Corso di Laurea in Matematica, Università di Roma "La Sapienza"

Corso di Data Mining

M. Falcone

Esercitazioni in Laboratorio

Foglio 5: Principal Component Analysis (PCA)

Per analizzare i dati contenuti in un file useremo la PCA. I dati vanno prima letti da un file (ad esempio `bodyfat_txt`) e vanno standardizzati. Il procedimento è iterativo e serve a eliminare progressivamente le osservazioni anomale.

Si considerino i dati in `bodyfat_txt` sullo studio della percentuale di grasso corporeo di un individuo.

1. Standardizza i dati mediante la funzione `zscore`;
2. Calcola la matrice di Correlazione ed individua in modo automatico gli elementi con correlazione significativa;
3. Determina le Componenti Principali (CP) per la matrice di correlazione dei dati. Valuta il numero minimo di CP per una buona rappresentazione della variabilità dei dati mediante i criteri visti;
4. Disegnare il diagramma di dispersione tra la prima componente principale e le variabili originarie standardizzate e confrontarle con il coefficiente di correlazione `rxi,yj`;
5. Eliminare eventuali osservazioni anomale e, in questo caso, ripetere l'analisi.

Funzioni Matlab(R) utili: `corrcoef, eig, cumsum, plot, text`