**Algoritmi Avanzati, A.A. 2016–2017**

# Prima prova scritta / Seconda prova parziale — Traccia della soluzione

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.

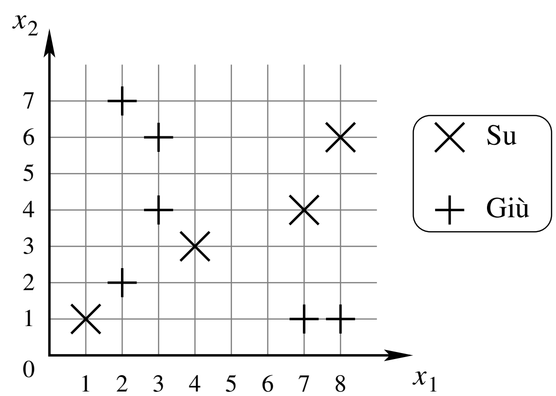
## Scelta fra prova completa e prova parziale

- 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.

**Soluzione 1**

**1.1)**

$$\Pr(Y = \text{Su}) = \frac{2}{5}, \quad \Pr(Y = \text{Giù}) = \frac{3}{5}; \qquad GI(Y) = 2 \cdot \frac{2}{5} \cdot \frac{3}{5} = \frac{12}{25}.$$

**1.2)** Per trovare la mediana di $X_1$ potremmo riordinare il dataset in base a $x_1$, ma è sufficiente individuare nel grafico fornito il valore di $x_1$ che separa il dataset a metà: $\theta_1 = 3.5$. A questo punto, a sinistra abbiamo un solo "Su" e quattro "Giù", il che comporta un'impurità di Gini pari a

$$GI(Y|X_1 < \theta_1) = 2 \cdot \frac{1}{5} \cdot \frac{4}{5} = \frac{8}{25}.$$

La parte di destra è meno "pura" (tre "Su" e due "Giù"), con un'impurità pari a

$$GI(Y|X_1 \geq \theta_1) = 2 \cdot \frac{3}{5} \cdot \frac{2}{5} = \frac{12}{25}.$$

L'impurità attesa è dunque

$$E[GI(Y|X_1)] = \frac{1}{2}\left(\frac{8}{25} + \frac{12}{25}\right) = \frac{2}{5}.$$

Anche la mediana di $X_2$ è $\theta_2 = 3.5$ (ne separa 5 sopra e 5 sotto). In entrambe le partizioni abbiamo due "Su" e tre "Giù", quindi le due impurità di Gini e anche la loro media sono pari a

$$E[GI(Y|X_2)] = 2 \cdot \frac{2}{5} \cdot \frac{3}{5} = \frac{12}{25}.$$

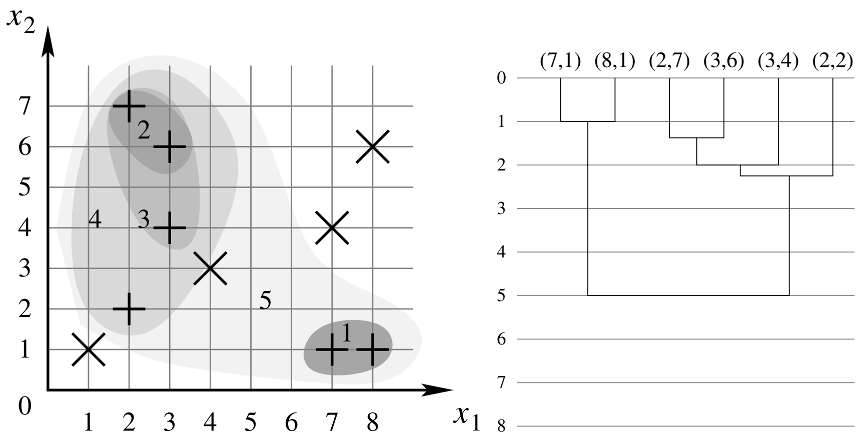
L'uso di $X_2$ alla radice non cambia l'impurità attesa rispetto all'intero dataset, mentre l'uso di $X_1$ la riduce, ed è quindi da preferire.

**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.

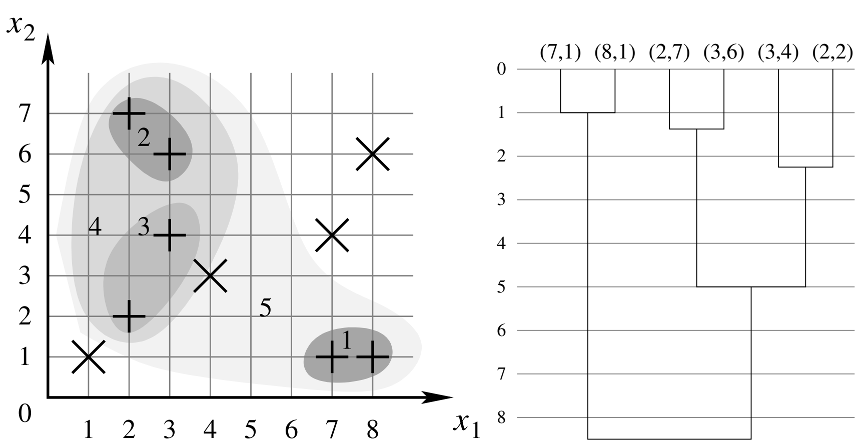
**Soluzione 2**

**2.1)** Fare riferimento alla figura e al corrispondente dendrogramma:




- I primi due elementi a essere uniti nel cluster 1 sono $(7, 1)$ e $(8, 1)$, a distanza $1$.
- In seguito, l'algoritmo unisce gli elementi $(2, 7)$ e $(3, 6)$, con distanza $\sqrt{2}$, nel cluster 2.
- A questo punto entra in gioco il single linkage: la distanza minore è quella fra l'elemento $(3, 4)$ e il cluster 2, a distanza 2, e si forma il cluster 3.
- Il cluster 3 viene poi unito all'elemento $(2, 2)$ a distanza $\sqrt{5}$ formando il cluster 4.
- Infine, i due cluster rimanenti, 1 e 4, sono uniti a distanza $5$.

**2.2)** Fare riferimento alla figura:




- I primi due passi (cluster 1 e 2) sono uguali a prima.
- La distanza fra il punto $(3, 4)$ e il cluster 2 è maggiore rispetto a prima (per via del complete linkage vale $\sqrt{10}$), quindi viene preferita l'unione fra $(3, 4)$ e $(2, 2)$, formando il cluster 3.
- Poi vengono uniti i due cluster 2 e 3, che hanno distanza $5$, nel cluster 4.
- Infine, i due cluster 1 e 4 vengono uniti a distanza $6\sqrt{2}$, ovvero la distanza fra i punti $(8, 1)$ e $(2, 7)$.

**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 $\ell$ 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?

**Soluzione 3**

Vedere gli appunti di lezione; in particolare:
**3.1)** Sezione 3.4.1, sezione sull'entropia; formula (3.8).
**3.2)** Sezione 4.1, formula (4.1).
**3.3)** Sezione 4.1, formula (4.2).

## 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.*

**Soluzione 4**

Riprendiamo la tabella del dataset e inseriamo per ogni elemento i primi tre vicini (metodologia leave-one-out: per l'addestramento si usa tutto il dataset tranne il campione su cui valutiamo, e si ripete per tutti i campioni):

| $i$ | 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 |
| 1. vicino | Giù | Su | Su | Giù | Giù | Giù | Giù | Giù | Su | Su |
| 2. vicino | Su | Giù | Giù | Giù | Su | Giù | Su | Giù | Giù | Giù |
| 3. vicino | Giù | Su | Giù | Su* | Su | Su | Su | Su | Su* | Su |

* — Alla pari con un "Giù", ho scelto arbitrariamente.

La matrice di confusione per $K = 1$, quindi utilizzando il primo vicino come predittore, risulta:

| | | Previsione | |
|---|---|---|---|
| | | Su | Giù |
| Classe | Su | 2 | 2 |
| corretta | Giù | 2 | 4 |

Un'altra scelta arbitraria: quale classe considerare "positiva". Nel mio caso scelgo "Su". Di conseguenza,

$$\text{accuratezza} = \frac{2+4}{10} = \frac{3}{5}; \quad \text{precisione} = \frac{2}{2+2} = \frac{1}{2}; \quad \text{sensibilità} = \frac{2}{2+2} = \frac{1}{2}.$$

La precisione e la sensibilità sono uguali, quindi lo è anche la loro media armonica:

$$F_1 = \frac{1}{2}.$$

La matrice di confusione per $K = 3$, quindi utilizzando i 3 primi vicini a maggioranza, risulta:

| | | Previsione | |
|---|---|---|---|
| | | Su | Giù |
| Classe | Su | 2 | 2 |
| corretta | Giù | 3 | 3 |

Di conseguenza la situazione peggiora:

$$\text{accuratezza} = \frac{2+3}{10} = \frac{1}{2}; \quad \text{precisione} = \frac{2}{2+3} = \frac{2}{5}; \quad \text{sensibilità} = \frac{2}{2+2} = \frac{1}{2}.$$

Infine,

$$F_1 = \frac{2 \cdot \frac{2}{5} \cdot \frac{1}{2}}{\frac{2}{5} + \frac{1}{2}} = \frac{4}{9}.$$

**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 $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \dots, m$ (regressione lineare unidimensionale).
**5.1)** Costruire la funzione da minimizzare.
**5.2)** Descrivere il metodo utilizzato per minimizzarla.

**Soluzione 5**

**5.1)** Vedere la sezione 3.2 degli appunti. La funzione da minimizzare è la somma dei quadrati degli scarti tra i valori corretti e previsti delle $y$:

$$\sum_{i=1}^{m} (\beta x_i - y_i)^2.$$

**5.2)** È sufficiente annullare la derivata prima rispetto a $\beta$, come nell'equazione (3.1). Si osservi che la domanda non chiedeva di effettuare i passaggi, ma solo di indicare il metodo.