Algoritmi Avanzati, A.A. 2016–2017

Seconda prova parziale

Venerdì 23 dicembre 2016

- 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.
- Non è consentito uscire prima della consegna, che può avvenire in qualunque momento. Una volta usciti, non sarà consentito il rientro.
- L'esercizio 1 vale 20 punti. Le 8 domande dell'esercizio 2 valgono 1.5 punti ciascuna ($+1.5$ se la risposta è corretta, -1 se è errata, 0 per le risposte non date).

Esercizio 1

È dato il seguente dataset di 12 campioni composti da due feature scalari $x_{i1} \in [0, 1]$ e $x_{i2} \in [0, 10]$, $i = 1, \ldots, 12$, come variabili indipendenti e una classe a due valori $y_i \in \{\text{Natale}, \text{Capodanno}\}$ come valore da prevedere:

i	x_{i1}	x_{i2}	y_i
1	0.52	7.3	Natale
2	0.22	2.3	Capodanno
3	0.72	4.4	Natale
4	0.98	3.6	Capodanno

i	x_{i1}	x_{i2}	y_i
5	0.81	5.8	Natale
6	0.92	8.6	Capodanno
7	0.88	0.6	Capodanno
8	0.03	8.2	Capodanno

i	x_{i1}	x_{i2}	y_i
9	0.61	1.5	Natale
10	0.43	9.4	Natale
11	0.37	6.5	Natale
12	0.13	9.7	Capodanno

1.1) Stimare l'entropia della variabile Y sulla base del dataset.
1.2) Costruire la radice dell'albero di decisione considerando per ciascuna delle due variabili le soglie del primo e del secondo quartile (25% e 50% delle distribuzioni), utilizzando l'impurità di Gini come criterio.
1.3) Costruire il secondo livello dell'albero di decisione considerando per ciascuna variabile la sola soglia della mediana.

Esercizio 2

Per ciascuna delle seguenti domande, riportare nel foglio protocollo il numero della risposta ritenuta corretta. Si prega di non segnare in alcun modo le domande e le risposte sul foglio.
In caso di incertezza è consentito motivare una risposta con una riga di testo.

1. Per quale dei seguenti motivi è spesso opportuno usare la mediana della distribuzione come soglia per binarizzare una variabile continua, invece della media?

   (a) Perché la mediana non risente molto della presenza di valori estremi (outliers).

   (b) Gli altri due motivi sono entrambi validi.

   (c) Perché il calcolo della mediana, non richiedendo somme e divisioni, è computazionalmente più efficiente del calcolo della media

2. Quale linkage criterion tende a generare dendrogrammi più bilanciati?

   (a) Il single linkage.

   (b) Il linkage criterion non ha influenza sul bilanciamento.

   (c) Il complete linkage.

3. Qual è la definizione dell'Impurità di Gini di una variabile casuale Y?

   (a) La probabilità di errore nel prevedere un esito $y \in Y$ se si sceglie un valore casuale $\tilde{y}$ con la stessa distribuzione di probabilità.

   (b) La probabilità di errore nel prevedere un esito $y \in Y$ se si sceglie un valore casuale $\tilde{y}$ con distribuzione di probabilità uniforme.

   (c) La probabilità di prevedere correttamente un esito $y \in Y$ se si sceglie un valore casuale $\tilde{y}$ con distribuzione di probabilità uniforme.

4. Quale tra i seguenti algoritmi visti a lezione **non** è greedy?

   (a) La costruzione di un albero di decisione.

   (b) La selezione iterativa delle feature sulla base dell'informazione mutua.

   (c) Il calcolo dell'entropia di una variabile casuale.

5. Ad ogni iterazione dell'algoritmo di clustering agglomerativo gerarchico su una matrice di distanze, come si scelgono i due cluster da aggregare?

   (a) Si scelgono i due cluster aventi distanza minima o massima, a seconda del linkage criterion scelto.

   (b) Si scelgono sempre i due cluster aventi distanza massima.

   (c) Si scelgono sempre i due cluster aventi distanza minima.

6. L'entropia di una variabile casuale Y...

   (a) ...dipende soltanto dalla distribuzione di probabilità di Y.

   (b) ...dipende soltanto dal dominio di Y.

   (c) ...dipende sia dalla distribuzione di probabilità, sia dal dominio di Y.

7. L'impurità di Gini di una variabile casuale Y...

   (a) ...dipende soltanto dalla distribuzione di probabilità di Y.

   (b) ...dipende soltanto dal dominio di Y.

   (c) ...dipende sia dalla distribuzione di probabilità, sia dal dominio di Y.

8. Se abbiamo una collezione di punti della forma (x, x^2), con $x \in [-1, 0]$ distribuito uniformemente, che valore ha il coefficiente di correlazione fra le due coordinate?

   (a) $\rho = 0$, perché la relazione non è lineare.

   (b) $\rho > 0$, perché x^2 è sempre positivo.

   (c) $\rho < 0$, perché la relazione fra la prima e la seconda coordinata è decrescente.