# CS145: INTRODUCTION TO DATA MINING

## 10: Vector Data: Density Estimation

**Instructor: Yizhou Sun**

yzsun@cs.ucla.edu

November 1, 2017

# Methods Learnt: Last Lecture

| | Vector Data | Set Data | Sequence Data | Text Data |
|---|---|---|---|---|
| **Classification** | **Logistic Regression; Decision Tree; KNN SVM; NN** | | | Naïve Bayes for Text |
| **Clustering** | **K-means; hierarchical clustering; DBSCAN;** Mixture Models | | | PLSA |
| **Prediction** | **Linear Regression** GLM* | | | |
| **Frequent Pattern Mining** | | Apriori; FP growth | GSP; PrefixSpan | |
| **Similarity Search** | | | DTW | |

# Vector Data: Density Estimation

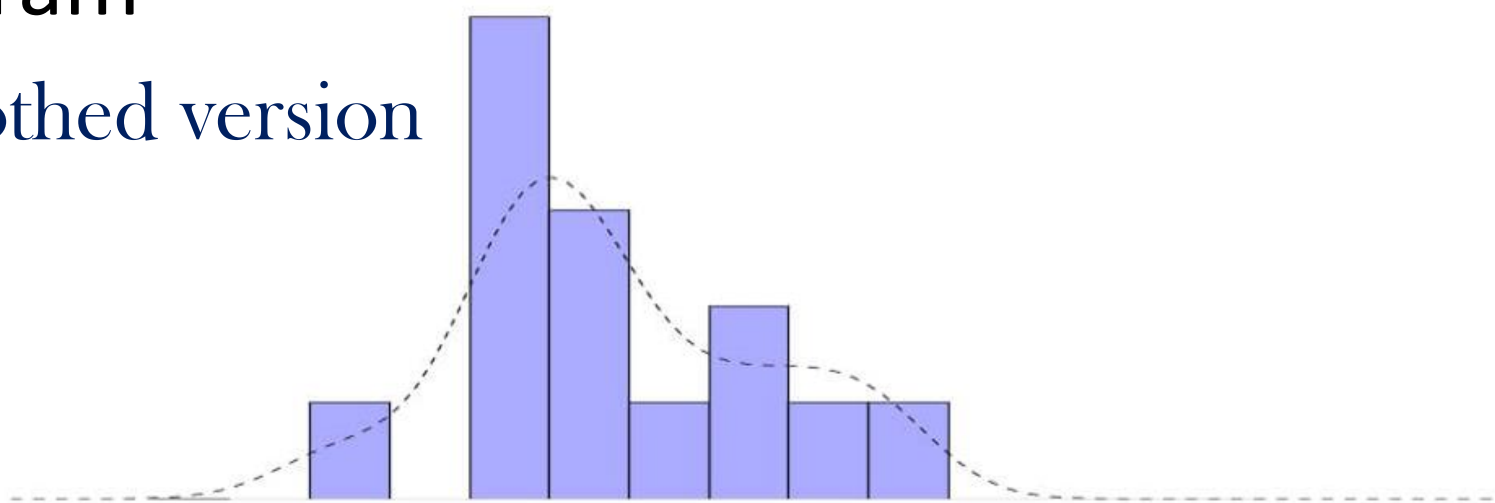
- Introduction
- Nonparametric Density Estimation
- Parametric Density Estimation
- Summary

# Vector Data: Density Estimation

- Introduction
- Nonparametric Density Estimation
- Parametric Density Estimation
- Summary

# Density Estimation from Data

- Goal
  - Estimate density function for a random variable from data
- Can be considered as an extension of histogram
  - Smoothed version



# Recall

- Density-based clustering can be viewed as identifying connected dense areas of a distribution
- Critical for many other mining functions
  - Classification
  - Outlier detection

# Nonparametric vs. parametric methods

- Nonparametric methods
  - No assumptions of the forms of the underlying densities
  - Can be used with arbitrary distributions
- Parametric methods
  - Have assumptions of the forms of the underlying densities
  - The densities are determined by fixed but unknown parameters

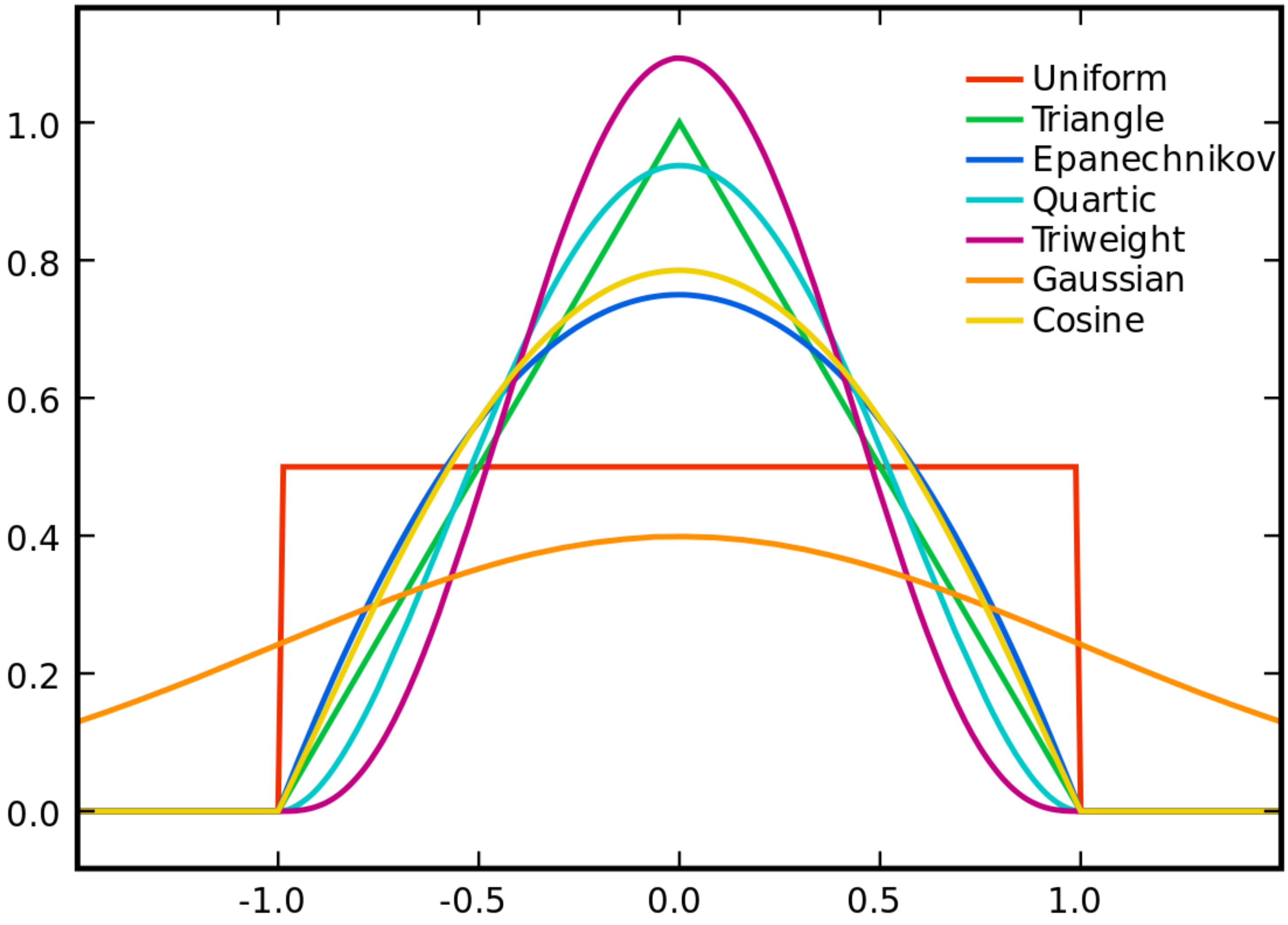
# Vector Data: Density Estimation

- Introduction
- Nonparametric Density Estimation
- Parametric Density Estimation
- Summary

# Kernel Density Estimation

- Given a dataset $D = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n)$, estimate its density function $f(\boldsymbol{x})$
- Kernel density estimator:
  - $\hat{f}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h})$
  - $h$: bandwidth, controlling the smoothness of $f$
  - $K$: a non-negative real-valued integrable function, serving as weighting function
    - $\int_{-\infty}^{+\infty} K(u)du = 1 \ (normalization)$
    - $K(u) = K(-u) \ for\ all\ u \ (symmetric)$

# Examples of Kernels



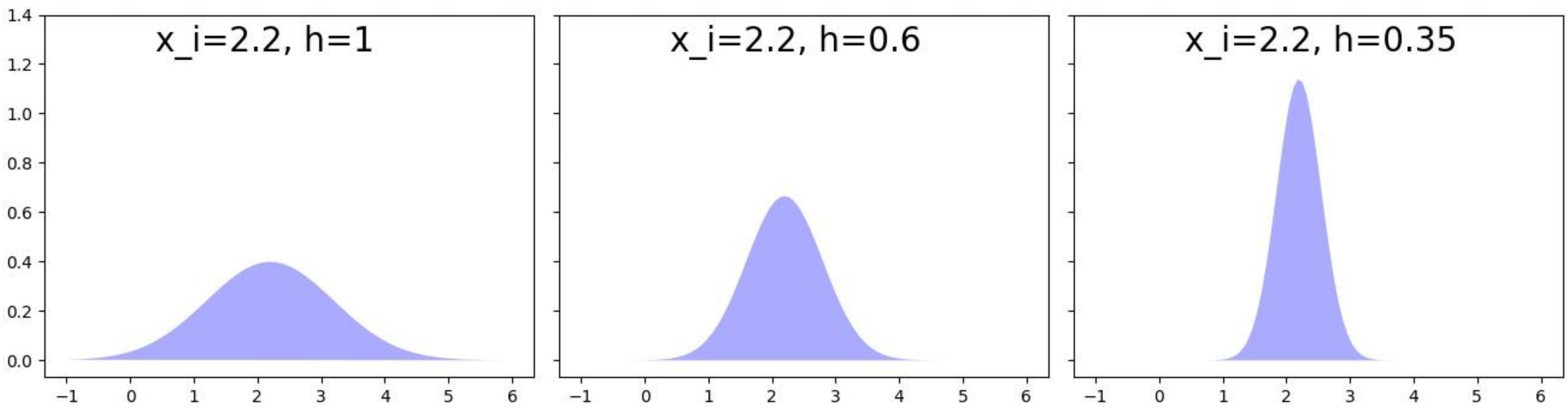
# Gaussian Kernel in 1-D case

- Example: Gaussian kernel
  - $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$
- Scaled kernel
  - $K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$
  - In the Gaussian kernel case: $K_h(u) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{u^2}{2h^2}}$

# Influence from one data point

- The influence of $x_i$ to $x$ can be considered as a weighting function centered at $x_i$

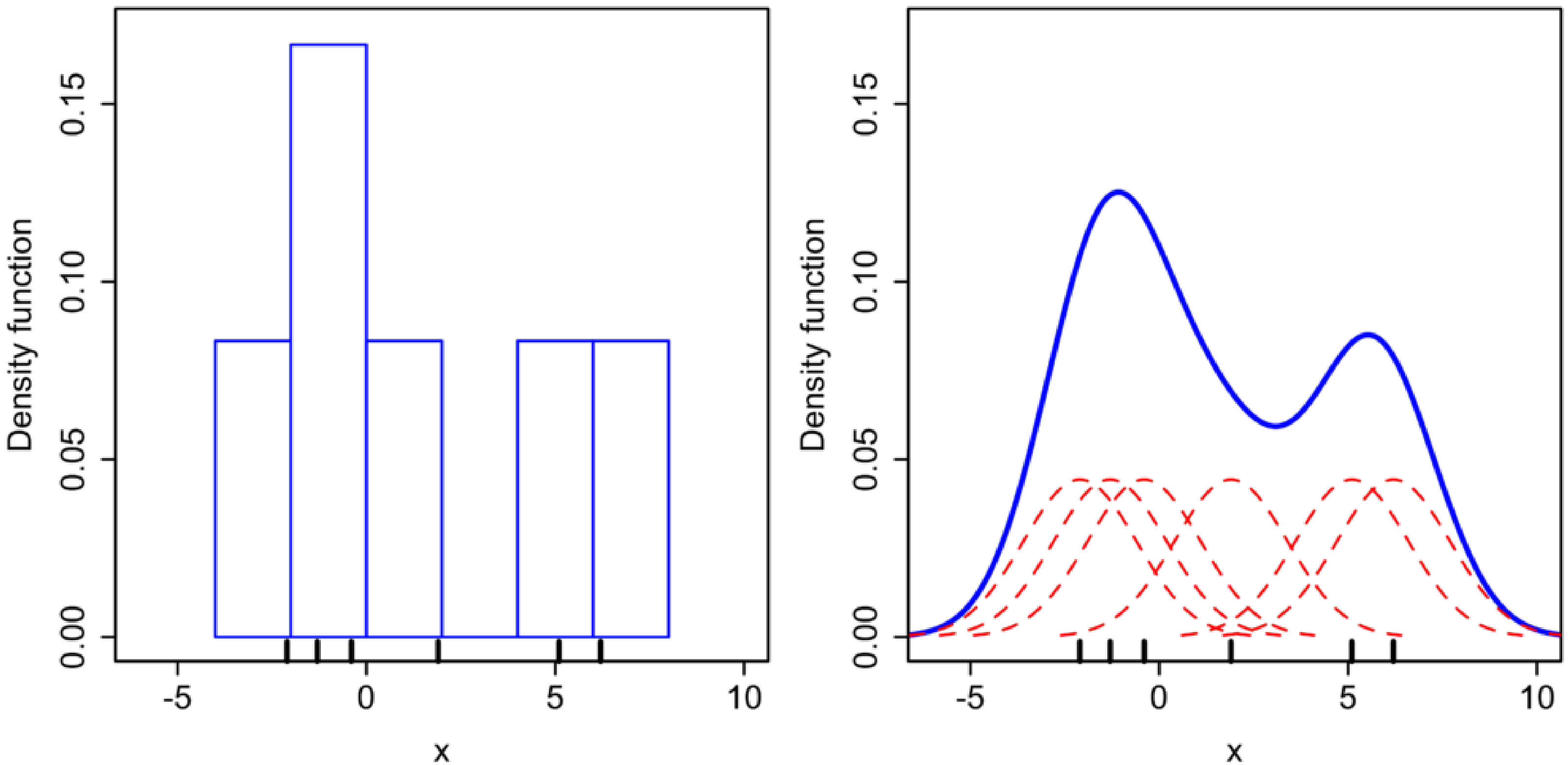
$$K_h(x - x_i) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}$$




Recall: $\hat{f}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h})$

# Influence from multiple data points

- Aggregate influence from multiple data points to $x$




histogram

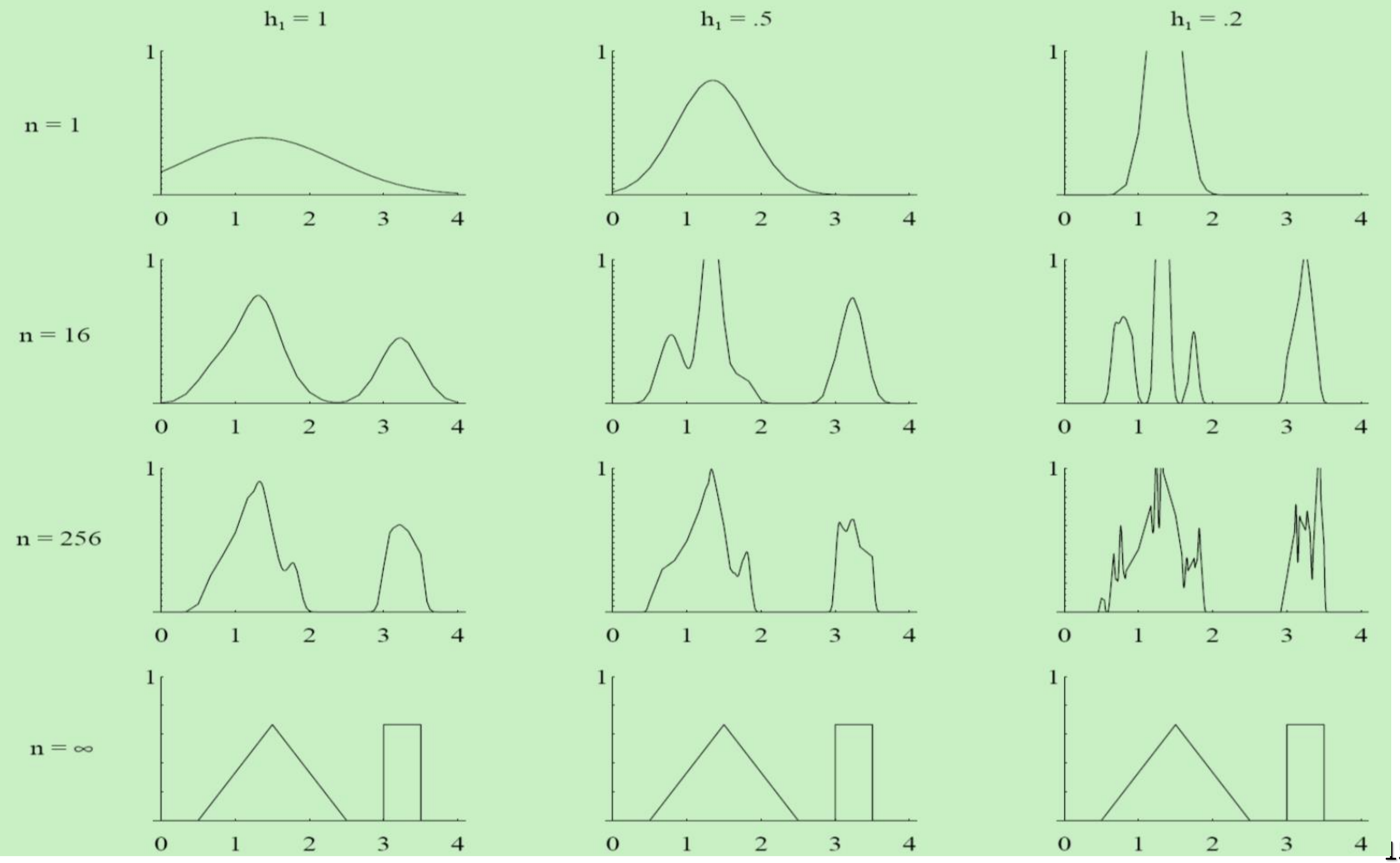
Density: Each red curve indicates $\frac{1}{\mathrm{n}} K_h(\boldsymbol{x} - \boldsymbol{x}_i)$

# Is it a density function?

- $\hat{f}_h(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} K_h(\boldsymbol{x} - \boldsymbol{x}_i)$
  - A density function has to integrate to 1
  - $K_h(\boldsymbol{x} - \boldsymbol{x}_i) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{(x-x_i)^2}{2h^2}}$ integrates to 1
  - Therefore, its average does so!

# Impact of bandwidth

higher h, smoother density function



# *DENCLUE: Using Statistical Density Functions for Clustering

- DENsity-based CLUstEring by Hinneburg & Keim (KDD'98)
- Using statistical density functions:

$$f_{Gaussian}(x,y) = e^{-\frac{d(x,y)^2}{2\sigma^2}}$$

influence of y on x

$$f_{Gaussian}^{D}(x) = \sum_{i=1}^{N} e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

total influence on x

$$\nabla f_{Gaussian}^{D}(x, x_i) = \sum_{i=1}^{N} (x_i - x) \cdot e^{-\frac{d(x,x_i)^2}{2\sigma^2}}$$

gradient of x in the direction of $x_i$

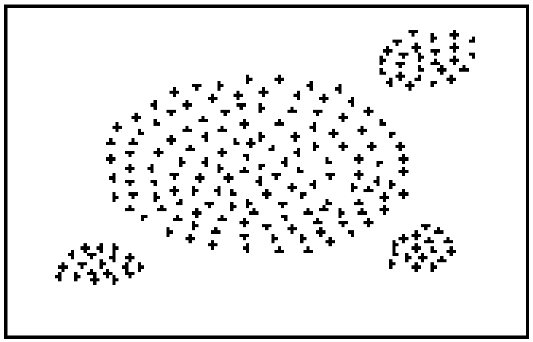
- Major features
  - Solid mathematical foundation
  - Good for data sets with large amounts of noise
  - Allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets
  - Significant faster than existing algorithm (e.g., DBSCAN)
  - But needs a large number of parameters

# *Denclue: Technical Essence

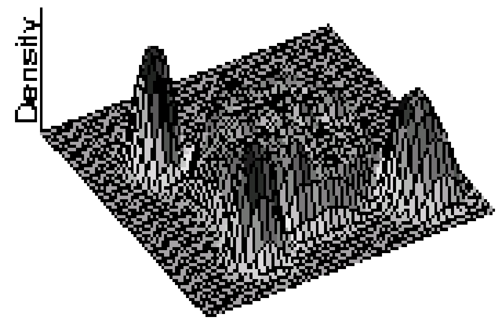
- Overall density of the data space can be calculated as the sum of the influence function of all data points
  - Influence function: describes the impact of a data point within its neighborhood
- Clusters can be determined mathematically by identifying density attractors
  - Density attractors are local maximal of the overall density function
  - Center defined clusters: assign to each density attractor the points density attracted to it
  - Arbitrary shaped cluster: merge density attractors that are connected through paths of high density (> threshold)

# *Density Attractor

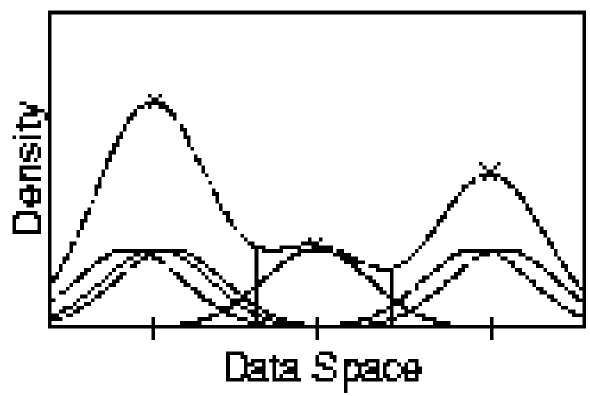
**Can be detected by hill-climbing procedure of finding local maximums**



(a) Data Set



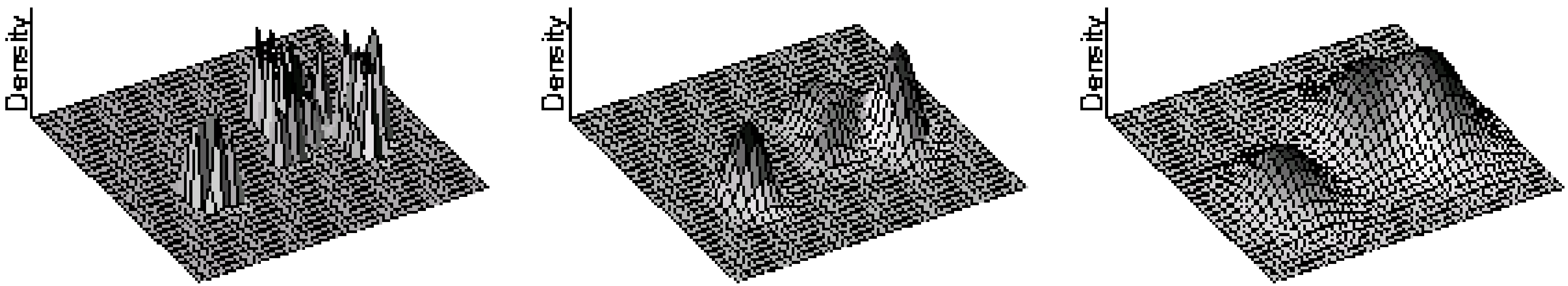

(c) Gaussian



# *Noise Threshold

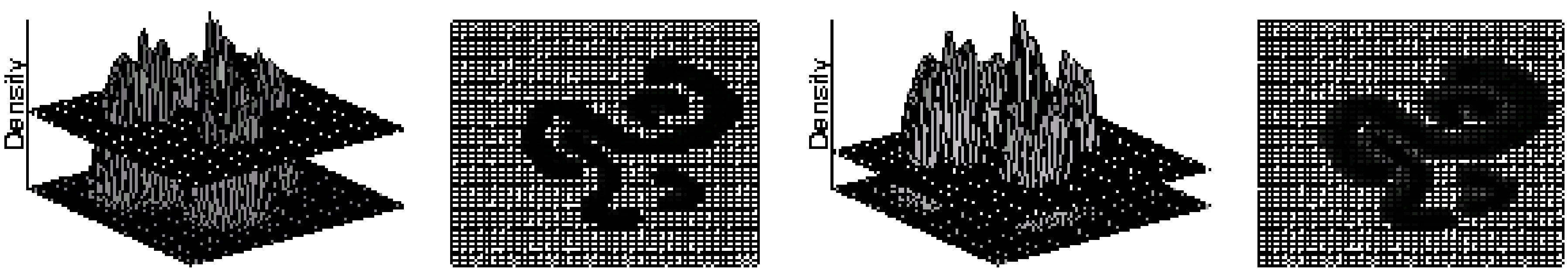
- Noise Threshold $\xi$
  - Avoid trivial local maximum points
  - A point can be a density attractor only if $\hat{f}(x) \geq \xi$

# *Center-Defined and Arbitrary




(a) $\sigma = 0.2$ (b) $\sigma = 0.6$ (d) $\sigma = 1.5$

Figure 3: Example of Center-Defined Clusters for different $\sigma$




(a) $\xi = 2$ (b) $\xi = 2$ (c) $\xi = 1$ (d) $\xi = 1$

Figure 4: Example of Arbitray-Shape Clusters for different $\xi$

# Vector Data: Density Estimation

- Introduction
- Nonparametric Density Estimation
- Parametric Density Estimation
- Summary

# Maximum-Likelihood Estimation

- Data: $D = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$
- Parameters: $\boldsymbol{\theta}$
- Model: $p(\boldsymbol{x}|\boldsymbol{\theta})$
- Likelihood of $\boldsymbol{\theta}$ with respective to a set of data samples

$$L(\boldsymbol{\theta}; D) = p(D|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\boldsymbol{x}_i|\boldsymbol{\theta})$$

- Maximum likelihood principle: find $\widehat{\boldsymbol{\theta}}$ that maximizes $L$
  - Agrees the most with the observation of current dataset

# Log-likelihood function

- log-likelihood function

$$l(\boldsymbol{\theta}) \equiv \ln L(\boldsymbol{\theta}) = \ln p(D|\boldsymbol{\theta}) = \sum_i \ln p(x_i|\boldsymbol{\theta})$$

- Maximize likelihood function is equivalent to maximize log-likelihood function

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta})$$

$$\Rightarrow \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = 0$$

$$\nabla_{\boldsymbol{\theta}} \equiv \begin{bmatrix} \frac{\partial}{\partial \theta_1} \\ \vdots \\ \frac{\partial}{\partial \theta_p} \end{bmatrix}$$

# The Gaussian Case: Unknown Mean

- Consider 1-d Gaussian Distribution

$$x_i \sim N(\mu, \sigma^2)$$

where $\sigma^2$ is known, i.e., $\theta = \mu$

$$p(x_i|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- The log-likelihood is then

$$l(\mu) = \sum_i \ln p(x_i|\mu) = \sum_i (-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2})$$

- The MLE estimator for $\mu$ is then
  - $\nabla_\mu l(\mu) = 0 \Rightarrow \sum_i (x_i - \hat{\mu}) = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_i x_i$

# The Gaussian Case: Unknown Mean and Variance

- Consider 1-d Gaussian Distribution

$$x_i \sim N(\mu, \sigma^2)$$

where both $\mu$ and $\sigma^2$ are unknown, i.e., $\boldsymbol{\theta} = (\mu, \sigma^2)$

$$p(x_i|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- The log-likelihood is then

$$l(\mu, \sigma^2) = \sum_i \ln p(x_i|\mu, \sigma^2) = \sum_i (-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2})$$

- The MLE estimators for $\mu$ and $\sigma^2$ are then
  - $\frac{\partial l(\mu,\sigma^2)}{\partial \mu} = 0 \Rightarrow \sum_i (x_i - \hat{\mu})/\sigma^2 = 0 \Rightarrow \hat{\mu} = \frac{1}{n}\sum_i x_i$
  - $\frac{\partial l(\mu,\sigma^2)}{\partial \sigma^2} = 0 \Rightarrow \sum_i (-\frac{1}{2\widehat{\sigma^2}} + \frac{(x_i - \hat{\mu})^2}{2(\widehat{\sigma^2})\text{^}2}) = 0 \Rightarrow \widehat{\sigma^2} = \boxed{\frac{1}{n}\sum_i (x_i - \hat{\mu})^2}$

Note it is biased

# Vector Data: Density Estimation

- Introduction
- Nonparametric Density Estimation
- Parametric Density Estimation
- Summary

# Summary

- Nonparametric Density Estimation
  - Kernel density estimation
- Parametric Density Estimation
  - Maximum likelihood estimation