Python per il Calcolo Scientifico

Angelo Cardellicchio

Clustering

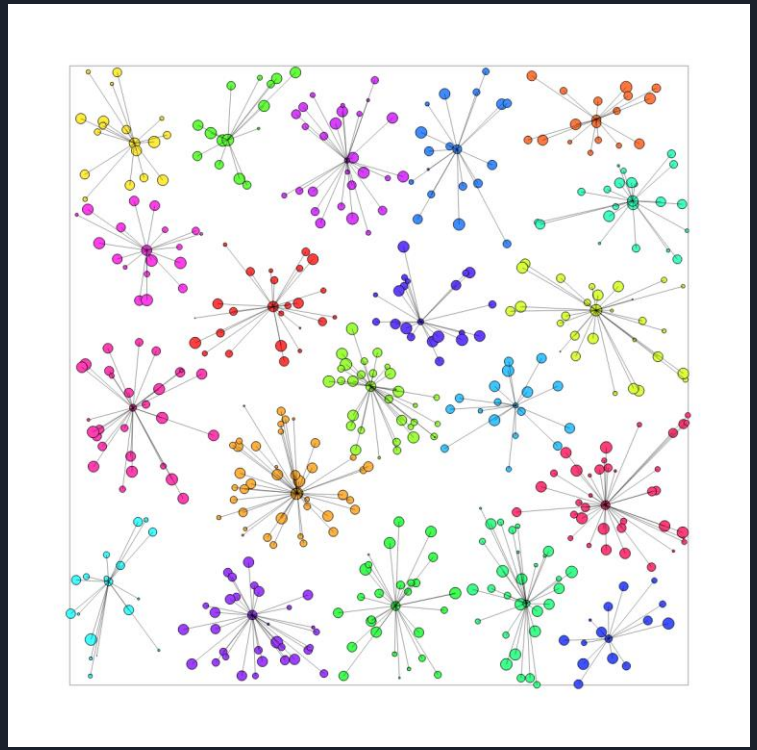
- Cosa è il clustering?
- Un approccio non supervisionato
- Metriche (alcuni esempi)
- Algoritmi (giusto un paio)

Cosa è il clustering?

- **Gli algoritmi di clustering inferiscono suddividono i dati in gruppi omogenei, detti appunto** *cluster*.
- **L'obiettivo di questi algoritmi è** *raggruppare* **i dati più "simili",** *separando* **quelli più complessi.**
- **Sono algoritmi che si basano fortemente sulle nozioni di distanza dei dati nell'iperspazio delle feature; ognuno ha sia pregi, sia difetti.**

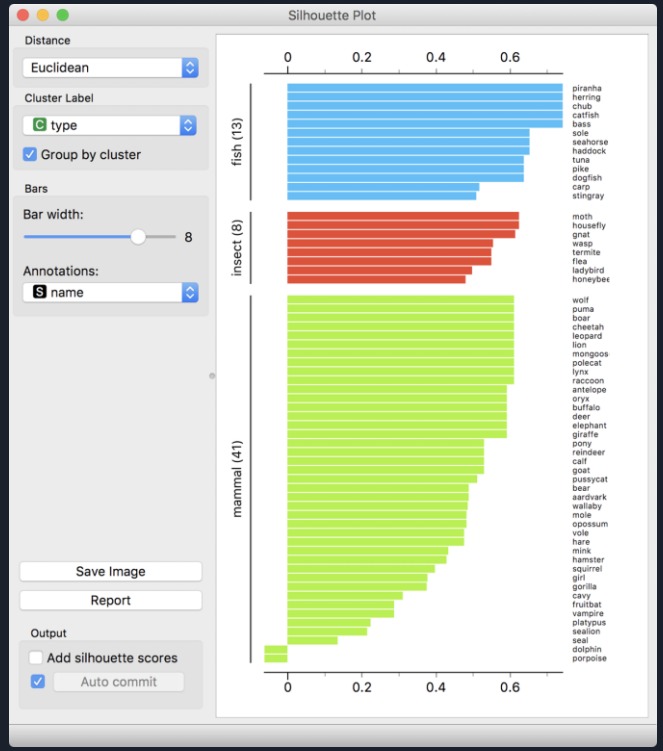
Un approccio *non* supervisionato

- Quelli di clustering sono tipicamente degli algoritmi *non* supervisionati.
- Ciò significa che i cluster vengono inferiti direttamente dai dati, e non è necessario alcun intervento da parte dell'esperto di dominio.



Metriche (alcuni esempi)

- **Silhouette Score:** è una misura di quanto ogni campione in un cluster assomiglia agli altri campioni nello stesso cluster, e di quanto contestualmente diverge da quelli presenti negli altri.
- Viene calcolata a partire dalla distanza tra i due campioni nell'iperspazio delle feature.
- **Non prevede la conoscenza pregressa di alcuna label.**



Metriche (alcuni esempi)

- **Rand Index:** a differenza del silhouette score, prevede la conoscenza del ground truth (ovvero, del valore ‘vero’ dei cluster)
- È una misura di similarità tra i risultati ottenuti dall’algoritmo di clustering e quelli considerati ‘veri’
- Analiticamente:

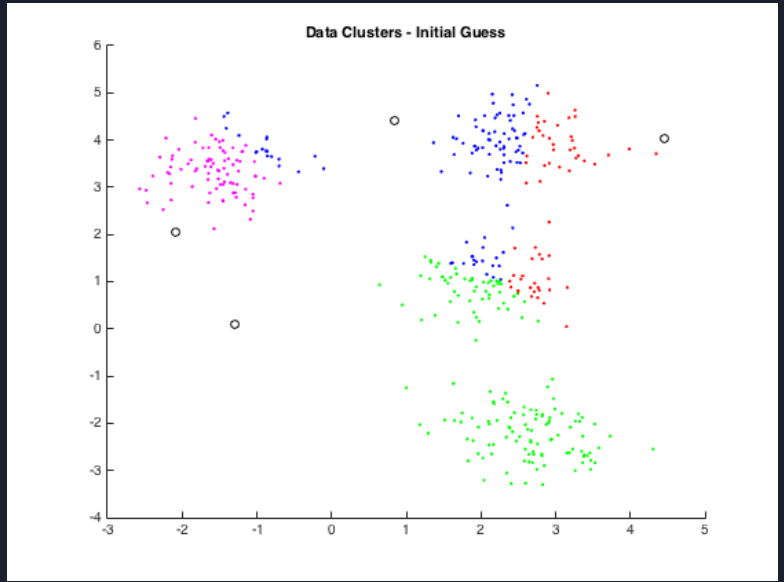
$$RI = \frac{n_a}{n_p}$$

- dove n_a è il numero di elementi il cui cluster è correttamente assegnato, ed n_p il numero totale di elementi.
- Ne esiste una versione normalizzata, chiamata **adjusted rand index:**

$$ARI = \frac{(RI - E[RI])}{(\max(RI) - E[RI])}$$

Algoritmi (giusto un paio)

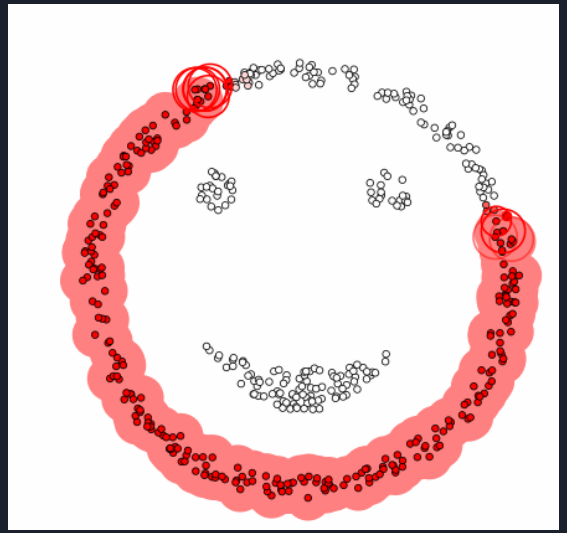
- **KMeans**
  - Si basa sul concetto di distanza nello spazio delle feature
  - Numero di cluster determinato a priori
  - Cluster convessi
  - Diversi presupposti (isotropia e gaussianità dei dati)



Fonte

Algoritmi (giusto un paio)

- **DBSCAN**
  - Supera molti dei limiti del KMeans
  - Clustering agglomerativo
  - Molto sensibile ai parametri **epsilon** e **min_samples**



Fonte