CLASTER QUALITY

Bisogna distinguere due scenari:

1) EXTRINSIC: Metodo supervisionato dalla presenza del Ground Truth.

2) INTRINSIC: Metodo non-supervisionato. Il Ground Truth non c'è.

EXTRINSIC

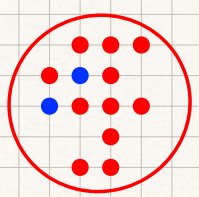
1) B-CUBED PRECISION e RECALL

Semplicemente valutando precision e recall per ogni oggetto nel cluster, ma prima dobbiamo introdurre il concetto di correttezza:

$$\text{CORRECTNESS}(o_i, o_j) = \begin{cases} 1 & \text{se } \begin{matrix}\text{Classe}(o_i) = \text{classe}(o_j) \\ \text{Cluster}(o_i) = \text{cluster}(o_j)\end{matrix} \\ \emptyset & \text{altrimenti} \end{cases}$$

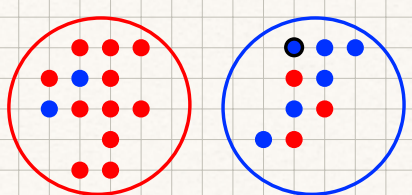
Nella corretta notazione $\text{Classe}(o) \Rightarrow L(o)$ / $\text{Cluster}(o) \Rightarrow C(o)$

PRECISION: la precisione di un oggetto o indica quanti altri oggetti nello stesso cluster di o hanno la stessa classe di o:



per gli oggetti 🔴 abbiamo una precision di 10 su 12, mentre per 🔵 di 2 su 12.

RECALL: la recall di un oggetto o ci dice quanti altri oggetti della stessa classe di o sono assegnati a uno stesso cluster



Analizzando 🔵, abbiamo che 6 oggetti, aventi stessa classe, su 8 sono assegnati allo stesso cluster di 🔵

$$PRECISION = \frac{\sum_{i=1}^{n} \frac{\sum_{C(o_i)=C(o_j)} CORRECTNESS(o_i, o_j)}{||\{o_j \mid i \neq j,\ C(o_i)=C(o_j)\}||}}{n}$$

$$RECALL = \frac{\sum_{i=1}^{n} \frac{\sum_{L(o_i)=L(o_j)} CORRECTNESS(o_i, o_j)}{||\{o_j \mid i \neq j,\ L(o_i)=L(o_j)\}||}}{n}$$

Dove con $L(o_i) = L(o_j)$ si intende oggetti con la stessa classe mentre $C(o_i) = C(o_j)$ stesso cluster.

I metodi di tipo extrinsic, e quindi supervisionati vengono usati in ambito sperimentale, usando dataset supervisionati per cercare di capire se un nuovo algoritmo di cluster funziona o no, e quanto è buono. Anche perché difficilmente in campo pratico avremo classe etichette.

INTRINSIC

1) SILHOUETTE COEFFICIENT

Per ogni oggetto $o \in D$ valutiamo:

- $a(o)$: Distanza media fra o e gli altri oggetti dello stesso cluster di o

- $b(o)$: Distanza media fra o e gli altri cluster diversi rispetto a quello di o, dove si prende o, si fa la distanza media fra o e i punti del cluster diverso da o.

Una buona divisione in cluster tende ad avere $a(o)$ tendente a 0 mentre $b(o)$ tendente a valori grandi.
Formalmente:

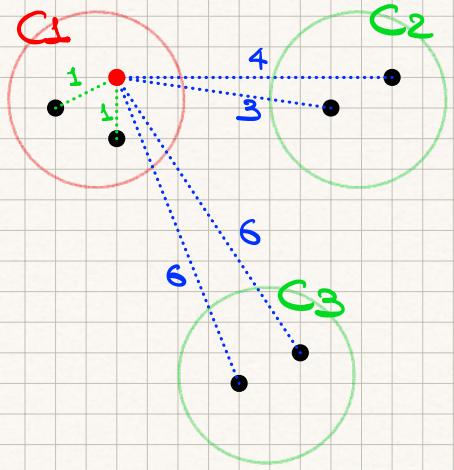
$$a(o) = \frac{\sum_{\substack{o' \in C_i \\ o' \neq o}} dist(o, o')}{|C_i| - 1}$$

$$b(o) = \underset{C_j : 1 \leq j \leq K, j \neq i}{\text{Minimo}} \left\{ \frac{\sum_{o' \in C_j} dist(o, o')}{|C_j|} \right\}$$

$$\text{SILHOUETTE} = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

Il silhouette coefficient sta fra $[-1, 1]$.

ESEMPIO GRAFICO



Considerando $o =$ ●

$$a(o) = \frac{1+1}{2} = 1$$

$$b(o) = \min\left(\frac{4+3}{2}\ ;\ \frac{6+6}{2}\right)$$

$$= \min(3.5;\ 6) = 3.5$$

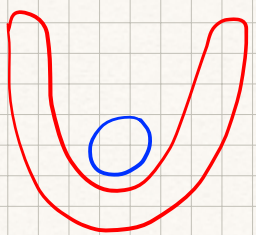
$$\text{SILHOUETTE} = \frac{3.5 - 1}{3.5} = \frac{2.5}{3.5} = 0.714$$

COME VALUTARE IL SILHOUETTE

Per avere il quadro generale di tutti i cluster trovati basta semplicemente fare la media di tutti i silhouette per ogni punto del dataset.

Il valore di $s(o) \in [-1; 1]$, dove 1 rappresenta il valore ottimale, dove quindi il punto o è compatto coi punti del suo cluster e lontano dagli altri cluster. Quando $s(o)$ è **negativo** vuol dire che $b(o) < a(o)$, ovvero che o è in media più vicino al cluster diverso dal suo che ai punti del suo cluster. BRUTTA SITUAZIONE NO.

Il valore -1 non è sempre negativo:



Se consideriamo la seguente situazione $a(o)$ è alto rispetto a $b(o)$ perché la distanza di un punto o rispetto al

cluster opposto è più piccola rispetto alla media della distanza da o ai suoi compagni di cluster.
Ma i cluster in figura sono ben distinti!
Il silhouette coefficient non funziona per cluster concavi.