

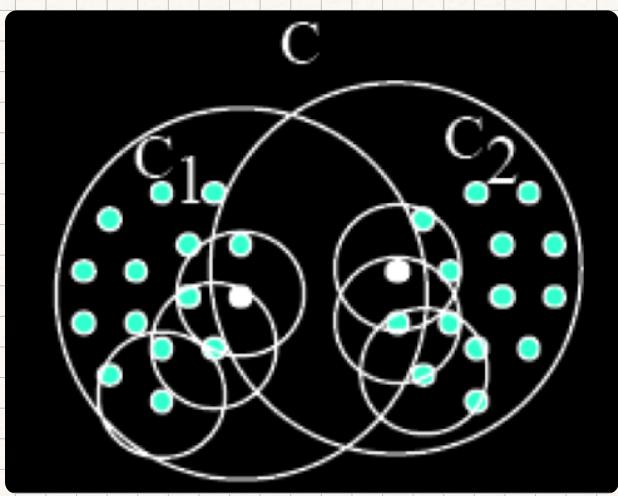
OPTICS

Ordering Points To Identify the Clustering Method

Difatti riordina il DB sulla base di informazioni di densità dei veri cluster presenti. Questo ordinamento è fatto sulla base della descrizione dei parametri di densità che ogni cluster ha.

In DBSCAN, fissato il valore di Minpoints, un cluster C_1 creato a partire da E_1 molto denso è completamente contenuto da un cluster C_2 con E_2 tale per cui la densità sia minore ($E_1 > E_2$)

Esempio



I cluster C_1 e C_2 ad alta densità sono completamente contenuti dal cluster C costituito con termini di densità inferiori.
 C_1 e C_2 con E_x
 C con E_y
 $E_x < E_y$

OPTICS non assegna un punto a un cluster bensì traccia l'ordine in cui i punti sono analizzati e le informazioni che sarebbero uscite da un DBSCAN avanzato per assegnare un punto a un cluster. Queste informazioni sono la **core-distance** e la **reachability-distance**:

CORE-DISTANCE

È il più piccolo ϵ che rende il punto p sotto processo, un CORE-OBJECT

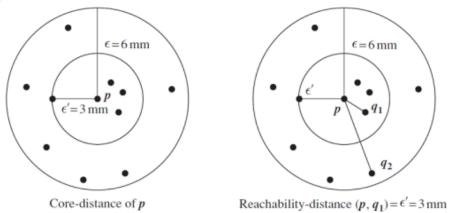
REACHABILITY-DISTANCE

È il minimo raggio tale per cui un punto p è directly density-reachable da un punto q .

Il punto p deve essere un core-object e il punto q deve stare nel vicinato di p

$$\text{reachability-distance}(p, q) = \max(\text{core-distance}(p) \text{ e } \text{dist}(p, q))$$

Se p non è un core-object la reachability-distance non è definita.



PASSI

1) Inizia partendo da un oggetto random in D : p .

Trova poi:

- E-vicinato(p)
- core-distance(p)
- reachability-distance(p)

2) If (p non è un CORE-OBJECT)

Va avanti al prossimo oggetto p nella lista ORDER-SEEDS. Se quest'ultima è vuota segue l'ordine del DB.

else ($p == \text{CORE-OBJECT}$)

$\forall q \in E\text{-vicinato}(p)$ aggiorna la reachability-dist

da q .

Inserisci q dentro la lista ORDER-SEEDS se q non è stato ancora processato.

- 3) La lista ORDER-SEEDS è ordinata per la metrice $\text{rechability-distance}$. (Ordine crescente)
- 4) Una volta determinato ϵ -vicinato (\bar{r}) e la core-distance (r), \bar{r} viene messo in un file detto ORDERED-FILE con la sua core-distance e rechability distance.
- 5) Finisce quando non ci sono più oggetti da processare.

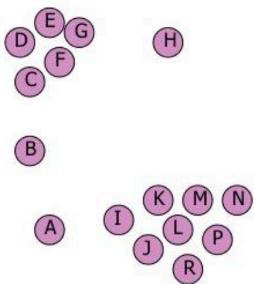
Nota

Alla fine abbiamo una lista di di oggetti ordinata nell'ordine in cui i punti vengono processati.

COMPLESSITÀ: $O(n^2)$

δ sempio

$\epsilon = 44$, MinPoints = 3

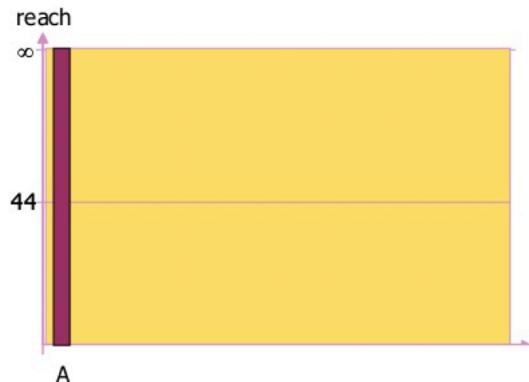
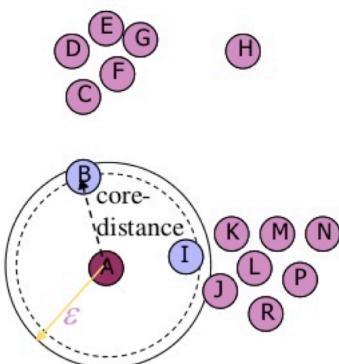


Partiamo in modo casuale da
A. Calcoliamo

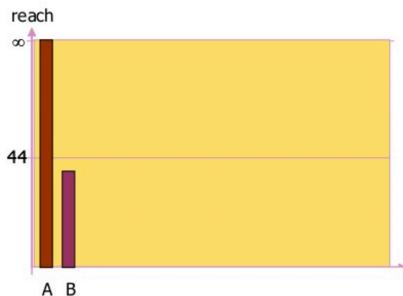
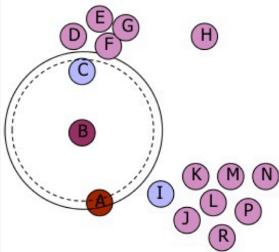
- 1) ϵ -vicinato(A) = {B, I}
- 2) core-distance(A) = $\text{dist}(A, B) = 40$
- 3) reachability- $d(A \rightarrow B) = 40$
- 4) reachability- $d(A \rightarrow I) = 40$

La reachability-distance $(A \rightarrow I) = 40$ poiché bisogna considerare il massimo fra $\text{dist}(A, I)$ e la core-distance(A).

$$\text{SeedList} = \{(B, 40), (I, 40)\}$$

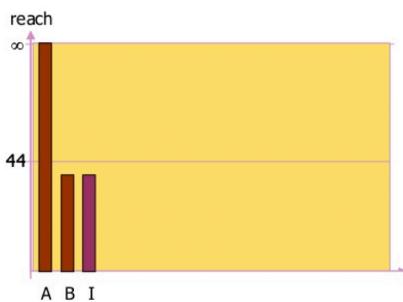
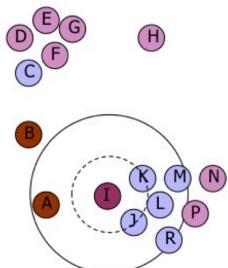


Il passo successivo è estrarre il primo punto da SeedList e analizzarlo: B



seedlist: (I, 40) (C, 40)

Ottieniamo ora I della SeedList



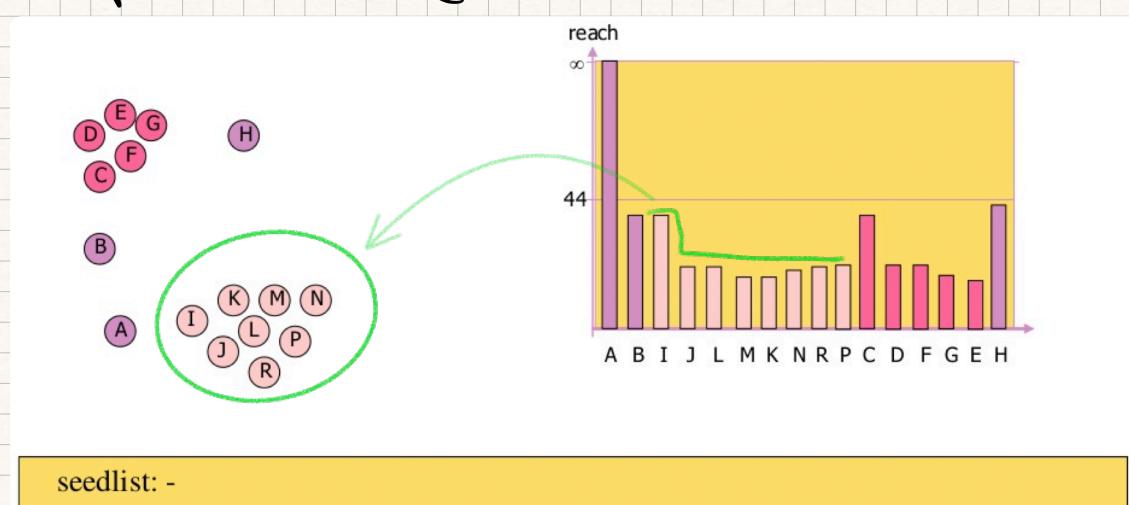
seedlist: (J, 20) (K, 20) (L, 31) (C, 40) (M, 40) (R, 43)

La SeedList è ordinata in ordine crescente di reachability-distance dell'oggetto vicino analizzato nella iterazione precedente:

$$\text{reachability-distance } (I \rightarrow I) = 20$$

Se nell'ingresso un punto nella SeedList, esso è già presente, si mantiene quello con reachability-d minore.

Alla fine dell'algoritmo:

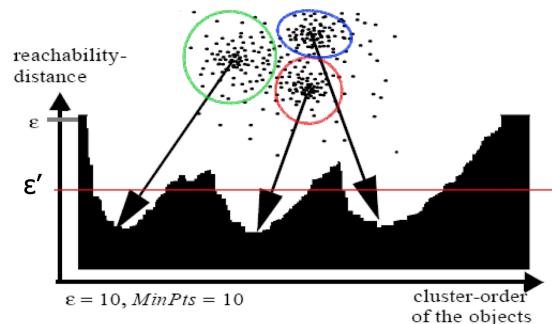


I cluster più densi tendono ad avere reachability-distance minore, e con altezza costante.

Obiettivo di OPTICS

L'algoritmo non produce dei cluster, ma una file dove i punti sono ordinati nell'ordine in cui optics li analizza, ovvero in ordine crescente di reachability distance.

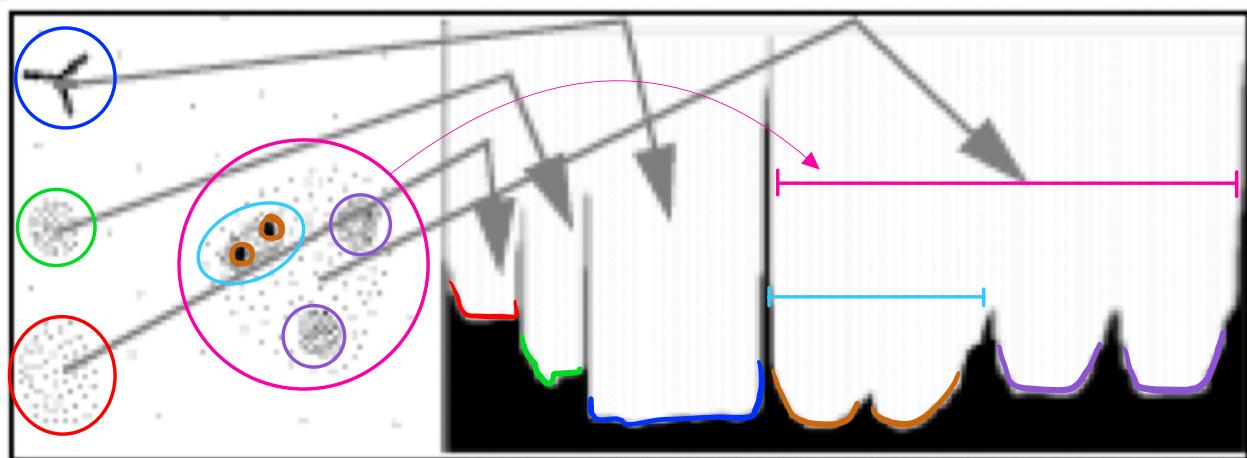
Stiamo cercando un ottimo ϵ fissando MinPoints.



Nell'esempio possiamo fissare $\epsilon = \epsilon'$ per poi usarlo in DBSCAN, così da trovare i tre cluster naturali presenti in questo esempio. Funziona perché D ha densità costante.

Cosa succede se D non ha densità costante?

Siamo in una situazione dove diversi cluster possono avere diverse densità. Se applichiamo OPTICS in questo contesto possiamo notare che nel grafico le "valle" fra le varie "cime" cambia per dimensione, in particolare per cluster meno densi le "valle" ha una estensione minore. Possiamo distinguere regioni ad alta densità e regioni a bassa densità.



Per dataset con densità non costanti è **impossibile** settare i parametri per l'intero D . Possiamo però ceprice con OPTICS, come fare un'analisi in una sotto regione del dataset.