Insegnamento di "Data mining"
Prova d'esame del 9 luglio 2013

1. Spiegare che cosa si intende per "correlazione parziale", e si specifichi in quali situazioni è necessario considerarla.

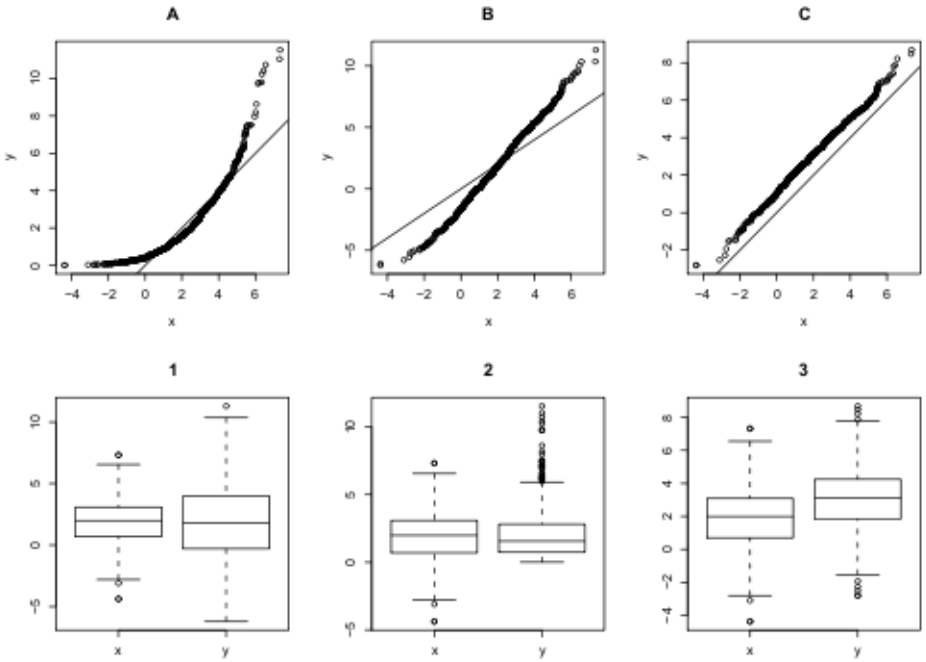
2. Si descrivano differenze e similitudini tra la regressione effettuata tramite *projection pursuit* e utilizzando le neti neurali.

3. Si presenti l'algoritmo *apriori* e si specifichi a quale problema fornisce una risposta.

4. Si spieghi che cosa è un intervallo di confidenza di livello 95% e si presenti come un tale intervallo potrebbe essere scelto.

5. Si considerano tre coppie di variabili, per ciascuna delle quali si dispone di 1000 osservazioni. Nella figura che segue si confrontano mediante *qq-plot* e boxplot le distribuzioni delle due variabili di ciascuna coppia.

   Si associno i *qq-plot* alle coppie di boxplot corrispondenti.



6. (facoltativo)

   Si consideri una variabili Y che può assumere solo due modalità, 0 o 1, distribuita come una Bernoulli. Da questa variabile si osservano n unità indipendenti, per cui la funzione di probabilità del numero di 1 osservati ($X = \sum_{i=1}^{n} Y$) è

$$f(x;p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}.$$

   (a) Si scriva la funzione di log-verosimiglianza rispetto al parametro p

   (b) si ottenga la stima di massima verosimiglianza per p, cioè si trovi il valore di p corrispondente al massimo della funzione di log-verosimiglianza.

   (c) si ottenga il valore della funzione di log-verosimiglianza in corrispondenza del suo massimo.

Soluzioni

Esercizio 5 Gli abbinamenti corretti sono i seguenti.: A2. È evidente dal qqplot che le variabili differiscono per la forma della distribuzione, corrispondentemente nel boxplot marcato con 2 si nota che la variabile X è simmetrica mentre Y presenta una marcata asimmetria positiva.

B1. È evidente dal qqplot che le due distribuzioni differiscono per la loro variabilità (Y ha una variabilità maggiore), alla stessa conclusione si perviene guardando al boxplot 1.

C3. I punti si dispongono parallelamente alla bisettrice del I-III quadrante, per cui forma e variabilità sono simili, il fatto che i punti siano discosti dalla retta si deve alla diversa media (posizione) delle due distribuzioni. La stessa caratteristica appare dal boxplot 3: i due diagrammi sono infatti molto simili ma quello relativo alla Y è spostato verso lalto di una quantit'a pari circa a uno.

Esercizio 6

(a)