

Analisi Statistica Multivariata. A.A. 2011-2012

Progetto n.1. Gruppo II. Consegna: 30/5/2012.

I dati per i problemi sono reperibili sul sito del corso

Problema 1.

Si considerino i dati **Prawn Trawling** (in *DASL and More Data*). (Sostituire NA con -1 nei dati per permettere la lettura in matlab)

Studiare gli oggetti e le variabili mediante un'analisi di clustering usando single e complete linkage, eventualmente considerando varie distanze. Interpretare i risultati indicando le eventuali eccezioni.

Valutare i raggruppamenti con anche un metodo di Multidimensional scaling, e confrontare i risultati.

Problema 2.

Si considerino i dati **Crude Oil** (in *T11-7.DAT di JW*), costituiti da 3 campioni di olio da tre diverse zone: Wilhelm, Sub-Mulinia e Upper.

1. Per la matrice di osservazioni relativa ad ogni popolazione considerata, fare uno studio della normalità univariata, bivariata e multivariata. Verificare la presenza di outliers ed eliminarli dall'analisi successiva giustificando la scelta. Trasformare eventualmente le variabili, o alcune di esse, per migliorare la normalità dei dati.
2. Determinare intervalli simultanei di confidenza (99% e 95%) per la media, relativamente alla famiglia di Upper.
3. Per le osservazioni provenienti dalle diverse popolazioni, valutare l'ipotesi di uguale media delle popolazioni, con livello di significatività $\alpha = 0.05$ e $\alpha = 0.01$. Commentare i risultati. In caso di rifiuto, valutare quale delle variabili è più responsabile del rifiuto, facendo il confronto di medie sulle singole variabili.
4. Determinare intervalli simultanei di confidenza (95%) e commentare su eventuali differenze tra i risultati ottenuti rispetto al test del quesito precedente.
5. Sfruttare il metodo di discriminanza principale di Fisher per una valutazione grafica delle famiglie e confrontare con i precedenti risultati.
6. Impostare un test di discriminanza per allocare la nuova osservazione

$$\mathbf{x} = [7.1, 30.0, 0.05, 4.01, 2.80]$$