Ch 05. Non-parametric Method

Part 1 Parzen Window Estimation

Approaches to Pattern Classification

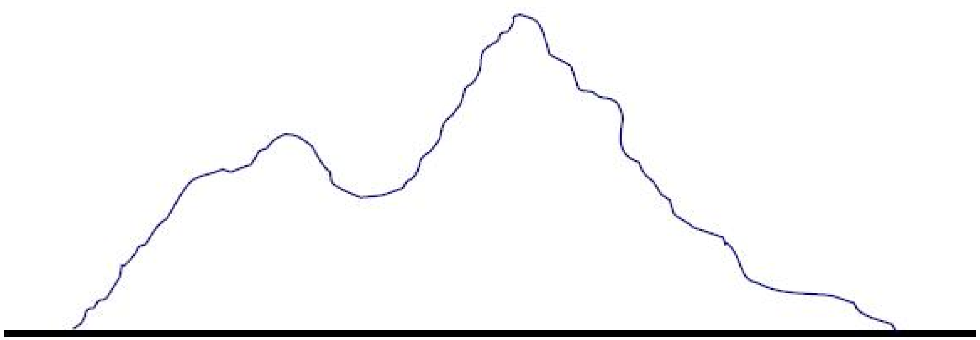
- **Approach 1**： Estimate class-conditional probability density $p(\mathbf{x} \mid \omega_i)$
  - Through $p(\mathbf{x} \mid \omega_i)$ and $P(\omega_i)$, calculate posterior probability $P(\omega_i \mid \mathbf{x})$ with Bayes' rule, then make decisions with maximum posterior probability
  - Two Methods
    - **Method 1a**： Parameter estimation of probability density

      Based on parametric description of $p(\mathbf{x} \mid \omega_i)$
    - **Method 1b**： Non-parametric estimation of probability density

      Based on non-parametric description of $p(\mathbf{x} \mid \omega_i)$
- **Approach 2**： Estimate posterior probability $P(\omega_i \mid \mathbf{x})$
  - Don't have to estimate $p(\mathbf{x} \mid \omega_i)$ in advance
- **Approach 3**： Compute discrimination function
  - Don't have to estimate $p(\mathbf{x} \mid \omega_i)$ or $P(\omega_i \mid \mathbf{x})$

Possible Problems of Parameter Estimation

- The form of the probability density function is unknown
- The classical density function cannot describe the real data well
  - The parametric form of the classical density function is generally unimodular
  - Real data is often multimodal
  - Some complex data are difficult to model in parametric form



- Solution： **non-parametric method** （非参数方法）

Non-parametric Method

- can handle any probability density
- don't have to assume the parametric form of the density function
- No Free Lunch!
  - The training samples required by non-parametric methods to obtain better results are generally much larger than those of parametric methods

Non-parametric Density Estimation

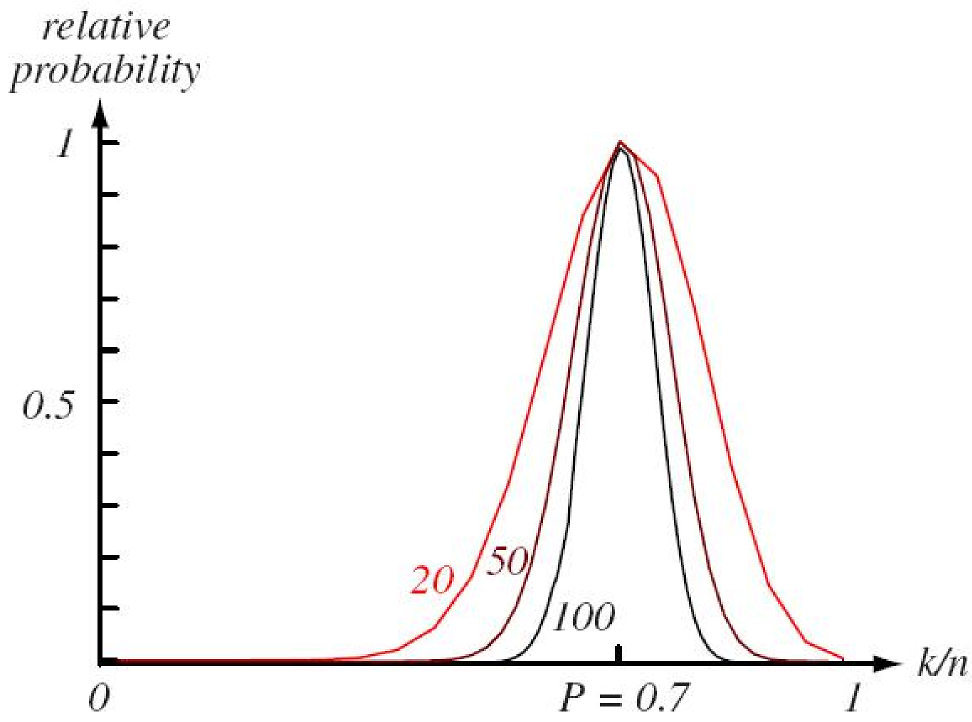
- Suppose the probability density of x is $p(x)$, then the probability of any x falling into region **R** is $P = \int_{\mathcal{R}} p(\mathbf{x})\, d\mathbf{x}$
- The basic idea of non-parametric density estimation
  - Estimate p of x by estimating the probability of a small region R around x
- Suppose there are n i.i.d. samples, and the probability of k samples falling into R is

$$P_k = \binom{n}{k} P^k (1-P)^{n-k}$$

binomial distribution

- The expectation value of k $E[k] = nP$
- When the amount of data n is large, P_K has a very significant peak near nP, and E(k) can be replaced by the observed value of k $P = \frac{E[k]}{n} \approx \frac{k}{n}$

Non-parametric Density Estimation



Non-parametric Density Estimation

- Mean value theorem of integrals

$$P = \int_{\mathcal{R}} p(\mathbf{x})\, d\mathbf{x} = p(\mathbf{x}')\, V$$

  - $V = \int_{\mathcal{R}} d\mathbf{x}$ is the measure of the size of region R (length, area, volume, etc.)
  - x' is a certain point of R

- If R is small enough so that the change in p of x is small in R, then $P \approx p(\mathbf{x})\, V$

  x is any point in R

- Put $P = \frac{E[k]}{n} \approx \frac{k}{n}$ into, get

$$\frac{k}{n} \approx p(\mathbf{x})\, V \quad \text{or} \quad p(\mathbf{x}) \approx \frac{k/n}{V}$$

The Choice of V

- In the case of a finite number of samples n
  - **V** is too large

    p(x) is smoothed
  - **V** approaches 0
    - If there are no sample points in R, then $p(\mathbf{x}) = \dfrac{k/n}{V} = 0$
    - If there happens to be a sample in R,then $p(\mathbf{x}) = \dfrac{k/n}{V} \approx \infty$
- Suppose number of samples can be unlimited
  - Construct a series of regions containing **x:** $R_1, R_2, \ldots$
    - R_1 uses one sample
    - R_2 uses two samples
    - ……

The number of samples that fall into R_n

Estimate by R_n

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$$

The volume of R_n

The Choice of V

- If $p_n(x) \rightarrow p(x)$, the following conditions must be satisfied
  - $\lim_{n\to\infty} V_n = 0$

    The smoothed P/V can converge to p(x)
  - $\lim_{n\to\infty} k_n = \infty$

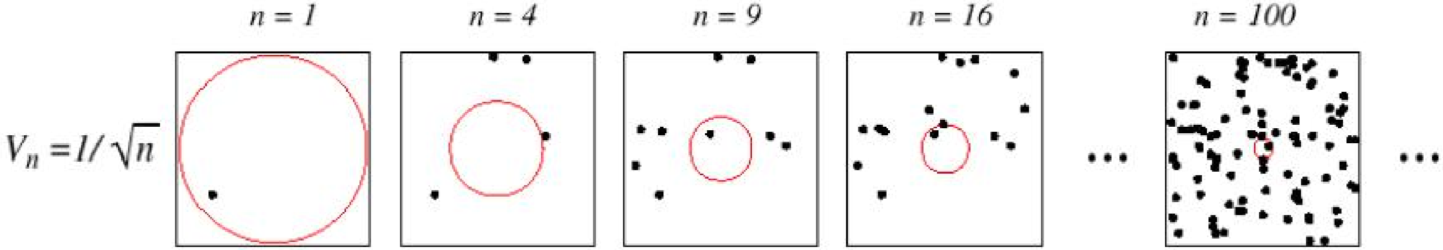
    The ratio of frequency k/n can converge to P
  - $\lim_{n\to\infty} k_n/n = 0$

    Even if the samples falling in R tend to infinity, their proportion in the entire data set is still small

The Choice of V

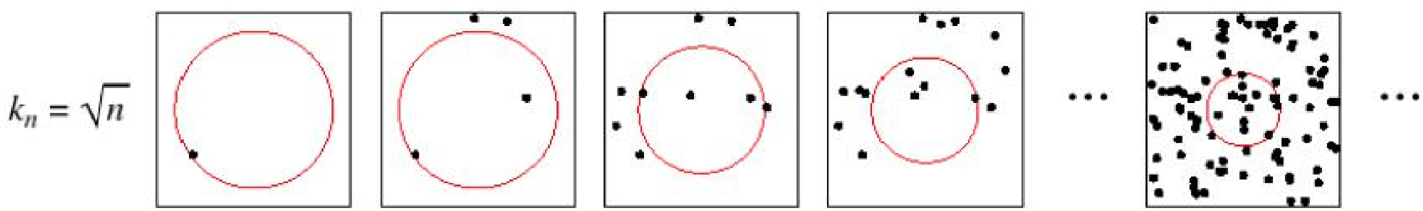
- Two approaches
  - To gradually shrink a given initial interval according to a given volume function, e.g $V_n = 1/\sqrt{n}$

Parzen window method



  - Identify k_n as some function of n $k_n = \sqrt{n}$

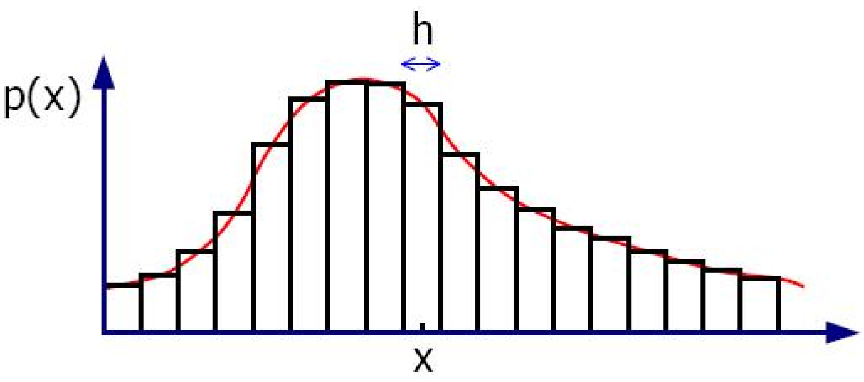
k_n-nearest neighbor method



A simple example

- Given a data set containing n samples $D = \{x_1, x_2, ..., x_n\}$
- Use histograms to simulate p(x)

$$P(|x - x_j| \le \frac{h}{2}) \approx p(x)\,h$$



- Suppose k samples fall into a small bar (width h) with x as the midpoint. If n is large enough, then

$$P(|x - x_j| \le \frac{h}{2}) \approx \frac{k}{n}$$

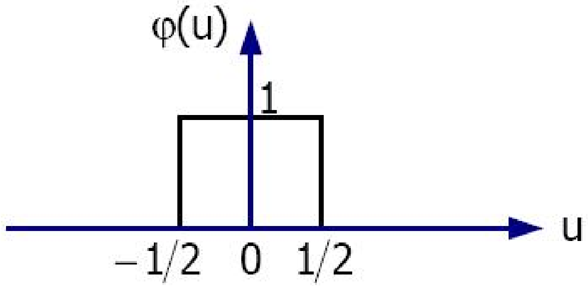
- According to the above two approximations, get

$$p(x) \approx \frac{k/n}{h}$$

A simple example

- Define window function (kernel function, potential function)

$$\varphi(u) = \begin{cases} 1 & |u| \le 1/2 \\ 0 & \text{otherwise} \end{cases}$$



- The number of samples that fall into bars of width h and midpoint x

$$k = \sum_{j=1}^{n} \varphi\left(\frac{x - x_j}{h}\right)$$

- Non-parametric simulation of p(x)

$$p(x) \approx \frac{k/n}{h} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h}\varphi\left(\frac{x - x_j}{h}\right)$$

The mean of a certain function of x_j

Parzen Window Method

- Suppose R is a d-dimensional hypercube with side length h, then the volume of R is $V = h^d$
- Define window function

$$\varphi(\mathbf{u}) = \begin{cases} 1 & |u_i| \le 1/2 \quad \text{for all } 1 \le i \le d \\ 0 & \text{otherwise} \end{cases}$$

- The number of samples that fall into R

$$k = \sum_{j=1}^{n} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_j}{h}\right)$$

- Approximate p(x) with n samples

$$p(\mathbf{x}) \approx \hat{p}_n(\mathbf{x}) = \frac{k/n}{V} = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{V} \varphi\left(\frac{\mathbf{x} - \mathbf{x}_j}{h}\right)$$

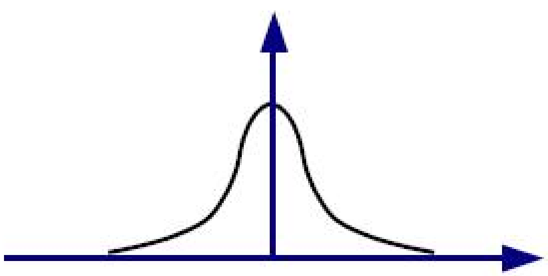
Parzen Window Method

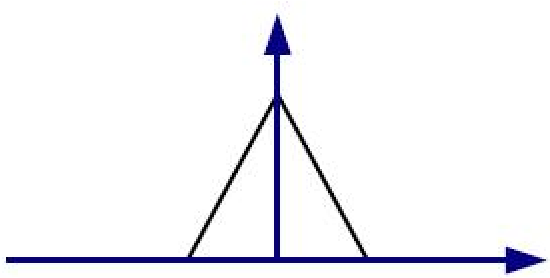
- Define $\delta(\mathbf{x}) \equiv \frac{1}{V}\varphi\left(\frac{\mathbf{x}}{h}\right)$

  Then the approximation to p(x) can be rewritten as

$$p(\mathbf{x}) \approx \frac{1}{n}\sum_{j=1}^{n} \boxed{\delta\left(\mathbf{x}-\mathbf{x}_j\right)} \longrightarrow \text{interpolation function}$$

- Basic idea
  - Each sample x_j makes a contribution to the estimation of p(x), which is represented by some form of interpolation function based on the distance from x_j to x
- Generalization： examples of interpolation functions





Parzen Window Method

- Because $\varphi(\mathbf{u}) \geq 0$ and $\int \varphi(\mathbf{u})\, d\mathbf{u} = 1$, so

$$\hat{p}_n(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{V}\varphi\left(\frac{\mathbf{x}-\mathbf{x}_j}{h}\right) \geq 0$$

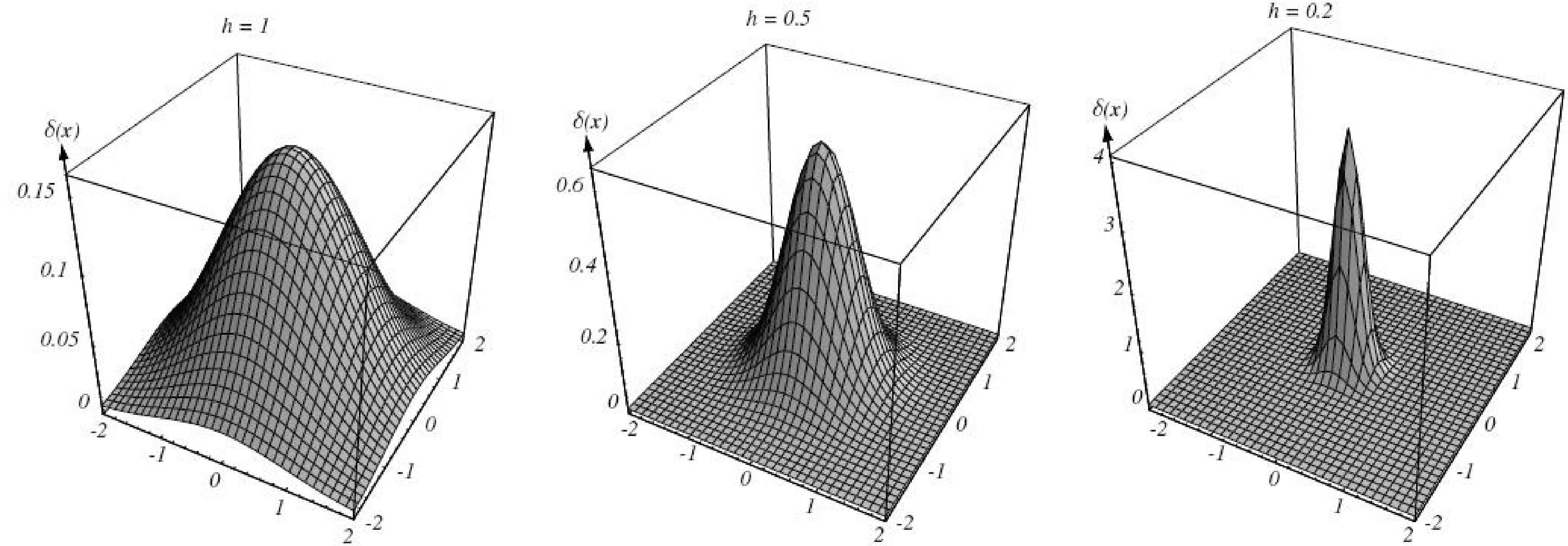
$$\int \hat{p}_n(\mathbf{x})\, d\mathbf{x} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{V}\int \varphi\left(\frac{\mathbf{x}-\mathbf{x}_j}{h}\right) d\mathbf{x} = \frac{1}{n}\sum_{j=1}^{n}\frac{h^d}{V}\int \varphi(\mathbf{u})\, d\mathbf{u} = 1$$

It shows that the estimated $\hat{p}_n(\mathbf{x})$ is a reasonable probability density function

- Discreteness
  - if $\varphi(\mathbf{u})$ is discrete， then $\hat{p}_n(\mathbf{x})$ is discrete
  - if $\varphi(\mathbf{u})$ is continuous， then $\hat{p}_n(\mathbf{x})$ is continuous

Parzen Window Method

$$\delta(\mathbf{x}) \equiv \frac{1}{V}\varphi\left(\frac{\mathbf{x}}{h}\right) \qquad p(\mathbf{x}) \approx \frac{1}{n}\sum_{j=1}^{n}\delta\left(\mathbf{x}-\mathbf{x}_j\right)$$



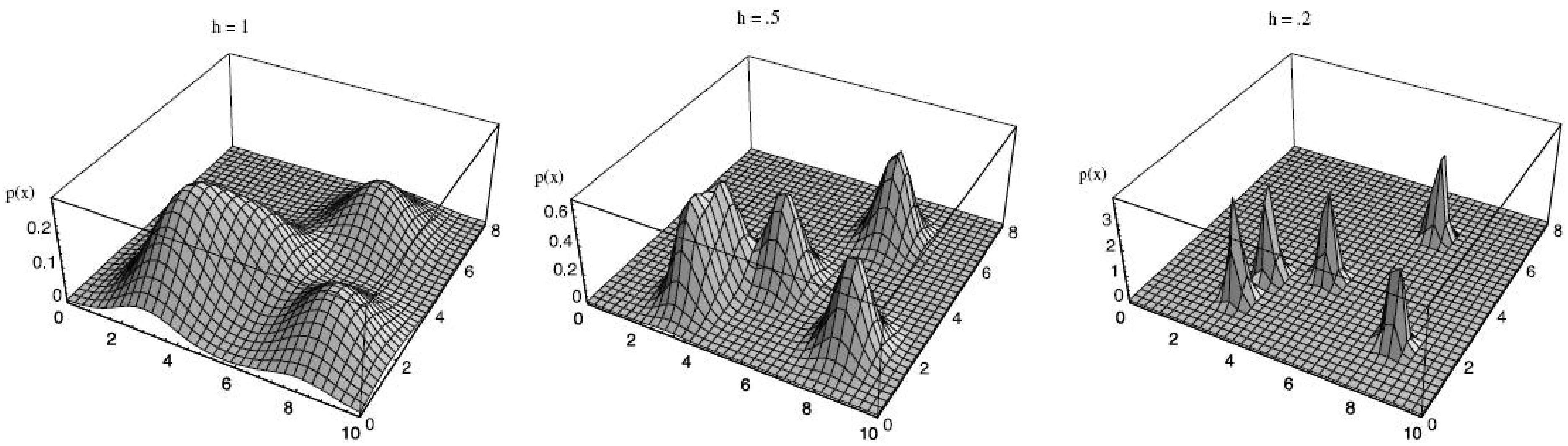
Parzen Window Method

$$\delta(\mathbf{x}) \equiv \frac{1}{V}\varphi\left(\frac{\mathbf{x}}{h}\right) \qquad p(\mathbf{x}) \approx \frac{1}{n}\sum_{j=1}^{n}\delta(\mathbf{x}-\mathbf{x}_j)$$

- The effect of h(or V) on $\hat{p}_n(\mathbf{x})$
  - When **h is very large**, the distance between x_j and x has little effect on $\delta(\mathbf{x}-\mathbf{x}_j)$
    - $\hat{p}_n(\mathbf{x})$ is the sum of n wide, slowly varying functions
    - $\hat{p}_n(\mathbf{x})$ is a very smooth estimate of p(x) -- the defocus estimate
    - The resolution of the estimated results is low
  - When **h is very small**, the peak of $\delta(\mathbf{x}-\mathbf{x}_j)$ is very sharp
    - $\hat{p}_n(\mathbf{x})$ is the superposition of n sharp pulses centered on sample points
    - $\hat{p}_n(\mathbf{x})$ is a noisy estimate of p(x)
    - The statistical stability of the estimated results is insufficient

Parzen Window Method

- Under the constraint of finite number of samples n, h (or V) should take some acceptable compromise
- With the increase of n, h (or V) should be reduced gradually to make the estimate of P (x) more accurate



Parzen Window Method

- Generalization of the **window function**

$$p(\mathbf{x}) \approx \hat{p}_n(\mathbf{x}) = \frac{k/n}{V} = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{V}\boxed{\varphi\left(\frac{\mathbf{x}-\mathbf{x}_j}{h}\right)}$$

  - Don't have to specify R as a hypercube, but rather some generalized form defined by the window function, which satisfies the conditions

    $$\varphi(\mathbf{u}) \geq 0 \qquad \int \varphi(\mathbf{u})\, d\mathbf{u} = 1$$

    to make $\hat{p}_n(\mathbf{x})$ be a reasonable probability density function
  - h： **the width of window**

Example One

- p(x) is a normal distribution of zero mean value, unit variance and univariate
- window function is

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}$$

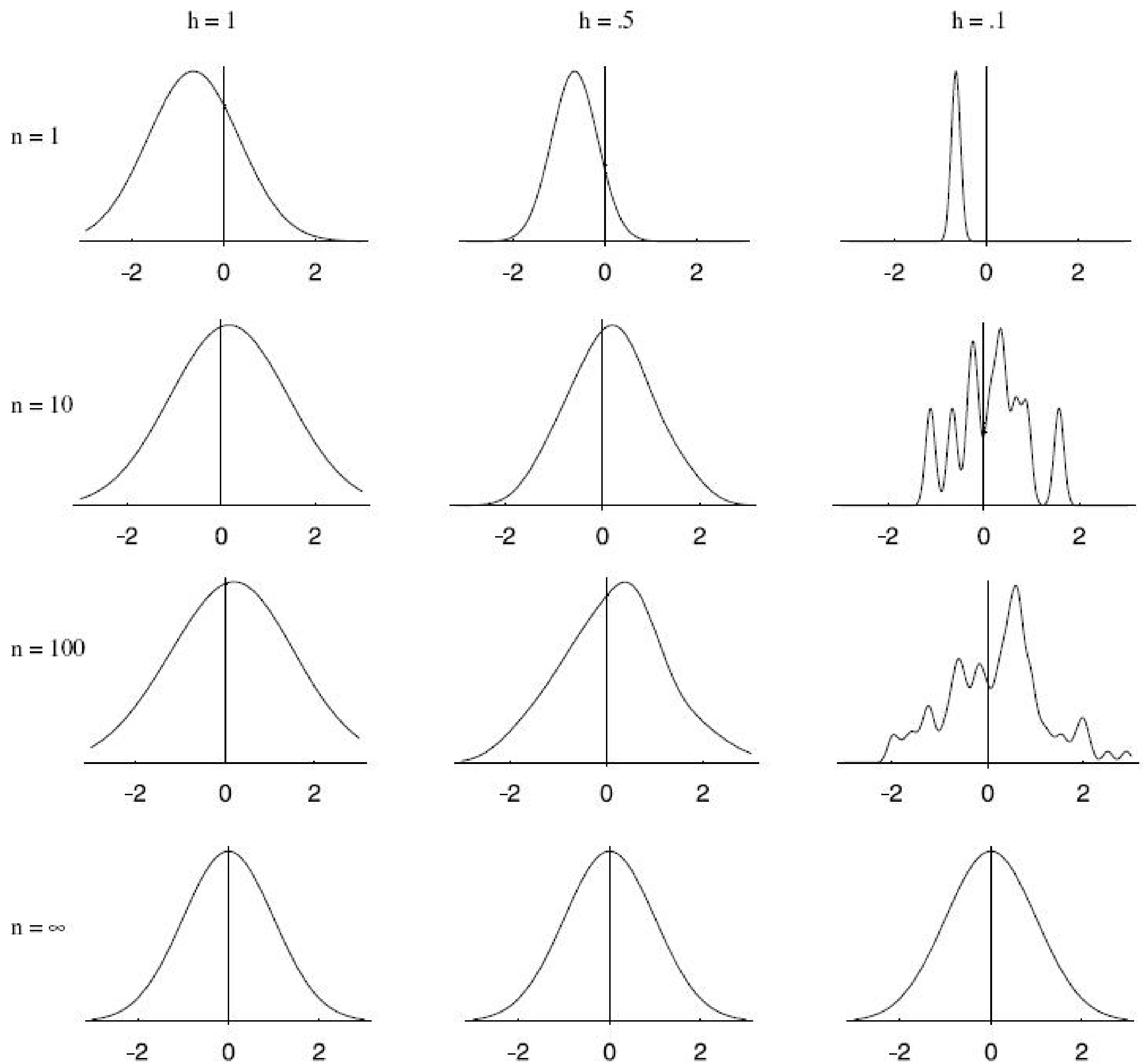
- the volume

$$V_n = h_n = h_1 / \sqrt{n}$$

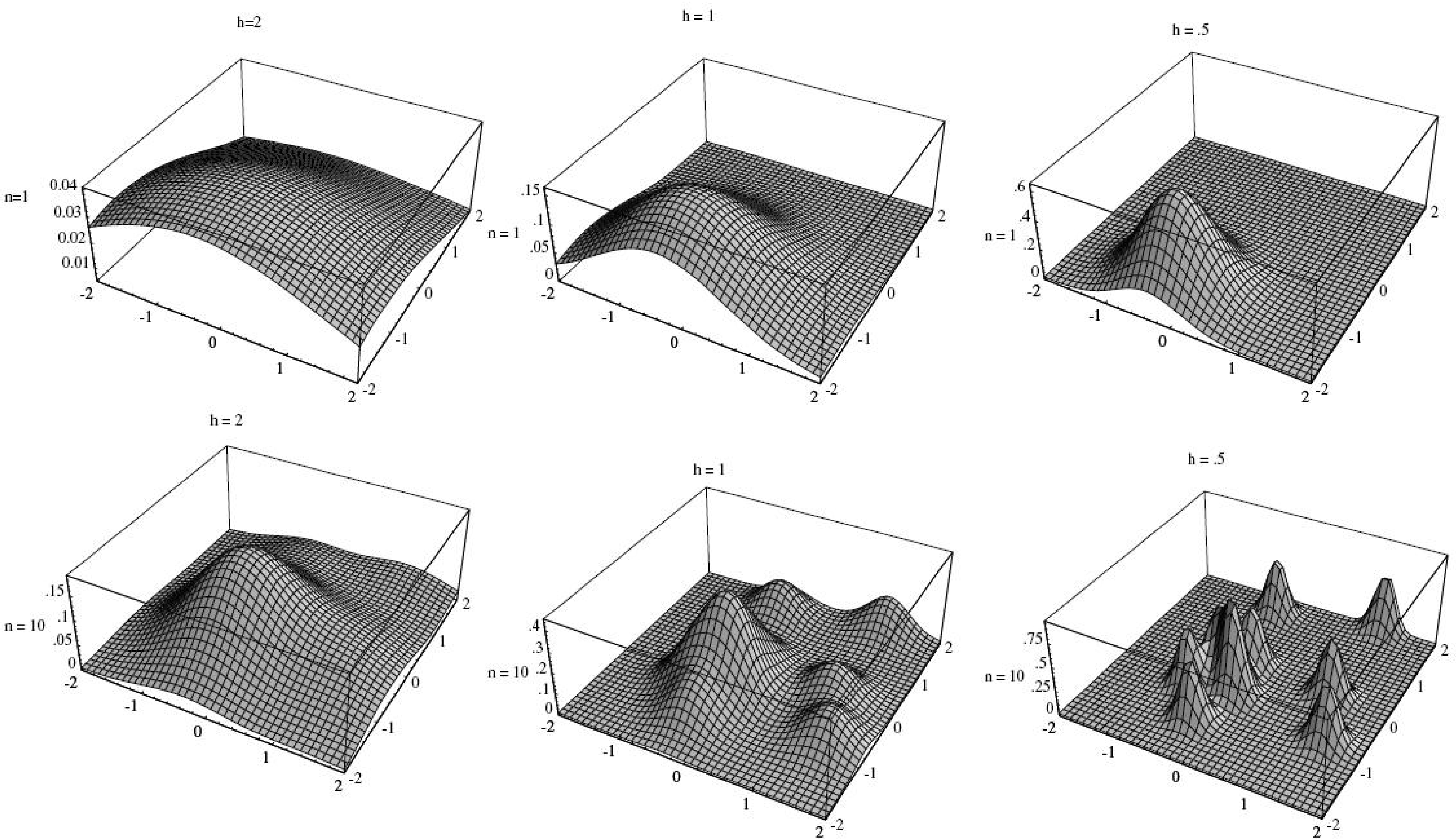
- Parzen window estimation

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right)$$

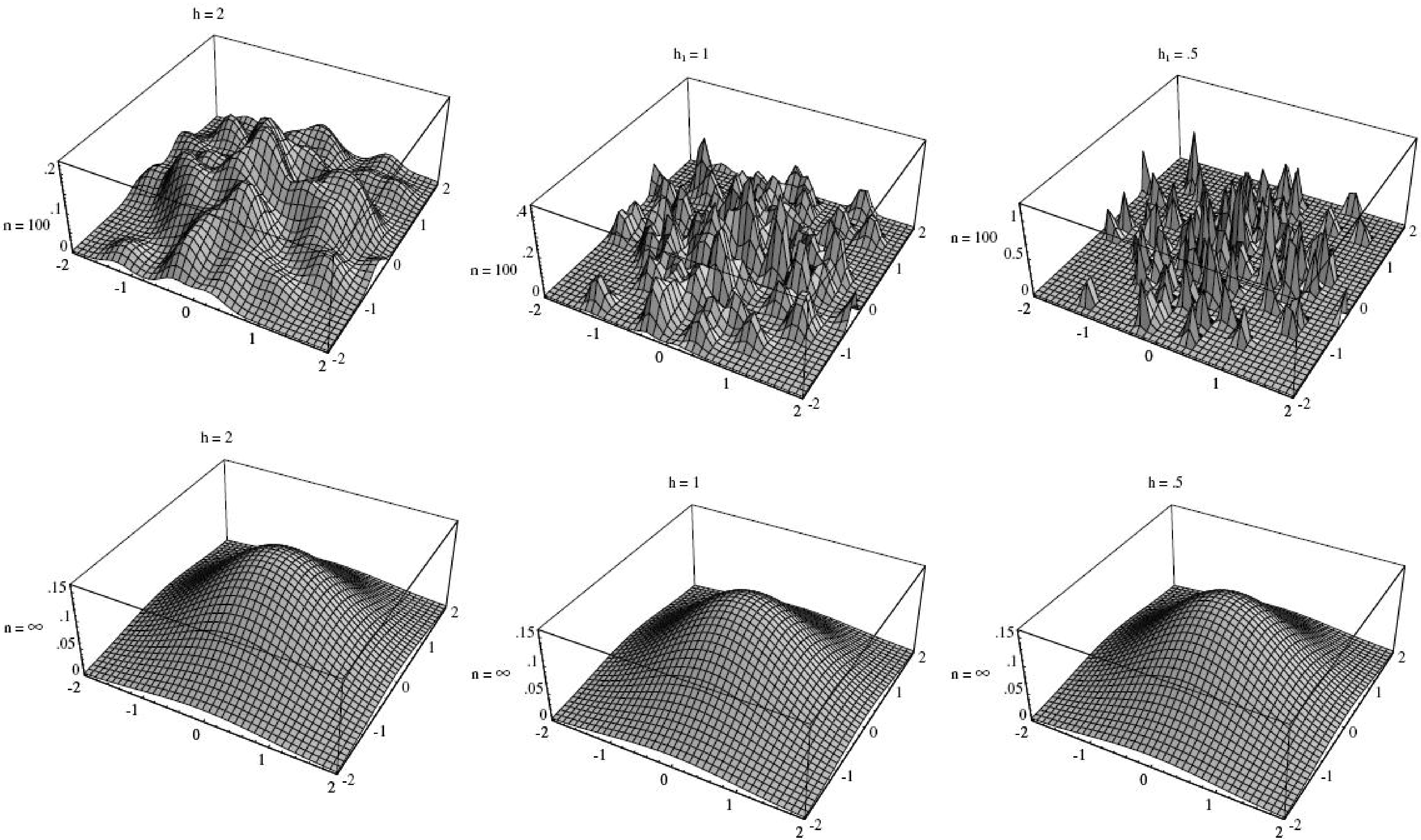
Example One



Example One : The Two-dimensional Case

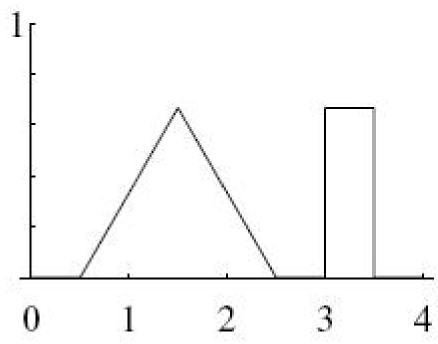


Example One: The Two-dimensional Case



Example Two

- p(x) is a mixed distribution of a uniform distribution and a triangular distribution



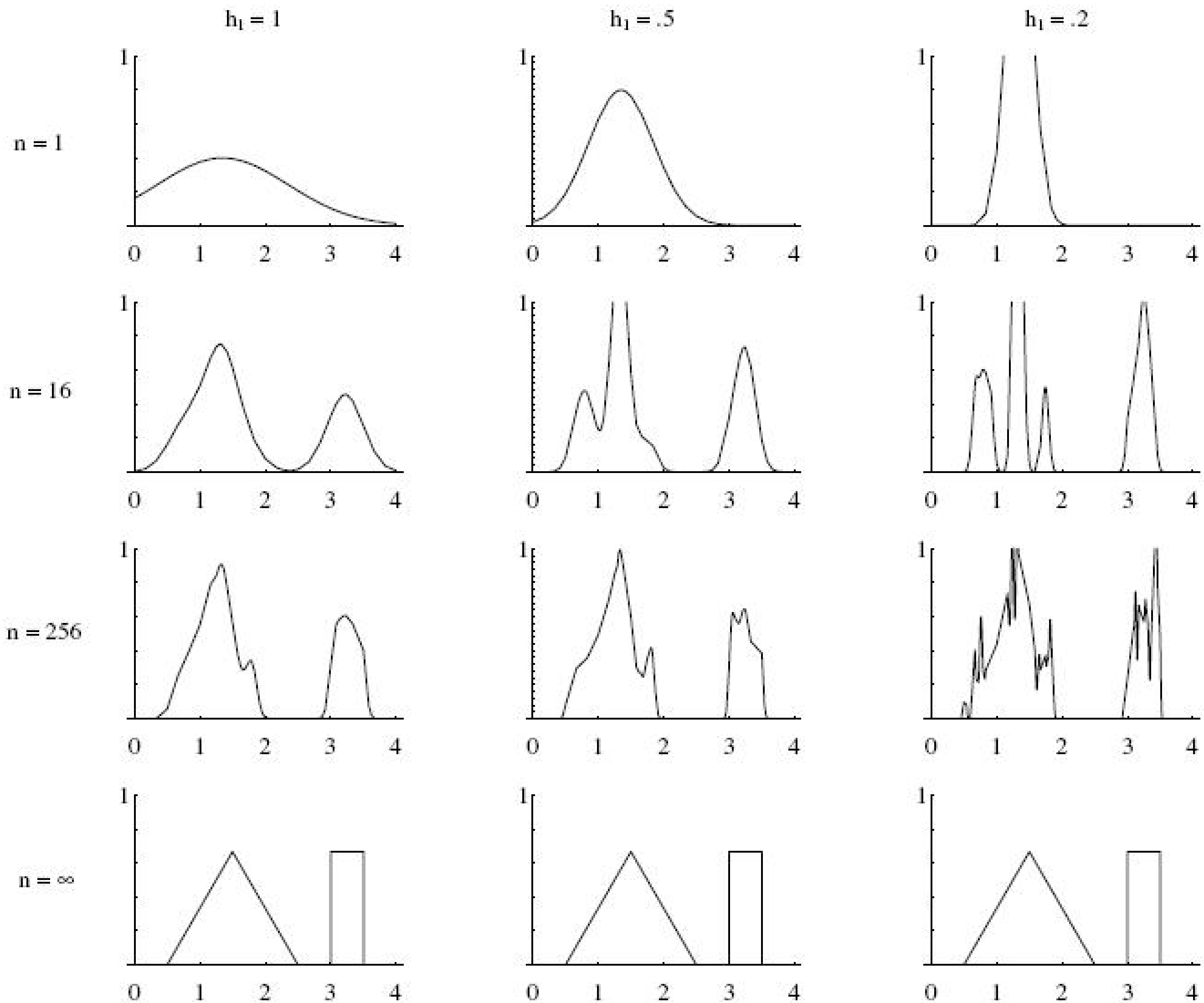
- the window function

$$\varphi(u) = \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}$$

- the width of window

$$h_n = h_1 / \sqrt{n}$$

Example Two

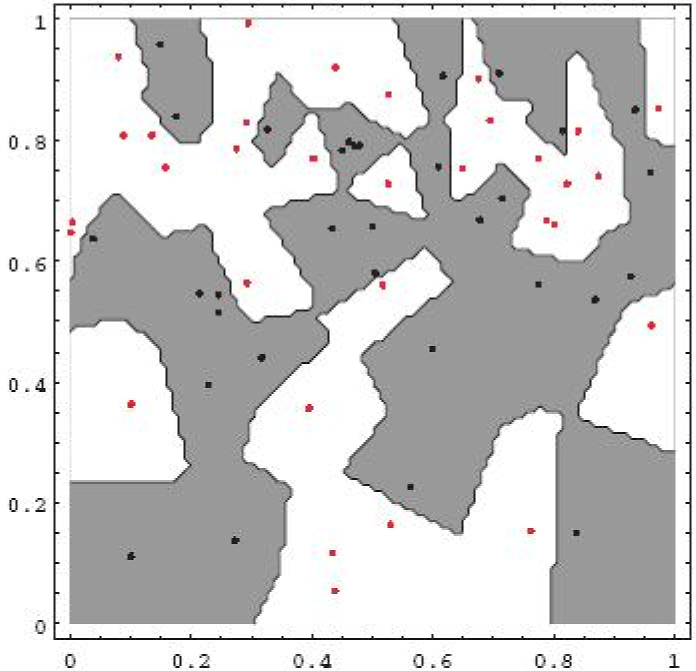


Classification

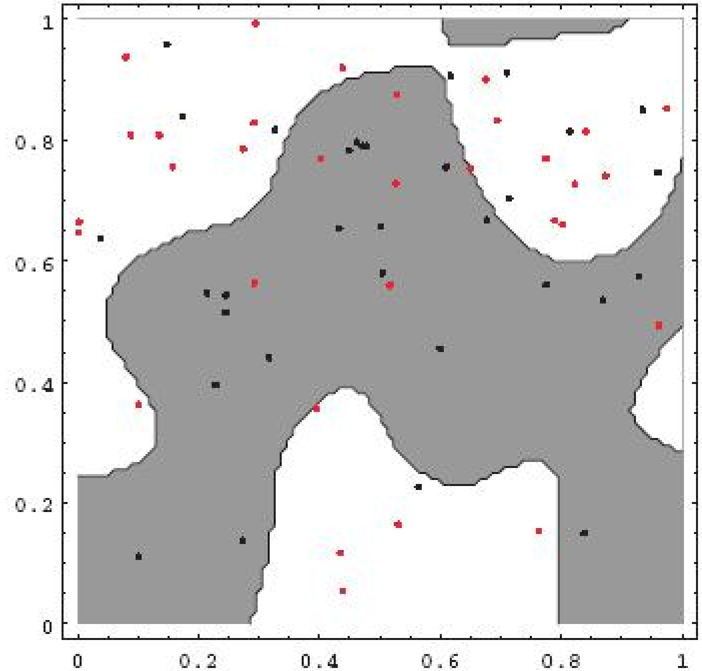
- The classifier based on Parzen window estimation
  - for each class, use the Parzen window to estimate the class-conditional probability density $p(\omega_i \mid \mathbf{x})$
  - then, calculate the posterior probability $p(\mathbf{x} \mid \omega_i)$ by using Bayes formula
  - Classification according to the principle of "maximum posterior probability" (MAP)

Classification

- Parzen window classifier's **decision domain** and **window function**



small window width



large window width

Advantages vs. Disadvantages

- Non-parametric method
  - **Advantages**
    - Generality: It is possible to estimate distributions without knowing their forms
    - When training samples are sufficient, no matter what the form of the actual probability density function is, a reliable convergence result will definitely be obtained in the end
  - **Disadvantages**
    - A large number of training samples are required, which are often much larger than the number of training samples required for parameter estimation given the parametric form of the distribution
    - **The curse of dimensionality** (**维数灾难**) for the number of training samples increases exponentially with the dimensionality of the feature space