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

Corso di Data Mining

M. Falcone

Esercitazioni in Laboratorio

Foglio 6: Algoritmi di clustering: K-means, Lloyd's

L'obiettivo dei metodi di clustering (agglomerazione) è quello di separare un insieme di dati sulla base di un criterio di similitudine definito attraverso una particolare funzione distanza. Evidentemente l'insieme dei dati di partenza può avere caratteristiche molto diverse così come i criteri che si vogliono utilizzare per suddividere l'insieme di partenza. Il procedimento è iterativo e serve a costruire progressivamente K gruppi di oggetti separati sulla base della loro "similitudine" rispetto ad una metrica, si noti che il valore intero K dei clusters è scelto a priori e che anche la distanza utilizzata può essere scelta molto liberamente (abbiamo visto molti esempi nei lucidi). Una caratteristica dei metodi K-means e Lloyds è che un singolo elemento del data set può appartenere ad un solo cluster.

Si considerino, ad esempio, i dati contenuti nel data base IRIS di Fisher (è presente su MATLAB e scaricabile da elearning)

1. Leggere i dati ;
2. Definire una funzione distanza;
3. Definire il numero K dei clusters;
4. Iterare l'algoritmo fino a convergenza;
5. Visualizzare il risultato (ad esempio con colori diversi per i vari cluster);
6. Variare l'analisi cambiando K ed eventualmente la funzione distanza.

Funzioni Matlab(R) utili: le funzioni distanza disponibili `euclidean, seuclidean, cityblock, minkowski, jaccard,....`