Lecture 22 - Apr 8

Density-based Clustering
DBSCAN
OPTICS

<u>References</u>

Data Mining and Machine Learning
Ch 15 - Density-based Clustering

Where K-Means and EM Fails…

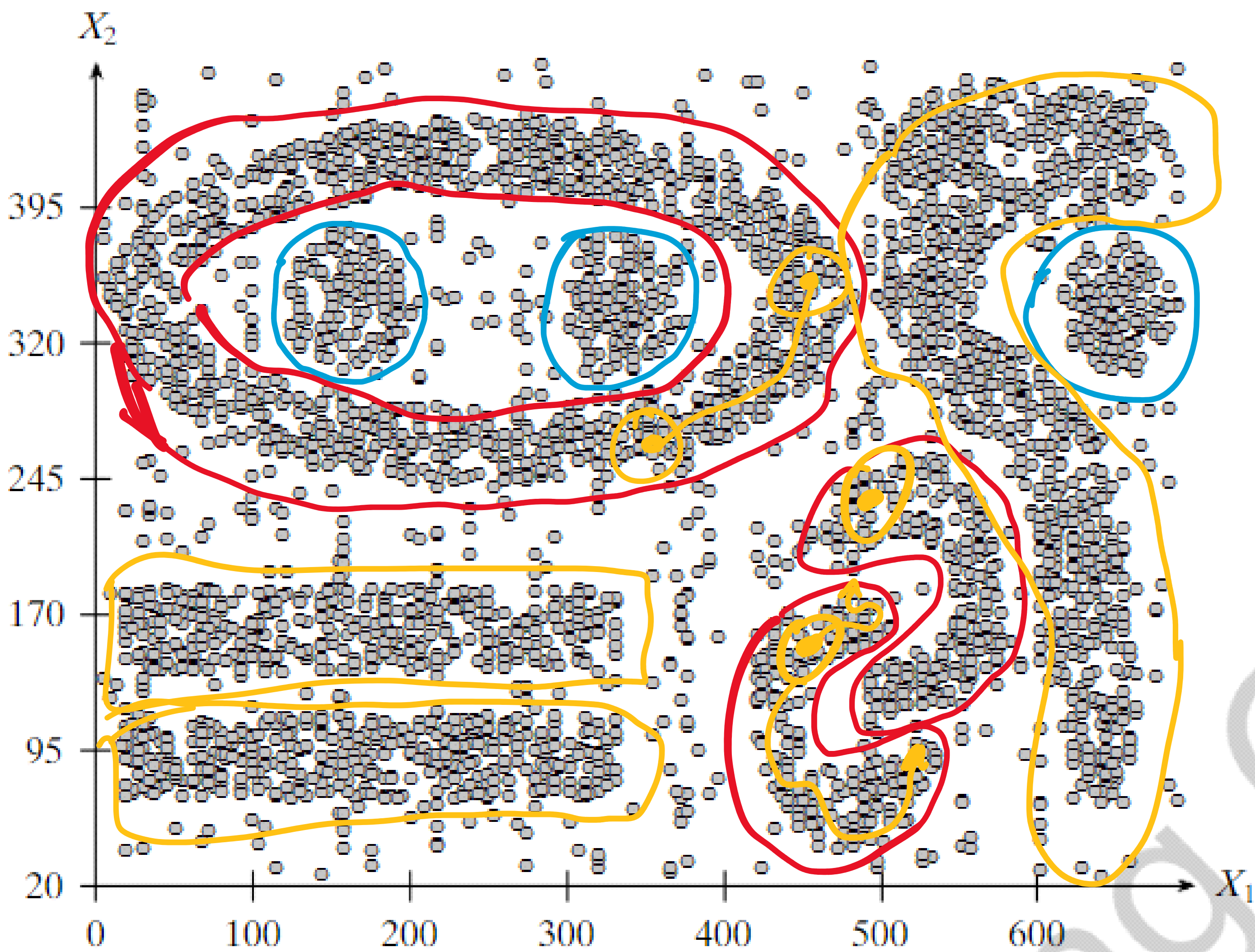


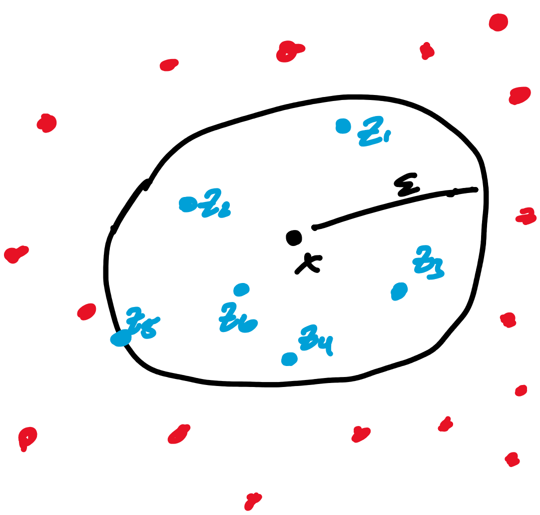
Figure 15.1. Density-based dataset.

  $X = \{x_1, \dots, x_n\}$ = data points.

Density-based clustering uses local density of points to determine clusters rather than only distance between points.

We define a ball of radius ε around a point $x \in \mathbb{R}^d$, called an ε-neighborhood of x as

$$N_\varepsilon(x) = B_d(x, \varepsilon) = \{ z \in X \mid \underbrace{\|x - z\|_2}_{\text{Euclidean distance, but others may be used}} \leq \varepsilon \}$$



$N_\varepsilon(x) = \{z_1, z_2, \dots, z_6\}$

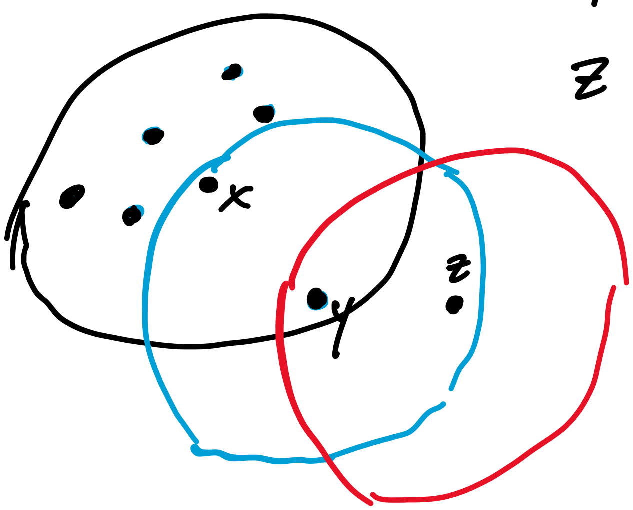
$x \in X$ is a core point if $|N_\varepsilon(x)| \geq \underbrace{m}_{\text{user-defined local density threshold}}$

$x \in X$ is a border point if $|N_\varepsilon(x)| < m$ and if it belongs to an ε-neighborhood of some core point [i.e. for some $z \in X$ with sufficiently dense neighborhood s.t. $x \in N_\varepsilon(z)$]

... border point. it

If $x \in X$ is not a core point nor a border point, it is a <u>noise point</u> (or outlier)

<u>Example</u>: Let $m=6 \Rightarrow$ x is a core point
y is a border point
z is a noise point

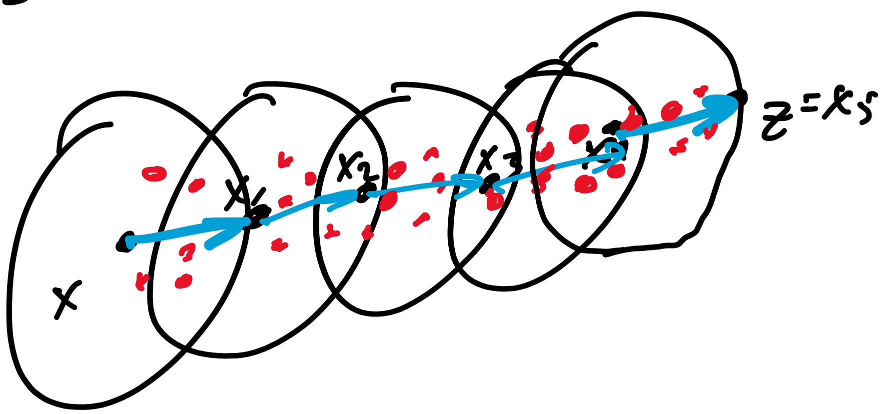


x is <u>directly density reachable</u> from y if $x \in N_\varepsilon(y)$ and y is a core point

x is <u>density reachable</u> from z if there exists a chain of points $x_0, x_1, \ldots, x_\ell$ with $x = x_0$ and $z = x_\ell$ such that x_i is directly density reachable from x_{i-1} for all i.

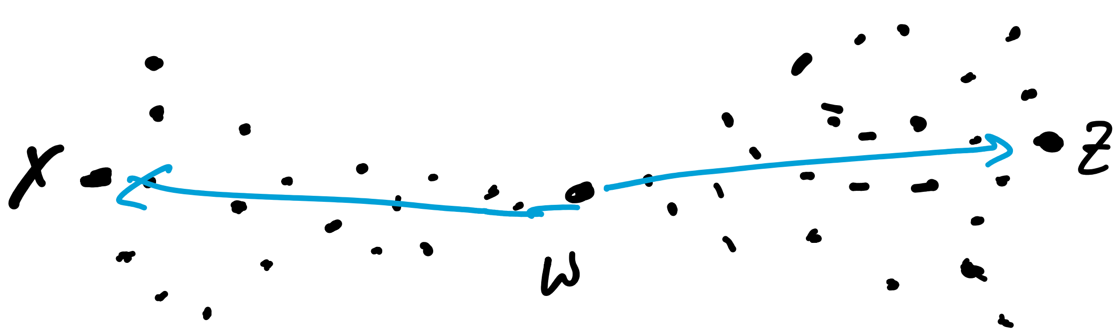
Such that x_i is directly density [illegible]

↳ i.e there is a set of core points leading from z to x.



Density reachability is asymmetric, i.e. it is possible for x to be D.R from z but z not be D.R. from x.

x and z are density connected if there exists a point w s.t. x and z are both D.R from w.



A density-based cluster is a maximal set of density connected points

↳ note it can be any shape, not just convex shapes

DBSCAN Algorithm

inputs: X, ε, m

radius of neighborhoods

local density threshold

$Core \leftarrow \emptyset$

for $x_i \in X$:

  compute $N_\varepsilon(x_i)$

  $id(x_i) \leftarrow \emptyset$ ← set cluster id to $\emptyset$

  if $|N_\varepsilon(x_i)| \geq m$: $Core \leftarrow Core \cup \{x_i\}$

Compute ε-neighborhood for each point and find core points

$k \leftarrow 0$

for $x_i \in Core$ with $id(x_i) = \emptyset$:

  $k \leftarrow k+1$

  $id(x_i) \leftarrow k$

  DensityConnected(x_i, k)

for each unassigned core point, recursively find all of its DR points + assign them all to cluster k

$\mathcal{E} \leftarrow \{E_1, \ldots, E_k\}$ where $E_i = \{x \in X \mid id(x) = i\}$

$Noise = \{x \in X \mid id(x) = \emptyset\}$ ← unassigned points are noise points

$Border = X \setminus (Core \cup Noise)$

Noise = ...

$$\text{Border} = X \setminus (\text{Core} \cup \text{Noise})$$

points that are not core points or noise points are border points

return ε, Core, Border, Noise

DensityConnected(x, k):

recursively follows neighborhoods of core points to assign to the same cluster

for $z \in N_\varepsilon(x)$:

$\quad id(z) \leftarrow k$

assign neighborhood points to cluster

$\quad$ if $z \in$ Core:

$\quad\quad$ DensityConnected(z, k)

explore the neighborhoods of DC core points to add to the cluster!

Limitation of DBSCAN: sensitive to ε

too small $\rightarrow$ sparse clusters seen as noise

too large $\rightarrow$ dense clusters merged

If there are clusters of different densities, a single ε may not suffice

OPTICS uses a <u>range of ε values</u>

<u>Complexity</u>: $O(n^2)$

Example

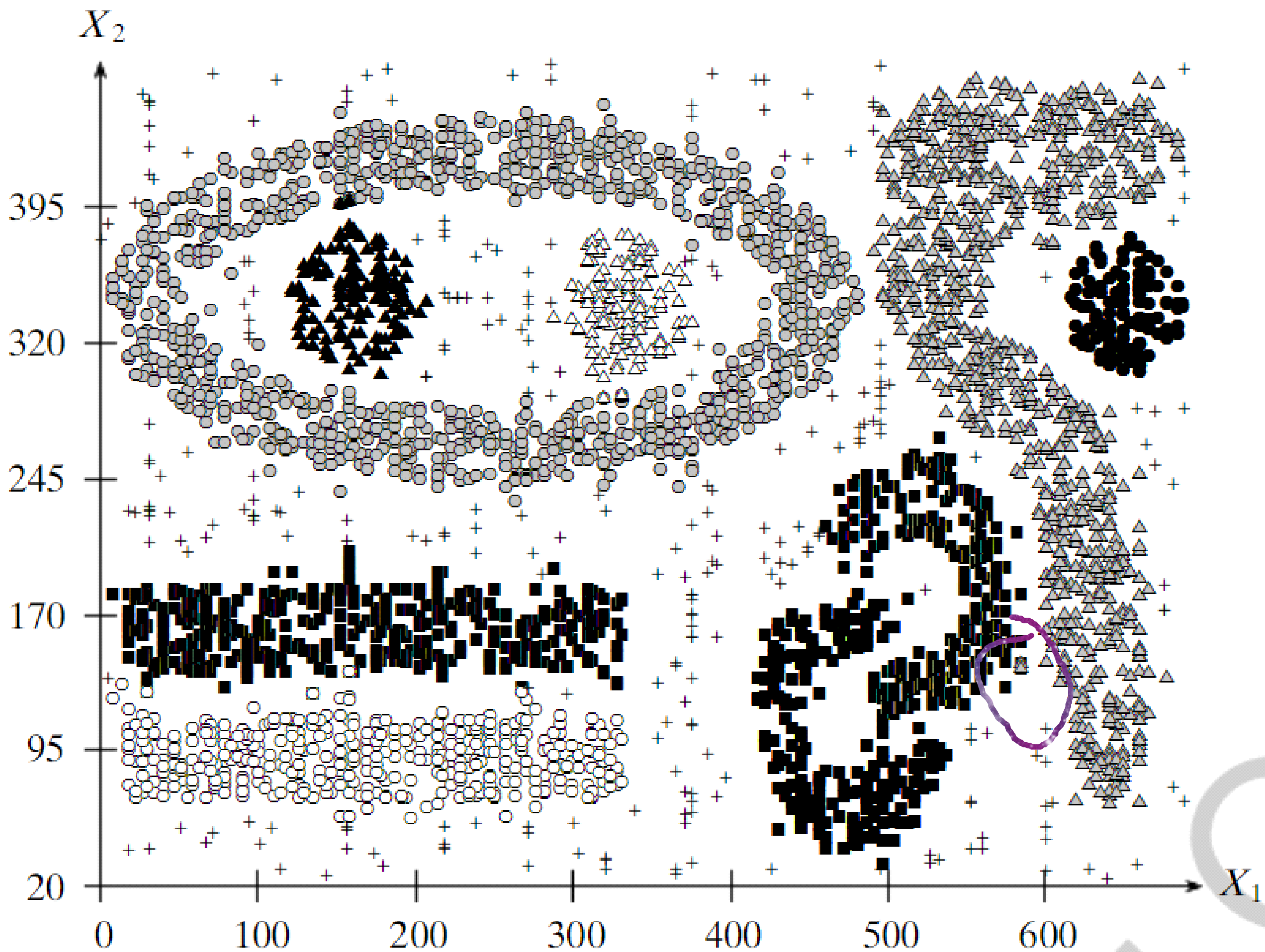


Figure 15.3. Density-based clusters.