Algoritmi Avanzati, A.A. 2016–2017

Seconda prova parziale — Traccia della soluzione

Venerdì 23 dicembre 2016

Esercizio 1

È dato il seguente dataset di 12 campioni composti da due feature scalari $x_{i1} \in [0,1]$ e $x_{i2} \in [0,10]$, $i = 1, \dots, 12$, 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.

Soluzione 1

1.1) (5 punti) È sufficiente osservare che i due valori di Y sono equiprobabili, quindi l'entropia è 1 bit, la massima possibile per una variabile a due valori.
Se si preferisce applicare la formula:

$$\begin{aligned} H(Y) &= -\big(\Pr(Y=\text{Natale})\log_2\Pr(Y=\text{Natale}) + \Pr(Y=\text{Capodanno})\log_2\Pr(Y=\text{Capodanno})\big) \\ &= -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1, \end{aligned}$$

dove le probabilità sono stimate sulla base delle frequenze.
1.2) (10 punti) L'esercizio ci chiede di calcolare, per ciascuna delle due variabili (X_1 e X_2), due possibili soglie (il primo e il secondo quartile).

Osservare che molti dei calcoli indicati in seguito possono essere omessi con semplici considerazioni pratiche. In particolare, i casi 2 e 4 (mediane) possono essere liquidati perché le due partizioni del dataset sono uniformi. Sporadici errori di calcolo non vengono penalizzati; la scelta di un "primo quartile" che non separa esattamente il 25% dei dati (quindi ne separa 3 su 12) è invece considerata un errore da 2 punti.

1. **Primo caso: variabile X_1, primo quartile**
   Il primo quartile di X_1 è il valore che separa il 25% dei valori di X_1 (cioè i tre valori più piccoli) dal resto, quindi si ottiene dalla media fra il terzo e il quarto valore di X_1 in ordine crescente:

   $$\theta_{1,25\%} = \frac{0.22 + 0.37}{2} = 0.295;$$

   ovviamente, ai nostri fini qualunque valore compreso fra 0.22 e 0.37 va bene. I tre elementi del dataset corrispondenti a $X_1 < \theta_{1,25\%}$ hanno tutti lo stesso output (Capodanno):

   $$\Pr(Y = \text{Natale}|X_1 < \theta_{1,25\%}) = 0; \qquad \Pr(Y = \text{Capodanno}|X_1 < \theta_{1,25\%}) = 1.$$

   La corrispondente impurità di Gini è dunque nulla:

   $$GI(Y|X_1 < \theta_{1,25\%}) = 0.$$

Per quanto riguarda il resto del dataset, per $X_1 \geq \theta_{1,25\%}$ abbiamo 6 istanze di Natale e 3 di Capodanno:

$$\Pr(Y = \text{Natale}|X_1 \geq \theta_{1,25\%}) = \frac{2}{3}; \qquad \Pr(Y = \text{Capodanno}|X_1 \geq \theta_{1,25\%}) = \frac{1}{3}.$$

La corrispondente impurità di Gini è dunque

$$GI(Y|X_1 \geq \theta_{1,25\%}) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = \frac{4}{9}.$$

L'impurità di Gini attesa (media pesata delle due impurità calcolate) è dunque

$$GI(Y) = \Pr(X_1 < \theta_{1,25\%})GI(Y|X_1 < \theta_{1,25\%}) + \Pr(X_1 \geq \theta_{1,25\%})GI(Y|X_1 \geq \theta_{1,25\%}) = \frac{1}{4}\cdot 0 + \frac{3}{4}\cdot\frac{4}{9} = \frac{1}{3}.$$

2. **Secondo caso: variabile X_1, secondo quartile**
   Il secondo quartile, ovvero la mediana, è il valore che divide in due parti uguali la distribuzione. Nel nostrocaso, è la media dei due valori centrali di X_1:

$$\theta_{1,50\%} = \frac{0.52 + 0.61}{2} = 0.565;$$

   Il dataset viene spezzato in due parti in cui gli output sono equiprobabili:

$$\Pr(Y = \text{Natale}|X_1 < \theta_{1,50\%}) = \Pr(Y = \text{Capodanno}|X_1 < \theta_{1,50\%}) = \frac{1}{2},$$

$$\Pr(Y = \text{Natale}|X_1 \geq \theta_{1,50\%}) = \Pr(Y = \text{Capodanno}|X_1 \geq \theta_{1,50\%}) = \frac{1}{2}.$$

   Di conseguenza, l'impurità di Gini dei due sotto-dataset è massima, e così pure la loro media:

$$GI(Y) = GI(Y|X_1 < \theta_{1,50\%}) = GI(Y|X_1 \geq \theta_{1,50\%}) = 1 - 2\left(\frac{1}{2}\right)^2 = \frac{1}{2}.$$

3. **Terzo caso: variabile X_2, primo quartile**
   Come prima, ma rispetto a X_2:

$$\theta_{2,25\%} = \frac{2.3 + 3.6}{2} = 2.95;$$

   Nel primo quartile di X_2, l'output vale una volta Natale e due Capodanno:

$$\Pr(Y = \text{Natale}|X_2 < \theta_{2,25\%}) = \frac{1}{3}; \qquad \Pr(Y = \text{Capodanno}|X_2 < \theta_{2,25\%}) = \frac{2}{3},$$

   di conseguenza l'impurità di Gini è

$$GI(Y|X_2 < \theta_{2,25\%}) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = \frac{4}{9}.$$

   Per $X_2 \geq \theta_{2,25\%}$ abbiamo 5 istanze pari a Natale e 4 di Capodanno:

$$\Pr(Y = \text{Natale}|X_2 \geq \theta_{2,25\%}) = \frac{5}{9}; \qquad \Pr(Y = \text{Capodanno}|X_2 \geq \theta_{2,25\%}) = \frac{4}{9},$$

   di conseguenza l'impurità di Gini è

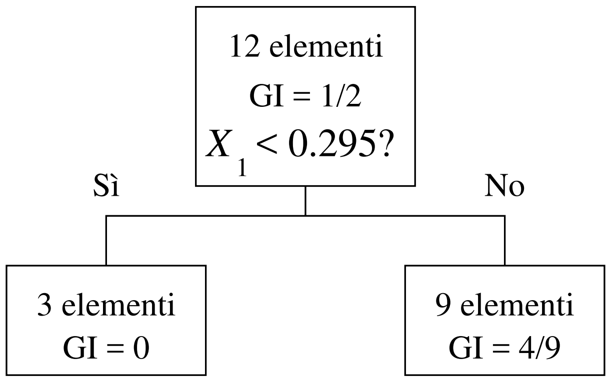
$$GI(Y|X_2 < \theta_{2,25\%}) = 1 - \left(\frac{5}{9}\right)^2 - \left(\frac{4}{9}\right)^2 = \frac{40}{81}.$$

   L'impurità di Gini attesa è

$$GI(Y) = \frac{1}{4}\cdot\frac{4}{9} + \frac{3}{4}\cdot\frac{40}{81} = \frac{13}{27}.$$

4. **Quarto caso: variabile X_2, secondo quartile**
   La mediana di X_2 è
   $$\theta_{2,50\%} = \frac{5.8 + 6.5}{2} = 6.15;$$
   In questo caso gli output delle due sotto-tabelle sono equidistribuiti (3 Natale e 3 Capodanno), quindi l'impurità di Gini risulta massima:
   $$GI(Y) = \frac{1}{2}.$$

Tirando le somme, il caso che risulta nella minore impurità attesa di Gini è il primo; di conseguenza, il primo livello dell'albero di decisione è il seguente:



1.3) (5 punti) Per calcolare il livello successivo dell'albero, osserviamo che il nodo di sinistra è puro, quindi ulteriori suddivisioni sono inutili. Per quanto riguarda il nodo di destra, ecco la sottotabella:

x_1	0.52	0.72	0.98	0.81	0.92	0.88	0.61	0.43	0.37
x_2	7.3	4.4	3.6	5.8	8.6	0.6	1.5	9.4	6.5
y	N	N	C	N	C	C	N	N	N

L'esercizio ci chiede di considerare solamente le mediane, quindi dobbiamo considerare solamente due casi. Dato che il dataset è composto da un numero dispari di casi, la scelta se il valore mediano debba cadere nel nodo di sinistra o di destra è lasciata allo studente. Nel seguito, verrà collocata nel nodo di destra.

1. **Primo caso: variabile X_1**
   La mediana di X_1 è $\theta'_{1,50\%} = 0.72$. Al di sotto della mediana, Y è sempre Natale. Di conseguenza,
   $$GI(Y|X_1 < \theta'_{1,50\%}) = 0.$$
   Per $X_1 \geq \theta'_{1,50\%}$, invece,
   $$\Pr(Y = \text{Natale}|X_1 \geq \theta'_{1,50\%}) = \frac{2}{5}; \qquad \Pr(Y = \text{Capodanno}|X_1 \geq \theta'_{1,50\%}) = \frac{3}{5},$$
   e la corrispondente impurità di Gini è
   $$GI(Y|X_1 \geq \theta_{1,50\%}) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = \frac{12}{25}.$$
   L'impurità attesa è
   $$GI(Y) = \frac{4}{9} \cdot 0 + \frac{5}{9} \cdot \frac{12}{25} = \frac{4}{15}.$$
2. **Secondo caso: variabile X_2**
   La mediana di X_2 è $\theta'_{2,50\%} = 5.8$. Al di sotto della mediana, Y è equidistribuita. Di conseguenza,
   $$GI(Y|X_2 < \theta'_{2,50\%}) = \frac{1}{2}.$$

Per $X_2 \geq \theta'_{2,50\%}$, invece,

$$\Pr(Y = \text{Natale}|X_2 \geq \theta'_{2,50\%}) = \frac{4}{5}; \qquad \Pr(Y = \text{Capodanno}|X_2 \geq \theta'_{2,50\%}) = \frac{1}{5},$$

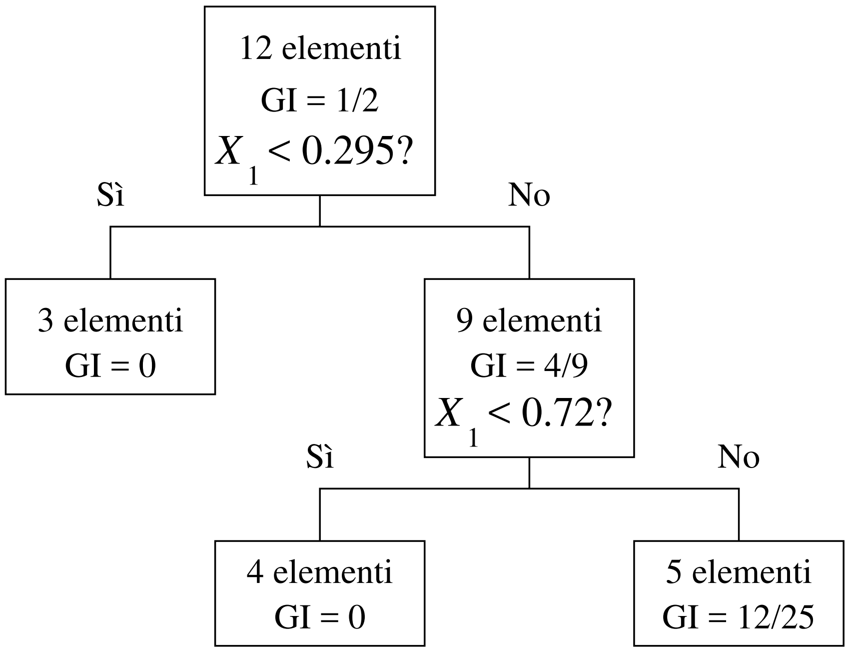
e la corrispondente impurità di Gini è

$$GI(Y|X_2 \geq \theta_{2,50\%}) = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = \frac{8}{25}.$$

L'impurità attesa è

$$GI(Y) = \frac{4}{9} \cdot \frac{1}{2} + \frac{5}{9} \cdot \frac{8}{25} = \frac{2}{5}.$$

Di conseguenza, la variabile che permette di discriminare meglio l'output è ancora X_1 e l'albero diventa:



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. 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.

2. 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.

3. 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.

4. Quale tra i seguenti algoritmi visti a lezione **non** è greedy?
   (a) Il calcolo dell'entropia di una variabile casuale.
   (b) La costruzione di un albero di decisione.
   (c) La selezione iterativa delle feature sulla base dell'informazione mutua.

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

6. Quale linkage criterion tende a generare dendrogrammi più bilanciati?
   (a) Il complete linkage.
   (b) Il single linkage.
   (c) Il linkage criterion non ha influenza sul bilanciamento.

7. 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) Perché il calcolo della mediana, non richiedendo somme e divisioni, è computazionalmente più efficiente del calcolo della media
   (c) Gli altri due motivi sono entrambi validi.

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 fra la prima e la seconda coordinata è decrescente.
   (b) $\rho = 0$, perché la relazione non è lineare.
   (c) $\rho > 0$, perché x^2 è sempre positivo.

Soluzione 2

Le risposte corrette sono le (a). Attenzione: nel testo originale sono rimescolate.

1. Le due risposte errate fanno riferimento all'uso di una distribuzione di probabilità uniforme, il che non corrisponde alla definizione dell'impurità di Gini.

2. L'entropia non dipende dai *valori* che la variabile assume, ma solo dalla loro probabilità.
   Nota bene: dato che la distribuzione di probabilità dipende a sua volta formalmente dal dominio, la formulazione della domanda è ambigua, ed è accettabile anche la risposta (c). La sola dipendenza dal dominio (b), invece, non è accettabile.

3. Le considerazioni precedenti valgono anche per l'impurità di Gini, quindi anche la risposta (c) è accettabile.

4. Gli algoritmi costruttivi visti a lezione (selezione iterativa di feature, costruzione dell'albero di decisione) effettuano la scelta che porta al massimo miglioramento immediato, senza più riconsiderarla. Sono dunque algoritmi greedy. Il calcolo dell'entropia è invece una semplice sommatoria.

5. Il clustering agglomerativo richede sempre e comunque l'aggregazione di cluster vicini (simili) fra loro, indipendentemente dal linkage criterion (che serve soltanto per il calcolo della distanza fra cluster). Non ha senso, in questo contesto, unire i due cluster più diversi.

6. Il complete linkage penalizza l'aggregazione di cluster grandi, perché ne sopravvaluta la distanza, favorendo dunque un albero bilanciato. Vedere ad esempio la figura 10.2 della dispensa.

7. La mediana dipende solo dai valori centrali, quindi non è sensibile alla magnitudine degli estremi, mentre la media può subire notevoli spostamenti in presenza di outlier particolarmente pesanti. Per contro, il calcolo della mediana di una serie di m dati ha complessità $O(m \log m)$ (richiede un ordinamento, almeno parziale), quindi è computazionalmente più pesante della media, che è ovviamente $O(m)$.

8. La relazione $x \mapsto x^2$ è decrescente in $[-1, 0]$, quindi il coefficiente di correlazione sarà negativo.