Algoritmi Avanzati, A.A. 2016–2017

Seconda prova scritta

Martedì 7 febbraio 2017

NOME:_____ COGNOME:_____ MATRICOLA:_____

Indicazioni generali

- Riportare il proprio nome, cognome e numero di matricola in cima a questo foglio e a tutti i fogli, di bella e di brutta copia.
- Al termine dello svolgimento della prova, è necessario riconsegnare *tutti* i fogli, comprese le brutte copie e il presente testo. In caso di riconsegna parziale la prova non verrà valutata.
- Il presente foglio non deve riportare alcuna scritta ad eccezione dei dati di identificazione.
- Durante lo svolgimento della prova non è consentito l'uso di libri, appunti, dispositivi elettronici.
- Riportare e commentare tutti i passaggi richiesti dai vari algoritmi, in particolare in tutte le situazioni di possibile ambiguità.
- Non è consentito uscire prima della consegna, che può avvenire in qualunque momento. Una volta usciti, non sarà consentito il rientro.
- I 5 esercizi valgono 6 punti ciascuno.
- La consegna invalida un eventuale risultato precedente.
- Un eventuale voto precedente può essere conservato ritirandosi.

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?

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.

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.

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.

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*.