Tesine di Sistemi di Elaborazione

Corso di Laurea in Informatica

Novembre 2015

Il *metodo dellle proiezioni random (Random Projection)* consiste nel proiettare lo spazio dei dati $\Re^d$ in $\Re^{d'}$ ($d' < d$) tramite una proezione lineare definita da una matrice di valori random $P = p_{ij}$ tale che:

$$p_{ij} = 1/\sqrt{d'} \times r_{ij} \text{ con } E(r_{ij}) = 0, Var(r_{ij}) = 1$$

.

Una tra le possibili matrici che soddisfano la condizione di sopra é la proiezione di Achlioptas, in cui r_{ij} é cosí definita:

$$r_{ij} = \sqrt{3} \begin{cases} 1 & \text{con probabitá } \frac{1}{6} \\ 0 & \text{con probabitá } \frac{2}{3} \\ \text{-1} & \text{con probabitá } \frac{1}{6} \end{cases}$$

Sia dato il dataset "Parkinsons" che rappresenta i valori numerici di 22 attributi estratti da 197 registrazioni vocali di 2 gruppi di persone, il primo affetto dal morbo di parkinson, l'altro no. Il dataset si presenta come una matrice V di 197 righe e 23 colonne.

Si intendono paragonare le due tecniche delle *Random Projection* e della *Principal Component Analysis* sul dataset in questione tramite un algoritmo di classificazione scelto a piacere capace di distinguere tra le due categorie di oggetti (malati e non). In particolare si richiede:

1) Di implemenrate la tecnica delle random projection, in modo da trovare m direzioni su cui proiettare i dati.

2) Di proiettare il dataset X sulle direzioni delle m componenti principali e delle m proiezioni delle random projections, ottenendo rispettivamente i dataset proiettati CX e PX.

3) Di misurare le performance del classificatore sui dataset CX e PX.

La comparazione di tali performance dovrá essere effettuata al variare di m, del training set T e della sua cardinalitá, ed eventualmente su altri parametri del classificatore, fornendo in output opportuni grafici esplicativi. Si precisa che le performance del classificatore devono essere mediate su piú scelte del training set T.