80250993 Machine Learning
@Tsinghua University

Chapter 11
Probability Density Estimation

Xuegong Zhang
November 11, 2021

11.1
Introduction

Recall: Bayesian Decision

- Bayesian Decision

  Given the number of classes (states) $\omega_i, i = 1, \cdots, c$, the prior and conditional densities

  $$P(\omega_i),\ P(\boldsymbol{x}|\omega_i),\ i = 1, \cdots, c\ ,$$

  we can make the best decision to minimize error or risk.

- Usual situations we face

  Given the number of classes (states) $\omega_i, i = 1, \cdots, c$ and a set of samples in each class $\mathcal{X}_i$

- Two steps:
  - Estimate $P(\omega_i)$ and $p(x|\omega_i)$ from the samples
  - and use the estimated $\hat{P}(\omega_i)$ and $\hat{p}(x|\omega_i)$ to do Bayesian Decision

—— "Once we know the density, we know everything."

Parametric Estimation

- Setting of the problem
  - Samples $\mathcal{X}$ are i.i.d. from density $p(x)$ of the form $p(x|\theta)$, in which only the parameters θ are unknown
  - Task: To estimate $p(x)$ from samples ➔ To estimate θ from samples
- Two approaches:
  - Maximum Likelihood Estimation
  - Bayesian Estimation

Non-parametric Estimation

- Setting of the problem
  - Samples $\mathcal{X}$ are i.i.d. from density $p(x)$ of unknown form $p(x|\theta)$
  - Task: To estimate $p(x)$ from samples

11.2 Maximum Likelihood Estimation

Maximum Likelihood Estimation

- Set of the problem

  ① Samples $\mathcal{X}_i, i = 1, \cdots, c$ are i.i.d. from density $p(x|\omega_i)$

  ② $p(\boldsymbol{x}|\omega_i)$ is of the form $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$, in which only the parameters $\boldsymbol{\theta_i}$ are unknown

  ③ Parameters $\boldsymbol{\theta}_i$ are unknown deterministic vectors

  ④ Samples only contain information of parameters of the same class.

  Problem:

  To estimate $\boldsymbol{\theta}_i$ from given samples

Parametric estimation

- Principle of Estimation: Maximum Likelihood

- Likelihood Function

$$l(\boldsymbol{\theta}) = p(\mathcal{X}|\boldsymbol{\theta}) = p(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N|\theta) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta)$$

$$H(\boldsymbol{\theta}) = \ln l\,(\boldsymbol{\theta})$$

- ML Estimation: $\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, H(\boldsymbol{\theta})$
- Solution
  - If the likelihood function is continuous and differentiable, then the estimate is the solution of the set of equations:

$$\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}) = 0$$

$$\nabla_{\boldsymbol{\theta}} = \left[\frac{\partial}{\partial \boldsymbol{\theta}_1}, \frac{\partial}{\partial \boldsymbol{\theta}_2}, \cdots, \frac{\partial}{\partial \boldsymbol{\theta}_s}\right]^T$$

- For Gaussian Distribution, the ML estimation of mean and variance is

$$\hat{\mu} = \hat{\theta}_1 = \frac{1}{N}\sum_{k=1}^{N} x_k$$

$$\hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{N}\sum_{k=1}^{N} (x_k - \hat{\mu})^2$$

- However, it can be shown that the estimate $\hat{\sigma}^2 = \hat{\theta}_2 = \frac{1}{N}\sum_{k=1}^{N}(x_k - \hat{\mu})^2$ is biased. The unbiased (but not minimum variance) estimator is

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{k=1}^{N} (x_k - \hat{\mu})^2$$

R

11.3
Bayesian Estimation

Bayesian Estimation

- Set of the problem

  ① Samples $\mathcal{X}_i, i = 1, \cdots, c$ are i.i.d. from density $p(\boldsymbol{x}|\omega_i)$

  ② $p(\boldsymbol{x}|\omega_i)$ is of the form $p(\boldsymbol{x}|\boldsymbol{\theta}_i)$, in which only the parameters $\boldsymbol{\theta}_i$ are unknown

  ③ Parameters $\boldsymbol{\theta}_i$ are ~~unknown deterministic vectors~~ **random vectors** with *a prior* density of $p(\boldsymbol{\theta}_i)$

  ④ Samples only contain information of parameters of the same class.

  Problem:

  To estimate $\boldsymbol{\theta}$ from given samples

- Principle of Estimation: Minimal Risk

- Loss Function of an estimation

$$\lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}), \text{ e.g., } \lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$$

- Expected Risk

$$R = \int_{E^d}\int_{\Theta} \lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{\theta}) d\boldsymbol{\theta} d\boldsymbol{x} = \int_{E^d}\int_{\Theta} \lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{\theta} d\boldsymbol{x} = \int_{E^d} R(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}$$

- Risk conditional on $\boldsymbol{x}$

$$R(\hat{\boldsymbol{\theta}}|\boldsymbol{x}) = \int_{\Theta} \lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{x}) d\boldsymbol{\theta}$$

- Empirical Risk on all training data

$$R(\hat{\boldsymbol{\theta}}|\boldsymbol{\mathcal{X}}) = \int_{\Theta} \lambda(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\mathcal{X}}) d\boldsymbol{\theta}$$

- Bayesian Estimation:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\, R(\boldsymbol{\theta}|\boldsymbol{\mathcal{X}})$$

Recall

For sample $\boldsymbol{x}$, how can we make the decision?

- Expected loss if we choose α_i:

$$R(\alpha_i|x) = E[\lambda(\alpha_i, \omega_j)|x] = \sum_{j=1}^{c} \lambda(\alpha_i, \omega_j) P(\omega_j|x), \quad i = 1, \cdots, k$$

- The expectation of loss for all samples under decision rule $\alpha(x)$:

$$R(\alpha) = \int R(\alpha(x)|x) p(x) dx$$

--- Expected risk, or average risk

$R(\alpha)$ is a functional: the function of decision function $\alpha(x)$

- Minimal Risk Decision: $\min_{\alpha(\cdot)} R(\alpha)$
- Decision Rule:

$$\text{If } R(\alpha_i|x) = \min_{j=1,\cdots,k} R(\alpha_j|x), \text{ then } \alpha = \alpha_i$$

- It can be proven that:

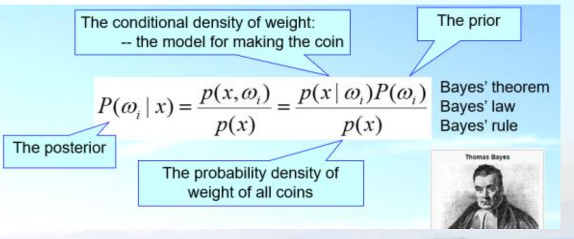
With lost function $\lambda(\hat{\theta},\theta) = \left(\theta - \hat{\theta}\right)^2$, the Bayesian estimate of θ given the sample x or the sample set $\boldsymbol{\mathcal{X}}$ is the conditional expectation of θ:

$$\hat{\theta} = E[\theta|x] = \int_{\Theta} \theta p(\theta|x) d\theta$$

or $$\hat{\theta} = E[\theta|\boldsymbol{\mathcal{X}}] = \int_{\Theta} \theta p(\theta|\boldsymbol{\mathcal{X}}) d\theta \ .$$

- Then the question is:
  - How to calculate the expectation with $p(\boldsymbol{\theta}|\boldsymbol{x})$ and $p(\boldsymbol{\theta}|\boldsymbol{\mathcal{X}})$ unknown?

Recall: How did we calculate $P(\omega_i|\boldsymbol{x})$?



- Method for Bayesian Estimation:

$P(\omega_i)$

① Set the prior $p(\theta)$

② Write down the joint density (conditional density):

$p(\boldsymbol{x}|\omega_i)$
$l(\theta)$

$$p(\boldsymbol{\mathcal{X}}|\theta) = p(x_1, x_2, \cdots, x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

③ Calculate the posterior

$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{\sum_j p(x|\omega_j)P(\omega_j)}$

$$p(\theta|\boldsymbol{\mathcal{X}}) = \frac{p(\boldsymbol{\mathcal{X}}|\theta)p(\theta)}{\int_{\Theta} p(\boldsymbol{\mathcal{X}}|\theta)p(\theta)d\theta}$$

④ Calculate the expectation:

$$\hat{\theta} = \int_{\Theta} \theta p(\theta|\boldsymbol{\mathcal{X}}) d\theta$$

D

- Relation with the maximum likelihood estimation?

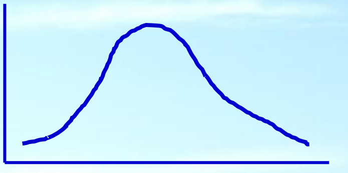
likelihood: $l(\theta) = p(\mathcal{X}|\theta) = p(x_1, x_2, \cdots, x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, l(\theta)$$

posterior：

$$p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta)p(\theta)}{\int_\Theta p(\mathcal{X}|\theta)p(\theta)d\theta}$$

$$\hat{\theta} = \int_\Theta \theta p(\theta|\mathcal{X})d\theta$$



- Step-wise (iterative) Bayesian Estimation

$$p(\theta|\mathcal{X}) = \frac{p(\mathcal{X}|\theta)p(\theta)}{\int_\Theta p(\mathcal{X}|\theta)p(\theta)d\theta}$$

$$p(\mathcal{X}^N|\theta) = p(x_N|\theta)p(\mathcal{X}^{N-1}|\theta)$$

$$p(\theta|\mathcal{X}^N) = \frac{p(x_N|\theta)p(\theta|\mathcal{X}^{N-1})}{\int_\Theta p(x_N|\theta)p(\theta|\mathcal{X}^{N-1})d\theta}$$

$$p(\theta|\mathcal{X}^0) = p(\theta)$$

$$p(\theta), p(\theta|x_1), p(\theta|x_1, x_2), \cdots$$

→ Bayesian Learning

Bayesian Estimation for Gaussian Distribution

- Simplest 1D case:
  - Data from $p(x|\mu) \sim N(\mu, \sigma^2)$ with μ unknown, σ^2 known
  - Set prior $p(\mu) \sim N(\mu_0, \sigma_0^2)$
  - The Bayesian estimation is:

$$\hat{\mu} = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

where $m_N = \frac{1}{N} \sum_{i=1}^{N} x_i$

➔Integrating **data** with **knowledge**

- The estimated density of μ:

(Data from $p(x|\mu) \sim N(\mu, \sigma^2)$ with μ unknown, σ^2 known)

$$p(\mu|\boldsymbol{\mathcal{X}}^N) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\left\{-\frac{1}{2}\left(\frac{\mu - \mu_N}{\sigma_N}\right)^2\right\} \sim N(\mu_N, \sigma_N^2)$$

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2 + \sigma^2} \mu_0$$

$$\sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}$$

- Q: Why did we want to estimate θ? ——To estimate the density of x
  - So we can estimate the density directly as

$$p(x|\boldsymbol{\mathcal{X}}) = \int_{\Theta} p(x|\theta)p(\theta|\boldsymbol{\mathcal{X}})d\theta$$

where

$$p(\theta|\boldsymbol{\mathcal{X}}) = \frac{p(\boldsymbol{\mathcal{X}}|\theta)p(\theta)}{\int_{\Theta} p(\boldsymbol{\mathcal{X}}|\theta)p(\theta)d\theta}$$

Note on the increase of variance due to the estimation.

- For the 1D Gaussian Distribution case:

$$p(x|\boldsymbol{\mathcal{X}}) = \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2+\sigma_N^2}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_N}{\sqrt{\sigma^2+\sigma_N^2}}\right)^2\right\} \sim N(\mu_N, \sigma^2+\sigma_N^2)$$

$$\mu_N = \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2} m_N + \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0, \quad \sigma_N^2 = \frac{\sigma_0^2\sigma^2}{N\sigma_0^2+\sigma^2}$$

- Why did we want to estimate θ? ——To estimate the density of x
- And, why did we wanted to estimate density?
  - To make decision based on the estimation

$$p(x|\omega_i) = \int_{\Theta} p(x|\theta_i)p(\theta_i|\boldsymbol{\mathcal{X}}_i)d\theta$$

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)}$$

Model → Prior → Model Estimation → Decision

Data

More Data

—— Bayesian Learning
Bayesian Inference

Bayesian Inference

$$\boldsymbol{P}(\boldsymbol{hypothesis}|\boldsymbol{data}) = \frac{\boldsymbol{P}(\boldsymbol{data}|\boldsymbol{hypothesis})\boldsymbol{P}(\boldsymbol{hypothesis})}{\sum_{\boldsymbol{h}} \boldsymbol{P}(\boldsymbol{h})\boldsymbol{P}(\boldsymbol{data}|\boldsymbol{h})}$$

- Bayes Rule

$$p(\theta|\boldsymbol{\mathcal{X}}) = \frac{p(\boldsymbol{\mathcal{X}}|\theta)p(\theta)}{\int_{\Theta} p(\boldsymbol{\mathcal{X}}|\theta)p(\theta)d\theta} \qquad P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{\sum_{j=1}^{2} p(x|\omega_j)P(\omega_j)}$$

$$\boldsymbol{P}(\boldsymbol{hypothesis}|\boldsymbol{data}) = \frac{\boldsymbol{P}(\boldsymbol{data}|\boldsymbol{hypothesis})\boldsymbol{P}(\boldsymbol{hypothesis})}{\sum_{\boldsymbol{h}} \boldsymbol{P}(\boldsymbol{h})\boldsymbol{P}(\boldsymbol{data}|\boldsymbol{h})}$$

- Machine Learning (from the view of Probabilistic Learning)
  - **Inference of hypothesis from data**
  - Bayesian learning (Bayesian Inference):
    - Estimate posteriors of all hypotheses based on given data
  - Maximum likelihood:
    - Find the hypothesis that best explains the data

1-minute break

11.4
Non-parametric estimation

Non-parametric Estimation

- Setting of the problem
  - Samples $\mathcal{X}$ are i.i.d. from density $p(x)$ of unknown form $p(x|\theta)$
  - Task: To estimate $p(x)$ from samples

~~Model → (Prior) →~~ Estimation → Decision

Data ↗

How to get the model?

To check whether data are normal

✓ Visual checking
- Histograms
- Boxplots
- QQ-plots

✓ Statistical tests
- χ^2 goodness-of-fit test
- Kolmogorov Smirnov test (KS test)
- Lilliefors test
- Shapiro Wilk test
- …

To make data normal

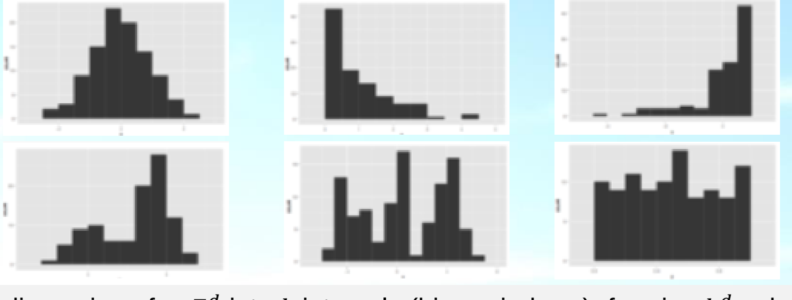
✓ Transformations, e.g.,
- log
- Box-Cox transformation
- …

If no good solution
→ abandon the model and do nonparametric estimation

11.4.1
Histograms

Histogram

- Graphical representation of a density
- Simplest non-parametric estimation of density



(1) Divide each dimension of $\boldsymbol{x} \epsilon E^d$ into k intervals (bins, windows), forming k^d cubes
(2) Count the number of samples q_i falling into cube i
(3) Estimate the probability density of the cube as $\frac{q_i}{NV}$, where N is the total number of samples, and V is the volume of the cube

Why histograms work?

Self-study

Task: Given sample set $\boldsymbol{\mathcal{X}} = \{x_1, \cdots, x_N\}$ i.i.d. sampled from unknown density $p(x)$, estimate a $\hat{p}(x)$ to approximate $p(x)$.

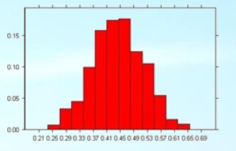
Solution:

The probability that sample x falls into region $\mathcal{R}$ is $P_{\mathcal{R}} = \int_{\mathcal{R}} p(x)dx$ if $p(x)$ is known.

If $p(x)$ is continuous and $\mathcal{R}$ is very small, and the volume of $\mathcal{R}$ is V, we can approximate $P_{\mathcal{R}}$ as $P_{\mathcal{R}} = \int_{\mathcal{R}} p(x)dx = p(x)V,\ x \in \mathcal{R}$.

- The probability that k samples in $\boldsymbol{\mathcal{X}}$ fall in the region is

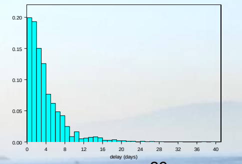
$$P_k = \binom{k}{N} P_{\mathcal{R}}^k (1 - P_{\mathcal{R}})^{N-k}$$



- The expectation of k is $E[k] = NP_{\mathcal{R}}$
- The estimation of $P_{\mathcal{R}}$ is $\hat{P}_{\mathcal{R}} = k/N$ when k samples are observed in $\mathcal{R}$.

Therefore, we have the estimate

$$\hat{p}(x) = \frac{k}{NV}$$



Self-study

So, $\hat{p}(x) = \frac{k/N}{V}$ if $\mathcal{R}$ is very small.

- What is "very small"? How do we choose the size of the small region?

Theoretical result:

If we have an unlimited number of samples, suppose we form a sequence of regions $\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_n, \cdots$ containing x to estimate the density. $\mathcal{R}_1$ contains k_1 samples, $\mathcal{R}_2$ contains $k_2, \ldots, \mathcal{R}_n$ contains k_n samples. Let the volume of $\mathcal{R}_n$ be V_n. $\hat{p}_n(x) = \frac{k_n}{NV_n}$ is the nth estimate of $p(x)$.

If $\hat{p}_n(x)$ is to converge to $p(x)$, three conditions are required:

$$(1)\ \lim_{n\to\infty} V_n = 0,\ (2)\ \lim_{n\to\infty} k_n = \infty,\ (3) \lim_{n\to\infty} \frac{k_n}{N} = 0.$$

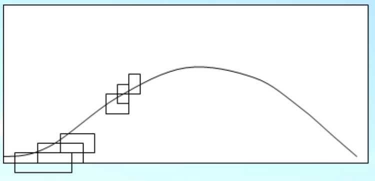
- Two ways to obtain this:
  - To specify the volume V_n → Parzen-window method
  - To enclose k_n samples → k_n-nearest neighbor method

Duda et al, *Pattern Classification* (2nd edition), p.163-164, 2006

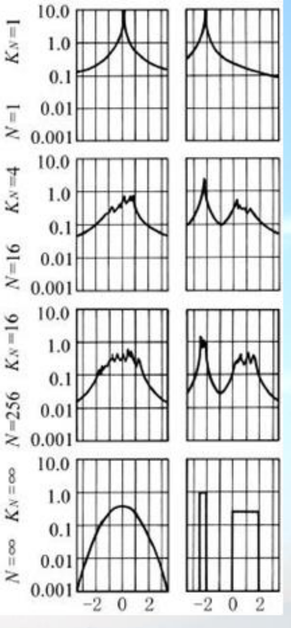
11.4.2
k_n-nearest neighbor method

k_n-nearest neighbor method

$$\hat{p}_n(x) = \frac{k_n/N}{V_n}$$



- Conditions $\lim_{n\to\infty} k_n = \infty$ and $\lim_{n\to\infty} k_n/n = 0$ are necessary and sufficient for $\hat{p}_n(x)$ to converge to $p(x)$ in probability at all points where $p(x)$ is continuous.
- A typical choice: $k = \sqrt{N}$



11.4.3
Parzen-window method

Parzen-window method
aka. Parzen-Rosenblatt Window, Kernel density estimation

- Assume region $\mathcal{R}_n$ is a d-dimensional hypercube with length h_n of each edge

$$V_n = h_n^d$$

- Define a window function

$$\varphi(u) = \begin{cases} 1 & |u_j| \leq 1/2 \quad j = 1, \ldots, d \\ 0 & \text{otherwise} \end{cases}$$

- The number of samples in hypercube centered at x is

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

- Therefore,

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

- The window $\varphi(u)$ can be defined in other forms as long as it satisfies

$$\varphi(u) \geq 0, \qquad \int \varphi(u)du = 1$$

$$\sup_u \varphi(u) < \infty$$

$$\lim_{\|u\|\to\infty} \varphi(u) \prod_{i=i}^{d} u_i = 0$$

- As a density
- Bounded
- Quickly shrinks to zero

- For a density function continuous at x, the following two extra conditions will assure the convergence of the estimate to $p(x)$:

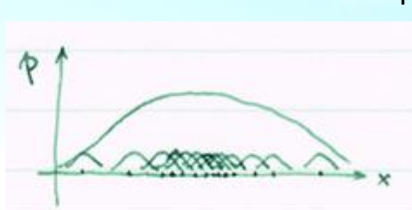
$$\lim_{n\to\infty} V_n = 0$$

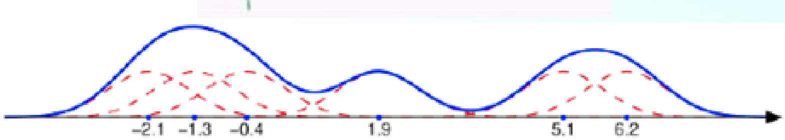
$$\lim_{n\to\infty} nV_n = \infty$$

- Parzen-window estimation can be written in the form of

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} k(\mathrm{x}, \mathrm{x}_i)$$

- to use a kernel $k(x, x_i)$ to control the small region
- $k(x, x_i)$ measures the contribution of sample x_i to $p(x)$,





- Popular kernels (window functions)
  - Hypercube

$$k(x,x_i)=\begin{cases}\frac{1}{h^d} & \text{if } \left|x^i-x_i^j\right|\le h/2, j=1,2,\cdots,d\\ 0 & \text{otherwise}\end{cases}$$

  - Gaussian

$$k(x,x_i)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(x-x_i)^2\right\}$$

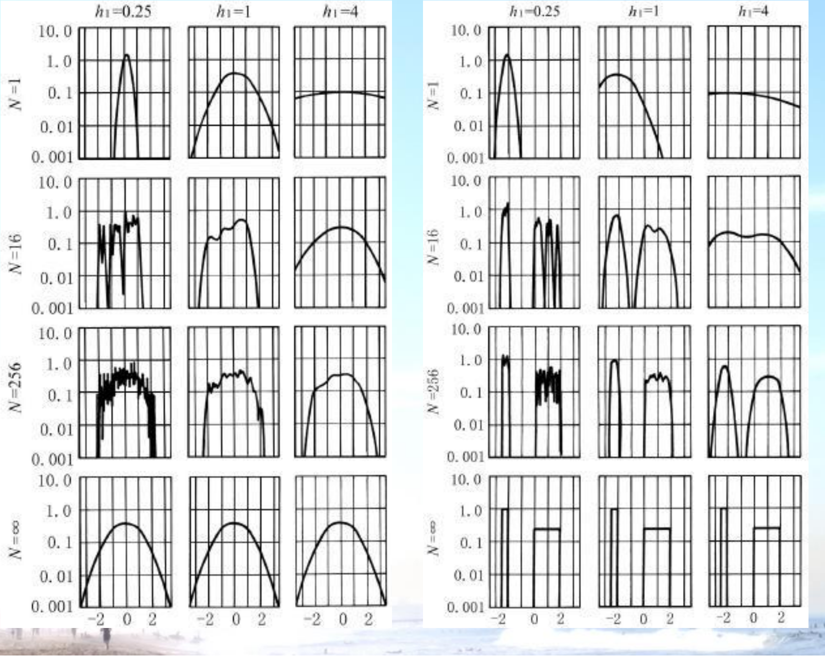
$$k(x,x_i)=\frac{1}{\sqrt{(2\pi)^d\rho^{2d}|Q|}}\exp\left\{-\frac{1}{2}\frac{(x-x_i)^TQ^{-1}(x-x_i)}{\rho^2}\right\}$$

  - Hypersphere

$$k(x,x_i)=\begin{cases}V^{-1} & \text{if } ||x-x_i||\le\rho\\ 0 & \text{otherwise}\end{cases}$$

- Choice of the window size
  - Large for small sample size; Small for large sample size. E.g., $\rho=N^{-\frac{\eta}{d}}$, $\eta\in(0,1)$

Examples



Homework

- Problem Set (Pr.6a)
  - Study methods for checking normal distribution, and for making data normal.
- Deadline:
  - Nov. 17 (Wednesday), 23:00

D

单选题 1分 设置

4-min break

A Yes, I'm back to my computer.

B Sorry, not yet.

提交