Analisi Statistica Multivariata. A.A. 2011-2012
Progetto n.1. Gruppo V. Consegna: 30/5/2012.
I dati per i problemi sono reperibili sul sito del corso

Problema 1.

I dati in **emissions_USA** (in *DASL and More Data*) si riferiscono alle emissioni totali di monossido di Carbonio di alcuni Stati americani, rispetto a varie sorgenti. Fare un'analisi di clustering usando single e complete linkage, sia per le variabili (sorgenti) che per le osservazioni (Stati), considerando più di una distanza. Interpretare i risultati.

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

Problema 2.

Si considerino i dati sui salmoni in T11-2 (in JW), corrispondenti ai valori di crescita (diametro degli anelli di crescita) di due popolazioni (dell'Alaska, famiglia 1 e Canadesi, famiglia 2), in due tempi diversi: in acqua dolce (primo anno) ed acqua salata(primo anno). In tutta l'analisi, si considerino **solo** i dati dei maschi (seconda colonna, genere=2).

1. Per la matrice di osservazioni relativa ad ogni popolazione considerata, fare uno studio della normalità univariata e bivariata. Verificare l'eventuale 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 la regione di confidenza (99% e 95%) per la media, relativamente alla famiglia Alaska, ed anche gli intervalli di confidenza. Riportare tutto sullo stesso grafico e commentare.

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. Impostare un test di discriminanza per allocare la nuova osservazione

$$\mathbf{x} = [140,\ 370]$$

anche mediante uno studio grafico.