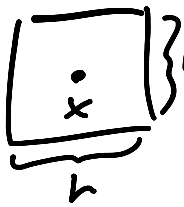
Lecture 23 - Apr 10

Density-based Clustering
DENCLUE

References

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



$\leftarrow x \in \mathbb{R}^2$

$\text{vol} = h^2$

If $x \in \mathbb{R}^d$, $H_d(h)$ = hypercube in dim d with side length h

$$\Rightarrow \text{vol}(H_d(h)) = h^d$$

We will estimate density at x in such a hypercube

$$\hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where the multivariate kernel function K satisfies $\int K(z)\,dz = 1$.

<u>Discrete kernel</u> $K(z) = \begin{cases} 1, & \text{if } |z_j| \le \frac{1}{2} \text{ for all } j = 1, \dots, d \\ 0, & \text{else} \end{cases}$

if $z = \frac{x - x_i}{h}$, this counts points in $H_d(h)$ since $K(z)$

is 1 iff $\left|\frac{x_j - x_{ij}}{h}\right| \le \frac{1}{2}$

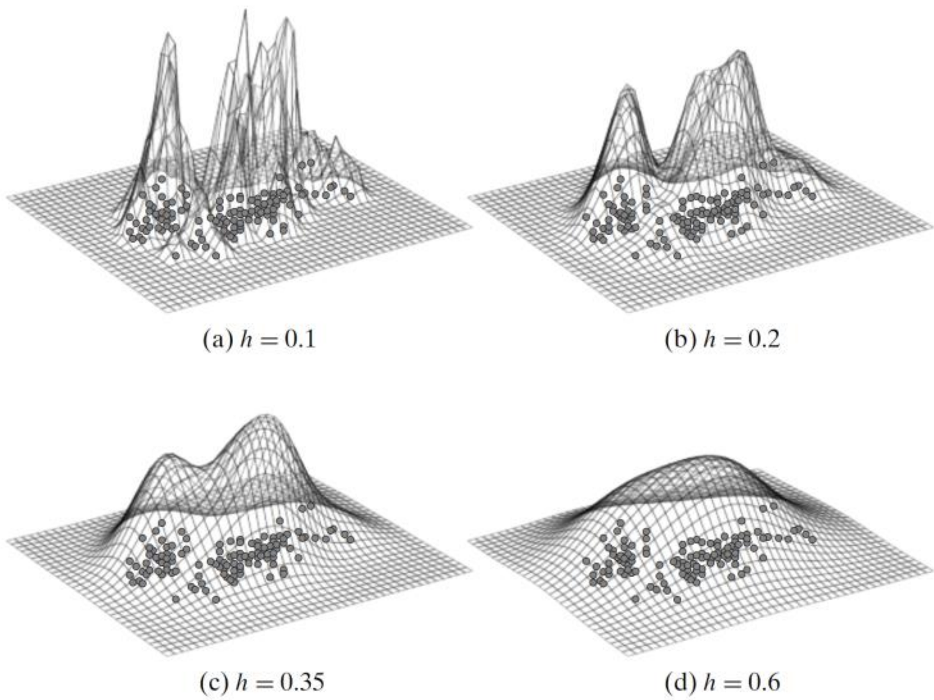
$$|x_j - x_{ij}| \le \frac{h}{2}$$

<u>Gaussian kernel</u>: $K(z) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{z^T z}{2}\right)$

(assuming covariance matrix $\Sigma = I_d$)

$$\Rightarrow K\left(\frac{x - x_i}{h}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{(x-x_i)^T(x-x_i)}{2h^2}\right)$$

Examples



(a) $h = 0.1$ (b) $h = 0.2$

(c) $h = 0.35$ (d) $h = 0.6$

Figure 15.7. Density estimation: 2D Iris dataset (varying h).

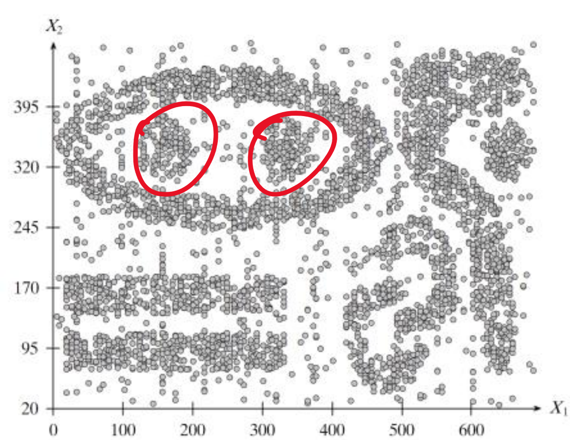


Figure 15.1. Density-based dataset.

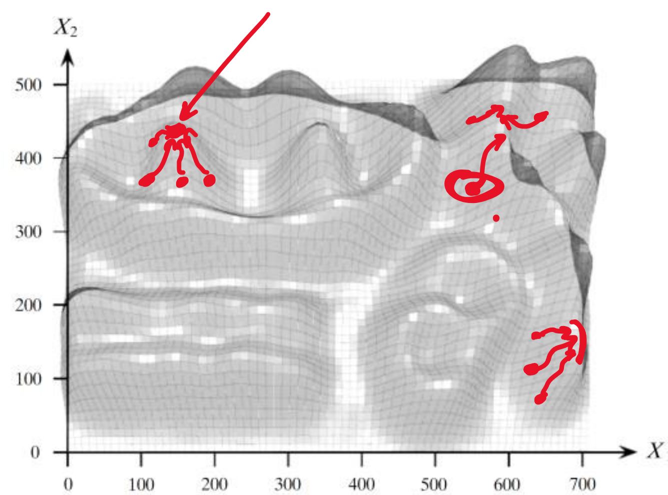


Figure 15.8. Density estimation: density-based dataset.

<u>Idea</u>: Use gradient-based optim. to find peaks in the density landscape to find regions with sufficiently high density

x^* is a <u>density attractor</u> f if is a local max of the estimated pdf $\hat{f}$

x is a ~~[illegible]~~
max of the estimated pdf $\hat{f}$

$$\nabla \hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \nabla K\left(\frac{x - x_i}{h}\right)$$

If K is Gaussian

$$K(z) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{z^T z}{2}\right)$$

$$\nabla K(z) = \underbrace{\frac{1}{(2\pi)^{d/2}} \cdot \exp\left(-\frac{z^T z}{2}\right)}_{K(z)} \cdot -z = -K(z) z \nabla z$$

$$\nabla \hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \cdot \left(\frac{x - x_i}{h}\right) \cdot \frac{-1}{h}$$

$$= \frac{1}{nh^{d+1}} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \underline{\underline{(x_i - x)}}$$

x^* is a <u>density attractor</u> for x if a gradient-ascent path converges to x^*

$$x_{t+1} = x_t + \alpha \nabla \hat{f}(x_t)$$

trick: try to solve $\nabla \hat{f}(x) = 0$

$$\cancel{\frac{1}{nh^{d+1}}} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \underline{\underline{(x_i - x)}} = 0$$

$$\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i = \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x$$

$$\sum_{i=1} K\left(\frac{\quad}{h}\right)x_i \qquad \sum_{i=1} K\left(\frac{\quad}{h}\right)x_i$$

$$\frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)x_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)} = x$$

$$x_{t+1} = \frac{\sum_{i=1}^{n} K\left(\frac{x_t-x_i}{h}\right)x_t}{\sum_{i=1}^{n} K\left(\frac{x_t-x_i}{h}\right)}$$

A cluster $C \subseteq X$ is a <u>center-defined cluster</u> if all $x \in C$ are attracted to a unique density attractor x^* such that $\hat{f}(x^*) > \xi$

user-defined min density threshold

A cluster $C \subseteq X$ is a <u>density-based cluster</u> if there a set of density attractors $x_1^*, \dots, x_k^*$ such that

1. Each $x \in C$ is density attracted to some x_j^*

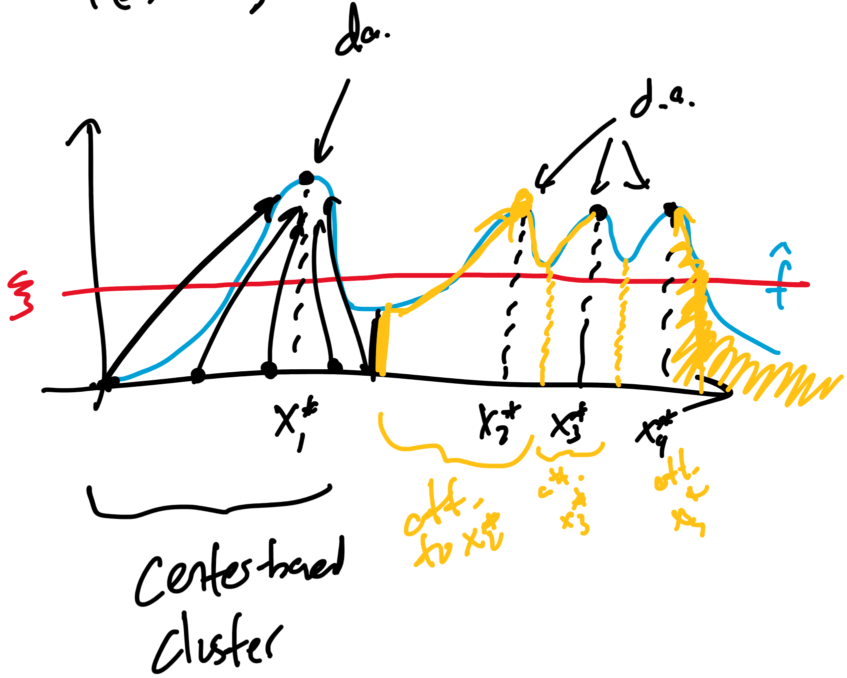
① Each $x \in C$ is density attracted ...

② Each x_j^* has $\hat{f}(x_j^*) \geq \xi$

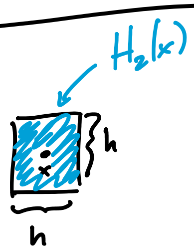
③ Any two attractors x_i^* & x_j^* are density reachable

i.e. $\exists$ path x_i^* to x_j^* s.t. all points on path z

has $\hat{f}(z) \geq \xi$



Multivariate Density Estimation



$x \in \mathbb{R}^2$

$Vol(H_2(x)) = h^2$

If $x \in \mathbb{R}^d$ ... $H_d(x)$ = hypercube of dim d centered at x w/ side lengths $= h$

$\Rightarrow Vol(H_d(x)) = h^d$

We estimate density as $\frac{1}{h^d} \cdot \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$

center of hypercube

pts in dataset

kernel function

Ex: Discrete kernel

$$K(z) = \begin{cases} 1, & \text{if } |z_j|^2 \le \frac{1}{2} \text{ for } j = 1, \dots, d \\ 0, & \text{else} \end{cases}$$

If $z = \frac{x - x_i}{h} \Rightarrow K(z) = 1$ if $\left| \frac{x_j - x_{ij}}{h} \right| \le \frac{1}{2}$

$|x_j - x_{ij}| \le \frac{h}{2}$

$\Rightarrow x_i \in H_d$

Gaussian kernel $K(z) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|z\|_2^2}{2}\right)$

(Assuming $\Sigma = I_d$)

$$K(x, x_i) = \frac{1}{\ldots} \exp\left(-\frac{\left\|\frac{x - x_i}{h}\right\|_2^2}{\ldots}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{(x - x_i)^T \Sigma^{-1} (x - x_i)}{2h^2}\right)$$

$$K\left(\frac{x-x_i}{h}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\left\|\frac{x-x_i}{h}\right\|_2^2}{2}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{(x-x_i)^T(x-x_i)}{2h^2}\right)$$

$$|\Sigma|$$

Use gradient-based optimization to find peaks in the density landscape to find regions with sufficently high density

Density Attractors + Gradient

x^* is a density attractor if it is a local max of pdf $\hat{f}$, which can be found by gradient ascent

$$\nabla \hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} \nabla K\left(\frac{x - x_i}{h}\right)$$

For the Gaussian kernel, $\nabla K(z) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{z^T z}{2}\right) \cdot -z \nabla z$

$$= K(z) \cdot -z \nabla z$$

$$\Rightarrow \nabla \hat{f}(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \cdot \left(\frac{x - x_i}{h}\right) \cdot \frac{-1}{h}$$

$$= \frac{1}{nh^{d+2}} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(x_i - x)$$

x^* is a <u>density attractor for x</u> if a gradient-ascent path converges to x^*

↳ Typically, $x_{t+1} = x_t + \alpha \nabla \hat{f}(x_t)$

This can be slow to converge, but we can optimize by solving

$$\nabla \hat{f}(x) = 0$$

$$\cancel{\frac{1}{nh^{d+2}}} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(x_i - x) = 0$$

$$\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \cdot x_i = x \cdot \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

$$\frac{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) x_i}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)} = x$$

$$\Rightarrow \quad x_{t+1} = \frac{\sum_{i=1}^{n} K\left(\frac{x_t - x_i}{h}\right) x_i}{\sum_{i=1}^{n} K\left(\frac{x_t - x_i}{h}\right)}$$

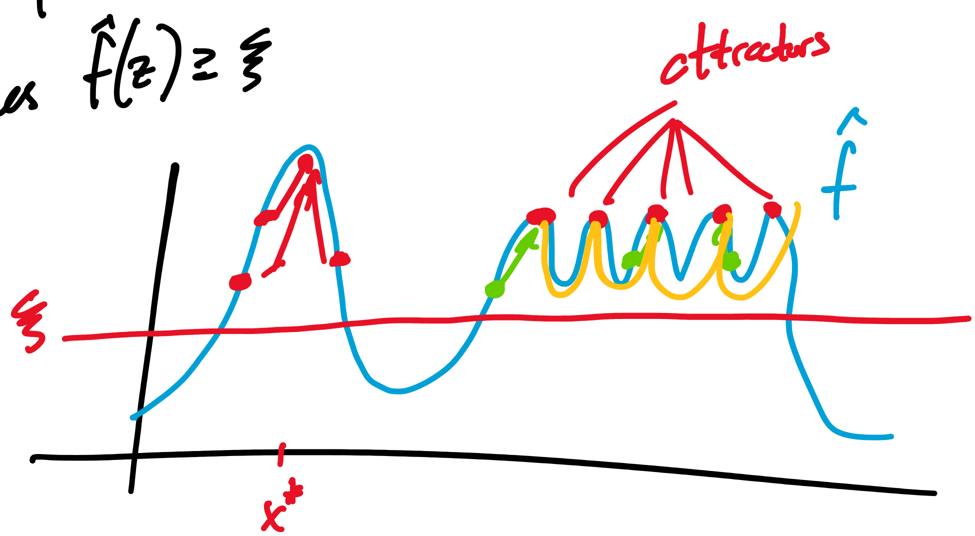
This results in much faster convergence.

This results in ...

A cluster $E \subseteq X$ is a <u>center-defined cluster</u> if all $x \in E$ are density attracted to a unique density attractor x^* such that $\hat{f}(x^*) > \xi$ user-defined min. density threshold

A cluster $E \subseteq X$ is a <u>density-based cluster</u> if there exists a set of density attractors $x_1^*, \ldots, x_m^*$ such that

1. Each $x \in E$ is attracted to some x_i^*
2. Each x_i^* has density above ξ, i.e. $\hat{f}(x_i^*) \geq \xi$
3. Any two $x_i^* + x_j^*$ are <u>density reachable</u>, i.e. there exists a path from x_i^* to x_j^* s.t. all points on the path z has $\hat{f}(z) \geq \xi$



Center-based

Center-based
clusters

inputs: X, h, ξ, ε

1. $\mathcal{A} \leftarrow \emptyset$

2. for $x \in X$:

   $x^* \leftarrow \text{FindAttractor}(x, \xi, h, \varepsilon)$ — X

   if $\hat{f}(x^*) \geq \xi$:

   $\mathcal{A} \leftarrow \mathcal{A} \cup \{x^*\}$

   $R(x^*) \leftarrow R(x^*) \cup \{x\}$

compute attractor for each x, x^*. If density is sufficient, add x^* to set of attractors $\mathcal{A}$ + add x to set of points attracted to x^*, $R(x^*)$

3. $\mathcal{E} \leftarrow \{\text{maximal } E \subseteq \mathcal{A} \mid \text{all } x_i^*, x_j^* \in E \text{ are density reachable}\}$?

$V = h^d$

$\xi = \frac{?}{h^d}$

find maximal subsets of $\mathcal{A}$ s.t. any pair are density-reachable

4. for $E \in \mathcal{E}$:

   for $x^* \in E$: $E \leftarrow E \cup R(x^*)$

add all points attracted to attractors in each cluster to the cluster

return $\mathcal{E}$

FindAttractor(x, X, h, ε):

$t \leftarrow 0$

$x_t \leftarrow x$

while $\|x_t - x_{t-1}\| \leq \varepsilon$:

hyperparam for stopping

$\sum K\left(\frac{x_t - x_i}{h}\right) x_i$

hill-climbing procedure using direct update rule while [illegible]

$$x_{t+1} \leftarrow \frac{\sum_{i=1}^{n} K\left(\frac{x_t - x_i}{h}\right) x_i}{\sum_{i=1}^{n} K\left(\frac{x_t - x_i}{h}\right)}$$

$$t \leftarrow t+1$$

direct update rule
(faster than gradient descent)

return x_t

Note: DENCLUE with $h = \varepsilon$, $\xi = m$ with discrete kernel
$= $ DBSCAN

Density-based approach can generate hierarchical clusterings using $\xi = 0$, $\xi = 0.1$, ..., $\xi = 1$, for example.

Examples

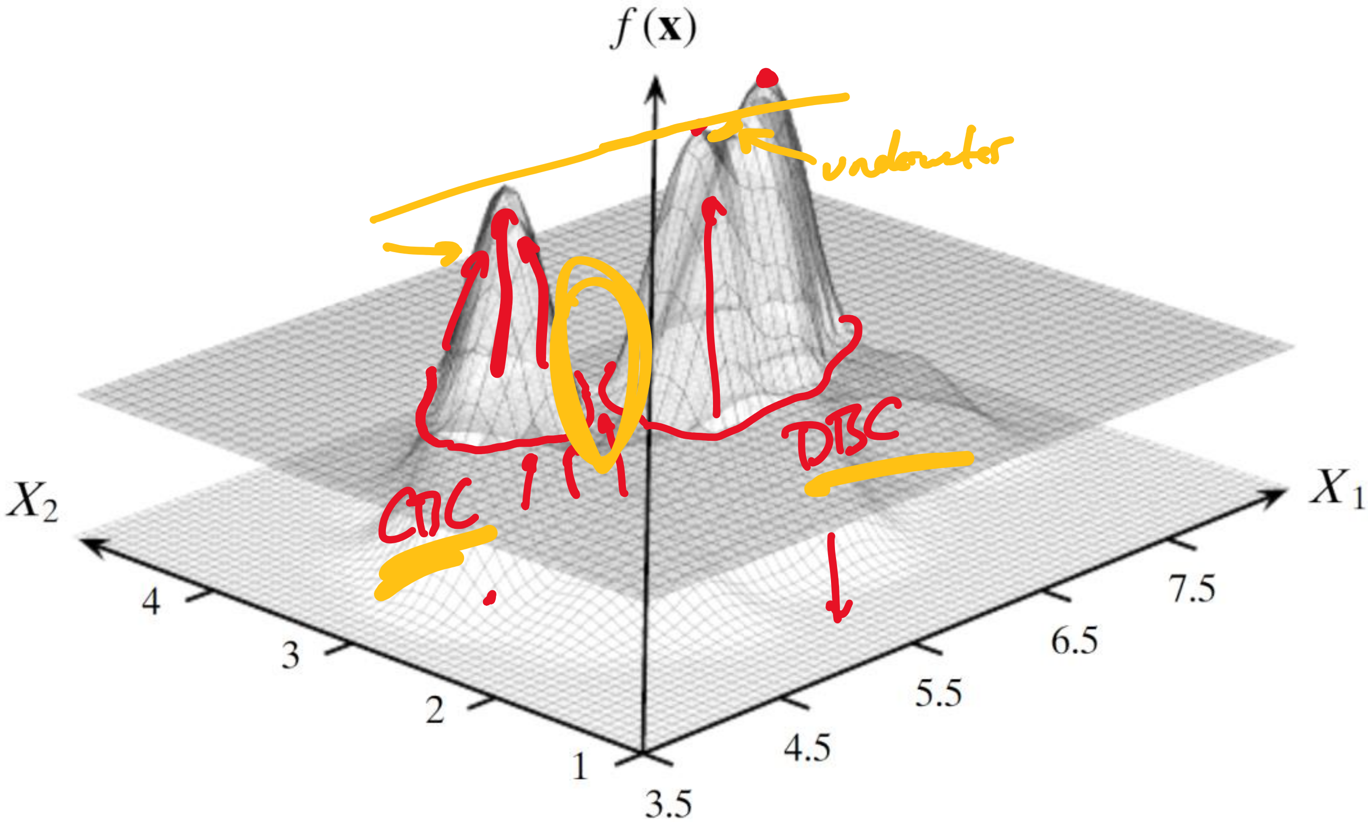


Figure 15.10. DENCLUE: Iris 2D dataset.

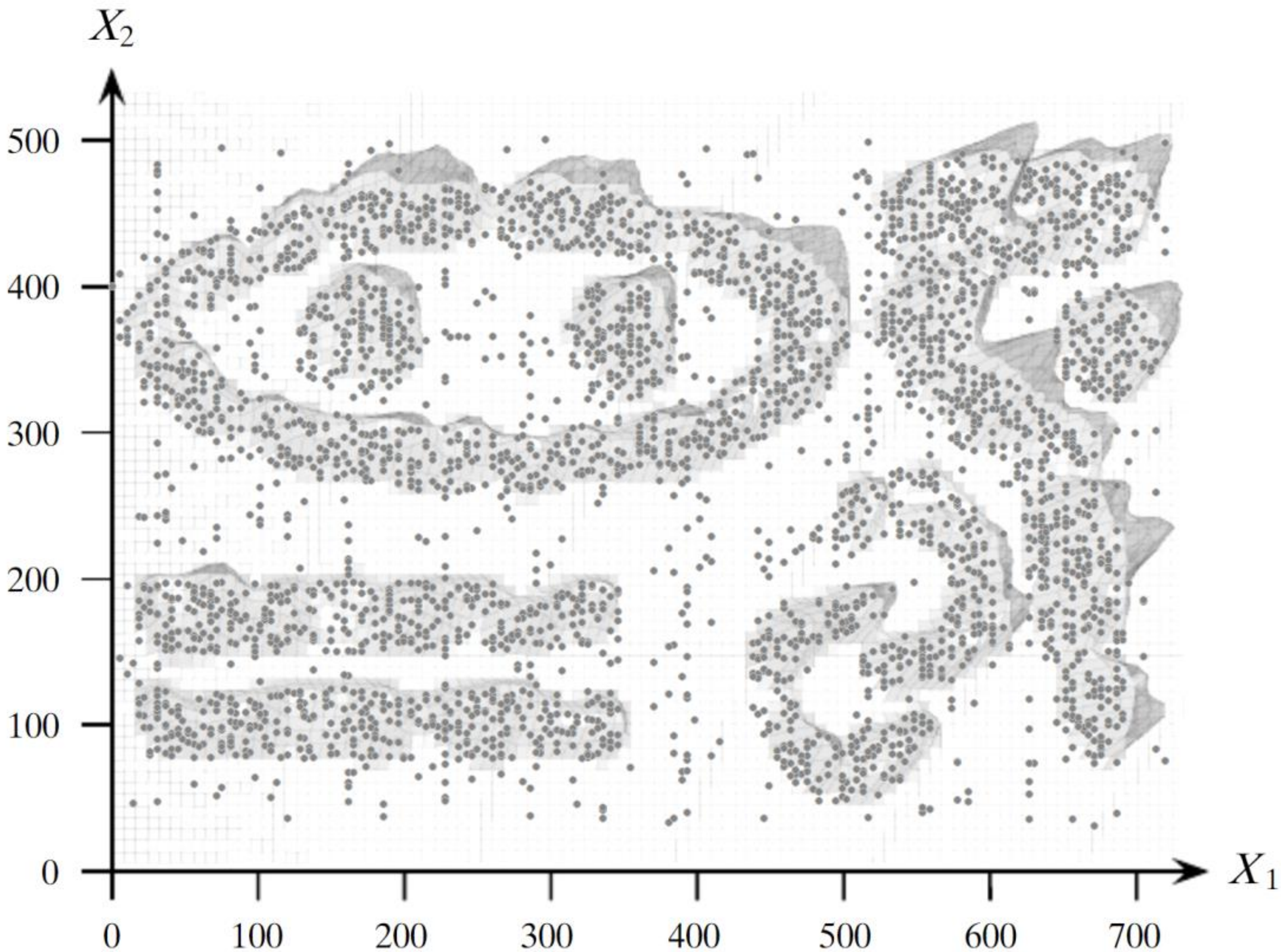


Figure 15.11. DENCLUE: density-based dataset.