北京交通大学硕士研究生《机器学习》课件

第3章 密度估计

桃李不言，下自成蹊。

——《史记·李将军列传论》

北京交通大学《机器学习》课程组

提要

1. 预备知识
2. 密度估计应用举例
3. 经典参数估计
4. 从机器学习公理出发的参数估计
5. 密度函数的非参数估计

- 条件概率、全概率、贝叶斯公式
- 密度函数
- 正态分布
- 均值、方差

预备知识

条件概率

①条件概率的定义．

定义 1.3 设 A、B 是任意两个事件且 $P(B)>0$，在事件 B 已发生的条件下，事件 A 发生的条件概率 $P(A|B)$ 定义为

$$P(A|B)=\frac{P(AB)}{P(B)}. \tag{1.12}$$

P(Height|男)

P(Height|女)

预备知识

全概率公式

(3)全概率公式．

若事件组 B_1、B_2、$\cdots$、B_n 满足

①B_1、B_2、$\cdots$、B_n 互不相容且 $P(B_i)>0 \quad (i=1,2,\cdots,n)$，

②$B_1 \cup B_2 \cup \cdots \cup B_n = S$，

则对任意事件 A，有

$$P(A)=\sum_{i=1}^{n} P(B_i)P(A \mid B_i).$$

式(1.22)称为全概率公式．

P(Height)

贝叶斯公式

(4)贝叶斯(Bayes)公式.

若 B_1、B_2、…、B_n 是一完备事件组,则对任意事件 $A(P(A)>0)$,有

$$P(B_j|A)=\frac{P(B_j)P(A|B_j)}{\sum_{i=1}^{n}P(B_i)P(A|B_i)}\qquad j=1,2,\cdots,n$$

式(1.23)称为贝叶斯公式.

P(男|Height)

P(女|Height)

概率密度函数

①设随机变量 X 的分布函数为 $F(x)$，若存在可积函数 $f(x)\geq 0$，使得对任意实数 x 都有：

$$F(x)=\int_{-\infty}^{x} f(t)\mathrm{d}t \tag{2.16}$$

则称 X 为连续型随机变量，而称 $f(x)$ 为 X 的概率(分布)密度函数．

②概率密度函数 $f(x)$ 有下述基本性质：

1° $f(x)\geq 0$ (2.17)

2° $\int_{-\infty}^{+\infty} f(x)\mathrm{d}x=1$. (2.18)

这两条是密度函数的特征性质．

3° 若 $x_1<x_2$，则

$$P\{x_1<X\leq x_2\}=\int_{x_1}^{x_2} f(x)\mathrm{d}x=F(x_2)-F(x_1). \tag{2.19}$$

正态分布

3° 正态分布：

若随机变量 X 的概率密度为

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\qquad -\infty<x<+\infty \tag{2.24}$$

其中 σ,μ 为常数，且 $\sigma>0$，则称 X 服从参数为 μ,σ 的正态分布，记为 $X\sim N(\mu,\sigma^2)$.

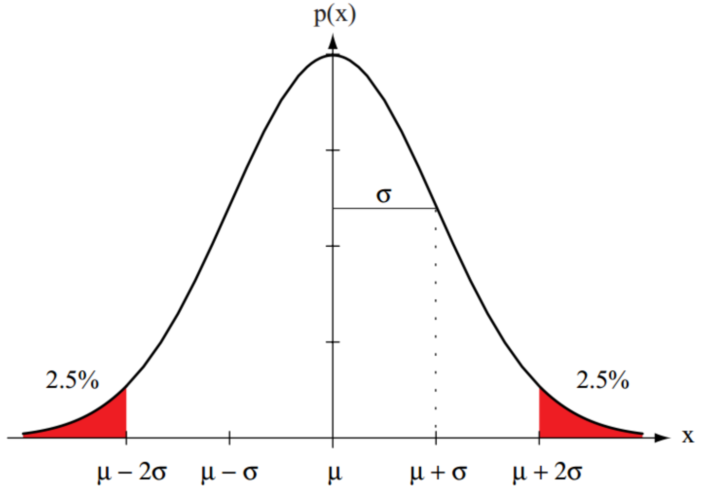


Figure 2.7: A univariate normal distribution has roughly 95% of its area in the range $|x-\mu|\le 2\sigma$, as shown. The peak of the distribution has value $p(\mu)=1/\sqrt{2\pi}\sigma$.

预备知识

数学期望

②设连续型随机变量 X 的概率密度为 $f(x)$，

若 $\int_{-\infty}^{\infty} x f(x)\mathrm{d}x$ 绝对收敛，则称它为 X 的数学期望，记为 $E(X)$，即

$$E(X) = \int_{-\infty}^{\infty} x f(x)\mathrm{d}x.$$

此时，称 X 的数学期望存在，或称 X 有有限的数学期望．

方差

①设随机变量 X 的数学期望 $E(X)$ 存在，若 $E[X-E(X)]^2$ 存在，则称它为 X 的方差，记为 $D(X)$ 或 $\mathrm{Var}(X)$，即

$$D(X)=E[X-E(X)]^2. \tag{4.16}$$

设连续型随机变量 X 的概率密度为 $f(x)$，其方差存在，根据(4.13)式则有

$$D(X)=\int_{-\infty}^{\infty}(x-E(X))^2 f(x)\,\mathrm{d}x.$$

密度估计

■ 密度估计问题：

已知n个观测值$x_1, x_2, \cdots, x_N$，且服从某个密度分布 $\mathbf{p}(x)$（未知），假设学到的密度函数为$\widehat{p(x)}$

■ 密度估计问题分类：

- 若知道p(x)的部分信息，比如p(x)属于某个概率分布（高斯），计算$\widehat{p(x)}$就是参数估计。
- 若除样本集X外，对p(x)一无所知，此时计算$\widehat{p(x)}$就是非参数估计。

内容提要

1. 预备知识
2. 密度估计应用举例
3. 经典密度估计
4. 从机器学习公理出发的密度估计
5. 密度函数的非参数估计

密度估计应用举例

$$P(男|Height) = \frac{p(height|男)P(男)}{p(Height)}$$

贝叶斯公式

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

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2,

Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2.

密度估计应用举例

正态密度： $P(\omega_1) = P(\omega_2)$

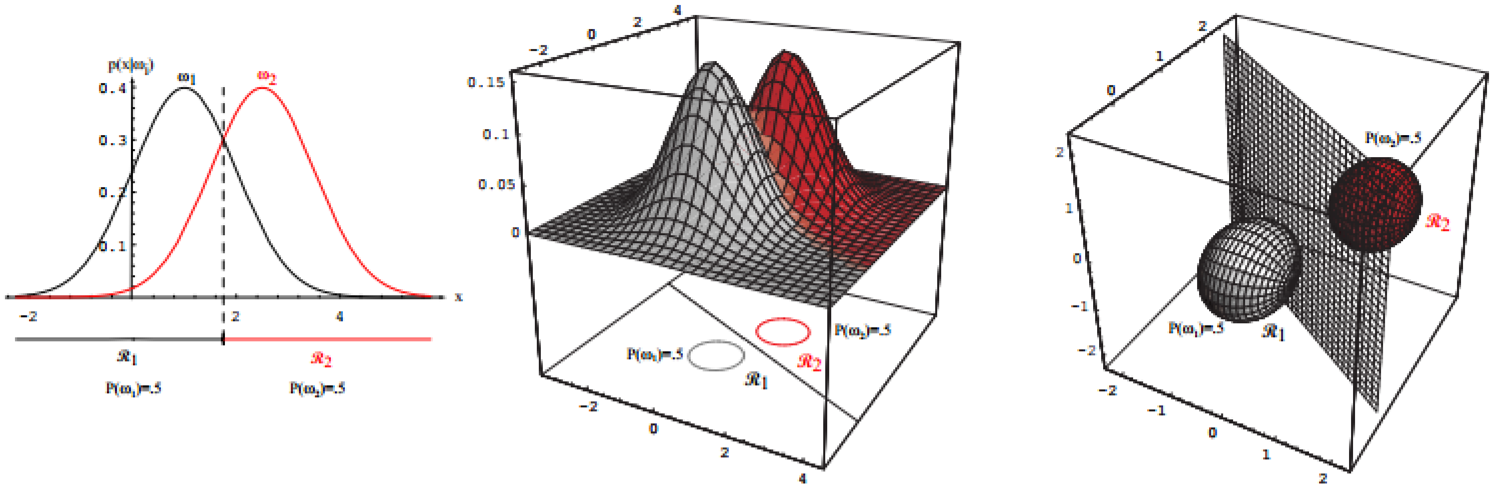


Figure 2.10: If the covariances of two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these 1-, 2-, and 3-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the 3-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$.

密度估计应用举例

正态密度：

$$P(\omega_i) \neq P(\omega_j)$$

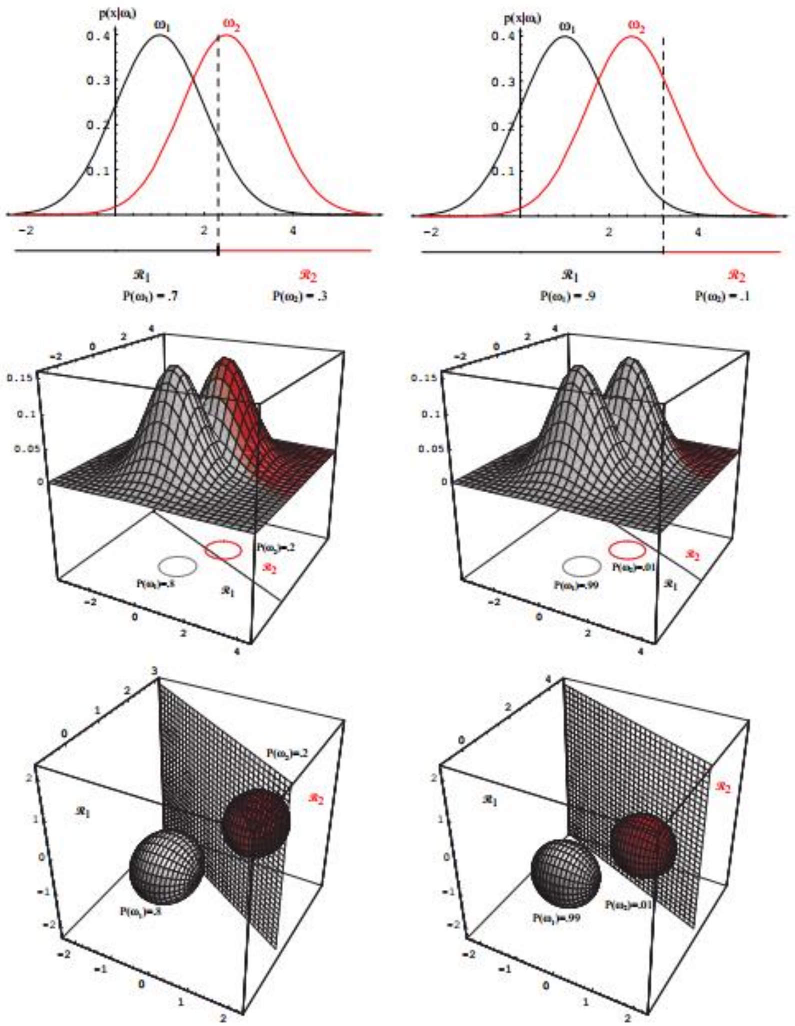


Figure 2.11: As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these 1-, 2- and 3-dimensional spherical Gaussian distributions.

密度估计应用举例

正态密度：一元情况

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)\, dx,$$

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx.$$

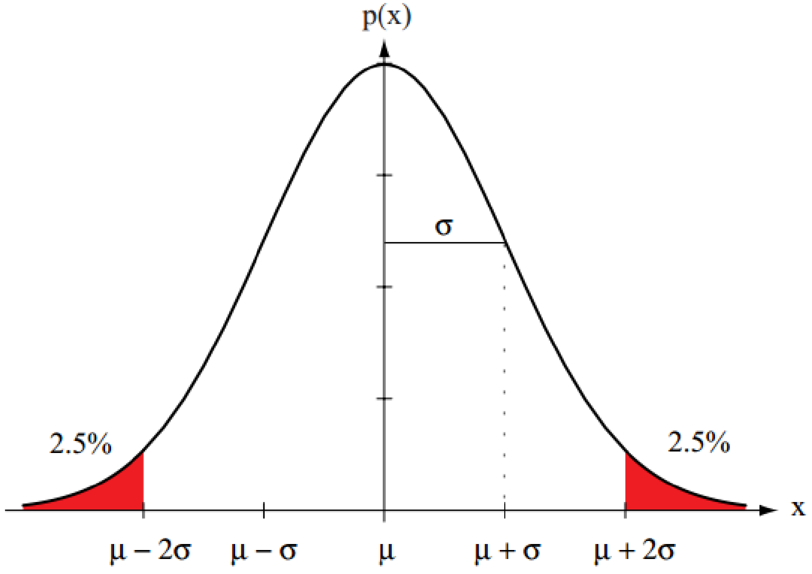


Figure 2.7: A univariate normal distribution has roughly 95% of its area in the range $|x-\mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$.

密度估计应用举例

正态密度：多元情况

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x}p(\mathbf{x})\, d\mathbf{x}$$

$$\boldsymbol{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})\, d\mathbf{x},$$

i是什么？

$$\mu_i = \mathcal{E}[x_i]$$

$$\sigma_{ij} = \mathcal{E}[(x_i-\mu_i)(x_j-\mu_j)].$$

密度估计应用举例

参数估计问题

Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2.

什么关系？

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} x p(x)\, dx,$$

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\, dx.$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x})\, d\mathbf{x}$$

$$\boldsymbol{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})\, d\mathbf{x},$$

内容提要

1. 预备知识

2. 贝叶斯决策论

3. 经典密度估计

   - 最大似然估计
   - 贝叶斯估计

4. 从机器学习公理出发的密度估计

5. 密度函数的非参数估计

最大似然估计

问题：

$$\text{Decide } \omega_1 \text{ if } p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2); \qquad \text{otherwise decide } \omega_2.$$

基于w1类样本估计 $p(x|\omega_1)$.

基于w2类样本估计 $p(x|\omega_2)$

最大似然估计

主要思想：

从参数为q（未知）的总体中独立抽取N个样本 x={$\mathbf{x}_1$,$\mathbf{x}_2$,…$\mathbf{x}_N$}

用**似然函数**表征抽中这N个样本的可能性：

$$l(\theta) = p(x|\theta) = p(x_1, x_2, \cdots x_N|\theta)$$

最大似然估计

最大似然假设：

最好的q应该是使得 p(x|q)达到极大值的q 。

独立同分布
假设

样本是独立从该类抽取的

$$\boldsymbol{l}(\theta)=\boldsymbol{p}(\boldsymbol{x}_1,\boldsymbol{x}_2\Lambda\ \boldsymbol{x}_N|\theta)=\boldsymbol{p}(\boldsymbol{x}_1|\theta)\,\boldsymbol{p}(\boldsymbol{x}_2|\theta)\Lambda\ \boldsymbol{p}(\boldsymbol{x}_N|\theta)$$

$$\ln(l(\theta))=\ln\left(\prod_{k=1}^{N}p(x_k|\theta)\right)=\sum_{k=1}^{N}\ln p(x_k|\theta)$$

最大似然估计

一个简单的例子

设一维样本服从正态分布$p(x|\theta) \sim N(\mu,\sigma^2)$，方差已知

通过抽出的样本集$x=\{0.1,0.2,L\ ,0.6\}$用极大似然法估计μ。

此时$\theta=\mu$。当θ自左向右取不同值时，计算x的概密：

$$p(x|\theta)=\prod_{k=1}^{6} p(x_k|\theta)$$

$p(x|\theta)$有不同值：

θ=0? θ=0.2? θ=0.35?

θ=0.35时$p(x|\theta)$达极大, 对应$\hat{\theta}$实际为均值；

最大似然估计

正态分布参数的极大似然估计

Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2.

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

最大似然估计

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\mathbf{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})$$

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}-\mathbf{s})^T\mathbf{W}(\mathbf{x}-\mathbf{s}) = 2\mathbf{W}(\mathbf{x}-\mathbf{s})$$

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}_k|\boldsymbol{\mu}) = \mathbf{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu}).$$

$$\ln(l(\theta)) = \sum_{k=1}^{N}\ln p(x_k|\theta)$$

$$\sum_{k=1}^{n}\mathbf{\Sigma}^{-1}(\mathbf{x}_k-\hat{\boldsymbol{\mu}}) = \mathbf{0},$$

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

最大似然估计

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\ln p(\mathbf{x}_k|\boldsymbol{\mu}) = -\frac{1}{2}\ln\left[(2\pi)^d|\mathbf{\Sigma}|\right] - \frac{1}{2}(\mathbf{x}_k-\boldsymbol{\mu})^t\mathbf{\Sigma}^{-1}(\mathbf{x}_k-\boldsymbol{\mu})$$

$$\frac{\partial \ln|\det(\mathbf{X})|}{\partial \mathbf{X}} = (\mathbf{X}^{-1})^T = (\mathbf{X}^T)^{-1}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^{-1}\mathbf{b}}{\partial \mathbf{X}} = -\mathbf{X}^{-T}\mathbf{a}\mathbf{b}^T\mathbf{X}^{-T}$$

$$(\mathbf{A}^T)^{-1} = (\mathbf{A}^{-1})^T$$

?

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k-\hat{\boldsymbol{\mu}})(\mathbf{x}_k-\hat{\boldsymbol{\mu}})^t$$

最大似然估计

应用：

两类训练样本

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k.$$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{k=1}^{n}(\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^t$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2.

贝叶斯估计

有时基于历史经验，不仅知道分布形式，甚至对θ的信息有所了解

- 柯洁　　围棋比赛成绩
- 烟台　　苹果
- 朋友交往　　第一印象
- 声誉　　对于事物的先验印象

贝叶斯估计

有时基于历史经验，**不仅知道分布形式，甚至对θ的信息有所了解**

- 对θ信息有所了解，但会随着观察的积累增多而改变，具有不确定性。

  $\theta \sim \boldsymbol{p(\theta|\theta_0)}$：对$\theta$信息的先验了解程度，$\theta_0$是事先确定的值；

  反映θ与固定值θ_0的相似度，即$Sim(\theta,\theta_0)=p(\theta|\theta_0)$。

  - 理论上，应选择与θ_0最相似的θ值。
  - 若无限相似，即变成信仰，观察改变不了θ的估计；
  - 若不是无限相似，则观察可改变对于θ的估计。

贝叶斯估计

基本原理：把参数q当作具有某种先验分布$p(q)$ 的随机变量, 对样本$\mathbf{x}$观测值将先验分布$p(\mathbf{x}|q)$转化为后验分布$p(q|\mathbf{x})$，据此再修正原先的估计：

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

贝叶斯估计

Bayes参数估计思路：

①确定q的先验概率密度函数p(q);

独立同分布
假设

由样本集**x** = {x_1,x_2,…,x_N }计算样本的联合分布

$$p(\mathrm{x}\,|\,\theta) = p(x_1, x_2, \ldots, x_N\,|\,\theta) = \prod_{k=1}^{N} p(x_k\,|\,\theta)$$

② 用Bayes公式求后验分布p(q |**x**)

$$p(\theta\,|\,x) = \frac{p(x\,|\,\theta)\,p(\theta)}{\int p(x\,|\,\theta)\,p(\theta)\,d\theta}$$

③求样本的估计量q:

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

贝叶斯估计

Bayes估计：一维正态分布

①样本为一维正态分布 $p(x|\theta)\sim N(\mu,\sigma^2)$，$\mu$未知

②μ是随机的，其先验概密 $p(\theta)\sim N(\mu_0,\sigma_0^2)$

③N个样本构成样本集 $x=\{x_1, x_2, \dots x_N\}$

求μ的估计量

贝叶斯估计

Bayes参数估计步骤:

①确定θ的先验概率密度函数p(θ)；

②由样本集$\boldsymbol{x}$ = {$x_1, x_2, \cdots, x_N$}计算样本的联合分布

$$p(\mathrm{x}|\theta)=p(x_1,x_2,\ldots,x_N|\theta)=\prod_{k=1}^{N}p(x_k|\theta)$$

③用Bayes公式求后验分布$p(\theta/\boldsymbol{x})$

$$p(\theta|x)=\frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$$

④求样本的估计量θ: $\hat{\theta}=E[\theta|x]=\int_{\Theta}\theta p(\theta|x)d\theta$

$$\hat{\theta}=\int\theta p(\theta|x)d\theta$$

$$p(\theta|x)=\frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta}$$

$$a=1/\int p(x|\theta)p(\theta)d\theta$$

$$p(\theta|x)=a\left\{\prod_{k=1}^{N}p(x_k|\theta)\right\}p(\theta)$$

贝叶斯估计

$$p(\theta \mid x) = a\left\{\prod_{k=1}^{N} p(x_k \mid \theta)\right\} p(\theta)$$

$p(x|\theta) \sim N(\mu, \sigma^2) \quad p(\theta) \sim N(\mu_0, \sigma_0^2)$

$$p(\theta|x) = a\left\{\prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left[-\frac{1}{2}\cdot\frac{(x_k-\mu)^2}{\sigma^2}\right]\right\}\cdot\frac{1}{\sqrt{2\pi}\cdot\sigma_0}\exp\left[-\frac{1}{2}\cdot\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right]$$

$$-\ln\left(p(\theta|x)\right) = a' - N\ln\left(\frac{1}{\sigma}\right) + \sum_{k=1}^{N}\frac{(x_k-\mu)^2}{2\sigma^2} + \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}$$

贝叶斯估计

$$L= -\ln\big(p(\mu|x)\big) = a' - N\ln(\tfrac{1}{\sigma})) + \sum_{k=1}^{N} \frac{(x_k-\mu)^2}{2\sigma^2} + \frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}$$

$$\begin{cases} \dfrac{\partial L}{\partial \mu} = \dfrac{\mu - \mu_0}{\sigma_0^2} - \displaystyle\sum_{k=1}^{N}\left(\frac{x_k-\mu}{\sigma^2}\right) \\ \dfrac{\partial L}{\partial \sigma} = N\sigma^{-1} - \displaystyle\sum_{k=1}^{N}(x_k-\mu)^2\,\sigma^{-3} \end{cases}$$

$$\begin{cases} \mu = \dfrac{\dfrac{\mu_0}{N} + \dfrac{\sigma_0^2}{\sigma^2}\dfrac{\sum_{k=1}^{N} x_k}{N}}{\dfrac{1}{N} + \dfrac{\sigma_0^2}{\sigma^2}} \\ \sigma^2 = \displaystyle\sum_{k=1}^{N}\frac{(x_k-\mu)^2}{N} \end{cases}$$

贝叶斯估计

应用：

$$\mu = \frac{\frac{\mu_0}{N} + \frac{\sigma_0^2}{\sigma^2}\frac{\sum_{k=1}^{N} x_k}{N}}{\frac{1}{N} + \frac{\sigma_0^2}{\sigma^2}}$$

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

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^t \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2.

内容提要

1. 预备知识
2. 密度估计应用举例
3. 经典参数估计
4. 从机器学习公理出发的参数估计
5. 密度函数的非参数估计

- 机器学习公理&密度估计
- 最大似然估计
- 贝叶斯估计

知识点回顾

- **类一致性准则**

  如果类表示唯一性公理不成立，一个好的归类结果应该使类表示唯一性公理在**逼近意义下尽可能成立**。

- **类紧致性准则**

  归类方法应该使其归类结果尽可能紧致。即每个对象的最相似类与其次相似类的相似度差别要大；表现为**最大化类内相似度或最小化类内方差**。

- **类分离准则**

  归类方法应该使得**类间的距离最大**。

- **奥卡姆剃刀准则**

  如非必要，勿增实体

从机器学习公理的角度看密度估计

X n个观测值$x_1, x_2, \cdots, x_N$	Y n个观测值$x_1, x_2, \cdots, x_N$
U $U = [1,1,\cdots,1]^T_{1\times N}$	V V$= [1,1,\cdots,1]^T_{1\times N}$
$\underline{\text{X}}$ $p(x)$	$\underline{\text{Y}}$ $\widehat{p(x)}$
Simx ?	Simy ?

类相似性映射

$X=\{x_1,x_2,\cdots,x_n\}$分成c类$\underline{X}_1,\underline{X}_2,\cdots,\underline{X}_c$，如果：

函数 $Sim_X(x_k,\underline{X}_i)$ 值增加表示 x_k 和 $\underline{X}_i$ 的相似性增大

函数 $Sim_X(x_k,\underline{X}_i)$ 值减小表示 x_k和 $\underline{X}_i$ 的相似性减小

$Sim_X: X\times\{\underline{X}_1,\underline{X}_2,\cdots,\underline{X}_c\}\mapsto R_+$ 是类相似性映射

假设输入类表示 $\underline{X}=\theta$：

$$Sim_X(x,\theta)=p(x|\theta) \quad ?$$

$$Sim_X(x,\theta)=r(x,\theta) \quad ?$$

$$Sim_X(x,\theta)=\frac{1}{d(x,\theta)} \quad ?$$

类紧致性准则

类紧致性准则

归类方法应该使其归类结果尽可能紧致。即每个对象的最相似类与其次相似类的相似度差别要大；表现为**最大化类内相似度或最小化类内方差**。

- $\max\limits_{\hat{\theta}} \prod_{k=1}^{N} Sim_Y\,(x_k, \hat{\theta})$
- $\max\limits_{\widehat{\theta}} \sum_{k=1}^{N} Sim_Y(x_k, \widehat{\theta})$

- $\min\limits_{\widehat{\theta}} \prod_{k=1}^{N} D_Y\,(x_k, \hat{\theta})$
- $\min\limits_{\widehat{\theta}} \sum_{k=1}^{N} D_Y(x_k, \widehat{\theta})$

最大似然估计

若已知p(x)所在的分布族$p(x|\theta)$，此时密度估计问题就变成**估计θ**。

此时，输入类表示 $\underline{X} = \theta$，相似性映射 $Sim_X(x,\theta) = p(x|\theta)$

假设对θ得到估计$\hat{\theta}$，则可设输出类表示 $\underline{Y} = \hat{\theta}$，相似性映射为 $Sim_Y(x,\hat{\theta}) = p(x|\hat{\theta})$

无需
独立同分布
假设

根据类紧致准则（希望最大类内相似度），得目标函数：

$$\max_{\hat{\theta}} \prod_{k=1}^{N} Sim_Y\left(x_k,\hat{\theta}\right) = \max_{\hat{\theta}} \prod_{k=1}^{N} p\left(x_k|\hat{\theta}\right) \qquad (3.1)$$

$$\min_{\hat{\theta}} \sum_{k=1}^{N} -\ln\left(Sim_Y\left(x_k,\hat{\theta}\right)\right) = \min_{\hat{\theta}} \sum_{k=1}^{N} -\ln\left(p\left(x_k|\hat{\theta}\right)\right) \qquad (3.2)$$

最大似然估计

例-高斯密度估计：假设N个R^p的样本点x_k服从高斯分布，试根据现有样本点估计高斯分布的具体参数。

假设$\forall k, x_k \in R^p, x \in R^p$,

$$p\left(x|\hat{\theta}\right) = \frac{1}{\sqrt{2\pi^p \hat{\sigma}^{2p}}} \exp[-\frac{1}{2}\frac{(x-\hat{\mu})^T(x-\hat{\mu})}{\hat{\sigma}^{2p}}]$$，其中$\hat{\theta}=\{\hat{\mu}，\hat{\sigma}^{2p}\}$

由公式（3.2），得如下**目标函数**:

$$\min \quad L = \sum_{k=1}^{N} -\ln(p\left(x_k|\hat{\theta}\right) = \sum_{k=1}^{N}\left(\frac{1}{2}(\frac{||x_k-\hat{\mu}||}{\hat{\sigma}^p})^2+\ln(\sqrt{(2\pi)^p\hat{\sigma}^{2p}})\right) \quad (3.3)$$

如何求参数的估计值，使得（3.3）最小?

最大似然估计

计算目标函数（3.3）的一阶导数，令其等于零可得到最优估计：

$$\begin{cases} \dfrac{\partial L}{\partial \hat{\mu}} = -\sum_{k=1}^{N} (\dfrac{x_k - \hat{\mu}}{\hat{\sigma}^{2p}}) = 0 \\ \dfrac{\partial L}{\partial \hat{\sigma}} = -p \sum_{k=1}^{N} ||x_k - \hat{\mu}||^2 \, \hat{\sigma}^{-2p-1} + Np\hat{\sigma}^{-1} = 0 \end{cases}$$

$$\boxed{L = \sum_{k=1}^{N} \left(\frac{1}{2} (\frac{||x_k - \hat{\mu}||}{\hat{\sigma}^p})^2 + \ln(\sqrt{(2\pi)^p \hat{\sigma}^{2p}}) \right)}$$

$$\Downarrow$$

$$\begin{cases} \hat{\mu} = \sum_{k=1}^{N} \dfrac{x_k}{N} \\ \hat{\sigma}^{2p} = \sum_{k=1}^{N} \dfrac{||x - \hat{\mu}||^2}{N} \end{cases} \qquad (3.5)$$

思考：

令$p\left(x|\hat{\theta}\right) = \frac{1}{\sqrt{(2\pi)^p \det(\hat{\Sigma})}} \exp[-\frac{1}{2} \frac{(x-\hat{\mu})^T (x-\hat{\mu})}{\hat{\Sigma}}]$，其中$\hat{\theta}=\{\hat{\mu}，\hat{\Sigma}\}$。如何估计$\hat{\theta}$?

贝叶斯估计

假设对θ得到估计$\hat{\theta}$，根据以上分析：

$$\underline{X} = \theta,\quad \underline{Y} = \hat{\theta}$$

$$Sim_Y(x,\hat{\theta}) = p(x|\hat{\theta}),\quad Sim(\hat{\theta},\theta_0) = p(\hat{\theta}|\theta_0)$$

无需
独立同分布
假设

- 根据 类紧致准则（希望最大类内相似度）：

$$\max_{\hat{\theta}} \prod_{k=1}^{N} Sim_Y\left(x_k,\hat{\theta}\right) = \max_{\hat{\theta}} \prod_{k=1}^{N} p\left(x_k|\hat{\theta}\right) \quad (3.1)$$

- 根据 类一致性准则：

$$\max_{\hat{\theta}} Sim\left(\hat{\theta},\theta_0\right) = \max_{\hat{\theta}} p(\hat{\theta}|\theta_0) \quad (3.10)$$

综合两准则，应最大化目标函数：

$$Sim(\hat{\theta},\theta_0) \prod_{k=1}^{N} Sim_Y\left(x_k,\hat{\theta}\right) = p(\hat{\theta}|\theta_0) \prod_{k=1}^{N} p\left(x_k|\hat{\theta}\right) \quad (3.11)$$

贝叶斯估计

高斯密度的贝叶斯估计：

假设$\forall k, x_k \in R^p, x \in R^p$,

$p\left(x|\hat{\theta}\right) = \frac{1}{\sqrt{2\pi^p \hat{\sigma}^{2p}}} \exp[-\frac{1}{2}\frac{(x-\hat{\mu})^T(x-\hat{\mu})}{\hat{\sigma}^{2p}}]$，其中$\hat{\theta}=\{\hat{\mu}, \hat{\sigma}^{2p}\}$；

$Sim(\hat{\theta}, \theta_0) = p(\hat{\theta}|\theta_0) = \frac{1}{\sqrt{2\pi^p \sigma_0^{2p}}} \exp[-\frac{1}{2}\frac{(\mu_0-\hat{\mu})^T(\mu_0-\hat{\mu})}{\sigma_0^{2p}}]$，其中$\theta_0=\{\mu_0, \sigma_0^{2p}\}$

根据公式（3.11），应该**最小化目标函数**：

$$L = -\ln\left(p(\hat{\theta}|\theta_0)\right) + \sum_{k=1}^{N} -\ln(p\,(x_k|\hat{\theta}))$$

$$= -\ln\left(\frac{1}{\sqrt{(2\pi)^p \hat{\sigma}_0^{2p}}}\right) + \frac{1}{2}\frac{(\mu_0-\hat{\mu})^T(\mu_0-\hat{\mu})}{\sigma_0^{2p}} + \sum_{k=1}^{N}(\frac{1}{2}\left(\frac{||x_k-\hat{\mu}||}{\hat{\sigma}^{2p}}\right)^2 + \ln(\sqrt{(2\pi)^p \hat{\sigma}^{2p}}))$$

贝叶斯估计

计算上式的一阶导数，令其等于零可得到最优估计$\hat{\theta}$：

$$\begin{cases} \dfrac{\partial L}{\partial \hat{\mu}} = -\dfrac{\mu_0 - \hat{\mu}}{\sigma_0^{2p}} - \displaystyle\sum_{k=1}^{N}\left(\dfrac{x_k - \hat{\mu}}{\hat{\sigma}^{2p}}\right) = 0 \\ \dfrac{\partial L}{\partial \hat{\sigma}} = -p\displaystyle\sum_{k=1}^{N} ||x_k - \hat{\mu}||^2\, \hat{\sigma}^{-2p-1} + Np\hat{\sigma}^{-1} = 0 \end{cases} \tag{3.13}$$

$$\begin{cases} \hat{\mu} = \dfrac{\dfrac{\mu_0}{N} + \dfrac{\sigma_0^{2p}}{\hat{\sigma}^{2p}}\dfrac{\sum_{k=1}^{N} x_k}{N}}{\dfrac{1}{N} + \dfrac{\sigma_0^{2p}}{\hat{\sigma}^{2p}}} \\ \hat{\sigma}^{2p} = \displaystyle\sum_{k=1}^{N} \dfrac{||x - \mu||^2}{N} \end{cases} \tag{3.14}$$

内容提要

1. 预备知识

2. 密度估计应用举例

3. 经典参数估计

4. 从机器学习公理出发的参数估计

5. 密度函数的非参数估计

- 直方图
- 核密度估计
- K近邻密度估计法

密度估计的非参数方法

除观测样本$x_1, x_2, \cdots, x_N$以外，若对p(x)一无所知但却需要估计p(x)，如何做?

最大似然估计和贝叶斯估计无法适用

非参数估计方法!

直方图

- 基本思想：利用极限思想，将空间划分成合适的区域，通过统计区域内的密度来得到$\widehat{p(x)}$。

- 计算过程：假设x所在区域含有l_x个观测样本，区域体积为V 。
  对于空间中的任意一点x，若其位于V_l区域内，得密度估计：

$$\widehat{p(x)} = \frac{l_x}{N \times V} \qquad (3.19)$$

直方图

优选方案：增加样本

区域体积小

同时保证区域内有充分多的样本

每个区域的样本是总样本数的很小的一部分

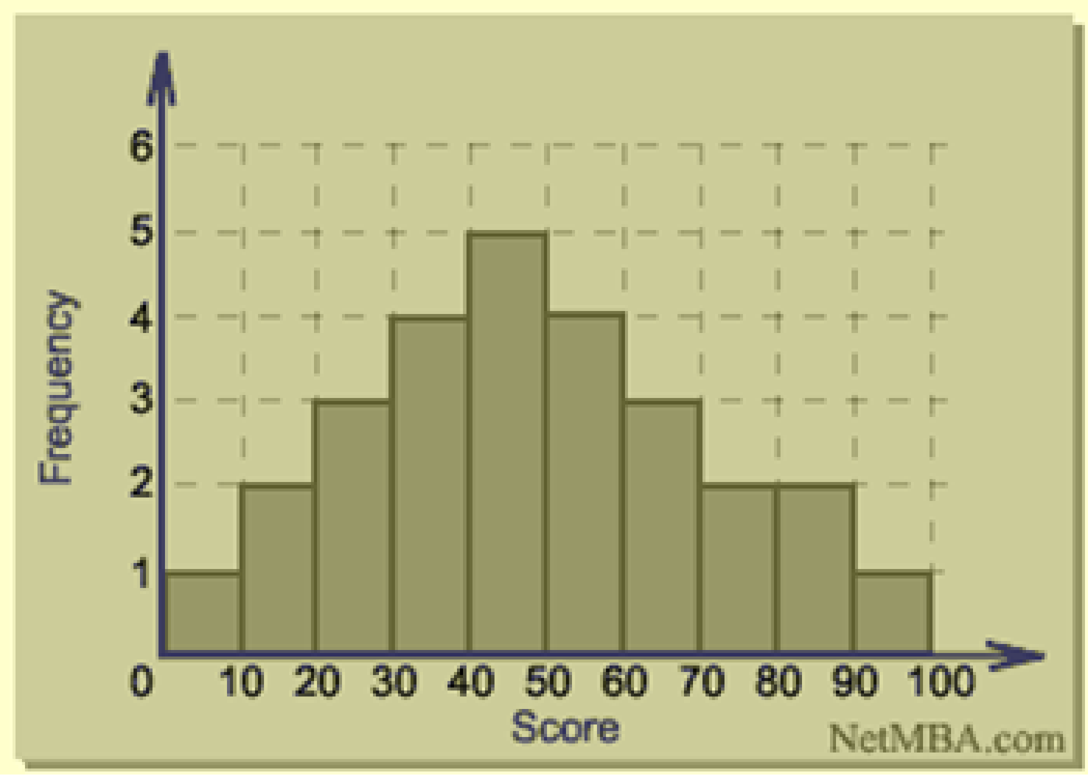
V的选择： V的大小选择与估计的效果是密切相连的

- **过大**：最终估计出来的**概率密度函数非常粗糙**；
- **过小：** 有些区域内根本没有样本或者样本非常少，这样会导致估计出来的**概率密度函数很不连续**。

直方图

例：统计某班的《机器学习》课程的期末成绩，画图直方图。

Group	Count
0 - 9	1
10 - 19	2
20 - 29	3
30 - 39	4
40 - 49	5
50 - 59	4
60 - 69	3
70 - 79	2
80 - 89	2
90 - 99	1



直方图

直方图方法优点：计算简单；适用于本身离散型的随机变量

直方图方法缺点：
- 密度函数不平滑；
- 直方图最多只能展示2维数据，如果维度更多则无法有效展示
- 密度函数受子区间（即每个直方体）宽度影响很大

■ **举例：**

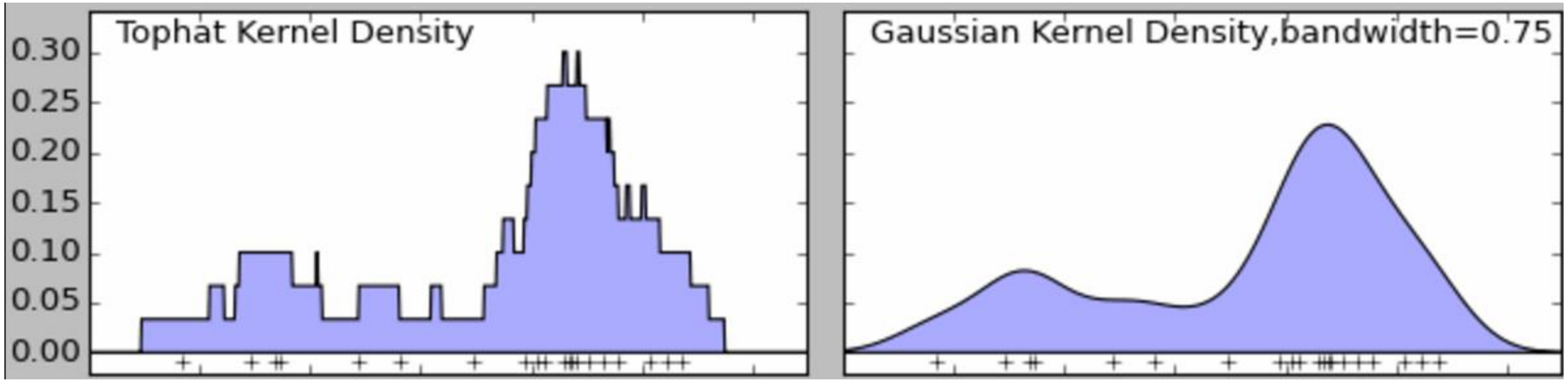
$x_1 = -2.1, x_2 = -1.3, x_3 = -0.6, \ x_4 = 1.9, \ x_5 = 5.1, x_6 = 6.2$

区间划分1：[-2.2, -1.2, -0.2, 0.8, 1.8, 2.8, 3.8, 4.8, 5.8, 6.8]

区间划分2：[-2.2, -1.2, -0.2, 0.8, 1.8, 2.8, 3.8, 4.8, 5.8, 6.8]

核密度估计

■ 直方图 vs 核密度估计方法



Kernel Density Estimation (KDE)

对于N个R^p的样本点x_k，假设其都服从概率密度函数p(x)，则**核密度估计公式**为：

$$\widehat{p(x)} = \frac{1}{N} K_h(x - x_k) = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{h} K\left(\frac{x - x_k}{h}\right) \qquad (3.21)$$

其中，h>0为一个平滑参数，称作带宽(bandwidth)，也叫窗口；

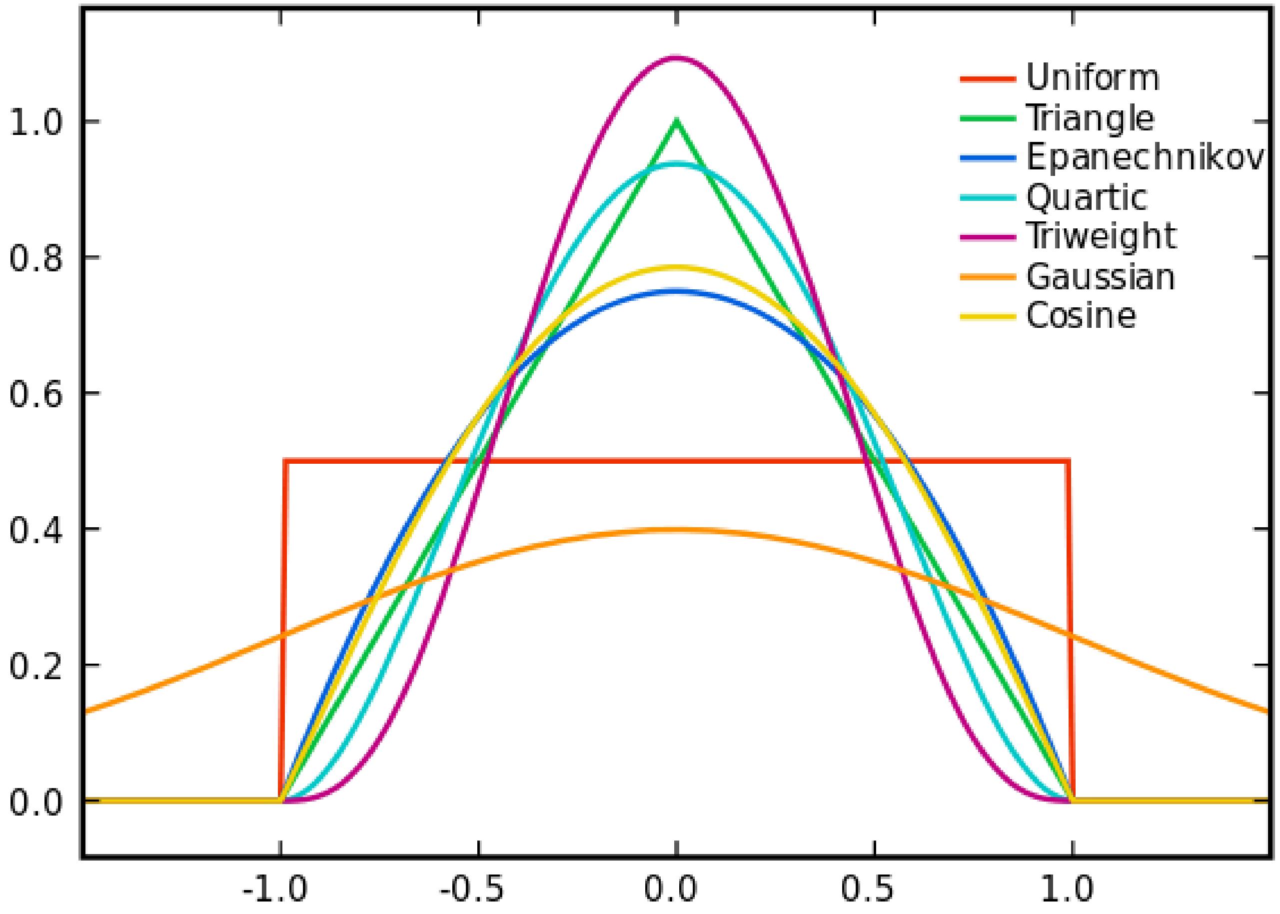
$K(\cdot)$为核函数，满足条件（非负、积分为0、均值为0）：

$$K(x) \geq 0, \int K(x)dx = 1 \qquad \int xK(x)dx = 0 \qquad \int x^2 K(x)dx > 0$$

核密度估计

常用的核函数：

- Gaussian
- Epanechnikov
- uniform
- triangular
- biweight
- triweight
- cosine



K近邻密度估计法

■ 基本思想：**固定划分区域内的样本点个数**为K，划分区域的体积大小自适应确定。

■ 计算过程：假设x所在的K近邻区域的区域体积为V_k，含有K个与其最近的样本。由此得K近邻密度估计：

参数K如何选择?

$$\widehat{p(x)} = \frac{K}{N \times V_k} \qquad (3.23)$$

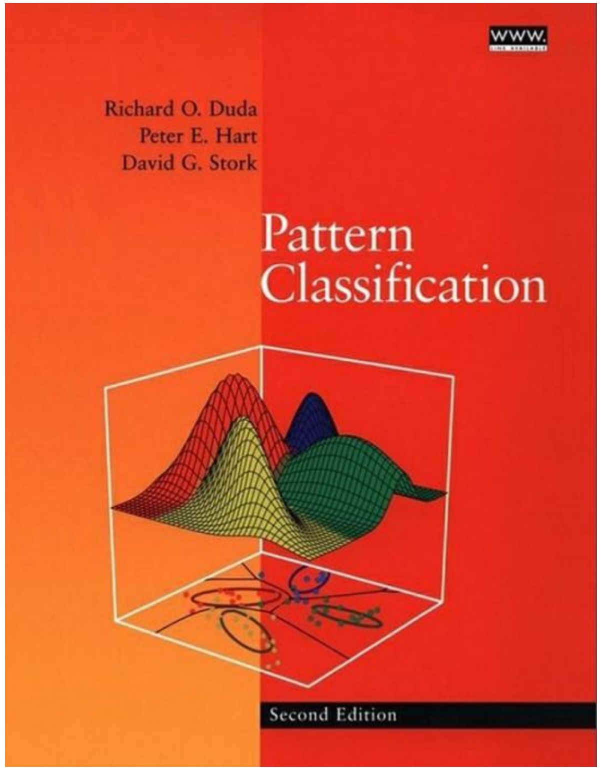
■ 方法评价：

在样本密度比较高的区域的体积就会比较小，而在密度低的区域的体积则会自动增大，这样就能够较好的兼顾在**高密度区域估计的分辨率**和在**低密度区域估计的连续性**。

内容提要

1. 预备知识
2. 密度估计应用举例
3. 经典参数估计
4. 从机器学习公理出发的参数估计
5. 密度函数的非参数估计

扩展阅读



vii

ONTENTS

北京交通大学《机器学习》课程组成员

于　剑：jianyu@bjtu.edu.cn，http://faculty.bjtu.edu.cn/6463/

景丽萍：lpjing@bjtu.edu.cn，http://faculty.bjtu.edu.cn/8249/

田丽霞：lxtian@bjtu.edu.cn，http://faculty.bjtu.edu.cn/7954/

黄惠芳：hfhuang@bjtu.edu.cn，http://faculty.bjtu.edu.cn/7418/

杨　凤：fengyang@bjtu.edu.cn，http://faculty.bjtu.edu.cn/8518/

吴　丹：wudan@bjtu.edu.cn，http://faculty.bjtu.edu.cn/8925/

万怀宇：hywan@bjtu.edu.cn，http://faculty.bjtu.edu.cn/8793/

王　晶：wj@bjtu.edu.cn，http://faculty.bjtu.edu.cn/9167/