Multivariate Analysis and Statistical Learning
Relazione Contest
Implementazione PC Algorithm

Alex Foglia
Tommaso Puccetti

December 17, 2018

Contents

1 Accenni di Teoria

Un esempio di modello grafico sono le reti bayesiane. Tali reti sono rappresentate attraverso l'utilizzo di un Grafo Aciclico Diretto, o DAG.
Un DAG è un grafo diretto in cui non compaiono cicli, dove per ciclo si intende un qualunque cammino finito che, a partire da un nodo iniziale v termini in v. Una rete bayesiana, rappresentata attraverso un DAG, ha delle applicazioni interessanti in contesti di Machine Learning e in particolare di analisi causale.
Sia $G = (V, E)$ un DAG su un insieme finito $X = \{X_v \forall v \in V\}$ di variabili casuali, allora:

$$\forall u, v \in V \text{ non adiacenti } | v \in nd(u) \Rightarrow u \perp\!\!\!\perp v | nd(u) - v$$

Dove $nd(u)$ è l'insieme dei nodi *non discendenti* di u, ossia tutti quei nodi u' per cui non esiste un cammino da u a u'.
Dato un insieme di variabili osservate, con distribuzione di probabilità congiunta gaussiana, è possibile imparare il DAG sottostante al campione osservato. A questo scopo è stato progettato un particolare algoritmo: il PC-Algorithm.
Esso è composto da due sotto-funzioni che risolvono due diversi problemi:

- La costruzione dello scheletro (o grafo morale)
- La costruzione del DAG a partire da un dato scheletro

1.1 Costruzione dello scheletro

La prima fase non produce esclusivamente lo scheletro, infatti in essa computiamo anche il separation set ossia uninsieme di variabili associato a ciascuna coppia di variabili indipendenti x, y. Gli elementi di tale insieme rappresentano tutte quelle variabili che condizionano l'indipendenza fra x e y, e che quindi si trovano nel cammino da x a y. Lo pseudo-codice dell'algoritmo, che prende in ingresso la $z-$trasformata delle correlazioni parziali stimate e il *tuning parameter* α è il seguente:

```
G = grafo_completo()
l = -1
repeat
  l = l + 1
  repeat
    seleziona una coppia ordinata di variabili adiacenti i,j in G
    se |adj(i,G)\{j}| >= l
      repeat TEST
        seleziona K tra i nodi adiacenti di i escluso j, con |K|=l
            se sqrt(n-|K|-3)|Z(i,j|K)| <= phi_inverse(1-alpha/2)
              cancella l'arco i,j da G
              salva K nel separation set di [i][j] e di [j][i]
              esci da TEST
    finchè tutti i K tali per cui |K| = l sono stati selezionati
  finchè tutte le coppie adiacenti sono state testate
finchè l > |adj(i,G)\{j}| per ogni i,j
```