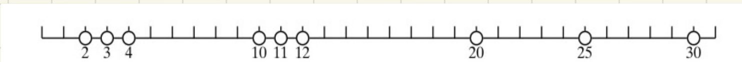
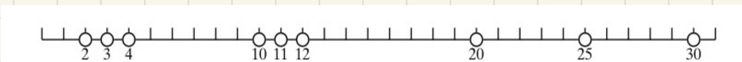
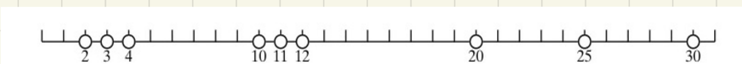
DBSCAN

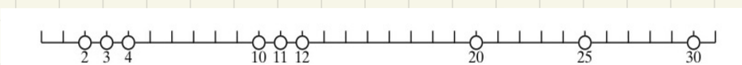
Clustering So Far

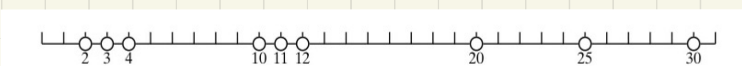
K-means

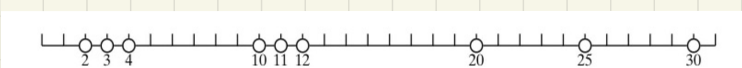




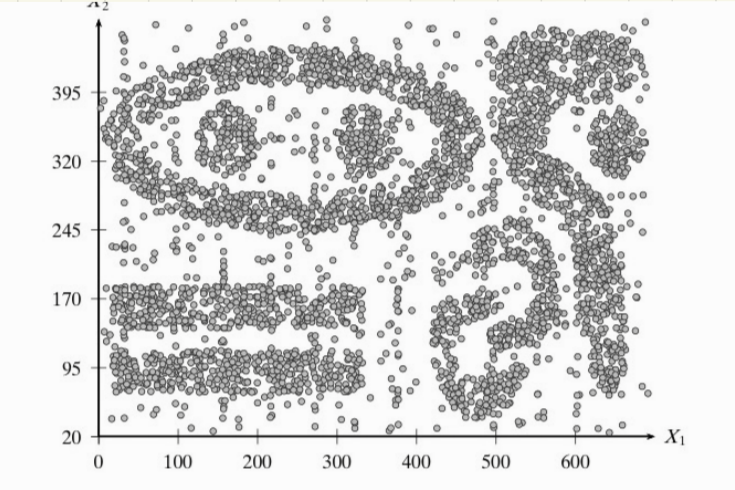








Even More Complex



Defs

Given Dataset D

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

ε-neighborhood of x

$$N_\varepsilon(x) = \{ y \in D : \|y - x\| \leq \varepsilon \}$$

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

Example minpts = 6

For any $x \in D$

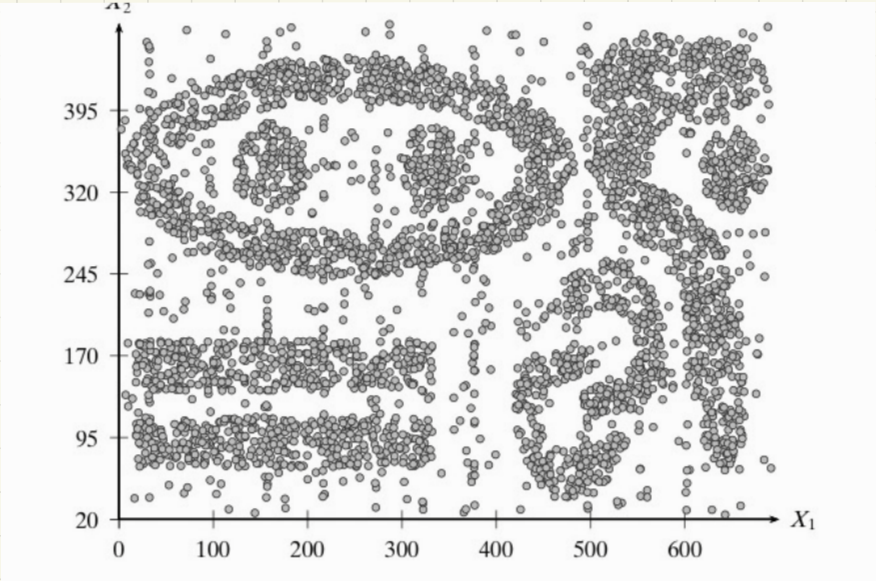
x is a core point if
there are at least
minpts in its ϵ-neighborhood

x is a border point if
not a core pt and in
ϵ-neighborhood of core pt

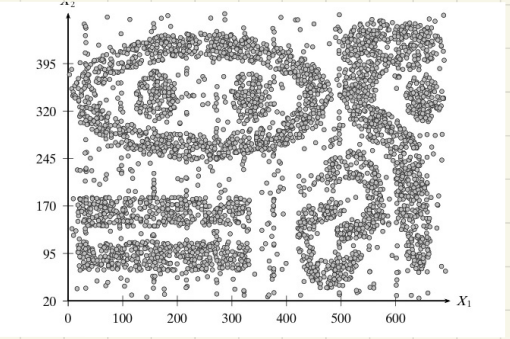
x is noise if !border & !core

DBSCAN algo (high level)

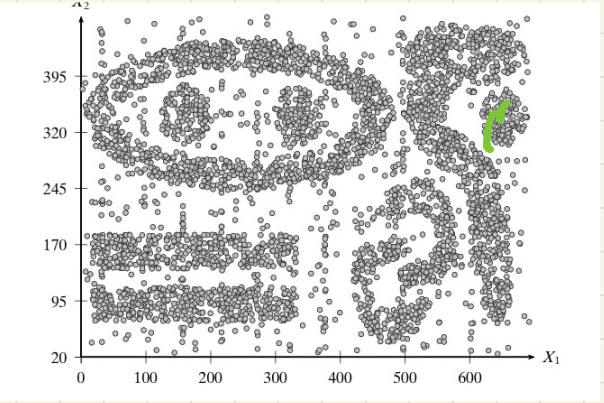
1. $\forall x \in D$
   - compute $N_\varepsilon(x)$
   - check if core pt
2. $\forall x \in$ Core Points
   - recursively find all "density connected" points



Pseudo Code (Find Concepts)



Pseudo code (find clusters)



Pseudo code (Density Connected)

Full Pseudocode

Algorithm 15.1: Density-based Clustering Algorithm

DBSCAN $(\mathbf{D}, \epsilon, minpts)$**:**

$Core \leftarrow \emptyset$
foreach $\mathbf{x}_i \in \mathbf{D}$ **do** // Find the core points
    Compute $N_\epsilon(\mathbf{x}_i)$
    $id(\mathbf{x}_i) \leftarrow \emptyset$ // cluster id for $\mathbf{x}_i$
    if $N_\epsilon(\mathbf{x}_i) \geq minpts$ **then** $Core \leftarrow Core \cup \{\mathbf{x}_i\}$
$k \leftarrow 0$ // cluster id
foreach $\mathbf{x}_i \in Core$, *such that* $id(\mathbf{x}_i) = \emptyset$ **do**
    $k \leftarrow k + 1$
    $id(\mathbf{x}_i) \leftarrow k$ // assign $\mathbf{x}_i$ to cluster id k
    DENSITYCONNECTED $(\mathbf{x}_i, k)$
$\mathcal{C} \leftarrow \{C_i\}_{i=1}^{k}$, where $C_i \leftarrow \{\mathbf{x} \in \mathbf{D} \mid id(\mathbf{x}) = i\}$
$Noise \leftarrow \{\mathbf{x} \in \mathbf{D} \mid id(\mathbf{x}) = \emptyset\}$
$Border \leftarrow \mathbf{D} \setminus \{Core \cup Noise\}$
return $\mathcal{C}, Core, Border, Noise$

DENSITYCONNECTED $(\mathbf{x}, k)$**:**

foreach $\mathbf{y} \in N_\epsilon(\mathbf{x})$ **do**
    $id(\mathbf{y}) \leftarrow k$ // assign $\mathbf{y}$ to cluster id k
    if $\mathbf{y} \in Core$ **then** DENSITYCONNECTED $(\mathbf{y}, k)$