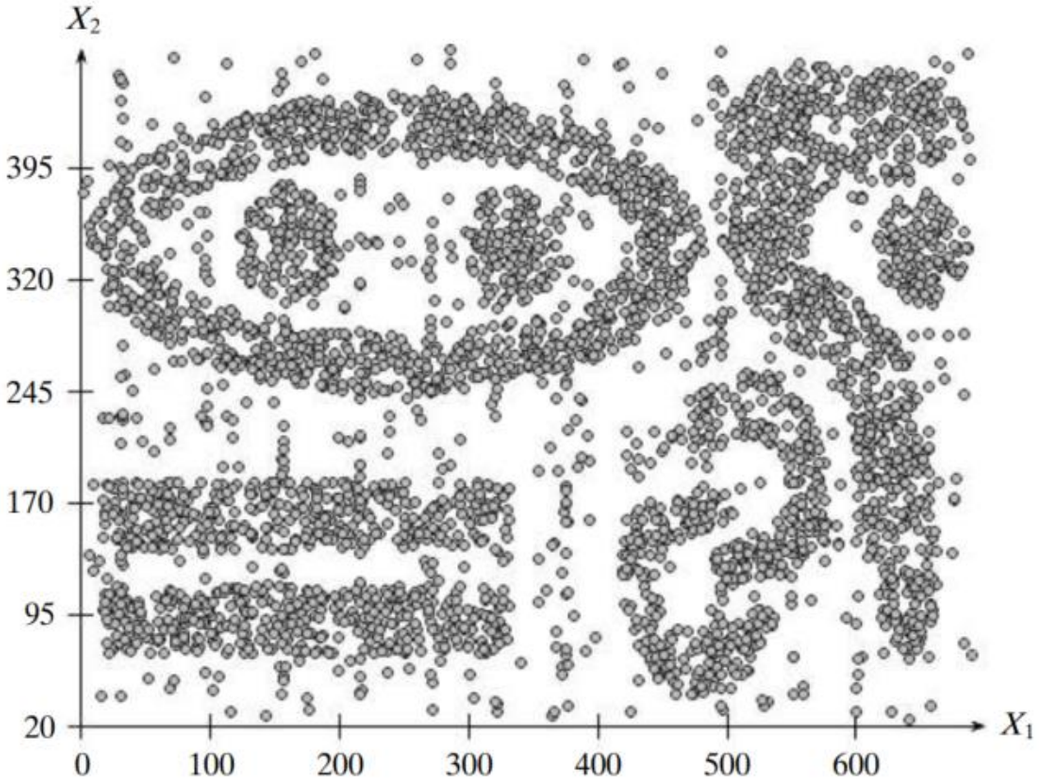
خوشه‌بندی مبتنی بر چگالی

Density-based Clustering

a.fahim@ut.ac.ir

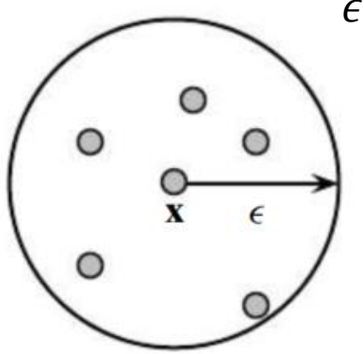
خوشه‌بندی‌های مبتنی بر نماینده برای خوشه‌های غیر محدب (Nonconvex Clusters) مناسب نیستند.
زیرا فاصله‌ی دو نقطه از دو خوشه‌ی متفاوت می‌تواند از فاصله‌ی آن‌ها از نماینده‌اشان کوتاه‌تر باشد.



Density-based dataset.

الگوریتم DBSCAN

ϵ–همسایگی (ϵ −neighborhood) $\boldsymbol{x}$ توپی به شعاع ϵ حول نقطه‌ی $\boldsymbol{x} \in \mathbb{R}^d$ است.



(a)

(a) Neighborhood of a point.

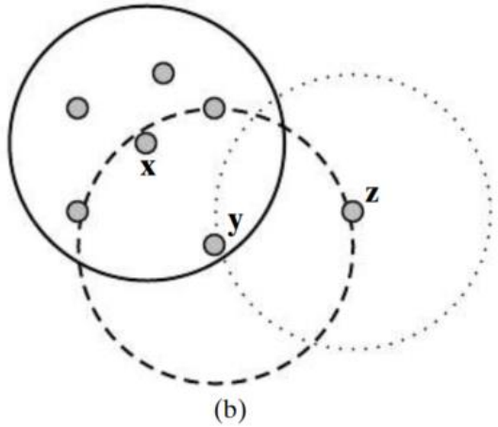
$$\boldsymbol{N}\epsilon(\boldsymbol{x}) = \boldsymbol{B}_d(\boldsymbol{x}, \epsilon) = \{\boldsymbol{y} | \delta(\boldsymbol{x}, \boldsymbol{y}) \leq \epsilon\}$$

که معمولا از فاصله‌ی اقلیدسی استفاده می‌شود.

$$\delta(x,y) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

نقطه‌ی اصلی (Core Point): $|N_\epsilon(\boldsymbol{x})| \geq$ minpts

نقطه‌ی مرزی (Border Point):
$|N_\epsilon(\boldsymbol{x})| <$ minpts
اما در همسایگی یک نقطه‌ی اصلی قرار دارد.



(b)

(b) Core, border, and noise points.

نقطه‌ی پرت یا نویز (Noise Point):
هر نقطه‌ای که نه اصلی و نه مرزی باشد.

اگر $\boldsymbol{x} \in N_\epsilon(\boldsymbol{y})$ باشد و $\boldsymbol{y}$ یک نقطه‌ی اصلی باشد،
آنگاه $\boldsymbol{x}$ مستقیما دسترسی چگالی به $\boldsymbol{y}$ دارد (Directly Density Reachable).

اگر یک زنجیره‌ای از نقاط اصلی $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_l$ وجود داشته باشد که
هر دو نقطه‌ی دنبال هم، مستقیما دسترسی چگالی به هم داشته باشند،
آنگاه دو نقطه‌ی ابتدایی و انتهایی x_0 و x_l نیز دسترسی چگالی به هم دارند.

اگر نقطه‌ی اصلی $\boldsymbol{z}$ وجود داشته باشد که دو نقطه‌ی $\boldsymbol{x}$ و $\boldsymbol{y}$ به نقطه‌ی $\boldsymbol{z}$
دسترسی چگالی داشته بشاند،
متصل چگالی (Density Directed) می‌باشند.

یک خوشه‌ی مبتنی بر چگالی، مجموعه بیشینه‌ی نقاط متصل چگالی است.

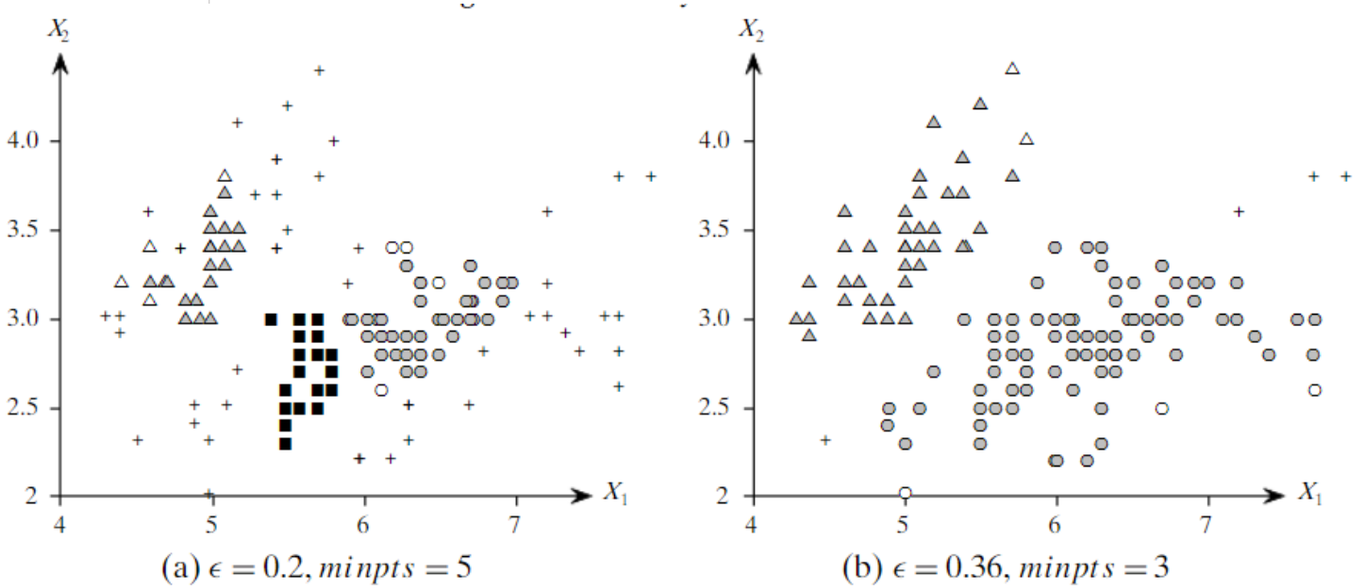
Algorithm 15.1: Density-based Clustering Algorithm

DBSCAN (D, ϵ, $minpts$):

```
Core ← ∅
foreach x_i ∈ D do // Find the core points
    Compute N_ε(x_i)
    id(x_i) ← ∅ // cluster id for x_i
    if N_ε(x_i) ≥ minpts then Core ← Core ∪ {x_i}
k ← 0 // cluster id
foreach x_i ∈ Core, such that id(x_i) = ∅ do
    k ← k + 1
    id(x_i) ← k // assign x_i to cluster id k
    DENSITYCONNECTED (x_i, k)
C ← {C_i}_{i=1}^k, where C_i ← {x ∈ D | id(x) = i}
Noise ← {x ∈ D | id(x) = ∅}
Border ← D \ {Core ∪ Noise}
return C, Core, Border, Noise
```

DENSITYCONNECTED (x, k):

```
foreach y ∈ N_ε(x) do
    id(y) ← k // assign y to cluster id k
    if y ∈ Core then DENSITYCONNECTED (y, k)
```



(a) $\epsilon = 0.2, minpts = 5$ (b) $\epsilon = 0.36, minpts = 3$

Figure 15.4. DBSCAN clustering: Iris dataset.

برآورد چگالی هسته (Kernel Density Estimation)

برآورد تابع توزیع تجمعی (Cumulative Distribution Function) بطور مستقیم از داده

فرض کنید X متغیر تصادفی پیوسته با تابع چگالی (نامشخص برای ما) $f(x)$ باشد.

هدف برآورد $f(x)$ به کمک n نمونه‌ی $x_1, \dots, x_n$ می‌باشد.

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \le x)$$

$$\hat{f}(x) = \frac{\hat{F}\left(x+\frac{h}{2}\right) - \hat{F}(x-\frac{h}{2})}{h} = \frac{k/n}{h} = \frac{k}{nh}$$

k تعداد نقاط درون پنجره‌ای به پهنای h و مرکز x است و تابع چگالی برآورد شده می‌شود، نسبت تعداد نقاط درون پنجره حول x، به تعداد کل نقاط به حجم پنجره.

دقت الگوریتم به پارامتر h بستگی دارد.

برآوردگر هسته‌ای (Kernel Estimator)

خواص تابع هسته‌ی چگالی (Density Kernel Function)

$$K(x) \ge 0, \qquad \int K(x)dx = 1, \qquad K(x) = K(-x)$$

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

هسته‌ی گسسته (Discrete Kernel)

$$K(x) = \begin{cases} 1 & |z| \le 1/2 \\ 0 & otherwise \end{cases}$$

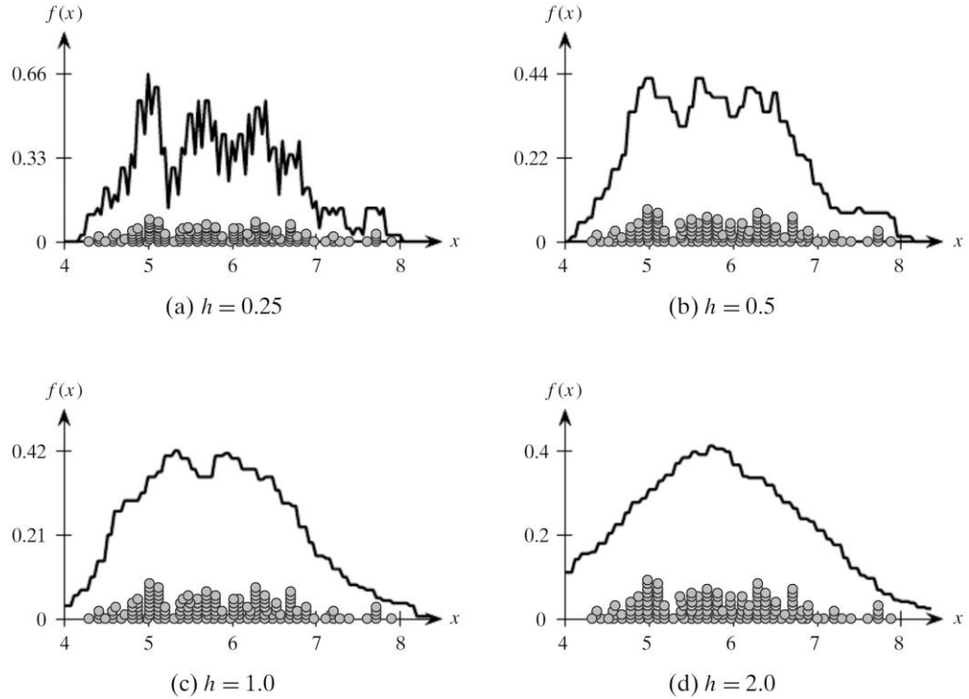


Figure 15.5. Kernel density estimation: discrete kernel (varying h).

هسته‌ی گاوسی (Gaussian Kernel)

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\}$$

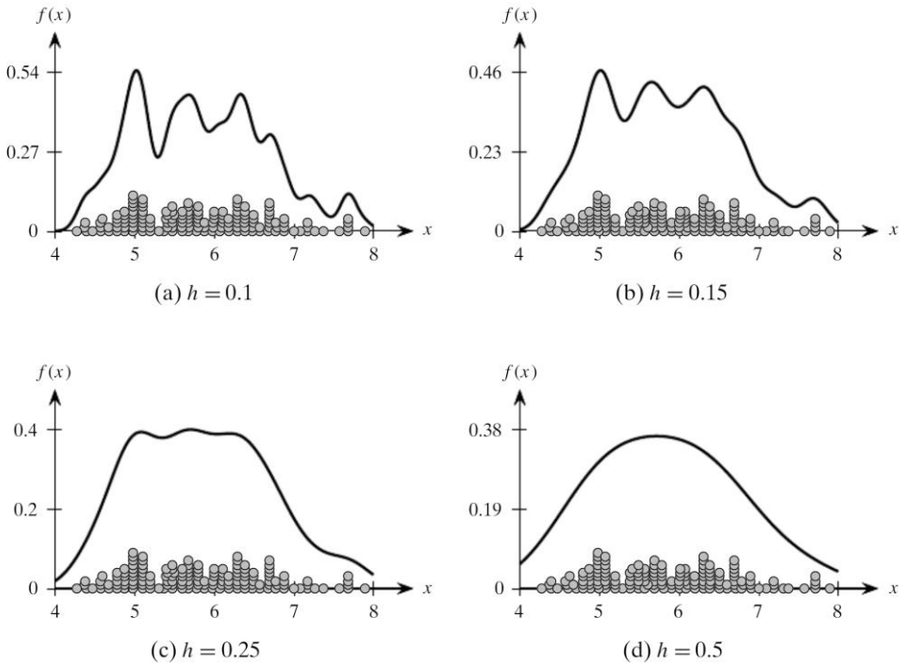


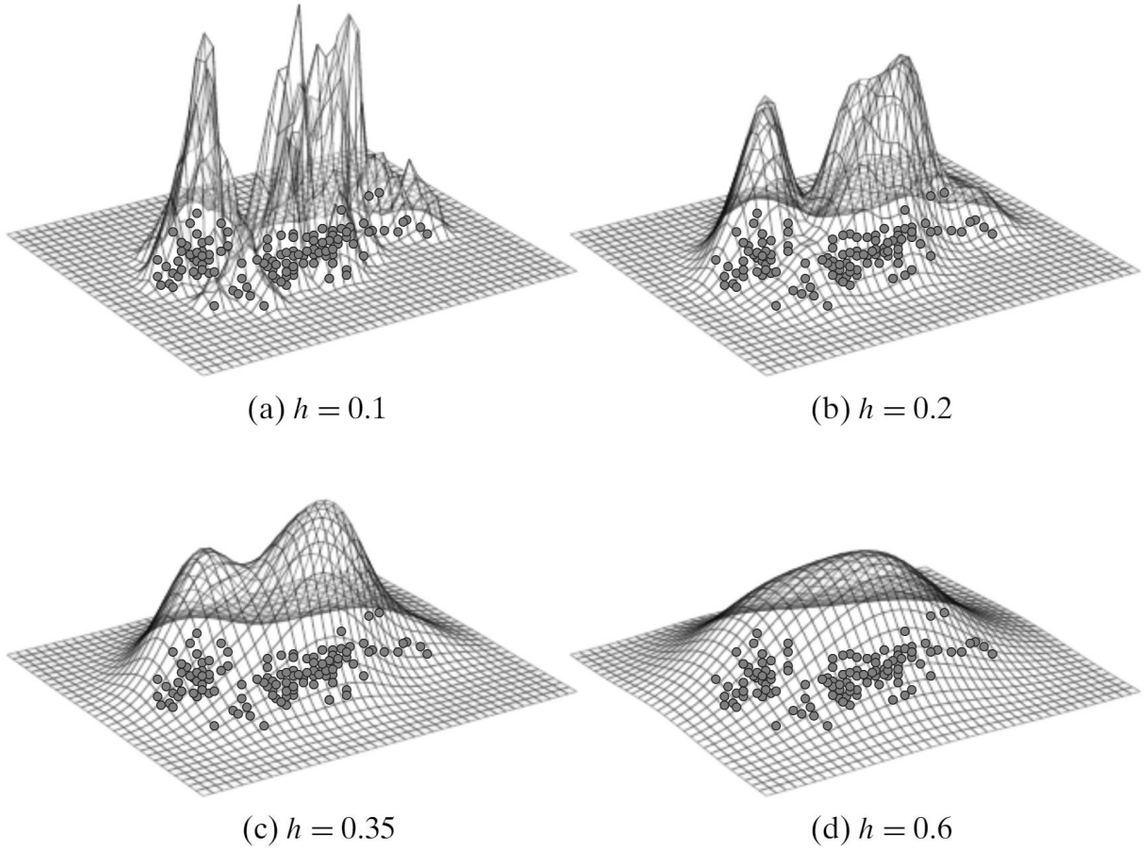
Figure 15.6. Kernel density estimation: Gaussian kernel (varying h).

برآورد چگالی چندمتغیره (Multivariate Density Estimation)

$$\boldsymbol{x} = (x_1, \dots, x_d)^T$$

$$K(\boldsymbol{x}) \geq 0, \qquad \int K(\boldsymbol{x}) d\boldsymbol{x} = 1, \qquad K(\boldsymbol{x}) = K(-\boldsymbol{x})$$

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



(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).

هسته‌ی گسسته (Discrete Kernel)

$$K(\mathbf{z}) = \begin{cases} 1 & \text{If } |z_j| \leq \frac{1}{2}, \text{ for all dimensions } j = 1, \ldots, d \\ 0 & \text{Otherwise} \end{cases}$$

هسته‌ی گاوسی (Gaussian Kernel)

$$K(\mathbf{z}) = \frac{1}{(2\pi)^{d/2}} \exp\left\{-\frac{\mathbf{z}^T\mathbf{z}}{2}\right\} \qquad \boldsymbol{\Sigma} = \mathbf{I}_d.$$

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

$$\hat{f}(\mathbf{x}) = \frac{k}{n\,\mathrm{vol}(S_d(h_{\mathbf{x}}))}$$

$$\mathrm{vol}(S_d(r)) = K_d \cdot r^d = \left(\frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}\right) r^d$$

$$\Gamma\left(\frac{d}{2}+1\right) = \begin{cases} \left(\frac{d}{2}\right)! & \text{if } d \text{ is even} \\ \sqrt{\pi}\left(\frac{d!!}{2^{(d+1)/2}}\right) & \text{if } d \text{ is odd} \end{cases}$$

در این رابطه h_x فاصله‌ی kامین نزدیکترین همسایه‌ی $\boldsymbol{x}$ و $\mathrm{vol}(S_d(h_x))$ حجم کره‌ی d بعدی به مرکز $\boldsymbol{x}$ و شعاع h_x است.
مقدار k از ابتدا مشخص می‌شود.

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx \qquad \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1) \qquad \Gamma(1) = 1 \quad \text{and} \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

هر نقطه‌ی x جذب یک x^* (Density Attractor) که چگالی بیشتر از آستانه دارد، می‌شود.

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta \cdot \nabla \hat{f}(\mathbf{x}_t)$$

$$\nabla \hat{f}(\mathbf{x}) = \frac{1}{nh^{d+2}} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h}\right) \cdot (\mathbf{x}_i - \mathbf{x})$$

$$\mathbf{x}_{t+1} = \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_t-\mathbf{x}_i}{h}\right) \mathbf{x}_i}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_t-\mathbf{x}_i}{h}\right)}$$

برای افزایش سرعت می‌توان جمع را روی همسایه‌ها با فاصله‌ی $r = zh$ انجام داد.

$$\mathbf{x}_{t+1} = \frac{\sum_{\mathbf{x}_i \in B_d(\mathbf{x}_t, r)} K\left(\frac{\mathbf{x}_t-\mathbf{x}_i}{h}\right) \mathbf{x}_i}{\sum_{\mathbf{x}_i \in B_d(\mathbf{x}_t, r)} K\left(\frac{\mathbf{x}_t-\mathbf{x}_i}{h}\right)}$$

ξ یک مقدار از پیش مشخص شده برای آستانه چگال است.

$$\hat{f}(\mathbf{x}^*) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x}^*-\mathbf{x}_i}{h}\right) \geq \xi$$

Algorithm 15.2: DENCLUE Algorithm

DENCLUE $(\mathbf{D}, h, \xi, \epsilon)$**:**

$\mathcal{A} \leftarrow \emptyset$
foreach $\mathbf{x} \in \mathbf{D}$ **do** // find density attractors
  $\mathbf{x}^* \leftarrow \text{FINDATTRACTOR}(\mathbf{x}, \mathbf{D}, h, \epsilon)$
  if $\hat{f}(\mathbf{x}^*) \geq \xi$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{\mathbf{x}^*\}$
    $R(\mathbf{x}^*) \leftarrow R(\mathbf{x}^*) \cup \{\mathbf{x}\}$
$\mathcal{C} \leftarrow \{\text{maximal } C \subseteq \mathcal{A} \mid \forall \mathbf{x}_i^*, \mathbf{x}_j^* \in C, \mathbf{x}_i^* \text{ and } \mathbf{x}_j^* \text{ are density reachable}\}$
foreach $C \in \mathcal{C}$ **do** // density-based clusters
  foreach $\mathbf{x}^* \in C$ **do** $C \leftarrow C \cup R(\mathbf{x}^*)$
return $\mathcal{C}$

FINDATTRACTOR $(\mathbf{x}, \mathbf{D}, h, \epsilon)$**:**

$t \leftarrow 0$
$\mathbf{x}_t \leftarrow \mathbf{x}$
repeat
  $\mathbf{x}_{t+1} \leftarrow \frac{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_t - \mathbf{x}_i}{h}\right) \cdot \mathbf{x}_i}{\sum_{i=1}^{n} K\left(\frac{\mathbf{x}_t - \mathbf{x}_i}{h}\right)}$
  $t \leftarrow t + 1$
until $\|\mathbf{x}_t - \mathbf{x}_{t-1}\| \leq \epsilon$
return $\mathbf{x}_t$

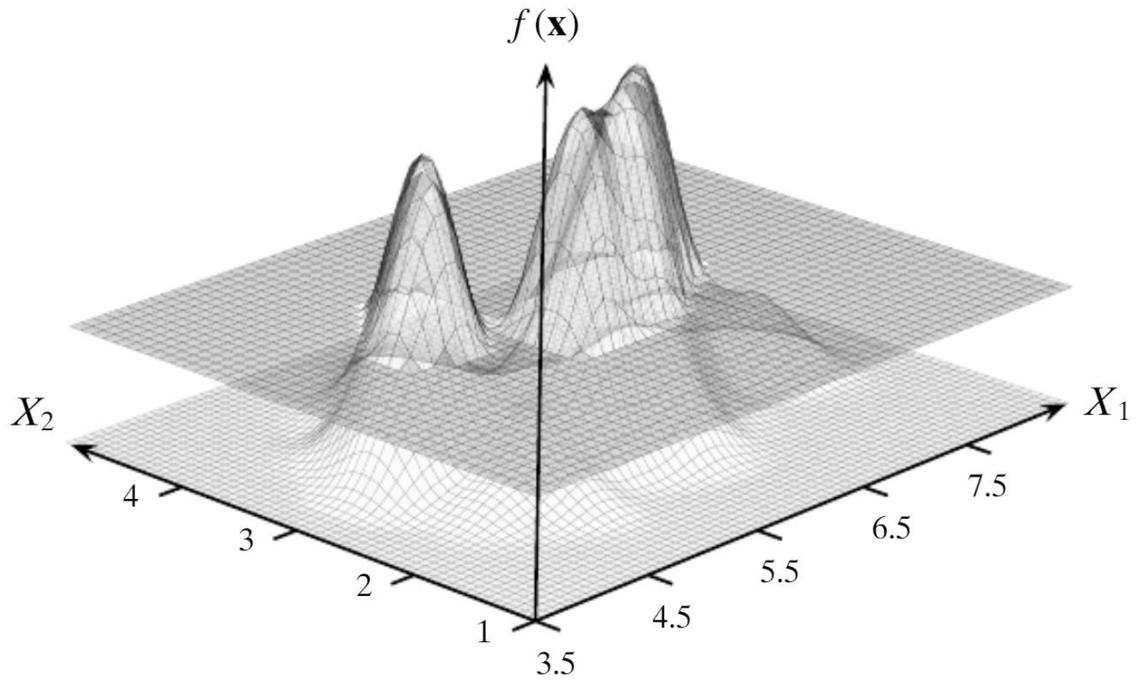


Figure 15.10. DENCLUE: Iris 2D dataset.

می‌توان نشان داد که DBSCAN حالت خاصی از DENCLUE (برآورد چگالی هسته) با انتخاب $h = \epsilon$ و minpts$= \xi$ است.

دو جاذب x_i^* و x_j^* از طریق چگالی به هم در دسترس می‌باشند اگر بتوان مسیری بین آن دو پیدا کرد که برای همه نقاط $\boldsymbol{y}$ روی مسیر $\hat{f}(\boldsymbol{y}) \geq \xi$