Riepilogo clustering globale

Normalizzazione dei dati

I dati sono stati prima normalizzati con la tecnica min max scaler che scala i dati fra zero e uno, colonna per colonna, con la seguente formula:

$$V_{scalato} = \frac{(V - min)}{(max - min)} \quad (1)$$

$$V = valore$$
$$min = valore\ minimo\ della\ colonna$$
$$max = valore\ massimo\ della\ colonna$$

Riduzione dimensionale

Successivamente è stato applicato un algoritmo di riduzione dimensionale PaCMAP (Pairwise Controlled Manifold Approximation) il quale è un metodo di riduzione della dimensionalità che può essere utilizzato per la visualizzazione, preservando la struttura *locale e globale* dei dati nello spazio originale. Sono state fatte due riduzioni dimensionali:

1. le *dimensioni di partenza sono state ridotte a 2*
2. le *dimensioni di partenza sono state ridotte a 3*

Questo per rendere visualizzabile il dataset, la riduzione dimensionale su due dimensioni comporta una maggiore distorsione ma una visualizzazione piu' semplice, quella a tre dimensioni permette di avere una distorsione minore ma una visualizzazione legermente meno agevole.

Applicazione algoritmo clustering

Infine è stato applicato un algoritmo di clustering sui dati scalati e ridotti dimensionalmente, e' stato utilizzato affinity propagation. Senza entrare in dettegli matematici, il principio di questo algoritmo e' quello di creare cluster attraverso lo scambio di messaggi tra coppie di data points (campioni), questi messaggi informano ognugno dell'*affinita'* con ogni altro. Ogni elemento del dataset "decide" man mano che i messaggi vengono scambiati chi dovrebbe essere il suo elemento piu' rappresentativo; questa informazione viene comunicata anch'essa tramite messaggi, la scelta del rappresentate viene infatti fatta anche rispetto alle valutazioni degli altri sul proprio rappresentante.
Questo avviene fino alla convergenza dell'algoritmo, in cui ogni elemento ha "deciso" il suo elemento rappresentante, tutti gli elementi con lo stesso rappresentante finisco nel medesimo cluster. Un set di dati viene quindi descritto utilizzando un piccolo numero di esemplari, che vengono identificati come quelli più rappresentativi di altri campioni.
Uno dei principali vantaggi di questo algoritmo e' il fatto che il numero di cluster viene identificato automaticamente, la scelta di questo algoritmo e' stata fatta basandosi sulla forma del dataset ridotto dimensionalmente. Si possono notare due *distinti* raggruppamenti, uno piu' grande e uno piu' piccolo, *non molto densi ma abbastanza sparsi ,* con al loro interno sotto raggruppamenti *di dimensione molto diversa*, questa e' una situazione in cui e' consigliabile utilizzare affinity propagation [Documentazione scikit-learn].