9

OPTICS

9. OPTICS

OPTICS (Ordering Points To Identify the Clustering Structure) est un algorithme de partitionnement de données, proposé par Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel and Jörg Sander.

Il s'agit d'un algorithme basé densité dont le principe est similaire à DBSCAN mais élimine son principal défaut : l'impossibilité de détecter des partitions de densités différentes.

Etroitement lié à DBSCAN, trouve un échantillon central de haute densité et étend les clusters à partir de celui-ci. Contrairement à DBSCAN, il conserve la hiérarchie des clusters pour un rayon de voisinage variable. Mieux adapté à l'utilisation sur de grands ensembles de données que l'implémentation actuelle de DBSCAN par Sklearn.

Les clusters sont ensuite extraits en utilisant une méthode similaire à DBSCAN (cluster_method = 'dbscan') ou une technique automatique proposée dans (cluster_method = 'xi').

9. OPTICS

Principe général [modifier | modifier le code]

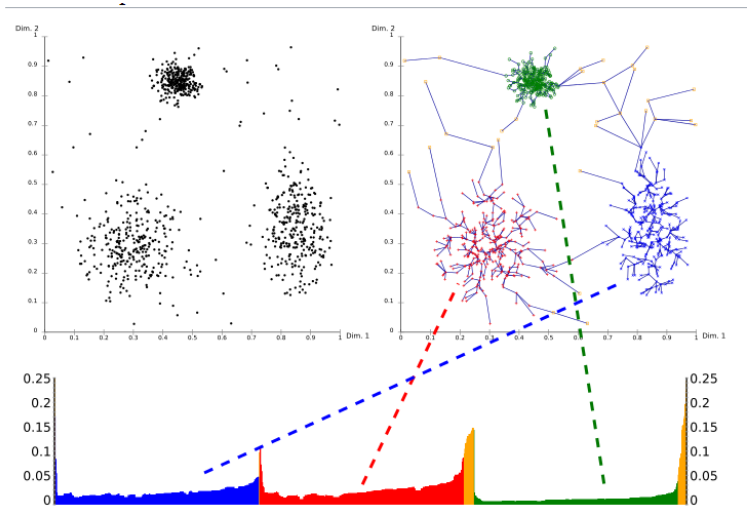
Comme DBSCAN, OPTICS demande deux paramètres : ε, définissant un rayon maximum à considérer, et $MinPts$, définissant un nombre de points minimum. Ces 2 paramètres définissent donc une densité minimale pour constituer un groupe de données. Un point p appartient à un groupe si au moins $MinPts$ points existent dans son ε-voisinage $N_\varepsilon(p)$. Par contre, à l'inverse de DBSCAN, le paramètre ε est optionnel. S'il est omis, il sera alors considéré comme infini. L'algorithme définit pour chaque point une distance, appelée *core distance*, qui décrit la distance avec le $MinPts$ème point le plus proche :

$$\text{core-distance}_{\varepsilon,MinPts}(p) = \begin{cases} \text{Indéfini} & \text{si } |N_\varepsilon(p)| < MinPts \\ \text{distance au } MinPts\text{-ème point le plus proche} & \text{sinon} \end{cases}$$

La *reachability-distance* du point p à un autre point o est la distance entre o et p, ou la *core-distance* de p:

$$\text{reachability-distance}_{\varepsilon,MinPts}(o,p) = \begin{cases} \text{Indéfini} & \text{si } |N_\varepsilon(p)| < MinPts \\ \max(\text{core-distance}_{\varepsilon,MinPts}(p), \text{distance}(p,o)) & \text{sinon} \end{cases}$$

La *core-distance* et la *reachability-distance* sont indéfinis si le groupe de points n'est pas suffisamment dense. Si ε est suffisamment grand, cela n'arrive jamais, mais toutes les requêtes d'ε-voisinage retourneront l'ensemble des points, la complexité étant alors en $O(n^2)$. Le paramètre ε est donc utile pour définir une densité minimale afin d'accélérer l'algorithme.




```
OPTICS(DB, eps, MinPts)
   pour chaque point p de DB
      p.reachability-distance = UNDEFINED
   pour chaque point non visité p de DB
      N = voisinage(p, eps)
      marquer p comme visité
      output p to the ordered list
      si (core-distance(p, eps, Minpts) != UNDEFINED)
         Seeds = priority queue vide
         update(N, p, Seeds, eps, Minpts)
         pour chaque q prioritaire dans Seeds
            N' = voisinage(q, eps)
            marquer q comme visité
            output q to the ordered list
            si (core-distance(q, eps, Minpts) != UNDEFINED)
               update(N', q, Seeds, eps, Minpts)
```

```
update(N, p, Seeds, eps, Minpts)
   coredist = core-distance(p, eps, MinPts)
   pour chaque o dans N
      si (o n'a pas été visité)
         new-reach-dist = max(coredist, dist(p,o))
         si (o.reachability-distance == UNDEFINED)
             o.reachability-distance = new-reach-dist
             Seeds.insert(o, new-reach-dist)
         sinon
             si (new-reach-dist < o.reachability-distance)
                o.reachability-distance = new-reach-dist
                Seeds.move-up(o, new-reach-dist)
```

9. OPTICS

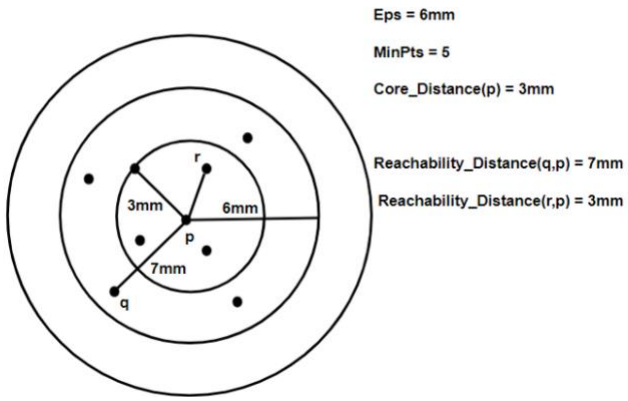
Comme DBSCAN, OPTICS requiert deux paramètres :

ε. qui décrit la distance maximale (rayon) à prendre en compte, et.
MinPts. qui décrit le nombre de points requis pour former un cluster.

Voici quelques définitions :

Point central. Un point p est un point central si au moins MinPts points sont trouvés dans son ε-voisinage,
Distance centrale. Il s'agit de la valeur minimale du rayon nécessaire pour classer un point donné comme point core. Si le point donné n'est pas un point central, alors sa distance centrale est indéfinie. Vous trouverez ci-dessous un exemple de Core Distance.

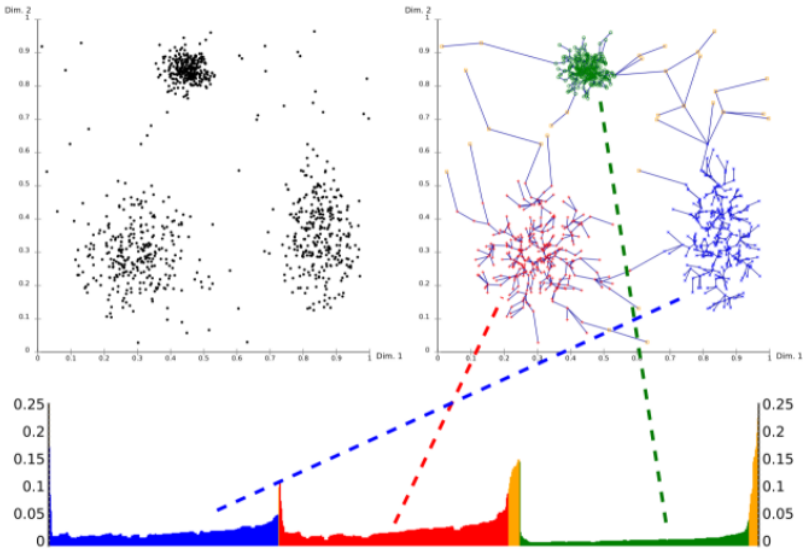
Distance d'atteignabilité. Elle est définie par rapport à un autre point de données q. La distance de joignabilité entre un point p et q est le maximum de la distance centrale de p et de la distance entre p et q. Notez que la distance de joignabilité n'est pas définie si q n'est pas un point central. Voici un exemple de la distance d'accessibilité.

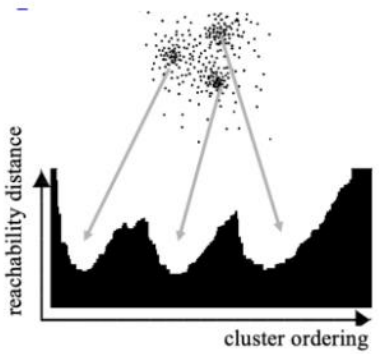


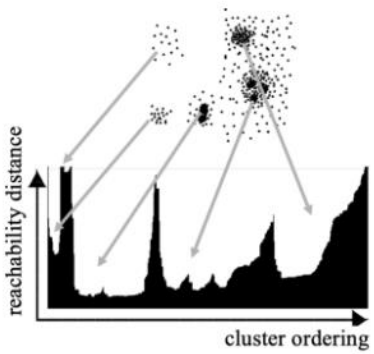
9. OPTICS

Le graphe d'accessibilité au clustering

En utilisant un reachability-plot (un type spécial de dendrogramme), la structure hiérarchique des clusters peut être facilement obtenue. Il s'agit d'un graphique en 2D, avec en abscisse l'ordre des points tels que traités par OPTICS et en ordonnée la distance d'accessibilité. Puisque les points appartenant à un cluster ont une faible distance d'accessibilité à leur voisin le plus proche, les clusters apparaissent comme des vallées dans le graphique d'accessibilité. Plus la vallée est profonde, plus le cluster est dense.

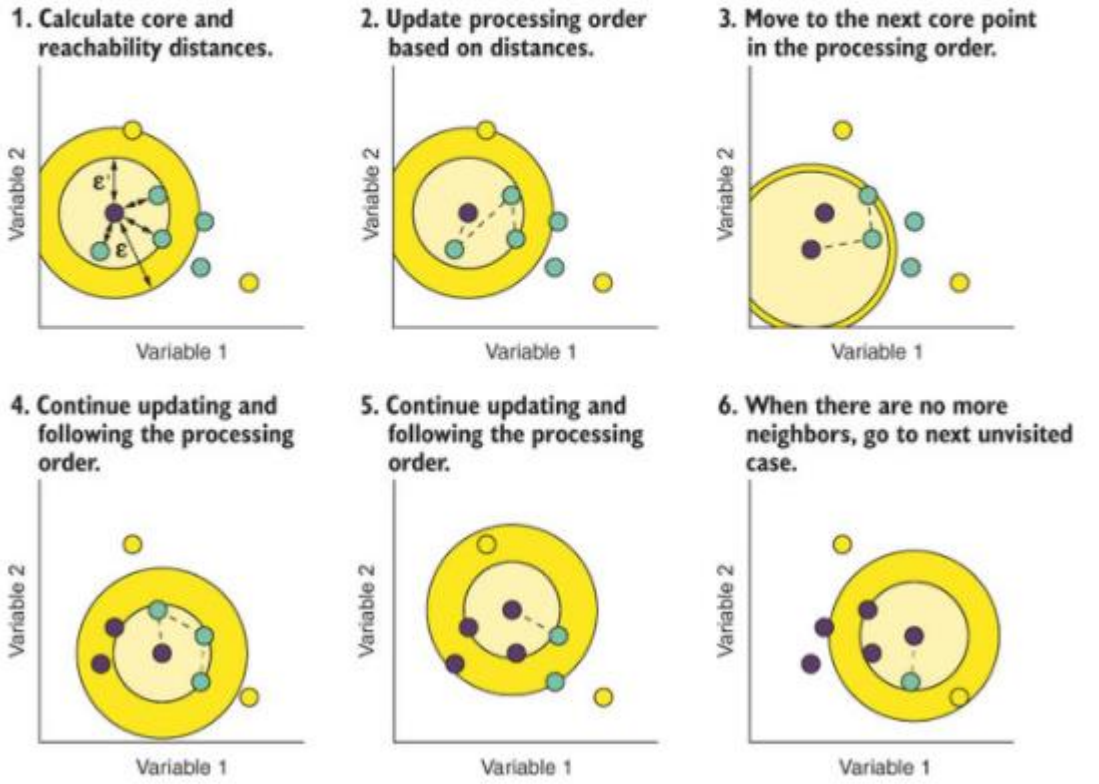






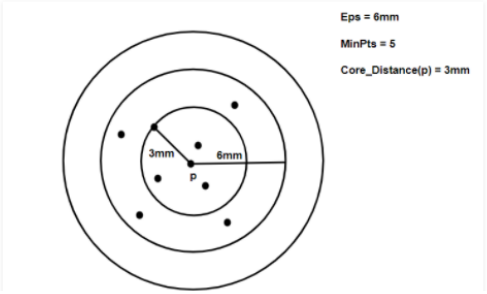
9. OPTICS

Figure 18.5. The OPTICS algorithm. A case is selected, and its core distance (ϵ') is measured. The reachability distance is calculated between this case and all the cases inside this case's maximum search distance (ϵ). The processing order of the dataset is updated such that the nearest case is visited next. The reachability score and the processing order are recorded for this case, and the algorithm moves on to the next one.

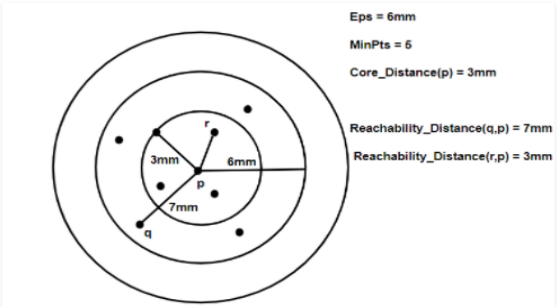


9. OPTICS

1. **Core Distance:** It is the minimum value of radius required to classify a given point as a core point. If the given point is not a Core point, then it's Core Distance is undefined.



2. **Reachability Distance:** It is defined with respect to another data point q(Let). The Reachability distance between a point p and q is the maximum of the Core Distance of p and the Euclidean Distance(or some other distance metric) between p and q. Note that The Reachability Distance is not defined if q is not a Core point.



	DBSCAN	OPTICS
Density	Boolean value (high/low)	Numerical value (core distance)
Density-connected	Boolean value (yes/no)	Numerical value (reachability distance)
Searching strategy	random	greedy

Clustering OPTICS vs Clustering DBSCAN :

Coût en mémoire : La technique de clustering OPTICS nécessite plus de mémoire car elle maintient une file d'attente prioritaire (Min Heap) pour déterminer le prochain point de données le plus proche du point en cours de traitement en termes de distance d'accessibilité. Elle nécessite également plus de puissance de calcul car les requêtes de plus proches voisins sont plus compliquées que les requêtes de rayon dans DBSCAN.

Moins de paramètres : La technique de clustering OPTICS n'a pas besoin de maintenir le paramètre epsilon qui n'est donné que dans le pseudo-code ci-dessus pour réduire le temps nécessaire. Cela conduit à la réduction du processus analytique de réglage des paramètres.
Cette technique ne sépare pas les données données en clusters. Elle produit simplement un graphique de distance d'accessibilité et c'est au programmeur d'interpréter le regroupement des points en conséquence.

Cette technique ne sépare pas les données données en clusters. Elle produit simplement un graphique de distance d'accessibilité et c'est au programmeur d'interpréter le regroupement des points en conséquence.

OPTICS est relativement insensible aux réglages des paramètres. Bon résultat si les paramètres sont juste "assez grands".

9. OPTICS

Hyperparamètres :

• **min_samples** : le nombre d'échantillons dans un voisinage pour qu'un point soit considéré comme un point central. De même, les régions à forte pente ascendante et descendante ne peuvent pas avoir plus de min_échantillons de points consécutifs non pentus. Exprimé comme un nombre absolu ou une fraction du nombre d'échantillons (arrondi pour être au moins 2).
• **xi** : détermine la pente minimale sur le graphe d'atteignabilité qui constitue une frontière de cluster. Par exemple, un point ascendant dans le graphe d'accessibilité est défini par le rapport d'un point à son successeur étant au plus 1-xi. Utilisé uniquement lorsque cluster_method='xi’.
• **min_cluster_size** : nombre minimum d'échantillons dans un cluster OPTICS, exprimé comme un nombre absolu ou une fraction du nombre d'échantillons (arrondi pour être au moins 2). Si None, la valeur de min_samples est utilisée à la place. Utilisé uniquement lorsque cluster_method='xi'.