Algoritmi Avanzati, A.A. 2017–2018

Prova scritta

Martedì 9 gennaio 2018

Esercizio 1

È 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	3	5	5	9	5	1	2	3	3	6
X_2	8	9	1	4	3	6	5	2	9	2
Y	Io	Io	Tu	Io	Tu	Io	Io	Tu	Tu	Io

1.1) Stimare l'impurità di Gini della variabile Y sulla base dei 10 campioni.
1.2) 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 2

2.1) Dato il dataset dell'esercizio 1, considerare i soli 6 elementi per cui $Y =$ Io. Tracciare il funzionamento dell'algoritmo di clustering agglomerativo gerarchico con single linkage criterion per questi 6 elementi, considerando la loro distanza euclidea nello spazio bidimensionale delle feature.
Disegnare il dendrogramma risultante.
2.2) Ripetere con il complete linkage criterion.

Suggerimento — *È estremamente utile rappresentare graficamente i dati su cui lavorare.*

Esercizio 3

Rispondere in modo conciso (massimo 3 righe di testo e formule) a ciascuna delle seguenti domande.
3.1) Scrivere la formula dell'entropia di Shannon di una variabile casuale V che assume ℓ valori $v_1, \ldots, v_\ell$ con rispettive probabilità $p_1, \ldots, p_\ell$.
3.2) Scrivere la formula che stima la covarianza tra due variabili X e Y sulla base di un insieme di m campioni $(x_i, y_i) \in X \times Y$, $i = 1, \ldots, m$.
3.3) Con le stesse definizioni del punto 3.2, scrivere la formula del coefficiente di correlazione di Pearson.
Perché la selezione delle feature rilevanti si basa sul coefficiente di correlazione e non sulla covarianza?

Esercizio 4

Valutare sul dataset fornito all'esercizio 1 il classificatore KNN con $K = 1$ e $K = 3$ rispetto ai principali indici di prestazione (accuratezza, precisione, sensibilità, F_1-score) utilizzando la metodologia leave-one-out.

Suggerimento — *Anche in questo caso la rappresentazione grafica del dataset è d'aiuto.*

Esercizio 5

Si vuole applicare il metodo dei minimi quadrati per determinare il coefficiente $\beta \in \mathbb{R}$ nel modello $y \sim \beta x$ sulla base delle coppie $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, m$ (regressione lineare unidimensionale).
5.1) Costruire la funzione da minimizzare.
5.2) Descrivere il metodo utilizzato per minimizzarla.