Algoritmi Avanzati, A.A. 2017–2018

Prova scritta

Martedì 6 febbraio 2018

Esercizio 1

È dato il seguente dataset di $m = 8$ campioni, con $n = 3$ attributi (X_1 numerico; X_2, X_3 categorici) e un output Y categorico (binario):

i	x_{i1}	x_{i2}	x_{i3}	y_i
1	3	`brutto`	`diritto`	`sì`
2	4	`bello`	`diritto`	`no`
3	4	`brutto`	`rovescio`	`sì`
4	4	`brutto`	`diritto`	`no`
5	3	`brutto`	`rovescio`	`no`
6	3	`bello`	`diritto`	`sì`
7	2	`bello`	`rovescio`	`sì`
8	5	`bello`	`rovescio`	`no`

1.1) Definire (a parole) l'impurità di Gini di una distribuzione di probabilità discreta e ricavarne la formula sulla base della definizione data.
1.2) Costruire l'albero di decisione sulla base dell'impurità di Gini, procedendo in modo greedy fino ad ottenere delle foglie pure. Per la variabile numerica, considerare la soglia data dalla mediana.

Esercizio 2

Considerato il dataset dell'esercizio 1, definiamo la seguente distanza nello spazio dei vettori di attributi:

$$\text{dist}(\boldsymbol{x}, \boldsymbol{x}') = |x_1 - x_1'| + \chi_{\neq}(x_2, x_2') + \chi_{\neq}(x_3, x_3'),$$

dove

$$\chi_{\neq}(a, b) = \begin{cases} 0 & \text{se } a = b \\ 1 & \text{se } a \neq b \end{cases}$$

è la cosiddetta "funzione caratteristica della disuguaglianza".
Ad esempio:

$$\begin{aligned} \text{dist}(\boldsymbol{x}_4, \boldsymbol{x}_5) &= \text{dist}\big((4, \texttt{brutto}, \texttt{diritto}), (3, \texttt{brutto}, \texttt{rovescio})\big) \\ &= |4-3| + \chi_{\neq}(\texttt{brutto}, \texttt{brutto}) + \chi_{\neq}(\texttt{diritto}, \texttt{rovescio}) \\ &= 1 + 0 + 1 = 2 \end{aligned}$$

2.1) Calcolare la distanza single-linkage fra i due insiemi

$$C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_7\}, \qquad C_2 = \{\boldsymbol{x}_4, \boldsymbol{x}_8\}$$

2.2) Calcolare la distanza complete-linkage fra gli stessi due insiemi.

Esercizio 3

3.1) Definire (a parole) l'entropia di Shannon di una distribuzione di probabilità discreta e ricavarne la formula sulla base della definizione data.
3.2) Scrivere la formula della funzione sigmoide e le sue principali proprietà (dominio, codominio, asintoti, derivata); motivarne l'uso nella regressione logistica.

Esercizio 4

Si consideri l'albero di decisione ricavato con l'esercizio 1, troncato a profondità 2 (dove la radice è a profondità 0); raggiunto il nodo di livello 2, l'albero risponde con il valore di Y più rappresentato nel nodo. In caso di parità, si supponga che vincano i `sì`.
4.1) Considerando il valore `sì` come classe positiva, stimare l'accuratezza, la precisione, la sensibilità e lo score F_1 del classificatore utilizzando lo stesso dataset utilizzato per l'addestramento.
4.2) Sulla base dell'analisi empirica appena svolta, discutere l'opportunità di utilizzare il classificatore per i seguenti compiti: (i) fungo velenoso (`sì`) / mangereccio (`no`); (ii) automobile usata in buone condizioni (`sì`) / catorcio (`no`).