HOPKINGS

Prima di valutare i cluster bisogna capire se il Dataset è propenso ad avere buoni cluster. Se i punti sono uniformemente distribuiti non riusciremo a trovare buoni cluster.

Hopkins è un test spaziale e statistico che testa la randomicità spaziale di una variabile distribuita nello spazio. Cerca di valutare la tendenza alla formazione di cluster del dataset, tramite la valutazione statistica che un set di dati abbia distribuzione uniforme.

PASSI

1) Campionare n punti da D, con $n \ll |D|$. Per ogni punto p_i trova il 1-NN in D. Definiamo x_i la distanza fra p_i e il suo vicino.

$$x_i = \min_{v \in D} \{ dist(p_i, v) \}$$

2) Genera un dataset randomico preso da una distribuzione uniforme una casuale con n punti $(q_1 \dots q_n)$ con la stessa variazione del set originale. Cercare ora y_i come la distanza di q_i e 1-NN in D

$$y_i = \min_{\substack{v \in D \\ v \neq q_i}} \{ dist(q_i, v) \}$$

3) Calcolare la statistica di Hopkins:

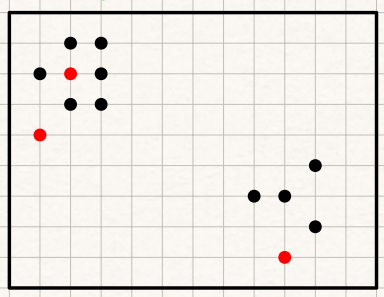
$$H = \frac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} x_i + \sum_{i=1}^{m} y_i}$$

Se D è uniformemente distribuito le sommatorie al denominatore dovrebbero essere simili, perciò H si aggira attorno a 1/2.

Altrimenti in presenza di cluster definiti il termine Σx_i dovrebbe essere più piccolo di Σy_i e ciò fa aumentare H.

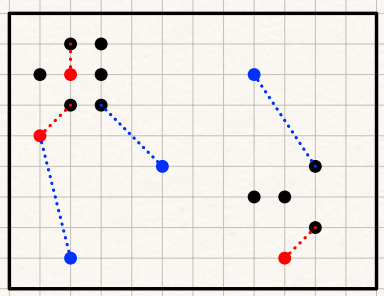
Esempio:

D:



Prendiamo $m = 3$, e selezioniamo 3 punti random. Quelli in rosso

RANDOM D



In blu i punti uniformemente distribuiti, ma randomicamente creati.
Se ci sono cluster ben definiti le distanza in blu (Σy_i) sono più grandi di quella in rosso (Σx_i).

NULL HYPOTESIS

Il dataset è uniformemente distribuito.

Solitamente Hopkins si valuta per diversi dataset $(q_1 \dots q_n)$ e poi si fa la media dei risultati.