Hierical Clustering Analysis & Density clustering

9/28

Analysis of agglomerative Hierarchial Clustering

Assume that dist comp takes const time

① init pwdm $O(n^2)$

② for each merge step 1 less cluster

⇒ at step t recompute $O(n-t)$ dists

⇒ # of dists

step	1	2	3	...	$n-1$
# of dists	$n-1$	$n-2$	$n-3$	...	1

⇒ total # of dists in alg $(n-1)+(n-2)+(n-3)\cdots+1 = O(n^2)$

③ at step i we need to:

1. find min dist
2. remove old dist (from individual clusters)
3. add new dist (for the merged cluster)

Lets use Priority Queue implemented w/ a heap (w/ arbitrary delete)

for heap of size k

1. find min: $O(1)$
2. update: $O(\log k)$

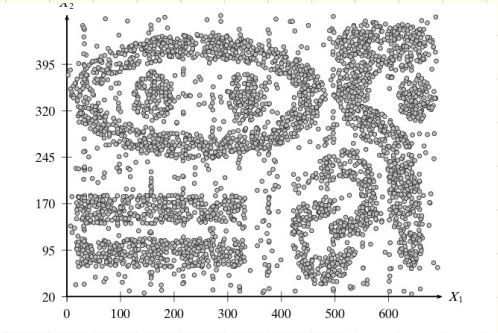
Heap Review

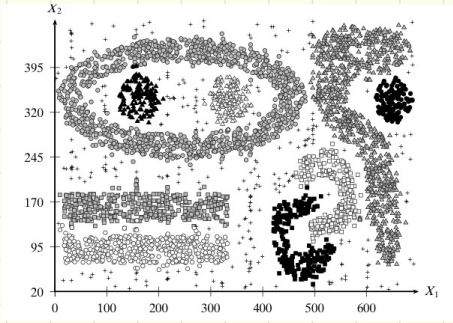
A. init the heap w/ $\binom{n}{2}$ elements - $O(n^2)$

B. perform $O(n^2)$ dist updates ⇒ $O(n^2 \log n^2) = O(n^2 \log n)$

⇒ total time is $O(n^2 \log n)$

Density based clustering





DBScan

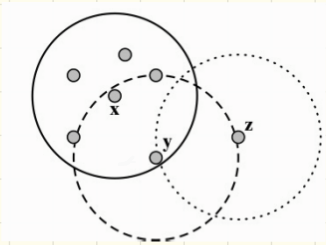
using Euclidean dist but in general can use any other metric

Let $\varepsilon > 0$ and $\vec{x} \in \mathbb{R}^d$

ε-neighborhood of $\vec{x}$

$$N_\varepsilon(\vec{x}) = B_d(\vec{x}, \varepsilon) = \{ \vec{y} \in D \mid \|\vec{y} - \vec{x}\| \leq \varepsilon \}$$

define $|N_\varepsilon(\vec{x})| = |\{\vec{y} \in D \mid \vec{y} \in N_\varepsilon(\vec{x})\}|$



minpts = 6
x is core pt
y is bdd
z is noise

Given param minpts $\in \mathbb{Z}^+$

for any $\vec{x} \in D$

$\vec{x}$ is a **core point** if there are at least minpts in its ε-neighborhood

$$|N_\varepsilon(\vec{x})| \geq \text{minpts}$$

$\vec{x}$ is a **border point** if it is not a core point and $\vec{x}$ is in ε-neighborhood of a core point

$$|N_\varepsilon(\vec{x})| < \text{minpts}$$

$$\exists \vec{y} \in D \ni \vec{x} \in N_\varepsilon(\vec{y}) \text{ and } |N_\varepsilon(\vec{y})| \geq \text{minpts}$$

$\vec{x}$ is a **noise point** if not core point nor a border point

Given $\vec{x}, \vec{y} \in D$

- $\vec{x}$ is **directly density reachable** from $\vec{y}$
  if $\vec{y}$ is a core point and $\vec{x} \in N_\epsilon(\vec{y})$

- $\vec{x}$ is **density reachable** from $\vec{y}$
  if $\exists$ seq $\vec{x} = \vec{x}_0, \vec{x}_1, \vec{x}_2, \dots, \vec{x}_\ell = \vec{y}$
  w/ $\vec{x}_i$ is directly reachable from x_{i-1} for $i = 1, \dots, \ell$

- $\vec{x}, \vec{y}$ **density connected** if $\exists$ a core point $\vec{z}$
  s.t. $\vec{y}$ & $\vec{x}$ are density reachable from $\vec{z}$

- **Density base clustering** is find maximal set of density connected pts