Politecnico di Milano
Facoltà di Ingegneria dei Sistemi
Appello di Statistica Applicata
9 Luglio 2008

©I diritti d'autore sono riservati. Ogni sfruttamento commerciale non autorizzato sarà perseguito.

Nome e cognome:
Numero di matricola:

Problema 1

Nel dataset `client.txt` sono contenuti i dati relativi a 150 clienti della PoliBank. Per ciascun cliente sono riportati età [anni], denaro investito a basso rischio [migliaia di €] (variabile safemoney) e denaro investito ad alto rischio [migliaia di €] (variabile riskymoney).

a) Utilizzando esclusivamente la variabile età, clusterizzate i clienti in tre gruppi e descriveteli in termini di età. Si utilizzi un algoritmo gerarchico agglomerativo basato sulla distanza euclidea ed il single linkage. Riportate il coefficiente cofenetico e le numerosità dei cluster.

b) Introducendo le opportune assunzioni sulle distribuzioni delle variabili safemoney e riskymoney all'interno dei tre gruppi, si esegua una MANOVA per vedere se vi è evidenza statistica di una differenza nelle distribuzioni congiunte delle variabili safemoney e riskymoney nei tre gruppi.

c) Si commenti il risultato della MANOVA per mezzo di opportuni intervalli di Bonferroni di confidenza globale 90%.

Problema 2

Nel dataset `temperature.txt` sono riportate le 12 temperature medie mensili misurate in 35 località canadesi. Supponendo le 35 misurazioni *iid* e congiuntamente normali:

a) si esegua un test al 10% per verificare l'ipotesi nulla secondo la quale la media delle temperature medie mensili non cambia al variare del mese.

Dette $T_1 = (T_{Jan} + T_{Feb} + \ldots + T_{Dec})/12$ la variabile aletaoria indicante la temperatura media annuale e $T_2 = T_{Jul} - T_{Jan}$ la variabile aleatoria indicante la differenza tra la temperatura media di Luglio e quella di Gennaio:

b) si costruiscano due intervalli di Bonferroni di confidenza globale 90% per le medie delle variabili T_1 e T_2;

c) si costruiscano due intervalli di Bonferroni di confidenza globale 90% per le varianze delle variabili T_1 e T_2.

Problema 3

Nel deposito nucleare di Sheffield arrivano quotidianamente barili contenenti sostanze debolmente radioattive (è noto che 1/3 dei barili emette particelle α mentre 2/3 particelle γ). Quando un barile giunge al deposito, il barile viene aperto in zona protetta ed un rivelatore misura la distanza d percorsa in aria dalla prima particella emessa in direzione del rivelatore. Secondo i modelli proposti dai fisici nucleari, la distanza percorsa in aria da una particella α segue approssimativamente una legge normale, mentre quella percorsa da una particella γ una legge esponenziale. Nei file `alpha.txt` e `gamma.txt` sono contenute rispettivamente le misurazioni della distanza $[dm]$ percorsa in aria da 100 particelle α e da 100 particelle γ.

a) Stimate le due densità di probabilità relative alla distanza percorsa dalle particelle α e dalle particelle γ.

b) Tenendo conto che il danno prodotto dall'errata assegnazione di un barile γ alla classe dei barili α è valutato essere 2 volte il danno dell'errore opposto, costruite un classificatore di barili che minimizzi il danno atteso e che utilizzi come predittore la distanza d misurata dal rivelatore.

c) Utilizzando le stime delle densità di probabilità ottenute al punto (a), calcolate la percentuale di barili α, di barili γ e di barili in generale che vengono misclassificati dalla regola calcolata al punto (b).

Problema 4

L'Associazione Panificatori Anglo-Italiani ha raccolto i dati relativi al prezzo (media nazionale rinormalizzata rispetto al prezzo del 2001) del pane in Italia e in Inghilterra (file `Italia.txt` e `UK.txt`) negli ultimi sei anni (2002-2007). Il modello utilizzato per analizzare i dati assume una crescita lineare del prezzo del pane sia in Italia che in Inghilterra:

$$P_{gi} = \alpha_g + \beta_g \cdot t_i + \epsilon_{gi} \ ,$$

con g lo stato; $t_i = 1, 2, 3, 4, 5, 6$ l'anno a partire dal 2002 ($t_i = 1$); P_{gi} il prezzo del pane nello stato g all'anno t_i; ed $\epsilon_{gi} \sim N(0, \sigma^2)$ *iid*.

a) Si stimino i parametri α_g, β_g e σ del modello completo.

b) Si individui un opportuno modello ristretto per descrivere i dati, si forniscano le stime dei suoi parametri e le si interpretino.

c) Il Times afferma che la crescita del prezzo del pane in Italia procede ad una velocità doppia rispetto all'Inghilterra. Si verifichi o smentisca tale affermazione mediante un opportuno test sui parametri del modello proposto al punto (b).

d) Si semplifichi ulteriormente il modello proposto al punto (b) nell'ottica di quanto affermato dal Times, si forniscano le stime dei parametri del modello semplificato e le si interpretino.

e) Tutti e tre i modelli forniscono stime intervallari non distorte della varianza σ^2 dell'errore. Si riportino i corrispondenti intervalli di confidenza ($\gamma = 95\%$) per la varianza σ^2 e se ne confrontino criticamente le ampiezze.