

Introduzione al Machine Learnig

Prof. Fabio Divino
Tutorial 3: Clustering

Considerare il microarray dataset **nci60_data.RData** contenente informazioni sull'espressione genica di 6830 geni relativi a 63 records riferiti a pazienti con patologie tumorali. In particolare i dati sono riportati nel dataframe **nci60_data** mentre le etichette delle patologie nell'array **nci60_labels**.

Dopo aver applicato il metodo del **clustering gerarchico**, tagliare l'albero nel dendrogramma dove si ritiene opportuno (usare la regola euristica della "presenza di spazio tra i rami") e determinare il numero di gruppi **k**. Utilizzando lo stesso numero di gruppi **k** applicare anche il metodo delle **k-means**.

Quesiti

- A) Utilizzando la funzione `table` confrontare la distribuzione dei records nei due clustering ottenuti (hierachical clustering e k-means) e commentare i risultati.
- B) Utilizzando la funzione `prcomp`, proiettare i clusters ottenuti con i due metodi sul piano delle componenti principali e commentare i risultati.

Risultati

Dopo aver svolto il tutorial, salvare la sessione di lavoro come file (rigorosamente tutto in minuscolo!)

nome_cognome_tutorial3.RData

insieme allo script

nome_cognome_tutorial3.R

Inviare i file a

fabio.divino@unimol.it

riportando nel testo del messaggio i commenti relativi ai due quesiti.