Politecnico di Milano
II Scuola - Ingegneria dei Sistemi (MI)
Appello di Statistica Applicata
16 luglio 2012

Nome e cognome:
Numero di matricola:

Problema 1

La Poli-Cruise organizza da diversi anni crociere scientifico-culturali per i neolaureati del Politecnico. Nel file `cruise.txt` sono riportati per le ultime 150 crociere il numero di passeggeri suddiviso in quattro diverse categorie: ingegnere-uomo, ingegnere-donna, architetto/designer-uomo e architetto/designer-donna. Assumendo indipendenti i dati relativi alle diverse crociere ma dipendenti i dati relativi alla stessa crociera:

a) si forniscano 5 intervalli T^2-simultanei di confidenza globale 90% per il numero medio di passeggeri per ognuna delle 4 categorie e per il totale di passeggeri;

b) si esegua un test di livello 5% per confermare o smentire l'ipotesi secondo la quale il numero medio di ingegneri è uguale al numero medio di architetti;

c) si esegua un test di livello 5% per confermare o smentire l'ipotesi secondo la quale il numero medio di uomini è il doppio del numero medio di donne;

d) si esegua un test di livello 10% per testare congiuntamente le ipotesi enunciate ai punti (b) e (c).

Problema 2

Il prof. Ålleniclüp, noto ornitologo dell'Isole Farøer, ha misurato apertura alare [cm], distanza becco-coda [cm] e peso [hg] di 50 esemplari maschio e di 50 esemplari femmina di pulcinella di mare (file `puffin-M.txt` e `puffin-F.txt`). Dopo aver scelto se fare riferimento all'indice di correlazione lineare o all'indice di covarianza lineare:

a) si esegua una PCA relativa agli esemplari maschio. In particolare si riporti e commenti il grafico della frazione di varianza totale spiegata e si riportino e commentino i loadings;

b) si esegua una PCA relativa agli esemplari femmina. In particolare si riporti e commenti il grafico della frazione di varianza totale spiegata e si riportino e commentino i loadings;

c) si confrontino i risultati delle due precedenti analisi mettendo in luce similarità e differenze.

Problema 3

L'ittiologo genetista Ozzülrem ritiene che nel fiordo di Geiranger vivano diverse specie di merluzzi simili alla vista ma geneticamente diversi. Nel file `gene.txt` sono riportati per 100 merluzzi pescati nel fiordo il livello di attivazione genica relativo a 1200 geni ritenuti di interesse.

a) Utilizzandi un algoritmo di clustering gerarchico agglomerativo basato sulla distanza di Manhattan e il linkage di Ward, si individuino eventuali gruppi di merluzzi. In particolare si riportino il numero di cluster e le relative numerosità.

b) Per ogni gene si esegua un'ANOVA one-way per il confronto dei livelli medi dell'attività genica tra i diversi gruppi individuati al punto precedente. Si riportino in particolare i geni associati a p-value inferiori all'1%.

c) Assumendo un FDR al più dell'1% si individuino i geni per i quali vi è evidenza statitica di una qualche differenza tra i livelli medi dell'attività genica dei diversi gruppi precedentemente individuati.

Problema 4

Uno dei pericoli ambientali più rilevanti a seguito dell'eruzione del vulcano islandese Eyjafjöll, avvenuta nel 2010, è la contaminazione da fluoruro, causata dal deposito delle ceneri liberate nell'atmosfera durante l'eruzione. Nel file `fluoruro.txt` sono riportate le coordinate di 50 siti di misurazione $\boldsymbol{s}_i$, $i = 1, ..., 50$, le corrispondenti misurazioni di fluoruro $F(\boldsymbol{s}_i)$, $i = 1, ..., 50$ [ppm], e le distanze $D_{\boldsymbol{s}_i}$, $i = 1, ..., 50$ [migliaia di km] di ciascun sito $\boldsymbol{s}_i$ dal cratere del vulcano. Indicando con δ un processo spaziale a media nulla, debolmente stazionario e isotropo:

a) Si stimino dai dati due variogrammi empirici, ipotizzando rispettivamente il modello di regressione $F(\boldsymbol{s}_i) = \beta_0 + \delta(\boldsymbol{s}_i)$ e il modello $F(\boldsymbol{s}_i) = \beta_0 + \beta_1 \cdot D_{\boldsymbol{s}_i} + \delta(\boldsymbol{s}_i)$ (si utilizzi ad esempio la funzione `variogram`); si scelga in particolare se inserire o meno il trend nel modello.

b) Si adatti al variogramma empirico scelto al punto (a) un modello gaussiano senza nugget, stimato con il metodo dei minimi quadrati pesati con parametri iniziali: sill=100, range=0.08 (si utilizzi ad esempio la funzione `fit.variogram`); si riportino in particolare sill e range stimati.

c) Si adatti al variogramma empirico scelto al punto (a) un modello sferico senza nugget, stimato con il metodo dei minimi quadrati pesati (si utilizzi ad esempio la funzione `fit.variogram`); si riportino in particolare sill e range stimati.

d) Confrontando tra di loro, e col variogramma empirico scelto al punto (a), i due variogrammi ottenuti ai punti (b) e (c) e sapendo che gli esperti ritengono che il fenomeno di deposito di cenere al suolo sia molto regolare, si scelga il modello di variogramma più appropriato;

e) Sulla base del modello scelto al punto (d), si stimi la concentrazione di fluoruro nella località di Raufarhöfn ($\boldsymbol{s}_0 = (0.3, 0.24)$, $D_{\boldsymbol{s}_0} = 0.1970$) dovuta all'eruzione del 2010 (si utilizzino ad esempio le funzioni `gstat` e `predict`);

f) Sulla base del modello scelto al punto (d), si stimi nella stessa località la concentrazione di fluoruro dovuta ad un'eventuale futura eruzione di pari intensità (si utilizzino ad esempio le funzioni `gstat` e `predict`).