Algoritmi Avanzati, A.A. 2016–2017

Prima prova scritta / Seconda prova parziale

Martedì 10 gennaio 2017

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.
- Non è consentito uscire prima della consegna, che può avvenire in qualunque momento. Una volta usciti, non sarà consentito il rientro.

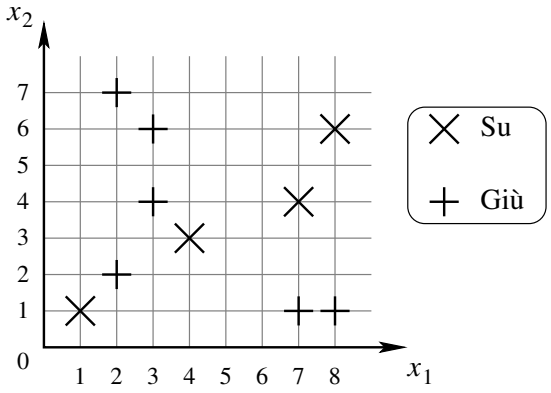
Scelta fra prova completa e prova parziale

- Chi ha conseguito la sufficienza nella prima prova parziale può scegliere di rispondere ai soli esercizi 1, 2 e 3. In tal caso:
  - l'elaborato varrà come seconda prova parziale;
  - ciascun esercizio sarà valutato 10 punti e il voto complessivo risulterà dalla media delle due prove;
  - la consegna invaliderà l'eventuale risultato della seconda prova parziale del 23 dicembre;
  - l'eventuale voto proposto in precedenza può essere conservato ritirandosi.
- Per chi deve (o sceglie di) sostenere la prova completa:
  - i 5 esercizi valgono 6 punti ciascuno;
  - anche in questo caso, la consegna invalida un eventuale risultato precedente;
  - un eventuale voto precedente può essere conservato ritirandosi.
- Non saranno possibili altri recuperi della seconda prova parziale: da febbraio in poi sarà possibile sostenere solamente la prova completa.

Gli esercizi 1, 2 e 4 fanno riferimento al 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_{i1}	1	2	3	4	7	2	8	3	7	8
x_{i2}	1	2	4	3	1	7	1	6	4	6
y_i	Su	Giù	Giù	Su	Giù	Giù	Giù	Giù	Su	Su

In forma grafica:



1 Parte comune (prova completa e seconda prova parziale)

Esercizio 1

1.1) Stimare l'impurità di Gini della variabile Y (output) del dataset.
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) Tracciare il funzionamento dell'algoritmo di clustering agglomerativo gerarchico con single linkage criterion per i soli 6 elementi del dataset per cui $y =$ Giù (i simboli "+" nel grafico), considerando la loro distanza euclidea nello spazio bidimensionale delle feature.
Disegnare il dendrogramma risultante.
2.2) Ripetere con il complete linkage criterion.

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 $p_1, \ldots, p_\ell$.
Perché l'entropia di una variabile casuale si può misurare in bit?
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?

2 Parte per la sola prova completa

Esercizio 4

Valutare sul dataset fornito 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 — *La maggior parte delle distanze può essere valutata a occhio, ricorrere a calcoli solo per i pochi casi dubbi.*
Non è necessario calcolare le radici quadrate.

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.