Studio e approfondimento del coefficiente Silhouette per la valutazione interna di risultati di clustering non-supervisionato

XXX YYY ZZZZZZ

February 10, 2025

Clustering

Da dizionario Merriam-Webster:

Cluster

"A number of similar things that occur together"

Cluster analysis

"A statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics"

Quanti sono i cluster?

- Gli algoritmi di clustering non sono in grado di fornire una metrica oggettiva per determinare quale sia il corretto numero di cluster
- Questo é particolarmente problematico negli algoritmi che hanno il numero di cluster come iperparametro (e.g. K-Means)
- Nel clustering non supervisionato non vi è a disposizione alcuna "ground truth"
- Un numero di cluster errato induce un raggruppamento non rappresentativo del dataset, bensí un raggruppamento artificioso

Introduzione di Silhouette

Silhouette si propone di rispondere alle seguenti domande:

- Il clustering è di buona qualità?
- Quali elementi sono stati ben classificati?
- Quali elementi sono inclassificabili?
- Il numero di cluster scelto è rappresentativo del dataset?

Situazione iniziale

- Si consideri un dataset di dimensione $N \times M$ (N elementi, M attributi)
- Si assuma che gli attributi siano tutti dati numerici (altezze, lunghezze, ecc...) e che siano tutti noti
- Da questo é possibile costruire una **matrice delle distanze** $N \times N$
- La cella (i, j) contiene il valore $d(i, j)$, che rappresenta il grado di "dissomiglianza" fra i e j

Funzione di distanza

Il grado di dissomiglianza fra i e j è calcolato mediante una **funzione di distanza** applicata ai loro attributi. Ad esempio:

Distanza Euclidea

$$d(i,j) = \sqrt{\sum_{m=1}^{M} (f_{i,m} - f_{j,m})^2}$$

Con $f_{i,m}$, $f_{j,m}$ valori del m-esimo attributo

Altri esempi: **distanza di Manhattan**, **distanza di Minkowski**.

K-Means

"1"	"2"	"3"	"4"	"5"	"6"
0	0.54	0.51	0.65	0.14	0.62
0.54	0	0.3	0.33	0.61	1.09
0.51	0.3	0	0.24	0.51	1.09
0.65	0.33	0.24	0	0.65	1.17
0.14	0.61	0.51	0.65	0	0.62
0.62	1.09	1.09	1.17	0.62	0

Table: Matrice delle distanze per il dataset `iris` (solo i primi 6 elementi).

Nota la matrice delle distanze, si applichi ad esempio K-Means per suddividere il dataset in K cluster

Fine

Domande?