. Per entrambe le feature considerare la sola soglia data dalla mediana.

Soluzione 4

Vedere sezione 3.4.1 delle dispense.
4.1)

$$\Pr(Y = \text{Vero}) = \frac{1}{2}, \quad \Pr(Y = \text{Falso}) = \frac{1}{2}; \qquad GI(Y) = 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{2}.$$

4.2) La mediana di X_1 è $\theta_1 = 5.5$. Inoltre abbiamo:

$$GI(Y|X_1 < \theta_1) = 2 \cdot \frac{2}{5} \cdot \frac{3}{5} = \frac{12}{25} \quad \text{e} \quad GI(Y|X_1 \geq \theta_1) = 2 \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{12}{25}.$$

L'impurità attesa è dunque:

$$E[GI(Y|X_1)] = \frac{1}{2}\left(\frac{12}{25} + \frac{12}{25}\right) = \frac{12}{25}.$$

La mediana di X_2 è invece $\theta_2 = 3.5$. In questo caso abbiamo:

$$GI(Y|X_2 < \theta_2) = 2 \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{12}{25} \quad \text{e} \quad GI(Y|X_2 \geq \theta_2) = 2 \cdot \frac{2}{5} \cdot \frac{3}{5} = \frac{12}{25}.$$

L'impurità attesa è dunque ancora uguale a $E[GI(Y|X_2)] = 12/25$.
Di conseguenza, è possibile usare indifferentemente X_1 o X_2 come discriminante alla radice dell'albero per ridurre l'impurità attesa rispetto all'intero dataset.

Esercizio 5

Sia dato il seguente dataset di $m = 6$ documenti testuali. Ogni documento comprende 3 lettere dell'alfabeto, corrispondenti alle 3 feature categoriche L_1, L_2, L_3, ognuna avente come dominio l'insieme $\{A, B, C, \ldots, X, Y, Z\}$:

i	1	2	3	4	5	6
L_1	A	D	A	Z	A	A
L_2	A	H	B	Z	A	Z
L_3	B	Z	D	B	A	D

Sia data la variabile aleatoria indicatrice $\mathcal{I}[E]$, che assume valore 1 quando l'evento E è vero e assume valore 0 quando l'evento E è falso. Si indichi inoltre con $L_j(i)$ il valore assunto dalla feature L_j in corrispondenza del documento i. Per valutare la similarità tra due documenti x e y, si utilizzi la seguente misura:

$$\text{sim}(x, y) = \sum_{j=1}^{3} \mathcal{I}[L_j(x) = L_j(y)], \tag{1}$$

che conta il numero di feature uguali tra due documenti.

5.1) Tracciare il funzionamento dell'algoritmo di clustering agglomerativo gerarchico con criterio *single linkage* per i 6 documenti del dataset, considerando la (1) come misura di similarità.
Disegnare il dendrogramma risultante.

5.2) Ripetere con il criterio *complete linkage*.

Soluzione 5

Vedere il capitolo 5 ed in particolare la sezione 5.2 delle dispense. Indichiamo con D_i l'i-esimo documento. Utilizzando la (1), la tabella delle similarità tra i vari documenti è la seguente:

	D_2	D_3	D_4	D_5	D_6
D_1	0	1	1	2	1
D_2		0	0	0	0
D_3			0	1	2
D_4				0	1
D_5					1

Occorre raggruppare i documenti che presentano la massima somiglianza. Nel primo passaggio è quindi possibile raggruppare D_1 e D_5, oppure D_3 e D_6.

5.1) Nel caso del criterio *single linkage* la somiglianza tra i vari cluster corrisponde alla *massima* somiglianza tra singole coppie di elementi. Riportiamo il primo passaggio, nel caso si scelgano D_1 e D_5 come elementi del primo cluster. La tabella risultante sarà:

	D_2	D_3	D_4	D_6
$\{D_1, D_5\}$	0	1	1	1
D_2		0	0	0
D_3			0	2
D_4				1

5.2) Nel caso del criterio *complete linkage* la somiglianza tra i vari cluster corrisponde alla *minima* somiglianza tra singole coppie di elementi. Riportiamo il primo passaggio, nel caso si scelgano D_1 e D_5 come elementi del primo cluster. La tabella risultante sarà:

	D_2	D_3	D_4	D_6
$\{D_1, D_5\}$	0	1	0	1
D_2		0	0	0
D_3			0	2
D_4				1