Density-based clustering

- Clusterings computed by Lloyd's algorithm or by agglomerative clustering often do not compute clusterings that practitioners find intuitive or useful.
- Hence, there are many other clustering algorithms that try to find intuitive clusterings.
- These algorithms usually do not try do optimize some objective function (like Lloyd's algorithm or agglomerative clustering.
- DBSCAN (=density based spatial clustering of applications with noise) is an important example.
- It computes geometrically well-defined clusterings.

Density-based clustering

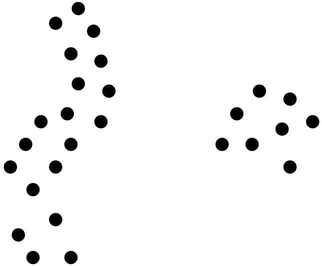


Figure: Geometric clusters defined by densities

Density-based clustering

$D : M \times M \to \mathbb{R}$ symmetric distance measure,
$P \subset M, \epsilon > 0, \mathrm{MinPts} \in \mathbb{N}$

Definition 7.1

1. *The ϵ-neighborhood $N_\epsilon(p)$ of a point p is defined as*

$$N_\epsilon(p) = \{q \in P : D(p, q) \leq \epsilon\}.$$

2. *$p \in P$ is called core point, if $|N_\epsilon(p)| \geq MinPts$.*
3. *$p \in P$ is called border point, if $|N_\epsilon(p)| < MinPts$.*

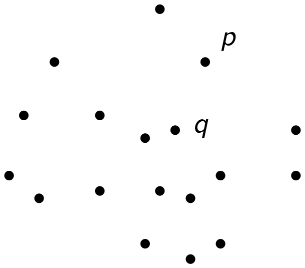
Density-based clustering

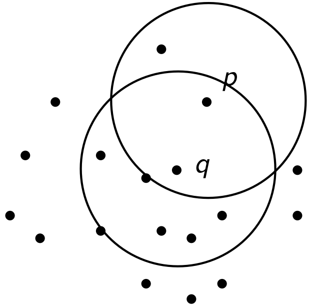
Definition 7.2

$P \subset M, \epsilon > 0, MinPts \in \mathbb{N}, p, q \in P$

1. *p is directly density reachable form q (wrt. $\epsilon, MinPts$), if*
   (a) *$p \in N_\epsilon(q)$ and*
   (b) *$|N_\epsilon(q)| \geq MinPts$, i.e. q is a core point.*
2. *p is density reachable from q (wrt. $\epsilon, MinPts$) , if there is a sequence of points $p_1, \ldots, p_n, p_1 = q, p_n = p$ such that p_{i+1} is directly density reachable from p_i.*
3. *p is density connected to point q (wrt. $\epsilon, MinPts$), if there is a point $r \in P$ such that p and q are density reachable from r.*

Density-based clustering





p border point
q core point

p directly density reachable from q
q not directly density reachable from p

Density-based clustering

Definition 7.3

$P \subset M, \epsilon > 0, MinPts \in \mathbb{N}$. A subset $C \subseteq P$ is called a density-based cluster (wrt. $\epsilon, MinPts$), if

(a) $\forall p, q \in P : (p \in C$ *and* q *density reachable from* $p) \Rightarrow q \in C$

(b) $\forall p, q \in C$*:* p *is density connected to* q.

Definition 7.4

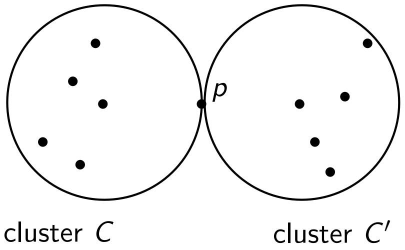
$P \subset M, \epsilon > 0, MinPts \in \mathbb{N}$. Let $C_1, \ldots, C_k$ be the clusters of P (wrt. $\epsilon, MinPts$). A point $q \notin \bigcup_{i=1}^{k} C_i$ is called a noise point.

Density-based clusters

Lemma 7.5

Let C, C' be two distinct density-based clusters of point set P wrt. to ϵ and MinPts. Then the intersection $C \cap C'$ of C, C' contains only border points.

p border point



Characterization of density-based clusters

Lemma 7.6

Let $p \in P$ such that $|N_\epsilon(p)| \geq MinPts$. Then the set

$$O := \{o \in P : o \text{ is density-reachable from } p \text{ w.r.t } \epsilon \text{ and } MinPts\}$$

is a cluster w.r.t. ϵ and MinPts.

Lemma 7.7

Let C be a cluster of P w.r.t. ϵ and MinPts and let p be any point in C with $|N_\epsilon(p)| \geq MinPts$. Then

$$C = \{o \in P : o \text{ is density-reachable from } p \text{ w.r.t } \epsilon \text{ and } MinPts\}.$$

Algorithm DBSCAN

DBSCAN(P)

$i := 0, U := P$ `/* U unclassified */`

repeat

  choose $p \in U$;

  if DENSREACH$(p, P) \neq \emptyset$ **then**

    $i := i + 1, C_i :=$ DENSREACH$(p, P), U := U \setminus C_i$

  else

    $U := U \setminus \{p\}$

  end

until $U = \emptyset$;

$N := P \setminus \bigcup C_j$;

return $C_1, \ldots, C_k$ as clusters and N as set of noise points

Algorithm DensReach

DensReach(p, P)

if $|N_\epsilon(p)| < MinPts$ **then**
  return $\emptyset$
else
  $C := \{p\}$, $C' := \{p\}, F := \emptyset$;
  /* C cluster, C'/F reached/finished corepoints */
  repeat
    choose $q \in C' \setminus F$;
    $C' := C' \cup \{r \in N_\epsilon(q) : |N_\epsilon(r)| \geq \text{MinPts}\}$;
    $C := C \cup N_\epsilon(q), F := F \cup \{q\}$;
  until $C' \setminus F = \emptyset$;
end
return C

Algorithm Dbscan

Theorem 7.8

On input a finite point set P, algorithm Dbscan *computes a partitioning of P into density-based clusters and noise points as defined in Definition 7.3 and in Definition 7.4.*