第九章 概率模型

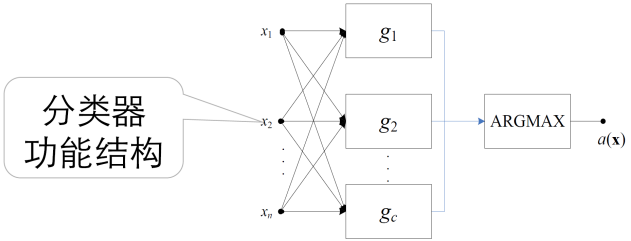
第 25 讲 概率密度函数的估计

黄定江

DaSE @ ECNU
djhuang@dase.ecnu.edu.cn

引言：基于样本的贝叶斯分类器设计

统计机器学习方法按其使用的技巧大致可以分为两类：核方法和贝叶斯学习。其中贝叶斯学习，又称为贝叶斯推断，其主要想法是在概率模型的学习和推理中，利用贝叶斯定理，计算在给定数据条件下模型的条件概率，即后验概率，并应用这个原理进行模型的估计以及对数据的预测。



- 在设计贝叶斯分类器时，需要已知先验概率和类条件概率密度，并按一定的决策规则确定判别函数和决策面。但实际工作中，类条件概率密度常常是未知的。

引言：基于样本的贝叶斯分类器设计

- 以鸢尾花分类任务为例，尽管在数据集中各种鸢尾花所占的比例是相等的，但是在实际中人们有可能根据所采集鸢尾花的地域大致判断其类别。例如，在中国的吉林省更有可能采集的是山鸢尾，而不是维吉尼亚鸢尾。这样，结合实际经验使得我们有可能推断先验概率 $P(\omega_i)$；
- 另外，通常我们可能获得各类鸢尾花的样本数据，但不可能给出类条件概率密度 $p(x|\omega_i)$。这就需要我们从所采集的各类鸢尾花样本中去估计出山鸢尾类概率密度和维吉尼亚鸢尾类概率密度。

引言：基于样本的贝叶斯分类器设计

在实际中，我们能收集到的是有限数目的样本，而未知的可能是：

- 先验概率 $P(\omega_i)$；
- 类条件概率密度 $p(x|\omega_i)$。

任务是利用样本集设计分类器，一个很自然的想法是把分类器设计分成两步：

1. 利用样本集估计先验概率 $P(\omega_i)$ 和类条件概率密度 $p(x|\omega_i)$，分别记为 $\hat{P}(\omega_i)$ 和 $\hat{p}(x|\omega_i)$
2. 然后利用估计的概率密度设计贝叶斯分类器。

这就是基于样本的两步 Bayes 分类器设计。

引言：基于样本的贝叶斯分类器设计

理想情况，希望当样本数 $N \to \infty$ 时，该方法设计的分类器收敛于理论上的最优解。为此目标，则需要

$$\hat{p}(\boldsymbol{x} \mid \omega_i) \overset{N\to\infty}{\longrightarrow} p(\boldsymbol{x} \mid \omega_i)$$

$$\hat{P}(\omega_i) \overset{N\to\infty}{\longrightarrow} P(\omega_i)$$

本讲主要内容：研究如何利用样本集估计概率密度函数。

类先验概率 $P(\omega_i)$ 的估计

- 依靠经验;
- 用训练数据中各类出现的频率来估计;
- 用频率估计概率的优点:
  - 无偏性
  - 相合性
  - 收敛速度快

类条件概率密度函数估计

概率密度函数可是满足下面条件的任何函数:

$$p(x) \geq 0, \quad \int p(x)dx = 1$$

估计的方法:

- 参数估计：已知类条件概率密度函数的形式，而参数未知。如已知样本总体符合正态分布，而正态分布的参数均值和方差未知。根据是否已知样本所在类别，参数估计又分为监督参数估计和非监督参数估计。根据是否使用先验信息，参数估计的方法又分为基于频率的的参数估计方法和基于贝叶斯的参数估计方法。
- 非参数估计：已知样本所在类别，未知类条件概率密度函数的形式，要求直接推断函数本身。一些常见典型的分布形式并不能总是满足实际需求。因此，有些实际问题需要根据样本推断总体分布。

概率密度函数的估计方法

	样本所属类别	总体概率密度函数的形式	推断	解决方法
监督参数估计	已知	已知	参数	矩估计、极大似然估计；贝叶斯估计
非监督参数估计	未知	已知		
非参数估计	已知	未知	概率密度函数	直方图、核密度估计、k近邻法

引言

考虑参数模型，其形式为：

$$\mathfrak{F} = \{f(x;\theta) : \theta \in \Theta\},$$

其中 $\Theta \subset \mathbb{R}^k$ 为参数空间，$\theta = (\theta_1, \cdots, \theta_k)$ 为参数。因此推断问题简化为 θ 的参数估计问题。

同学们在学习统计时经常会问：怎样能确定生成数据的分布是某种参数模型呢？实际上非常困难。但学习参数模型的方法仍然非常有用：

- 首先，根据有些案例的背景知识可以假定数据近似服从某种参数模型。例如，根据先验可以知道交通事故发生的次数服从近似泊松分布。
- 其次，参数模型的推断为理解非参方法提供了背景知识。

引言

本小节主要介绍基于频率观点（经典学派）的参数估计方法：

- 带头人：Pearson、Fisher、Neyman
- 观点：概率就是频率，参数就是参数，不会变化
- 主要方法：矩估计、极大似然估计等

25.2.1 矩估计

矩估计的基本思想：

- 上一讲提到，由大数定理，我们知道样本矩依概率收敛于总体矩，样本矩的连续函数依概率收敛于总体矩的连续函数；
- 又在许多分布中它们所含的参数都是矩的函数，例如正态分布 $N(\mu, \sigma^2)$ 中的参数 μ 和 σ^2 就是这个分布的的一阶原点矩和二阶中心矩；
- 因此很自然的会想到用样本矩作为相应总体矩的一种估计量。这种方法称为矩估计法。
- 矩估计不是最优的，但是最容易计算，它们也可以作为其他需要循环几次的算法的初始值。

接下来我们介绍矩估计的具体做法。

矩估计

假设总体 X 的概率函数为 $f(x;\theta_1,\cdots,\theta_k)$，其中 $\theta=(\theta_1,\cdots,\theta_k)$ 为待估参数，$X_1,\cdots,X_n$ 是来自 X 的样本。对于 $1\le j\le k$，定义总体 X 的 j 阶矩为

$$\alpha_j\equiv\alpha_j(\theta)=E_\theta(X^j)=\int x^j f(x;\theta_1,\cdots,\theta_k)dx,$$

一般来说，它们都是 $\theta_1,\cdots,\theta_k$ 的函数。而样本 $X_1,\cdots,X_n$ 的 j 阶样本矩定义为

$$\hat{\alpha_j}=\frac{1}{n}\sum_{i=1}^{n}X_i^j.$$

25.2.1 矩估计

定义 1

θ 的矩估计定义为 $\hat{\theta}$，使得

$$\begin{aligned} \alpha_1(\hat{\theta}) &= \hat{\alpha_1}, \\ \alpha_2(\hat{\theta}) &= \hat{\alpha_2}, \\ &\vdots \\ \alpha_k(\hat{\theta}) &= \hat{\alpha_k}. \end{aligned} \tag{1}$$

定义中的公式定义了带有 k 个未知参数 $\hat{\theta} = (\hat{\theta_1}, \cdots, \hat{\theta_k})$ 的 k 个方程的方程组，从中可以解出参数 $\theta = (\theta_1, \cdots, \theta_k)$ 的矩估计量。

矩估计举例

例 1

令 $X_1, \cdots, X_n \sim Bernoulli(p)$. 则 $\alpha_1 = E_p(X) = p$ 且 $\hat{\alpha}_1 = n^{-1} \sum_{i=1}^{n} X_i$. 让它们相等可以得到估计值

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

矩估计举例

例 2

令 $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. 则 $\alpha_1 = E_\theta(X) = \mu$ 且
$\alpha_2 = E_\theta(X_1^2) = D(X_1) + (E_\theta(X))^2 = \sigma^2 + \mu^2$. 现在需要解下述方程:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\hat{\sigma}^2 + \hat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

这是由两个方程组成含有两个未知参数的方程组。它的解为

$$\hat{\mu} = \bar{X},$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2.$$

矩估计的性质

定理 1

令 $\hat{\theta}$ 表示矩估计. 在适当的条件下, 下述成立:

1. 矩估计 $\hat{\theta}$ 以接近概率 1 存在.
2. 这个估计是相合的: $\hat{\theta} \xrightarrow{P} \theta$.
3. 这个估计是渐进正态的:

$$\sqrt{n}\left(\widehat{\theta}-\theta\right) \rightsquigarrow N(0, \Sigma)$$

其中,

$$\Sigma = gE_\theta\left(YY^{\mathrm{T}}\right)g^{\mathrm{T}},$$

$$Y = \left(X, X^2, \cdots, X^k\right)^{\mathrm{T}}, g = \left(g_1, \cdots, g_k\right), g_j = \partial\alpha_j^{-1}(\theta)/\partial\theta$$

定理最后一条可以用于求标准差和置信区间、然而，有比这更加简单的方法: Bootstrap 方法。本课程不做介绍。

25.2.2 极大似然估计（Maximum Likelihood Estimator，ML Estimator）

极大似然估计法是求估计的另一种方法，在参数模型中，它是最常用的参数估计方法。

- 极大似然估计法最早由高斯（G.F. Gauss) 提出，后来为费歇（R.A. Fisher）在 1912 年重新提出，并且证明了这个方法的一些性质。极大似然估计这一名称也是费歇给的.
- 它是建立在极大似然原理的基础上的一个统计方法。
- 极大似然原理的直观想法是：一个随机试验如有若干个可能的结果 $A, B, C, \cdots$。若在一次试验中，结果 A 出现，则一般认为试验条件对 A 有利，也即 A 出现的概率最大。

下面我们来介绍该方法。

极大似然估计（Maximum Likelihood Estimator，ML Estimator）

假设条件：

- 参数 θ 是确定而未知的量（非随机量）；
- 样本集按类别分开，样本集 $X^j(j=1,\cdots,c)$ 中样本都是从概率函数为 $f(x|\omega_j)$ 的总体中独立抽取出来（独立同分布，i.i.d.）；
- 概率函数 $f(x|\omega_j)$ 的形式已知，仅其参数 θ 未知，用 $f(x|\omega_j;\theta)$ 表示，对于同类别可简化为 $f(x;\theta)$；
- 各类样本只包含了本类分布的信息。

在上述假设前提下，可以分别处理 c 个独立的问题。独立地按照概率密度 $f(x;\theta)$ 抽取样本集 $X=\{X_1,X_2,\ldots,X_n\}$，用样本集 X 估计未知参数 θ。下面我们只考虑一类样本的极大似然估计。

似然函数

定义 2

设 $X_1, X_2, \ldots, X_n$ 为取自具有概率函数 $\{f(x;\theta) : \theta \in \Theta\}$ 的总体 X 的一个样本。样本 $X_1, X_2, \ldots, X_n$ 的联合概率函数在 X_i 取已知观测值 $x_i, i = 1, \cdots, n$ 时的值

$$L(\theta) = L(x_1, x_2, \ldots, x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta), \quad \theta \in \Theta,$$

称作这个样本的似然函数。对数似然函数为

$$H(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

极大似然估计

定义 3

极大似然估计 *MLE*，记为 $\hat{\theta}$，是使得 $L(\theta)$ 最大的 θ 值，也即满足

$$L(x_1, x_2, \ldots, x_n; \hat{\theta}) = \max_{\theta \in \Theta} L(x_1, x_2, \ldots, x_n; \theta),$$

称 $\hat{\theta}(x_1, x_2, \ldots, x_n)$ 为参数 θ 的极大似然估计值，其相应的统计量 $\hat{\theta}(X_1, X_2, \ldots, X_n)$ 为参数 θ 的极大似然估计量。

注：极大似然估计是典型的频率学派观点，它的基本意义是：待估计参数 θ 是客观存在的，只是未知而已，当 θ 满足 $\theta = \hat{\theta}$ 时，该组观测样本 $(X_1, X_2, \cdots, X_n) = (x_1, x_2, \cdots, x_n)$ 更容易被观测到，我们就说 $\hat{\theta}$ 是 θ 的极大似然估计值。也即，估计值 $\hat{\theta}$ 使得事件发生的可能性最大。

极大似然求解

由于 $\ln x$ 是单调递增函数，使得似然函数 $L(\theta)$ 最大的 $\hat{\theta}$ 也使得对数似然函数 $H(\theta)$ 最大，因此有时我们只要求对数似然最大即可！

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} H(\theta)$$

$$= \arg\max_{\theta} \sum_{i=1}^{n} \ln f(x_i; \theta)$$

必要条件：函数梯度（导数）为 0。

正态分布的极大似然估计

例 3

假设 $X \sim N(\mu, \sigma^2)$, μ, σ^2 为未知参数，$x_1, x_2, \ldots, x_n$ 是来自 X 的一个样本值，求 μ, σ^2 的极大似然估计量。

解

X 的概率密度为

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

对数似然函数为

$$H(\theta) = \ln L(\theta) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

正态分布的极大似然估计

解

由 $\nabla_\theta H(\theta) = 0$ 得

$$\begin{cases} \sum_{i=1}^{n} \frac{1}{\sigma^2}(x_i - \mu) = 0 \\ -\sum_{i=1}^{n} \frac{1}{\sigma^2} + \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{(\sigma^2)^2} = 0 \end{cases}$$

求解方程组得

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{N} (x_i - \hat{\mu})^2$$

因此得 μ, σ^2 的极大似然估计量分别为

$$\hat{\mu} = \bar{X}, \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2.$$

正态分布的极大似然估计

可以验证：

- 均值的极大似然估计量是无偏的

$$E(\hat{\mu}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mu$$

- 方差的极大似然估计量不是无偏的

$$E\left(\hat{\sigma}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2\right) = \frac{n-1}{n}\sigma^2 \neq \sigma^2$$

方差的无偏估计为样本方差

$$\frac{1}{n-1}\sum_{i=1}^{N}\left(x_i - \hat{\mu}\right)^2.$$

由上述分析可知：（1）正态总体均值的极大似然估计即为学习样本的算术平均；（2）正态总体方差的极大似然估计与样本的方差不同，当 N 较大的时候，二者的差别不大。

多元正态分布的极大似然估计

类似的，可以求解具有 n 个特征的多元正态分布的极大似然估计。

- 多元正态分布的估计值：

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

由上述估计值可以看出：（1）μ 的估计即为学习样本的算术平均；（2）估计的协方差矩阵是矩阵 $(x_i - \hat{\mu})(x_i - \hat{\mu})^T$ 的算术平均（$n \times n$ 阵列，$n \times n$ 个值）。

均匀分布的极大似然估计

例 4

令 $X_1, \cdots, X_n \sim \text{Uniform}(0, \theta)$，其概率密度函数为 $f(x;\theta) = \begin{cases} 1/\theta, & 0 \leqslant x \leqslant \theta \\ 0, & \text{其他} \end{cases}$，求未知参数 θ 的极大似然估计量。

解

考虑一个固定的 θ 值。假设对于某一个 i，有 $\theta < x_i$. 则 $f(x_i;\theta) = 0$，因此 $L(\theta) = \prod_i f(x_i;\theta) = 0$。对任意的 $x_i > \theta$, 则 $L(\theta) = 0$. 因此，如果 $\theta < x_{(n)}$，就有 $L(\theta) = 0$，这里 $x_{(n)} = \max\{x_1, \cdots, x_n\}$。现在考虑任意 $\theta \geqslant x_{(n)}$。对每一个 x_i，有 $f(x_i;\theta) = 1/\theta$，所以 $L(\theta) = \prod f(x_i;\theta) = \theta^{-n}$。总之

$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n, & \theta \geqslant x_{(n)} \\ 0, & \theta < x_{(n)} \end{cases}$$

在区间 $\left[x_{(n)}, \infty\right)$ 上，$L(\theta)$ 是严格递减的。因此 $\hat{\theta} = x_{(n)}$，其相应的估计量为 $\hat{\theta} = X_{(n)}$。

25.2.2 极大似然估计的性质

在某些条件下，极大似然估计有很多性质：

- 极大似然估计是相合估计：$\widehat{\theta} \xrightarrow{P} \theta_*$，其中，$\theta_*$ 表示参数 θ 的真实值。
- 极大似然估计是同变估计：如果 $\widehat{\theta}$ 是 θ 的极大似然估计，则 $g(\widehat{\theta})$ 是 $g(\theta)$ 的极大似然估计.
- 极大似然估计是渐近正态的：$\left(\widehat{\theta} - \theta_*\right) / \widehat{\mathrm{se}} \rightsquigarrow N(0,1)$。同时，估计的标准差 $\hat{se}$ 可以解出来。
- 极大似然估计是渐近最优或有效的：这表示，在所有表现优异的估计中，极大似然估计的方差最小，至少对大样本这肯定成立。
- 极大似然估计接近于贝叶斯估计。

25.2.3 非监督参数估计：极大似然法简介

考虑假设条件：

1. 样本集 $X = \{X_1, \ldots, X_N\}$ 中的样本分属于 c 个类别，但未知各样本所属类别；
2. 已知各类先验概率 $P(\omega_i), i = 1, \ldots, c$；（有时也可未知，一起估计）
3. 已知类条件概率密度形式 $p(x \mid \omega_i, \theta_i), i = 1, \ldots, c$；
4. 需估计未知的 c 个参数向量 $\theta_1, \theta_2, \ldots, \theta_c$。

似然函数

- 混合密度函数：分量密度的线性组合

$$p(x \mid \theta) = \sum_{i=1}^{c} \underbrace{p(x \mid \omega_i, \theta_i)}_{\text{分量密度}} \underbrace{P(\omega_i)}_{\text{混合参数}}$$

- 似然函数和对数似然函数：

$$L(\theta) = p(X \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta)$$

$$H(\theta) = \ln[L(\theta)] = \sum_{i=1}^{N} \ln p(x_i \mid \theta)$$

极大似然估计

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \prod_{k=1}^{N} p\left(x_k \mid \theta\right) = \arg\max_{\theta \in \Theta} \sum_{k=1}^{N} \ln p\left(x_k \mid \theta\right)$$

- 可识别性
  - 设 $\theta \neq \theta'$, 如对混合分布中每个 x 都有 $p(x \mid \theta) \neq p\left(x \mid \theta'\right)$, 则称密度 $p(x \mid \theta)$ 是可识别的;
  - 大部分常见连续随机变量的分布密度函数都是可识别的; 离散随机变量的混合概率函数往往是不可识别的。

极大似然估计

- 计算问题：求解微分方程组

$$\nabla_{\theta_i} H(\theta) = \sum_{k=1}^{N} \frac{1}{p\left(x_k \mid \theta\right)} \nabla_{\theta_i} \left[\sum_{j=1}^{c} p\left(x_k \mid \omega_j, \theta_j\right) P\left(\omega_j\right) \right]$$

$$= \sum_{k=1}^{N} \frac{1}{p\left(x_k \mid \theta\right)} \nabla_{\theta_i} \left[p\left(x_k \mid \omega_i, \theta_i\right) P\left(\omega_i\right) \right] \quad (\text{设} \theta_i, \theta_j \text{ 独立})$$

$$= \sum_{k=1}^{N} P\left(\omega_i \mid x_k, \theta_i\right) \nabla_{\theta_i} \ln p\left(x_k \mid \omega_i, \theta_i\right)$$

其中后验概率 $P\left(\omega_i \mid x_k, \theta_i\right) = \frac{p(x_k|\omega_i,\theta_i)P(\omega_i)}{p(x_k|\theta)}$

正态分布的非监督极大似然估计

例 5

求均值向量 $\boldsymbol{\mu}_i$ 未知，$\boldsymbol{\Sigma}_i, P(\omega_i), c$ 已知的正态分布的非监督极大似然估计。极大似然估计满足方程组

$$\sum_{k=1}^{N} \hat{P}\left(\omega_i \mid x_k, \hat{\theta}_i\right) \nabla_{\theta_i} \ln p\left(x_k \mid \omega_i, \hat{\theta}_i\right) = 0, \quad i = 1, \cdots c$$

代入正态分布

$$\sum_{k=1}^{N} P\left(\omega_i \mid x_k, \hat{\mu}_i \hat{\mu}_i(j+1)\right) = \frac{\sum_{k=1}^{N} P\left(\omega_i \mid x_k, \hat{\mu}_i(j)\right) x_k}{\sum_{k=1}^{N} P\left(\omega_i \mid x_k, \hat{\mu}_i(j)\right)} \frac{\omega_i \mid x_k, \hat{\mu}_i) \, x_k}{(\omega_i \mid x_k, \hat{\mu}_i)}.$$

25.3.1 贝叶斯理论体系

引言：回顾频率推断方法的基本思想

到目前为止我们讲述的方法都是频率论的估计方法（或经典方法）。频率论方法的观点基于下面的假设：

- 概率指的是相对频率，是真实世界的客观属性。
- 参数是固定的未知常数。由于参数不会波动，因此不能对其进行概率描述。
- 统计过程应该具有定义良好的频率稳定性。如：一个 95％的置信区间应覆盖参数真实值至少 95％的频率。

频率推断是根据样本信息对总体分布或总体的特征数进行推断，这里用到两种信息：

1. **总体信息**：总体分布提供的信息。
2. **样本信息**：抽取样本所得观测值提供的信息。

注：经典统计学更多关注频率推断。

频率推断方法的局限性

但是基于频率的估计方法，有其局限性：

- 基于频率估计方法的优良性，在大样本情况下有其理论上的保障。但在许多情况下，我们无法重复大量的试验，无法得到大量的试验结果，只能得到少量的试验结果。因此在小样本情况下，传统方法是否优良，是没有保障的。
- 设总体 X 的概率密度函数为 $f(x;\theta)$，$X_1, \cdots, X_n$ 为来自总体 X 的样本，当 n 较大时，用传统方法估计 θ，估计很准确。n 较小时，特别当 $n = 1$ 或 $n = 2$ 时，传统估计不是很可靠。

因而，人们一直在寻求小样本情况下的优良估计方法。解决的思路是：

- 用过去的经验，用人们过去对 θ 的了解 (或部分了解)，给出 θ 较可靠、较切合实际的估计。过去的看法、记忆或经验，常常支配着我们对事物的判断 (估计、评判)。

先验信息的重要性

- 例如，裁判打分，对知名运动员的评分总是要偏高，对新手的评分总是偏低。又如，对知名产品进行抽样检查，抽取了少量样品，如果全合格，便终止抽样，得出产品合格的结论。但对一个新厂生产的产品，则不会依据少量的抽样检查下结论。
- 因此，对 θ 的了解形成的一种先验信息，对估计可能是有帮助的。而在传统频率估计方法中，反映在数学上，我们把 X 当作随机变量，而把 θ 当作确定的未知常量，可能不一定恰当。$f(x;\theta)$ 提供的知识与信息是关于 X 的，它反映了 X 取值的规律性，但它没有反映 θ 的变化规律。

因此我们可以引入反映 θ 变化规律的信息，这种引入参数 θ 的先验信息的推断思想，就是贝叶斯推断。

贝叶斯推断的基本思想

贝叶斯方法基于下面的假设：

- 概率描述的是主观信念的程度，而不是频率。这样除了对从随机变化产生的数据进行概率描述外，我们还可以对其他事物进行概率描述。
- 可以对各个参数进行概率描述，即使它们是固定的常数。
- 为参数生成一个概率分布来对它们进行推导，点估计和区间估计可以从这些分布得到。

贝叶斯推断除了利用前面频率推断中提到的总体信息和样本信息，还使用第三种信息：

3. **先验信息**：即是抽样（试验）之前有关统计问题的一些信息。人们在试验之前对要做的问题在经验上和资料上总是有所了解的，这些信息对统计推断是有益的。一般说来，先验信息来源于经验和历史资料。先验信息在日常生活和工作中是很重要的。

注：机器学习和数据挖掘更偏爱贝叶斯推断。

贝叶斯统计学

基于上述三种信息进行统计推断的统计学称为**贝叶斯统计学**。它与经典统计学的差别就在于是否利用先验信息。贝叶斯统计通过对先验信息的收集、挖掘和加工，使它数量化，形成先验分布，参加到统计推断中来，以提高统计推断的质量。忽视先验信息的利用，有时是一种浪费，有时还会导出不合理的结论。上述利用先验信息形成先验分布的前提是：

1. 总体分布的参数是随机的，但有一定的分布规律
2. 参数是某一常数，但无法知道

目标是充分利用参数的先验信息对未知参数作出更准确的估计。所以贝叶斯方法就是把未知参数视为具有已知分布的随机变量，将先验信息数字化并利用的一种方法。

贝叶斯学派

贝叶斯学派：

- 带头人：Bayes,Laplace,Jeffreys,Robbins。
- 主要观点：频率不只是概率，存在主观概率，和实体概率可转化，参数作为随机变量。
- 主要方法：后验均值、贝叶斯估计、最大后验估计等。

贝叶斯

(Bayes，Thomas) (1702 1761)

- 贝叶斯是英国数学家。1702 年生于伦敦；1761 年 4 月 17 日卒于坦布里奇韦尔斯。
- 长期担任坦布里奇韦尔斯地方教堂的牧师。1742 年，贝叶斯被选为英国皇家学会会员。
- 如今在概率、数理统计学中以贝叶斯姓氏命名的有：贝叶斯公式、贝叶斯风险、贝叶斯决策函数、贝叶斯决策规则、贝叶斯估计量、贝叶斯方法、贝叶斯统计等等。

25.3.2 贝叶斯方法

贝叶斯推断通常用下面的方法来作：

- 选择一个概率密度函数 $\pi(\theta)$，用来表示在观察到数据之前我们对参数的信念（经验判断），称之为先验分布。
- 选择一个统计模型 $q(\mathbf{x};\theta)$（在此处记为 $q(\mathbf{x}|\theta)$），用来反映在给定参数 θ 情况下我们对 $\mathbf{x}$ 的信念（经验判断）。
- 当得到观察数据 $X_1, X_2, \cdots, X_n$ 后，改进我们原来的信念（经验判断），并且计算后验分布 $h(\theta|X_1, \cdots, X_n)$，从后验分布中得到点估计和区间估计。

下面我们介绍如何实现这些做法。

1. 先验分布和贝叶斯参数统计模型

Bayes 学派认为：样本分布族中的参数 θ 不是常量，而是随机变量，它可能取各种不同的值，取各种不同值的概率分布 $\prod(\theta)$ 也是确定的。

例 6

考虑某厂每天产品的次品率 p。关于 p 的算法：在当天的生产的产品中，进行产品全检，计算其次品率 p；或者抽取部分产品，估计其次品率 p。从当天看，p 是一个单纯的未知常数。但从较长的时间看，每天都有一个 p 值，其值因随机因素的作用，会产生波动，当天的 p 值可合理地视为随机变量 p 的一个可能值。如果我们有相当长一个时期的检验记录，则可以相当精确地定出 p 的概率分布。

形式上，把参数 θ 看成一个随机变量，并给出 θ 的概率分布 $\prod(\theta)$，或概率密度 $\pi(\theta)$，这个分布 $\prod(\theta)$ 在抽样前就给出了，把它称为 θ 的先验分布。

先验分布和贝叶斯参数统计模型

定义 4

参数 θ 的参数空间 Θ 上的一个概率分布称为 θ 的 **先验分布**，其密度族记为 $\{\pi(\theta) : \theta \in \Theta\}$；样本 $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ 的条件密度函数族 $\{q(\mathbf{x} \mid \theta) : \theta \in \Theta\}$ $(\mathbf{x} = (x_1, x_2, \ldots, x_n))$ 称为样本分布族；先验分布 $\{\pi(\theta) : \theta \in \Theta\}$ 与样本分布族 $\{q(\mathbf{x} \mid \theta) : \theta \in \Theta\}$ 构成 **贝叶斯参数统计模型**。

注 1：有时，把参数 θ 看成随机变量有其合理性，但把所有未知参数都视为随机变量则牵强。例，要估计某铁矿的含铁量 p，把 p 看成随机变量，就要设想这个铁矿是无穷多“类似”铁矿的一个样本，这是不自然的，不如把 p 看做一个独立的未知常数。

注 2：此外，虽然把参数 θ 看成随机变量有其合理性，但人们的先验知识没有确切到能用概率分布把 θ 表达出来。于是，引出了一系列先验分布的确定方法。

2、后验分布与贝叶斯公式的密度函数形式

- 设 θ 为随机变量，总体 X 依赖于参数 θ 的概率密度函数为 $f(x;\theta)$，在贝叶斯统计中记为 $f(x \mid \theta)$，它表示在随机变量 θ 取某个给定值时总体的条件概率密度函数；
- 根据参数 θ 的先验信息可确定先验分布，θ 的先验概率密度函数记为 $\pi(\theta)$；
- 从贝叶斯观点看，来自总体 X 的样本 $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ 的产生分两步进行：首先从先验分布 $\pi(\theta)$ 产生一个样本 θ_0，然后从 $f(\mathbf{x}|\theta_0)$ 中产生一组样本，这时样本的联合条件概率函数为

$$q(\mathbf{x} \mid \theta_0) = \prod_{i=1}^{n} f(x_i \mid \theta_0)$$

其中 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ 为样本观测值，这个分布综合了总体信息和样本信息。

贝叶斯公式的密度函数形式 (后验分布)

- 由于 θ_0 是未知的，它是按先验分布 $\pi(\theta)$ 产生的。为把先验信息综合进去，不能只考虑 θ_0，对 θ 的其它值发生的可能性也要加以考虑，故要用 $\pi(\theta)$ 进行综合。这样一来，样本 $X_1, \ldots, X_n$ 和参数 θ 的联合概率密度函数为：

$$g(\mathbf{x}; \theta) = q(\mathbf{x} \mid \theta)\pi(\theta), \tag{2}$$

其中 $q(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$，这个联合分布把总体信息、样本信息和先验信息三种可用信息都综合进去了。

- 在没有样本信息时，人们只能依据先验分布对 θ 作出推断。在有了样本观察值 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ 之后，则应依据 $g(\mathbf{x}; \theta)$ 对 θ 作出推断。由于联合概率密度函数等于条件概率密度函数和边际概率密度函数的乘积，也即

$$g(\mathbf{x}, \theta) = h(\theta \mid \mathbf{x})m(\mathbf{x}), \tag{3}$$

其中 $m(\mathbf{x}) = \int_{\Theta} g(\mathbf{x}, \theta)d\theta = \int_{\Theta} q(\mathbf{x} \mid \theta)\pi(\theta)d\theta$ 是 x 的联合边际概率密度函数，它与 θ 无关，不含 θ 的任何信息。

贝叶斯公式的密度函数形式 (后验分布)

- 联立公式(2)和(3)可知，能用来对 θ 作出推断的仅是条件概率密度函数 $h(\theta \mid \mathbf{x})$ ，它的计算公式是

$$h(\theta \mid \mathbf{x}) = \frac{g(\mathbf{x}, \theta)}{m(\mathbf{x})} = \frac{q(\mathbf{x} \mid \theta) \cdot \pi(\theta)}{\int_{\Theta} q(\mathbf{x} \mid \theta) \cdot \pi(\theta) d\theta}$$

其相应的条件分布称为 θ 的后验分布，它集中了总体、样本和先验中有关 θ 的一切信息。后验分布 $h(\theta \mid \mathbf{x})$ 的计算公式就是用密度函数表示的贝叶斯公式。它是用总体和样本对先验分布 $\pi(\theta)$ 作调整的结果，贝叶斯统计的一切推断都基于后验分布进行。

后验分布

定义 5

在 $\mathbf{X}=\mathbf{x}$ 的条件下, θ 的条件分布 (或条件概率密度) 称为**后验分布**, 后验分布由后验密度函数 $\{h(\theta \mid \mathbf{x}): \theta \in \Theta\}$ 描述, 其计算公式为:

$$h(\theta \mid \mathbf{x})=\frac{g(\mathbf{x}, \theta)}{m(\mathbf{x})}=\frac{q(\mathbf{x} \mid \theta) \pi(\theta)}{\int_{\Theta} q(\mathbf{x} \mid \theta) \cdot \pi(\theta) d \theta}$$

说明:
1) θ 的先验分布 $\pi(\theta)$ 概括了我们在试验前关于 θ 的认识;
2) 经过试验得到样本观测值 $\mathbf{x}=(x_1, x_2, \ldots, x_n)$ 后, 我们的认识起了变化, $h(\theta \mid \mathbf{x})$ 是我们重新认识 θ 的基础和根据;

3) 由于 $\int_\Theta q(\mathbf{x} \mid \theta) \cdot \pi(\theta) d\theta$ 不依赖于 θ, 在计算 θ 的后验分布中仅起到一个正则化因子的作用, 若把 $\int_\Theta q(\mathbf{x} \mid \theta) \cdot \pi(\theta) d\theta$ 省略, 可将后验密度函数改写为如下等价形式:

$$h(\theta \mid \mathbf{x}) \propto q(\mathbf{x} \mid \theta)\pi(\theta)$$

其中符号 "$\propto$" 表示两边仅相差一个不依赖于 θ 的常数因子。$q(\mathbf{x} \mid \theta)\pi(\theta)$ 称为后验分布 $h(\theta \mid \mathbf{x})$ 的核。

例 7

设 p 是某厂产品的合格率。在抽样前, 我们可以假定 p 在区间 $(0,1)$ 之间是均匀分布的 *(对 p 的认识不多, 不妨设 p 取各种值的可能性一样大)*。抽取了 n 个产品检查发现有 m 个废品, 这时我们会修正对 p 的认识, p 仍有可能取 *(0,1)* 区间的任何值, 但机会大小不处处一样了, 在 $p = \frac{m}{n}$ 这一点附近的可能性最大, 而接近 *0,1* 处则可能性很小。

3. 贝叶斯推断的原则

对贝叶斯统计而言，样本 $X_1, \cdots, X_n$ 的唯一作用在于把对 θ 的认识由先验分布转化成后验分布。因此，贝叶斯统计推断的原则就是：

- 对参数 θ 所作的任何推断 (参数估计、假设检验等) 必须基于且只能基 θ 的后验分布，即后验密度函数族 $\{h(\theta \mid \mathbf{x}) : \theta \in \Theta\}$。
- 一旦由样本 $X_1, \cdots, X_n$ 算出 θ 的后验分布，就设想我们除了这一后验分布外，其余的东西 (样本值、样本分布、先验分布) 全忘记了。这时，对 θ 的推断的唯一凭借就是这一后验分布。

传统的统计推断原则许多都不能用了。如无偏性原则，$\hat{\theta} = T(X_1, \cdots, X_n)$，$E(\hat{\theta}) = \theta$，完全利用样本 (样本的函数) 在进行推断没有利用后验分布，不符合贝叶斯统计推断的原则。

4. 先验分布的确定

贝叶斯推断涉及先验分布的确定。确定先验分布 $\pi(\theta)$ 的方法主要有客观法、主观概率法、同等无知原则以及共轭分布法。

(1) 客观法

以前的资料积累较多，对 θ 的先验分布能作出较准确的统计或估计。在这种情况下，分布的确定没有渗杂多少人的主观因素，故称之为客观法。

如果能用客观法确定 θ 的先验分布 $\pi(\theta)$，对贝叶斯学派持否定态度的统计学者也不反对用贝叶斯方法去作数据处理。

在不少情况下，以往积累的资料并不是直接给出了参数在当时的取值，而只是一种估计。例如，某厂产品的废品率，不可能是全检 (可能是破坏性检验)。有些资料不是直接关于 θ 取值分布的记录，但我们可以利用这些资料对 θ 的先验分布作出经验性的推断。

先验分布的确定

(2) 主观概率法

按照 Bayes 学派的说法，这是一种通过“自我反省”去确定先验分布的方法。就是说，对参数 θ 取某某值的可能性多大，通过思考，觉得该如何，而定下一个值。

主观先验分布反映了个人以往对 θ 的了解，包括经验知识和理论知识，其中有部分可能是通过他人获取的，也可能是他人对 θ 的了解。对过去的经验和知识，必须经过组织和整理。这样提出的先验分布，在主观上是正确的，但不能保证合乎某种客观标准。

先验分布的确定

(3) 同等无知原则

这一原则称为 Bayes 假定。以产品的废品率为例，当我们对 p 一无所知时，我们只好先验地认为，p 以同等机会取 $(0,1)$ 内各种值，因而以 $(0,1)$ 内均匀分布 $U(0,1)$ 作为 p 的先验分布。这一先验分布称为**无信息先验分布**。

注：这一原则会出现矛盾：如果我们对 p 无知，对 p^3 也同样无知。按同等无知原原则，可以取 U(0,1) 作为 p^3 的分布，但这时 p 的分布就不是 U(0,1) 了。

先验分布的确定

(4) 共轭分布方法

H.Raiffa, R.Schlaifer 提出了先验分布应取共轭分布才合适。

定义 6

设样本分布族为 $\{q(\mathbf{x} \mid \theta) : \theta \in \Theta\}$，若先验分布 $\pi(\theta)$ 与后验分布 $h(\theta \mid \mathbf{x})$ 属于同一分布类型，则先验分布 $\pi(\theta)$ 成为关于 $q(\mathbf{x} \mid \theta)$ 的共轭分布。确切地说，若 $\mathcal{F}$ 为 θ 的一个密度函数族，若任取 $\pi(\theta) \in \mathcal{F}$，得到样本观测值 $\mathbf{x}$ 后，由 $\pi(\theta)$ 及 $q(\mathbf{x} \mid \theta)$ 确定的后验密度函数 $h(\theta \mid \mathbf{x}) \in \mathcal{F}$，则称 $\mathcal{F}$ 是关于 $\{q(\mathbf{x} \mid \theta) : \theta \in \Theta\}$ 的共轭先验分布族，或称为参数 θ 的**共轭先验分布族**。

选取共轭先验分布有如下好处：a) 符合直观，先验分布和后验分布应该是相同形式的；b) 可以给出后验分布的解析形式；c) 可以形成一个先验链，即现在的后验分布可以作为下一次计算的先验分布，如果形式相同，就可以形成一个链条。

共轭先验分布

例 8

设 $X_1, X_2, \ldots, X_n$ 是来自正态分布 $N(\theta, \sigma^2)$ 的一个样本，其中 θ 已知，求方差 σ^2 的共轭先验分布.

解

$(X_1, X_2, \cdots, X_n)^T$ 的联合条件概率函数为

$$q(\mathbf{x} \mid \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right] \propto \left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right]$$

所以 σ^2 的共轭先验分布是

$$\pi(\sigma^2) = \frac{\lambda^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{\lambda}{\sigma^2}\right],$$

为倒 Γ 分布。

共轭先验分布

计算共轭先验分布的方法：由 $h(\theta \mid \mathbf{x}) = \pi(\theta)q(\mathbf{x} \mid \theta)/m(\mathbf{x})$，其中 $m(\mathbf{x})$ 不依赖于 θ，先求出 $q(\mathbf{x} \mid \theta)$，再选取与 $q(\mathbf{x} \mid \theta)$ 具有相同形式的分布作为先验分布，就是共轭分布。常见分布的共轭先验分布有：

- 二项分布 $b(n, \theta)$ 中的成功概率 θ 的共轭先验分布是贝塔分布 $Be(a, b)$ ；
- 泊松分布 $P(\theta)$ 中的均值 θ 的共轭先验分布是伽玛分布 $\Gamma(\alpha, \lambda)$ ；
- 指数分布中均值的倒数的共轭先验分布是伽玛分布 $\Gamma(\alpha, \lambda)$;
- 在方差已知时，正态均值 θ 的共轭先验分布是正态分布 $N\left(\mu, \tau^2\right)$;
- 在均值已知时，正态方差 σ^2 的共轭先验分布是倒伽玛分布 $I\Gamma(\alpha, \lambda)$ 。

5. 后验分布的计算

我们通过下面的例子来说明后验分布的计算：

例 9

假设总体 $X \sim N(\mu, \sigma^2)$ (σ^2 已知)，$X_1, X_2, \ldots, X_n$ 为来自总体 X 的样本，由过去的经验和知识，我们可以确定 μ 的取值范围在区间 $[-\mu_0, \mu_0]$ 之内，但无法得到关于 μ 的更多的信息，按同等无知的原则，我们假定 $\mu \sim U[-\mu_0, \mu_0]$，其概率密度为：

$$\pi(\mu) = \begin{cases} \frac{1}{2\mu_0} & |\mu| \le \mu_0 \\ 0 & |\mu| > \mu_0 \end{cases}$$

样本分布函数族为

$$q(\mathbf{x} \mid \mu) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=1}^{n} (x_i - \mu)^2 \right]$$

后验分布的计算

于是

$$h(\mu \mid \mathbf{x}) = \frac{q(\mathbf{x} \mid \mu) \cdot \pi(\mu)}{m(\mathbf{x})} = \frac{q(\mathbf{x} \mid \mu) \cdot \pi(\mu)}{\int_{-\infty}^{+\infty} q(\mathbf{x} \mid \mu) \cdot \pi(\mu) \, d\mu}$$

$$\propto \begin{cases} \exp\left[-\frac{1}{2\sigma^2} \sum_{n=1}^{n} (x_i - \mu)^2\right] & |\mu| \leq \mu_0 \\ 0 & |\mu| > \mu_0 \end{cases}$$

消去分子分母中的公共部分，得:

$$h(\mu \mid \mathbf{x}) = \begin{cases} \frac{1}{c(\mathbf{x})} \cdot \exp\left[-\frac{n}{2\sigma^2} \cdot (\bar{x} - \mu)^2\right] & |\mu| \leq \mu_0 \\ 0 & |\mu| > \mu_0 \end{cases}$$

其中

$$c(\mathbf{x}) = \int_{-\mu_0}^{+\mu_0} \exp\left[-\frac{n}{2\sigma^2} \cdot (\bar{x} - \mu)^2\right] d\mu$$

6. 后验分布的应用：基于后验分布的点估计和区间估计

- 首先，可以通过集中后验的中心得到点估计．通常，使用后验的均值或众数．后验均值为

$$\bar{\theta}_n = \int \theta h(\theta \mid \mathbf{x}) \, \mathrm{d}\theta = \frac{\int \theta q(\mathbf{x} \mid \theta) \cdot \pi(\theta) \mathrm{d}\theta}{\int_\Theta q(\mathbf{x} \mid \theta) \cdot \pi(\theta) \mathrm{d}\theta}.$$

- 也可以得到贝叶斯区间估计．可以求出 a 和 b，使得

$$\int_{-\infty}^{a} h(\theta \mid \mathbf{x}) \, \mathrm{d}\theta = \int_{b}^{\infty} h(\theta \mid \mathbf{x}) \, \mathrm{d}\theta = \alpha/2.$$

令 $C = (a, b)$．则

$$\mathbb{P}(\theta \in C \mid \mathbf{x}) = \int_a^b h(\theta \mid \mathbf{x}) \, \mathrm{d}\theta = 1 - \alpha,$$

所以 C 是 $1 - \alpha$ 后验区间。

例 10

令 $X_1, \cdots, X_n \sim \text{Bernoulli}(p)$. 假设把均匀分布 $\pi(p) = 1$ 或 *Beta* 分布作为 p 的先验分布. 考虑其后验估计和贝叶斯区间估计。

解

根据贝叶斯定理, 后验的形式为

$$h(p \mid \mathbf{x}) \propto \pi(p) q(\mathbf{x} \mid p) = p^s (1-p)^{n-s} = p^{s+1-1} (1-p)^{n-s+1-1},$$

其中, $s = \sum_{i=1}^{n} x_i$ 是成功的次数. 回想起如果一个随机变量服从参数为 α 和 β 的 *Beta* 分布, 其密度为

$$f(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}.$$

可以求出 p 的后验分布是参数为 $s+1$ 和 $n-s+1$ 的 *Beta* 分布, 即

$$h(p \mid \mathbf{x}) = \frac{\Gamma(n+2)}{\Gamma(s+1)\Gamma(n-s+1)} p^{(s+1)-1} (1-p)^{(n-s+1)-1}.$$

解

（续）将其记为

$$p \mid \mathbf{x} \sim \mathrm{Beta}(s+1, n-s+1).$$

注意到并没有真正做积分 $\int_p q(\mathbf{x} \mid p) \cdot \pi(p)\mathrm{d}p$ 就求出了归一化系数. 由于 *Beta* (α, β) 的均值为 $\alpha/(\alpha+\beta)$, 所以贝叶斯估计为

$$\bar{p} = \frac{s+1}{n+2}$$

可以把这个估计改写为

$$\bar{p} = \lambda_n \widehat{p} + (1-\lambda_n)\,\widetilde{p},$$

其中, $\widehat{p}$ 是极大似然估计, $\widetilde{p} = 1/2$ 是先验均值, $\lambda_n = n/(n+2) \approx 1$. 通过计算 $\int_a^b h(p \mid \mathbf{x})\,\mathrm{d}p = 0.95$ 得到 a 和 b, 从而得到一个 95% 的后验区间.

解

(续) 假设先验分布不是用均匀分布, 而是用 $p \sim Beta\ (\alpha, \beta)$. 如果重复上述的计算, 可以得到 $p \mid \mathbf{x} \sim \mathrm{Beta}(\alpha + s, \beta + n - s)$. 扁平先验 (均匀分布) 仅仅是 $\alpha = \beta = 1$ 时的一个特例. 后验均值为

$$\bar{p} = \frac{\alpha + s}{\alpha + \beta + n} = \left(\frac{n}{\alpha + \beta + n}\right)\widehat{p} + \left(\frac{\alpha + \beta}{\alpha + \beta + n}\right)p_0,$$

其中, $p_0 = \alpha/(\alpha + \beta)$ 是先验均值.

从上述例子中可以发现先验是 Beta 分布，后验也是 Beta 分布，因此先验是关于模型共轭的。

关于贝叶斯统计推断还包括贝叶斯假设检验等内容，本课程不作详细介绍！下一小节我们将从统计方法的优良性角度讨论贝叶斯估计。

引言

- 前面已经考虑了几种点估计，如矩估计、极大似然估计和后验均值。事实上，还有许多其它的估计方法。如何选择这些方法呢？可以通过决策理论来评价这些方法，统计决策理论是比较统计过程的正规理论。
- 20 世纪 40 年代末，瓦尔德（Wald）建立了统计决策理论。1950 年发表了《统计决策函数》一书，系统地论述了他的理论。
- 这一理论对参数估计、区间估计、假设检验等统计问题在统计决策的观点下统一处理。它通过将统计问题表示成数学最优化问题的解，引进了各种优良性准则。这个理论的一些基本观点现在已经不同程度地渗透到各个统计分支，对数理统计学的发展产生了重大的影响。统计决策理论是二战后数理统计学发展的重大事件。

引言

- 统计决策与统计推断是既有联系又有区别的。统计决策问题要考虑到决策的损失，而统计推断问题，一般是指解决一类统计问题的方法，但不考虑决策的损失问题。如在参数的点估计中，矩估计与极大似然估计是进行点估计的统计方法，属于统计推断的范围，而讨论点估计的优良性，则与统计决策有关。统计决策是统计推断研究的深化。统计决策方法可以作为产生优良统计推断的手段。
- 贝叶斯（Bayes）估计是贝叶斯统计的主要部分，它是运用统计决策理论研究参数估计问题。

本小节，我们先简要介绍统计决策理论，然后引出贝叶斯估计。

25.4.1 统计决策的基本概念

1. 统计决策问题的三个要素

统计决策问题的三个要素是：样本空间和样本分布族、决策（行动）空间、损失函数。

1. 样本空间和分布族
   设总体 X 的分布函数为 $F(x;\theta)$，$\theta \in \Theta$ 是未知参数，Θ 是参数空间。若设 $X_1, \ldots, X_n$ 是来自总体 X 的一个样本，则样本所有可能值组成的集合称为样本空间，记为 $\mathcal{X}$。由于 X_i 的分布函数为 $F(x_i;\theta)$，$i = 1, \cdots, n$，则 $X_1, \ldots, X_n$ 的联合分布函数为

$$F(x;\theta) = \prod_{i=1}^{n} F(x_i;\theta), \theta \in \Theta$$

   若记

$$F^* = \left\{ \prod_{i=1}^{n} F(x_i;\theta), \theta \in \Theta \right\},$$

   则称 F^* 为样本 $X_1, \ldots, X_n$ 的概率分布族，简称样本分布族。

所谓给定了一个参数统计模型，实质上是指给定了样本空间和样本分布族。

统计决策问题的三个要素

2. 决策空间（判决空间）
   对于一个统计问题，如参数的点估计、区间估计以及参数的假设检验问题，我们常常要给予适当的回答。对参数的点估计，一个具体的估计值就是一个回答。在假设检验中，它是一个决定，即是接受还是拒绝原假设。在统计决策中，每个具体的回答称为一个决策（或行动），一个统计问题中可能选取的全部决策组成的集合称为决策空间，记为 $\mathcal{A}$。一个决策空间至少应有两个决策，假如 $\mathcal{A}$ 中只含有一个决策，那么人们就无需选择，从而也形成不了一个统计决策问题。

本课程我们讨论的决策主要集中在点估计。

统计决策问题的三个要素

3. 损失函数

统计决策的一个基本观点是假设：每采取一个决策，必然有一定的的后果，所采取的决策不同，后果就不同。这种后果必须以某种方式通过损失函数的形式表示出来。这样，每一决策有优劣之分。统计决策的一个基本思想就是把决策的优劣性以数量的形式表现出来，其方法是引入一个依赖参数值 $\theta \in \Theta$ 和决策 $d \in \mathcal{A}$ 的二元实值非负 $L(\theta, d) \geq 0$，称之为损失函数，它表示当参数真值为 θ 而采取决策 d 时所造成的损失，决策越正确，损失就越小。

由于在统计问题中人们总是利用样本对总体进行推断，所以误差是不可避免的，因而总会带来损失，这就是损失函数定义为非负函数的原因。

常见损失函数

对于不同的统计问题，可以选取不同的损失函数，对于参数的点估计问题常见的损失函数有如下几种：

当$\theta \geq d$ 时，$L(\theta, d) = k_0(\theta - d)$，当$\theta < d$ 时为$k_1(d - \theta)$ 线性损失，其中 k_0, k_1 是两个非负常数；

$L(\theta, d) = |\theta - d|$ 绝对损失；

$L(\theta, d) = (\theta - d)^2$ 平方损失；

$L(\theta, d) = |\theta - d|^p$ L_p 损失；

当$\theta = d$ 时，$L(\theta, d) = 0$，当$\theta \neq d$ 时为1 $0-1$ 损失；

$L(\theta, d) = \lambda(\theta)\, W(|\theta - d|)$ 凸损失；

$L(\theta, d) = (d - \theta)^T A (d - \theta)$ 多元二次损失；

$L(\theta, d) = \int \log\left(\frac{f(x;\theta)}{f(x;d)}\right) f(x;\theta)\mathrm{d}x$ Kullback-Leibler 损失。

2. 统计决策函数及其风险函数

统计决策函数

假设给定了一统计决策问题的三要素：样本空间 $\mathcal{X}$ 和样本分布族，决策空间 $\mathcal{A}$ 及损失函数 $L(\theta, d)$。我们的问题是对每一样本观测值 $\mathbf{x} = (x_1, \cdots, x_n)$，即对每一个 $\mathbf{x} \in \mathcal{X}$，有一个确定的法则，在 $\mathcal{A}$ 中选取一个决策 d。这样一个对应关系是定义在样本空间 $\mathcal{X}$ 上，取值于决策空间 $\mathcal{A}$ 的一个函数（即由 $\mathcal{X}$ 到 $\mathcal{A}$ 的一个映射）$d(\mathbf{x})$。

定义 7

定义在样本空间 $\mathcal{X}$ 上，取值于决策空间 $\mathcal{A}$ 内的函数 $d(\mathbf{x})$，称为统计决策函数，简称决策函数。

易见，决策函数 $d(\mathbf{x})$ 就是一个行动方案，当有了样本观测值 $\mathbf{x}$ 后，按既定的方案采取行动(决策)$d(\mathbf{x})$；因此 $d(\mathbf{X})$ 本质上就是一个统计量。决策函数 $d(\mathbf{x})$ 就是所给定的统计决策问题的一个解。

风险函数

给定一个统计决策问题，若使用决策函数 $d(\mathbf{x})$，由于样本 $\mathbf{X}=(X_1,\ldots,X_n)$ 是随机的，从而 $d(\mathbf{X})$ 也是随机的，因而 $L(\theta,d(\mathbf{X}))$ 也是随机的，它是样本 $\mathbf{X}$ 的函数。当样本取不同的值 $\mathbf{x}$，决策 $d(\mathbf{x})$ 可能不同，所以损失函数值 $L(\theta,d)$ 也不同。因此为了判断一个决策的好坏，一般从总体上来评价比较决策函数，也即用 $L(\theta,d(\mathbf{X}))$ 关于样本的数学期望，代表了取决策函数 $d(\mathbf{x})$ 时在概率意义下的平均风险或损失，这个平均风险就是统计决策理论中非常重要的风险函数的概念。

定义 8

设样本空间和样本分布族分别为 $\mathcal{X}$ 和 $F^*=\{F(\mathbf{x};\theta):\theta\in\Theta\}$，决策空间为 $\mathcal{A}$，损失函数为 $L(\theta,d)(\theta\in\Theta,d\in\mathcal{A})$，则统计决策函数 $d(\mathbf{x})$ 的风险函数定义为

$$R(\theta,d)=E_\theta[L(\theta,d(\mathbf{X}))]=\int_{\mathcal{X}}L(\theta,d(\mathbf{x}))dF(\mathbf{x};\theta),$$

$R(\theta,d)$ 是 θ 的函数，当 θ 取定值时，$R(\theta,d)$ 称为决策函数 $d(\mathbf{x})$ 在参数值 θ 时的风险。

点估计问题的常用风险函数

风险函数 $R(\theta, d)$ 是统计决策问题当采取决策函数 d 时统计意义下的平均损失。风险函数是 Wald 统计决策理论的基本概念。评价一个决策函数 d 的依据就是其风险函数。
点估计问题在各种损失函数下的风险：

- 设决策函数为 $d(\mathbf{x}) = \hat{\theta}(\mathbf{x})$（即 θ 的点估计），则对应平方损失函数的风险函数为

$$R(\theta, \hat{\theta}) = E_\theta[(\theta - \hat{\theta}(\mathbf{X}))^2] = \int_{\mathcal{X}} (\theta - \hat{\theta}(\mathbf{x}))^2 dF(\mathbf{x}; \theta),$$

  即为估计量 $\hat{\theta}$ 的均方误差；

- 对应绝对值损失函数的风险函数为

$$R(\theta, \hat{\theta}) = E_\theta[|\theta - \hat{\theta}(\mathbf{X})|] = \int_{\mathcal{X}} |\theta - \hat{\theta}(\mathbf{x})| dF(\mathbf{x}; \theta),$$

  即为估计量 $\hat{\theta}$ 的平均绝对误差。

此外，针对区间估计和假设检验问题，也可以给出在相应损失函数下的风险函数。

风险函数的比较：优良性准则

Wald 理论引进统计决策函数及其风险函数，将各类统计推断问题用统一的观点与方法处理。若要论及统计推断方法的优良性，必须考虑统计推断所采取决策的损失，即要考虑风险函数。按照 Wald 的理论，风险函数越小，决策函数就越优良。但是对于给定的决策函数，风险函数仍是参数 θ 的函数。所以，两个决策函数风险大小的比较，情况比较复杂，因此就产生了种种优良性准则。

定义 9

设 d_1, d_2 是统计问题中的两个决策函数，若其风险函数满足不等式

$$R(\theta, d_1) \leq R(\theta, d_2), \forall \theta \in \Theta$$

则称决策函数 d_1 优于 d_2。若不等号严格成立，则称决策函数 d_1 一致优于 d_2；若 $R(\theta, d_1) = R(\theta, d_2), \forall \theta \in \Theta$，则称 d_1, d_2 等价。

风险函数的比较：一致最小风险决策函数

定义 10

设 $D=\{d(\mathbf{x})\}$ 是一切定义在样本空间 $\mathcal{X}$ 上，取值于决策空间 $\mathcal{A}$ 上的决策函数全体，若存在一个决策函数 d^*，使对任意一个 $d(X)$ 都有

$$R(\theta, d^*) \leq R(\theta, d), \forall \theta \in \Theta, d, d^* \in D$$

则称 d^* 为一致最小风险决策函数，或一致最优决策函数。

风险函数的比较：两个例子

例 11

设总体 $X \sim N(\mu, 1), \mu \in (-\infty, +\infty)$，估计未知参数 μ。

解

选取损失函数为：$L(\mu, d) = (d-\mu)^2$ 则对 μ 的任一估计 $d(\mathbf{X})$，风险函数为

$$R(\mu, d) = E_\mu[L(\mu, d)] = E_\mu(d-\mu)^2$$

若要求 $d(\mathbf{X})$ 是无偏估计，即 $E_\mu(d(\mathbf{X})) = \mu$，则风险函数为：

$$R(\mu, d) = E_\mu(d - Ed)^2 = D_\mu(d(\mathbf{X}))$$

即风险函数为估计量 $d(\mathbf{X})$ 的方差。
若取 $d(\mathbf{X}) = \bar{\mathbf{X}}$，则 $R(\mu, d) = D\bar{\mathbf{X}} = \frac{1}{n}$
若取 $d(\mathbf{X}) = \mathbf{X}_1$，则 $R(\mu, d) = D\mathbf{X}_1 = 1$
显然，当 $n > 1$ 时，后者的风险比前者大，$\bar{\mathbf{X}}$ 优于 $\mathbf{X}_1$。

风险函数的比较：两个例子

例 12

设总体 $X \sim P(x;\lambda)$，估计未知参数 λ。

解

选取损失函数为：$L(\lambda, d) = (d-\lambda)^2$
则对 λ 的任一估计 $d(\mathbf{X})$，风险函数为

$$R(\lambda, d) = E_\lambda[L(\lambda, d)] = E_\lambda(d-\lambda)^2$$

若要 $d(\mathbf{X})$ 是无偏估计，即 $E_\lambda(d(\mathbf{X})) = \lambda$，则风险函数为：

$$R(\lambda, d) = E_\lambda(d - Ed)^2 = D_\lambda(d(\mathbf{X}))$$

若取 $d(\mathbf{X}) = \bar{\mathbf{X}}$，则 $R(\lambda, d) = D\bar{\mathbf{X}} = \frac{\lambda}{n}$
若取 $d(\mathbf{X}) = \mathbf{X}_1$，则 $R(\lambda, d) = D\mathbf{X}_1 = \lambda$
显然，当 $n > 1$ 时，风险不同。

风险函数的比较：存在的问题

在一个统计决策问题中，可供选择的决策函数往往很多，自然希望寻找使风险最小的决策函数，然而在这种意义下的最优决策函数往往是不存在的。这是因为：

1. 风险函数是二元函数，极值往往不存在或不唯一
2. 在某个区间内的逐点比较不现实（麻烦）
3. 对应不同参数的，同一决策函数，风险值不相等
4. 由统计规律的特性决定不能点点比较

因此必须由一个整体指标来代替点点比较。要解决这个问题，就要建立一个整体指标的比较准则。贝叶斯方法通过引进先验分布把两个风险函数的点点比较转化为用一个整体指标的比较来代替，从而可以决定优劣。贝叶斯风险和最大风险就是采用这种形式定义的。

25.4.2 贝叶斯估计

首先我们考虑贝叶斯风险。

定义 11

对于给定的统计决策问题，设 $d(\mathbf{x})$ 为该统计问题的决策函数，又设 $d(\mathbf{x})$ 的风险函数为 $R(\theta, d)(\theta \in \Theta)$，设参数 θ 的先验密度函数为 $\pi(\theta)(\theta \in \Theta)$。风险函数 $R(\theta, d)$ 的关于 θ 的期望

$$R_B(d) = E(R(\theta, d)) = \int_{\Theta} R(\theta, d)\pi(\theta)d\theta$$

称为决策函数 $d(\mathbf{x})$ 在给定先验分布 $\pi(\theta)$ 下的贝叶斯风险，简称 $d(\mathbf{x})$ 的贝叶斯风险。

贝叶斯规则：贝叶斯风险最小的决策规则

使贝叶斯风险最小的决策规则称为贝叶斯规则，相应的决策函数称为贝叶斯规则或贝叶斯决策。

定义 12

对于给定的统计决策问题，设总体 X 的分布函数 $F(x,\theta)$ 中参数 θ 为随机变量，$\pi(\theta)$ 为 θ 的先验分布，若在决策函数类 $\mathcal{A}$ 中存在一个决策函数 $d^*(\mathbf{x})$，使得对决策函数类 $\mathcal{A}$ 中的任一决策函数 $d(\mathbf{x})$，均有

$$R_B\left(d^*\right)=\inf_d R_B(d), \forall d \in \mathcal{A}$$

则称 $d^*(\mathbf{x})$ 是统计决策问题在先验分布 $\pi(\theta)$ 下的**贝叶斯规则或贝叶斯决策**。

贝叶斯风险的密度函数表达式

当总体 X 和 θ 都是连续型随机变量时，设 X 的概率密度函数为 $f(x;\theta)$，θ 的先验概率密度函数为 $\pi(\theta)$，记 $q(\mathbf{x} \mid \theta) = \Pi_{i=1}^{n} f(x_i;\theta)$（此即为样本密度），则

$$
\begin{aligned}
R_B(d) &= \int_{\Theta} R(\theta, d)\pi(\theta)\,d\theta \\
&= \int_{\Theta}\int_{\mathcal{X}} L(\theta, d(\mathbf{x}))\,q(\mathbf{x} \mid \theta)\pi(\theta)\,d\mathbf{x}\,d\theta \\
&= \int_{\Theta}\int_{\mathcal{X}} L(\theta, d(\mathbf{x}))\,m(\mathbf{x})h(\theta \mid \mathbf{x})\,d\mathbf{x}\,d\theta \\
&= \int_{\mathcal{X}} m(\mathbf{x}) \left\{ \int_{\Theta} L(\theta, d(\mathbf{x}))h(\theta \mid \mathbf{x})\,d\theta \right\} d\mathbf{x}
\end{aligned}
$$

其中 $m(\mathbf{x}) = \int_{\Theta} q(\mathbf{x} \mid \theta)\pi(\theta)\,d\theta$ 为 $(\mathbf{X}, \theta)$ 关于 $\mathbf{X}$ 的边缘联合密度函数。对于离散型随机变量：$R_B(d) = \sum_{\mathbf{x}} m(\mathbf{x}) \{\sum_{\theta} L(\theta, d(\mathbf{x}))h(\theta \mid \mathbf{x})\}$。

由上式可见，贝叶斯风险可以看作是对随机损失函数 $L(\theta, d(\mathbf{X}))$ 求两次数学期望而得到的，第一次先对 θ 的后验分布求数学期望，第二次是关于样本的边缘分布求数学期望。

后验风险

定义 13

设 $L(\theta, d)(\theta \in \Theta, d \in \mathcal{A})$ 为某一统计决策问题的损失函数，则称 $L(\theta, d)$ 对后验分布 $h(\theta \mid \mathbf{x})$ 的数学期望，记作

$$R(d \mid \mathbf{x}) = E(L(\theta, d)) = \int_{\Theta} L(\theta, d(\mathbf{x}))h(\theta \mid \mathbf{x})d\theta$$

为样本观测值为 $\mathbf{x}$ 时，决策 d 的**后验风险**。

贝叶斯估计

定理 2

任给 $\mathbf{x} \in \mathcal{X}$，若对任一 $d \in \mathcal{A}$，$R(d \mid \mathbf{x}) < +\infty$，又存在决策函数 $d_{\mathcal{X}}$，使得后验风险达到最小，即

$$R\left(d_{\mathcal{X}} \mid \mathbf{x}\right) = \min_{d \in \mathcal{A}} R(d \mid \mathbf{x})$$

则由下式定义的决策函数

$$d^*(\mathbf{x}) = d_{\mathcal{X}}, \quad \mathbf{x} \in \mathcal{X}$$

是在后验风险准则下的最优决策函数，称为贝叶斯决策函数或贝叶斯估计。

贝叶斯估计

证明.

设样本的分布为 $\{q(\mathbf{x} \mid \theta) : \theta \in \Theta\}$，参数 θ 的先验密度为 $\pi(\theta)$，$d(\mathbf{x})$ 为一决策函数，则 $d(\mathbf{x})$ 的贝叶斯风险为

$$\begin{aligned} R_B(d) = E_\pi[R(\theta, d)] &= \int_\Theta R(\theta, d)\pi(\theta)\,d\theta \\ &= \int_{\mathcal{X}} m(\mathbf{x}) \int_\Theta L(\theta, d(\mathbf{x}))h(\theta \mid \mathbf{x})\,d\theta\,d\mathbf{x} \\ &= \int_{\mathcal{X}} R(d \mid \mathbf{x})m(\mathbf{x})\,d\mathbf{x} \end{aligned} \tag{4}$$

由于，对任意的 $\mathbf{x} \in \mathcal{X}$，有

$$R\left(d^*(\mathbf{x}) \mid \mathbf{x}\right) = \min_{d \in \mathcal{A}} R(d \mid \mathbf{x}) \leq R(d(\mathbf{x}) \mid \mathbf{x})$$

从而有，$R_B\left(d^*\right) = \int_{\mathcal{X}} R\left(d^*(\mathbf{x}) \mid \mathbf{x}\right) m(\mathbf{x})\,d\mathbf{x} \leq \int_{\mathcal{X}} R(d(\mathbf{x}) \mid \mathbf{x})m(\mathbf{x})\,d\mathbf{x} = R_B(d)$

即 d^* 贝叶斯决策函数。 □

基于特定损失函数的贝叶斯估计：平方损失函数下的贝叶斯估计

下面给出各种损失函数下的贝叶斯估计。

定理 3

设 θ 的先验分布为 $\pi(\theta)$, 损失函数为 $L(\theta, d) = (\theta - d)^2$, 则 θ 的贝叶斯估计是

$$d^*(\mathbf{x}) = E(\theta \mid \mathbf{X} = \mathbf{x}) = \int_{\Theta} \theta \cdot h(\theta \mid \mathbf{x}) d\theta$$

其中 $h(\theta \mid \mathbf{x})$ 为参数 θ 的后验概率密度函数。

证明.

由于最小化

$$R_B(d) = \int_{\mathcal{X}} m(\mathbf{x}) \left\{ \int_{\Theta} [\theta - d(\mathbf{x})]^2 h(\theta \mid \mathbf{x}) d\theta \right\} d\mathbf{x}$$

与最小化 $\int_{\Theta} (\theta - d(\mathbf{x}))^2 h(\theta \mid \mathbf{x}) d\theta$ 等价。 □

平方损失函数下的贝叶斯估计

证明（续）.

而

$$\begin{aligned}\int_\Theta (\theta - d(\mathbf{x}))^2 h(\theta \mid \mathbf{x})\, d\theta &= \int_\Theta (\theta - E(\theta \mid \mathbf{x}) + E(\theta \mid \mathbf{x}) - d(\mathbf{x}))^2 h(\theta \mid \mathbf{x})\, d\theta \\ &= \int_\Theta (\theta - E(\theta \mid \mathbf{x}))^2 h(\theta \mid \mathbf{x})\, d\theta + \int_\Theta (E(\theta \mid \mathbf{x}) - d(\mathbf{x}))^2 h(\theta \mid \mathbf{x})\, d\theta. \\ &\quad + 2\int_\Theta (\theta - E(\theta \mid \mathbf{x}))(E(\theta \mid \mathbf{x}) - d(\mathbf{x})) h(\theta \mid \mathbf{x})\, d\theta\end{aligned}$$

其中 $E(\theta \mid \mathbf{x}) = \int_\Theta \theta \cdot h(\theta \mid \mathbf{x})\, d\theta$，又

$$\begin{aligned}\int_\Theta (\theta - E(\theta \mid \mathbf{x}))(E(\theta \mid \mathbf{x}) - d(\mathbf{x})) h(\theta \mid \mathbf{x})\, d\theta &= (E(\theta \mid \mathbf{x}) - d(\mathbf{x})) \int_\Theta (\theta - E(\theta \mid \mathbf{x})) h(\theta \mid \mathbf{x})\, d\theta \\ &= (E(\theta \mid \mathbf{x}) - d(\mathbf{x}))(E(\theta \mid \mathbf{x}) - E(\theta \mid \mathbf{x})) = 0\end{aligned}$$

□

平方损失函数下的贝叶斯估计

证明（续）.

所以

$$\begin{aligned}&\int_{\Theta}(\theta-d(\mathbf{x}))^2h(\theta\mid\mathbf{x})d\theta\\=&\int_{\Theta}(\theta-E(\theta\mid\mathbf{x}))^2h(\theta\mid\mathbf{x})d\theta+\int_{\Theta}(E(\theta\mid\mathbf{x})-d(\mathbf{x}))^2h(\theta\mid\mathbf{x})d\theta\end{aligned}$$

显然，当 $d^*(\mathbf{x})=E(\theta\mid\mathbf{x})$ 时，$R_B(d)$ 达到最小。 □

注：我们常说的贝叶斯估计是指平方损失函数下用后验分布的均值作为 θ 的点估计，也称为后验期望估计。

贝叶斯估计的一般求解步骤

求贝叶斯估计的一般步骤：

1. 根据总体 X 的分布，求得条件概率 $q(\mathbf{x} \mid \theta)$
2. 在已知 θ 的先验分布 $\pi(\theta)$ 下，求得 $\mathbf{X}$ 与 θ 的联合分布密度 $g(\mathbf{x}, \theta) = \pi(\theta) q(\mathbf{x} \mid \theta)$
3. 求得 X 的边缘分布 $m(\mathbf{x})$
4. 计算 $h(\theta \mid \mathbf{x}) = \pi(\theta) q(\mathbf{x} \mid \theta) / m(\mathbf{x})$
5. 求数学期望 $\hat{\theta} = \int_{\Theta} \theta \cdot h(\theta \mid \mathbf{x}) d\theta$
6. 求得贝叶斯风险（如果需要的话）

$$R_B(d) = \int_{\Theta} \int_{\mathcal{X}} L(\theta, d(\mathbf{x})) q(\mathbf{x} \mid \theta) \pi(\theta) d\mathbf{x} d\theta$$

正态分布的贝叶斯估计

例 13

$X_1, X_2, \ldots, X_n$ 来自正态分布 $N(\theta, \sigma_0^2)$ 的一个样本，其中 σ_0^2 已知，θ 未知，假设 θ 的先验分布为正态分布 $N(\mu, \tau^2)$，其中先验均值 μ 和先验方差 τ^2 均已知，试求 θ 的贝叶斯估计。

解

样本 $\mathbf{X}$ 的联合分布和 θ 的先验分布分别为

$$q(\mathbf{x} \mid \theta) = \left(2\pi\sigma_0^2\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}$$

$$\pi(\theta) = \left(2\pi\tau^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\tau^2}(\theta-\mu)^2\right\}$$

正态分布的贝叶斯估计

解

（续）由此可以写出 $\mathbf{x}$ 与 μ 的联合分布

$$f(\mathbf{x},\theta)=k_1\cdot\exp\left\{-\frac{1}{2}\left[\sigma_0^{-2}\left(n\theta^2-2n\theta\bar{\mathbf{x}}+\sum_{i=1}^{n}x_i^2\right)+\frac{\theta^2-2\theta\mu+\mu^2}{\tau^2}\right]\right\}$$

其中 $k_1=(2\pi)^{-(n+1)/2}\tau^{-1}\sigma_0^{-n}$

若记 $A=\frac{n}{\sigma_0^2}+\frac{1}{\tau^2},B=\frac{n\bar{x}}{\sigma_0^2}+\frac{\mu}{\tau^2},C=\sigma_0^{-2}\sum_{i=1}^{n}x_i^2+\frac{\mu^2}{\tau^2}$

则有

$$\begin{aligned}f(\mathbf{x},\theta)&=k_1\exp\left\{-\frac{1}{2}\left[A\theta^2-2B\theta+C\right]\right\}\\&=k_1\exp\left\{-\frac{(\theta-B/A)^2}{2/A}-\frac{1}{2}\left(C-B^2/A\right)\right\}\end{aligned}$$

正态分布的贝叶斯估计

解

（续）注意到 A, B, C 均与 θ 无关，样本的边际密度函数

$$m(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x}, \theta) d\theta = k_1 \exp\left\{-\frac{1}{2}\left(C - B^2/A\right)\right\} \cdot \sqrt{\frac{2\pi}{A}}$$

应用贝叶斯公式即可得到后验分布

$$h(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{m(\mathbf{x})} = \sqrt{\frac{A}{2\pi}} \exp\left\{-\frac{1}{2/A}(\theta - B/A)^2\right\}$$

这说明在样本给定后，θ 的后验分布为 $N(B/A, 1/A)$ ，即 $\theta \mid x \sim N(B/A, 1/A)$

正态分布的贝叶斯估计

解

(续) 记作 $\theta \mid \mathbf{x} \sim N\left(\mu_1, \sigma_1^2\right)$, 其中

$$\mu_1 = \frac{B}{A} = \frac{n\sigma_0^{-2}\bar{\mathbf{x}} + \tau^{-2}\mu}{n\sigma_0^{-2} + \tau^{-2}}, \sigma_1^2 = \frac{1}{A} = \frac{\sigma_0^2\tau^2}{\sigma_0^2 + n\tau^2}$$

后验均值即为其贝叶斯估计:

$$\hat{\theta} = \frac{n\tau^2}{n\tau^2 + \sigma_0^2}\bar{\mathbf{x}} + \frac{\sigma_0^2}{n\tau^2 + \sigma_0^2}\mu$$

它是样本均值 $\bar{\mathbf{x}}$ 与先验均值 μ 的加权平均。

基于特定损失函数的贝叶斯估计

定理 4

设 θ 的先验分布为 $\pi(\theta)$, 取损失函数为加权平方损失函数

$$L(\theta, d) = \lambda(\theta)(d-\theta)^2$$

则 θ 的贝叶斯估计为 $d^*(\mathbf{x}) = \frac{E[\lambda(\theta)\theta|\mathbf{x}]}{E[\lambda(\theta)|\mathbf{x}]}$

定理 5

设 $\theta\,(\theta_1, \theta_2, \ldots, \theta_p)^T$ 的先验分布为 $\pi(\theta)$ ， 损失函数为 $L(\theta, d) = (d-\theta)^T Q(d-\theta)$, Q 正定

则 θ 的贝叶斯估计为 $d^*(\mathbf{x}) = E(\theta \mid \mathbf{x}) = \begin{bmatrix} E(\theta_1 \mid \mathbf{x}) \\ \vdots \\ E(\theta_p \mid \mathbf{x}) \end{bmatrix}$

基于特定损失函数的贝叶斯估计

定理 6

设 θ 的先验分布为 $\pi(\theta)$, 在线性损失函数

$$L(\theta, d) = \begin{cases} k_0(\theta - d), d \le \theta \\ k_1(d - \theta), d > \theta \end{cases}$$

下, 则 θ 的贝叶斯估计 $d^*(\mathbf{x})$ 为后验分布 $h(\theta|\mathbf{x})$ 的 $k_1/(k_0 + k_1)$ 上侧分位数。

定理 7

设的先验分布为 $\pi(\theta)$ ，损失函数为绝对值损失 $L(\theta, d) = |d - \theta|$，则 θ 的贝叶斯估计 $d^*(\mathbf{x})$ 为后验分布 $h(\theta \mid \mathbf{x})$ 的中位数。

最大后验估计

定义 14

设 θ 的后验密度函数为 $h(\theta \mid \mathbf{x})$, 若 $\hat{\theta} = \hat{\theta}(x_1, x_2, \ldots, x_n)$ 使得

$$h(\hat{\theta} \mid \mathbf{x}) = \max_{\theta \in \Theta} h(\theta \mid \mathbf{x})$$

则称 $\hat{\theta}$ 为 θ 最大后验估计。

最大后验估计

例 14

设总体 $X \sim E(\theta), X_1, X_2, \ldots, X_n$ 为来自总体 X 的样本, θ 的先验分布为指数分布 $E(\lambda)$ (λ 已知), 求 θ 的最大后验估计。

解

因为先验概率密度函数为:

$$\pi(\theta) = \begin{cases} \lambda e^{-\lambda\theta} & \theta > 0 \\ 0 & \theta \leq 0 \end{cases}$$

样本 $(X_1, X_2, \ldots, X_n)$ 的联合概率密度为:

$$q(\mathbf{x} \mid \theta) = \prod_{n=2}^{n} f(x_i \mid \theta) = \begin{cases} \theta^n e^{-\theta \sum_{i=1}^{n} x_i} & x_1, x_2, \ldots, x_n > 0 \\ 0 & \text{其它} \end{cases}$$

最大后验估计

解

所以 θ 的后验分布密度

$$h(\theta \mid \mathbf{x}) \propto \theta^n e^{-\left(\lambda+\sum_{i=1}^n x_i\right)\theta}$$

$$\ln h(\theta \mid \mathbf{x}) = n \ln \theta - \left(\lambda + \sum_{i=1}^n x_i\right)\theta + \ln c(\mathbf{x})$$

$$\frac{\partial \ln h(\theta \mid \mathbf{x})}{\partial \theta} = \frac{n}{\theta} - \left(\lambda + \sum_{i=1}^n x_i\right) = 0$$

求得 θ 的最大后验估计为 $\hat{\theta} = \frac{1}{\bar{x}+\lambda/n}$。当 $n \to \infty$ 时，$\hat{\theta} \to \frac{1}{\bar{x}}$，与传统意义下的极大似然估计是一致的。

常用贝叶斯估计

基于后验分布 $h(\theta \mid \mathbf{x})$ 的贝叶斯估计，常用如下三种:

1. 用后验分布的密度函数最大值作为 θ 的点估计，称为最大后验估计;
2. 用后验分布的中位数作为 θ 的点估计，称为后验中位数估计;
3. 用后验分布的均值作为 θ 的点估计，称为后验期望估计。

用得最多的是后验期望估计，简称为贝叶斯估计，记为 $\hat{\theta}_B$

贝叶斯估计的误差

定义 15

设 θ 的后验分布为 $h(\theta \mid \mathbf{x})$，贝叶斯估计为 $\hat{\theta}$，则 $(\hat{\theta}-\theta)^2$ 的后验期望

$$\mathrm{MSE}(\hat{\theta} \mid \mathbf{x}) = E_{\theta/\mathbf{x}}(\hat{\theta}-\theta)^2$$

称为 $\hat{\theta}$ 的**后验均方差**，其平方根称为后验标准误差，$\hat{\theta}$ 的后验均方差越小，贝叶斯估计的误差就越小，当 $\hat{\theta}$ 为 θ 的后验期望 $\hat{\theta}_B = E(\theta \mid \mathbf{x})$ 时，

$$\mathrm{MSE}\left(\hat{\theta}_B \mid \mathbf{x}\right) = E_{\theta/\mathbf{x}}\left(\hat{\theta}_B-\theta\right)^2 = D(\theta \mid \mathbf{x})$$

称为**后验方差**，其平方根称为后验标准差。

贝叶斯估计的误差

后验均方差与后验方差，有如下关系：

$$\begin{aligned}\mathrm{MSE}(\hat{\theta} \mid \mathbf{x}) &= E_{\theta/\mathbf{x}}(\hat{\theta}-\theta)^2 \\ &= E_{\theta/\mathbf{x}}\left[\left(\hat{\theta}-\hat{\theta}_B\right)+\left(\hat{\theta}_B-\theta\right)\right]^2 \\ &= E_{\theta/\mathbf{x}}\left(\hat{\theta}_B-\hat{\theta}\right)^2 + D(\theta \mid \mathbf{x}) \\ &= \left(\hat{\theta}_B-\hat{\theta}\right)^2 + D(\theta \mid \mathbf{x})\end{aligned}$$

上面的关系式表明，当 $\hat{\theta}$ 取后验期望时，可使后验均方差达到最小，所以取后验期望作为 θ 的贝叶斯估计是合理的.

25.4.3 最小最大估计

在统计决策理论中，风险函数提供了一个衡量决策函数好坏的尺度。贝叶斯决策是根据贝叶斯风险最小的原则而取的最优决策，如果将“最优性”的准则改变，就可以得到另一种“最优”决策。接下来我们介绍基于最大风险的最小最大决策规则。与基于贝叶斯风险的贝叶斯决策相比，最小最大决策不依赖于参数的先验信息。

定义 16

对于给定的统计决策问题，设 $d(\mathbf{x})$ 为该统计问题的决策函数，又设 $d(\mathbf{x})$ 的风险函数为 $R(\theta, d)(\theta \in \Theta)$，称

$$M(d) = \sup_{\theta} R(\theta, d),$$

为决策函数 $d(\mathbf{x})$ 的最大风险。

最小最大规则

使最大风险最小的决策称为最小最大规则，相应的决策函数称为最小最大决策。

定义 17

对于给定的统计决策问题，若在决策函数类 $\mathcal{A}$ 中存在一个决策函数 $d^*(\mathbf{x})$，使得对决策函数类 $\mathcal{A}$ 中的任一决策函数 $d(\mathbf{x})$，均有

$$\sup_{\theta} R(\theta, d^*) = \inf_{d} \sup_{\theta} R(\theta, d), \forall d \in \mathcal{A}$$

则称 d^* 为最小最大（*Minmax*）决策，或称 d^* 为该统计问题的**最小最大规则**或**最小最大解**。

当问题为估计或检验时，称 d^* 为最小最大估计或最小最大检验。后面我们主要关注最小最大估计。

最小最大规则

- 最小最大规则从风险函数的整体性质来确定决策风险的优良性。使决策函数的最大风险达到最小是考虑到最不利的情况，要求最不利的情况尽可能地好。也就是人们常说的从最坏处着想，争取最好的结果。因此最小最大规则是比较保守的规则。
- 通常，如果对参数 θ 的先验信息有所了解，则利用贝叶斯为好；若对参数 θ 的信息毫无了解，则可使用最小最大准则。
- 寻求最小最大决策函数的一般步骤是：
  (1) 对 $\mathcal{A}$ 中所有决策函数求最大风险 $\max_{\Theta}(R(\theta, d)), \forall d \in D$
  (2) 在所有最大风险值中选取最小值 $\min_d(\max_{\Theta}(R(\theta, d)))$
  此最小值所对应的决策函数就是最小最大决策函数。

伯努利分布的最小最大估计

例 15

设总体 $X \sim B(1,p)$, **即**

$$P(X=x)=p^x(1-p)^{1-x}(x=0,1)$$

其中 $p \in \Theta=\left\{\frac{1}{4},\frac{1}{2}\right\}$, **试求参数** p **的最小最大估计量。**

伯努利分布的最小最大估计

解：决策空间为 $\mathcal{A}=\left\{\frac{1}{4},\frac{1}{2}\right\}$，设损失函数 $L(p,a)$ 为下表所示

表：损失函数 $L(p,a)$ 取值表

L(p, a) \ a / p	a_1	a_2
$p_1=\frac{1}{4}$	1	4
$p_2=\frac{1}{2}$	3	2

如果我们选取容量为 1 的样本为 X_1，由于 X_1 仅取两个可能值及 $\mathcal{A}$ 中只有两个元素，因而决策函数的集合 D 是由 4 个元素所组成，其分别记为 d_1,d_2,d_3,d_4，即有

伯努利分布的最小最大估计

表：决策函数表

X \ D	d_1	d_2	d_3	d_4
0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
1	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$

由上表计算可得

$$R(p_1, d_1) = L(p_1, a_1)\, P(X=0) + L(p_1, a_1)\, P(X=1) = 1 \times \frac{3}{4} + 1 \times \frac{1}{4} = 1$$
$$R(p_1, d_2) = L(p_1, a_2)\, P(X=0) + L(p_1, a_2)\, P(X=1) = 4 \times \frac{3}{4} + 4 \times \frac{1}{4} = 4$$
$$R(p_1, d_3) = L(p_1, a_1)\, P(X=0) + L(p_1, a_2)\, P(X=1) = 1 \times \frac{3}{4} + 4 \times \frac{1}{4} = \frac{7}{4}$$
$$R(p_1, d_4) = L(p_1, a_2)\, P(X=0) + L(p_1, a_1)\, P(X=1) = 4 \times \frac{3}{4} + 1 \times \frac{1}{4} = \frac{13}{4}$$

同理计算可得

$$R(p_2, d_1) = 3, \quad R(p_2, d_2) = 2, \quad R(p_2, d_3) = \frac{5}{2}, \quad R(p_2, d_4) = \frac{5}{2}$$

伯努利分布的最小最大估计

表：风险函数值与最大值

$d_1\,(x_1)$	d_1	d_2	d_3	d_4
$R\,(p_1, d_i)$	1	4	$\frac{7}{4}$	$\frac{13}{4}$
$R\,(p_2, d_i)$	3	2	$\frac{5}{2}$	$\frac{5}{2}$
$\max_{\theta\in\Theta} R\,(p, d_i)$	3	4	$\frac{5}{2}$	$\frac{13}{4}$

于是 p 的最小最大值估计为：

$$\hat{p}\,(X_1) = d_3 = \begin{cases} \frac{1}{4} & X_1 = 0 \\ \frac{1}{2} & X_1 = 1 \end{cases}$$

验证最小最大决策函数

寻找最小最大决策函数通常是较困难的，然而贝叶斯决策函数与最小最大决策函数有一定的联系。以下定理可以作为验证某一决策 d 为最小最大决策函数的方法。

定理 8

设 $d^*(\mathbf{x})$ 为某一先验分布 $\pi(\theta)$ 下的贝叶斯决策函数，且对任意的 $\theta \in \Theta$，$d^*(\mathbf{x})$ 的风险函数 $R(\theta, d^*) = c$ 为常数，则 $d^*(\mathbf{x})$ 为该统计决策问题的最小最大决策函数。

证明.

用反证法。若 $d^*(\mathbf{x})$ 不是最小最大决策函数，则存在决策函数 $d(\mathbf{x})$，使得 $M(d) < M(d^*) = \sup_{\theta\in\Theta} R(\theta, d^*) = c$，此时有

$$R_\pi(d) = \int_\Theta R(\theta, d)\pi(\theta)\,d\theta \le \int_\Theta M(d)\pi(\theta)\,d\theta < c = R_\pi(d^*)$$

这与 $d^*(\mathbf{x})$ 为先验分布 $\pi(\theta)$ 下的贝叶斯决策函数相矛盾。因此，$d^*(\mathbf{x})$ 必定是一个最小最大决策函数。 □

最小最大决策函数定理

定理 9

设给定一个贝叶斯决策问题，在先验分布 $\pi_k(\theta)$ 下的贝叶斯决策函数为 d_k，而 d_k 的贝叶斯风险为 $B_{\pi_k}(d_k)\,(k=1,2,\ldots)$ 。若

$$\lim_{k\to\infty} B_{\pi_k}(d_k) = \rho < +\infty$$

且 d^* 为一决策函数，满足

$$\sup_{\theta\in\Theta} R(\theta, d^*) \le \rho$$

则 d^* 为该统计决策问题的最小最大决策函数。

最小最大决策函数定理

证明.

用反证法。若 d^* 不是最小最大决策函数，则存在决策函数 d，使得

$$\sup_{\theta\in\Theta} R(\theta, d) < \sup_{\theta\in\Theta} R(\theta, d^*) \leq \rho$$

此时，存在 $\varepsilon > 0$，使得 $R(\theta, d) \leq \rho - \varepsilon, \forall\theta \in \Theta$，因此，对一切 k，有

$$B_{\pi_k}(d) = \int_\Theta R(\theta, d)\pi_k(\theta)d\theta \leq \rho - \varepsilon$$

由于 d_k 为在先验分布 $\pi_k(\theta)$ 下的叶斯决策函数，所以有

$$B_{\pi_k}(d) \geq B_{\pi_k}(d_k) > \rho - \varepsilon$$

矛盾，故 d^* 为最小最大决策函数。 □

正态分布的最小最大估计

例 16

设总体 X 服从正态分布 $N(\theta,1)$, $X_1,X_2,\ldots,X_n$ 为来自总体 X 的样本, 损失函数为 $L(\theta,d)=(\theta-d)^2$, 求 θ 的最小最大估计。

解

选取一列先验分布 $\{\pi_k\}$, $\pi_k(\theta)\sim N\left(0,k^2\right)$, 在 π_k 下, θ 的贝叶斯估计为

$$d_k=E(\theta\mid\mathbf{x})=\frac{n\bar{x}}{n+\frac{1}{k^2}}=\frac{k^2n\bar{x}}{k^2n+1}$$

由于

$$\begin{aligned}R\left(\theta,d_k\right)&=E_\theta L\left(\theta,d_k\right)=E_\theta\left[\frac{k^2n\bar{X}}{k^2n+1}-\theta\right]^2\\&=\frac{E_\theta\left[k^2n(\bar{X}-\theta)-\theta\right]^2}{\left(k^2n+1\right)^2}=\frac{k^4n+\theta^2}{\left(k^2n+1\right)^2}\end{aligned}$$

正态分布的最小最大估计

解

（续）所以 d_k 的贝叶斯风险为

$$B_{\pi_k}(d_k) = E_{\pi_k}\left[\frac{k^4 n + \theta^2}{(k^2 n + 1)^2}\right] = \frac{k^2}{k^2 n + 1}$$

因为 $\lim_{k\to\infty} B_{\pi_k}(d_k) = \frac{1}{n}$, 而取决策函数 $d^* = \bar{x}$, 则 d^* 的风险函数

$$R(\theta, d^*) = E_\theta L(\theta, d^*) = E_\theta[\bar{X} - \theta]^2 = \frac{1}{n}$$

从而 $\sup_{\theta\in\Theta} R(\theta, d^*) = \frac{1}{n}$, 于是 θ 的最小最大估计为 $d^* = \bar{x}$ 。

小结：极大似然估计、贝叶斯估计

- 极大似然估计：是把参数 θ 看成为确定的未知参数。然后求似然函数 $L(\theta)$ 为最大的 $\hat{\theta}$ 作为极大似然估计量

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \prod_{i=1}^{n} f(x_i \mid \theta)$$

- 贝叶斯估计：是把参数 θ 看成为随机的未知参数，一般 θ 具有先验分布 $\pi(\theta)$，然后通过似然函数 $q(\mathbf{x}|\theta)$ 和贝叶斯公式将 θ 的先验分布 $\pi(\theta)$ 转化为后验分布 $h(\theta|\mathbf{x})$。利用公式

$$h(\theta \mid \mathbf{x}) = \frac{q(\mathbf{x} \mid \theta)\pi(\theta)}{\int g(\mathbf{x} \mid \theta)\pi(\theta)d\theta}, \qquad \hat{\theta}_{\mathrm{B}} = E(\theta \mid \mathbf{x}) = \int_{\Theta} \theta h(\theta \mid \mathbf{x})d\theta$$

  求出贝叶斯估计量 $\hat{\theta}_{\mathrm{B}}$.

贝叶斯估计可用于贝叶斯学习：是利用 θ 的先验分布及样本提供的信息求出 θ 的后验分布 $h(\theta|\mathbf{x})$，然后直接求总体分布。

小结：极大似然估计与最大后验估计的关系

MAP 和 MLE 的关系：

- 当样本数趋于无穷时，最大后验概率估计一般趋向于极大似然估计。
- 极大似然估计也可看作参数的先验概率密度函数服从均匀分布（相当于没有先验知识）的最大后验概率估计。
- 当参数的先验概率密度函数比较准确时，最大后验概率估计的小样本性质大大优于极大似然估计。

小结：极大似然法和贝叶斯方法选择标准

- 标准一：方法的计算复杂度。此标准下选择极大似然法，因为 MLE 仅涉及一些微分运算或梯度搜索技术，而 Bayesian 要计算非常复杂的多重积分。
- 标准二：可理解性。MLE 比 Bayesian 更易理解和掌握，因为 MLE 结果是基于设计者所提供的训练样本的一个最佳答案，而 Bayesian 得到的结果则是许多可行解的加权平均，反映出对各种可行解的不确定程度。
- 标准三：对初始先验知识的信任程度，比如对概率密度函数 $q(\mathbf{x}|\theta)$ 的形式。

小结：极大似然法和贝叶斯方法选择标准

- 总之，通过使用全部 $h(\theta|\mathbf{x})$ 中的信息，Bayesian 方法比 MLE 法能够利用更多有用的信息。如果这些信息可靠，有理由认为 Bayesian 比 MLE 能够得到更准确的结果。
- 在没有特别先验知识（如均匀分布）情况下，二种方法比较相似。
- 若有非常多的训练样本，使 $h(\theta|\mathbf{x})$ 形成一个非常显著的尖峰，而先验概率 $\pi(\theta)$ 又是均匀分布，从本质上来说，MLE 和 Bayesian 相同。
- 若 $h(\theta|\mathbf{x})$ 波形比较宽，或者在 $\hat{\theta}$ 附近是不对称的（此不对称由问题本身决定），MLE 和 Bayesian 产生的结果就不相同。非常明显的不对称性显然表示了分布本身的某些特点。Bayesian 能够利用这些特点，而 MLE 却忽略这些特点。

小结：极大似然估计、贝叶斯估计和最小最大估计

- 在绝大多数大样本参数模型中，MLE 近似最小最大估计。
- 常值风险函数的贝叶斯估计就是最小最大估计。

除此之外，

- 概率模型有时既含有观测变量 (observable variable)，又含有隐变量或潜在变量 (latent variable)。
- 如果概率模型的变量都是观测变量，那么给定数据，可以直接用极大似然估计法或贝叶斯估计法估计模型参数。但是，当模型还有隐变量时，就不能简单地使用这些估计方法。
- 可以通过使用 EM 算法 (期望极大法)，也就是含有隐变量的概率模型参数的极大似然估计法或极大后验概率估计法，来进行估计。

引言

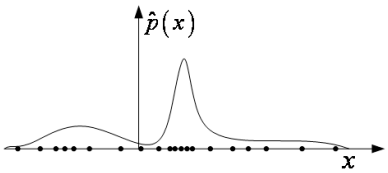
参数估计要求总体的密度函数的形式已知，但这种假定有时并不成立；常见的一些函数形式很难拟合实际的概率密度，实际中样本维数较高，且关于高维密度函数可以表示成一些低维密度函数乘积的假设通常也不成立；经典的密度函数都是单峰的，而在许多实际情况中却是多峰的，即有多个局部极大值。

但是为了设计贝叶斯分类器，仍然需要总体分布的知识，于是提出一些直接用样本来估计总体分布的方法，称之为：估计分布的非参数方法。

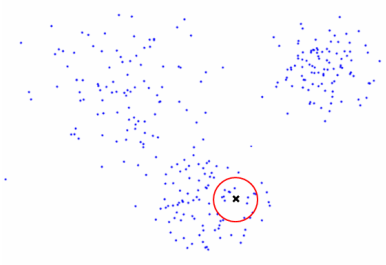
- 非参数估计：密度函数的形式未知，也不作假设，利用训练数据（样本）直接对任意的概率密度进行估计。又称作模型无关方法。

回顾：概率密度估计问题

概率密度估计问题：
给定 i.i.d. 样本集：$X = \{X_1, X_2, \cdots, X_n\}$
估计概率分布：$p(X)$



(a)



(b)

图 1: 概率密度估计

非参估计方法

非参概率密度估计方法：

- 直方图密度估计
- 核密度估计
- k 近邻估计

25.5.1 直方图密度估计

- 在经典的统计学中，直方图主要用于描述数据的频率。本节将介绍如何使用直方图估计一个随机变量的密度。
- 直方图密度估计与用直方图估计频率的差别在于：在直方图密度估计中，我们需要对频率估计进行归一化，使其成为一个密度函数的估计。
- 直方图估计是非参数概率密度估计最简单的方法。

1. 一元函数的直方图密度估计

先讨论一元函数的直方图密度估计。

- 假定有数据 $x_1, x_2, \cdots, x_n \in [a, b)$。对区间 $[a, b)$ 做如下划分，即 $a = a_0 < a_1 < a_2 < \cdots < a_k = b, I_i = [a_{i-1}, a_i), i = 1, 2, \cdots, k.$ 我们有 $\bigcup_{i=1}^{k} I_i = [a, b), I_i \cap I_j = \varnothing, i \neq j.$ 令 $n_i = \#\{x_i \in I_i\}$ 为落在 I_i 中数据的个数.
- 定义直方图密度估计为：

$$\hat{p}(x) = \begin{cases} \frac{n_i}{n(a_i - a_{i-1})}, & \text{当} x \in I_i; \\ 0, & \text{当} x \notin [a, b), \end{cases}$$

带宽或窗宽

- 在实际操作中，经常取相同的区间，即 $I_i(i=1,2,\cdots,k)$ 的宽度均为 h，在此情况下，有

$$\hat{p}(x)=\begin{cases}\frac{n_i}{nh}, & \text{当} x\in I_i;\\ 0, & \text{当} x\notin[a,b).\end{cases}$$

  上式中 h 既是归一化参数，又表示每一组的组距，称为带宽或窗宽。

- 另外，我们可以看到

$$\int_a^b \widehat{p}(x)\mathrm{d}x=\sum_{i=1}^{k}\int_{I_i} n_i/(nh)\mathrm{d}x=\sum_{i=1}^{k} n_i/n=1.$$

- 由于位于同一组内所有点的直方图密度估计均相等，因而直方图所对应的分布函数 $\hat{F}_h(x)$ 是单调增的阶梯函数。这与经验分布函数形状类似。实际上，当分组间隔 h 缩小到每组中最多只有一个数据时，直方图的分布函数就是经验分布函数，即 $h\to 0$，有 $\hat{F}_h(x)\to\hat{F}_n(x)$。

期望和方差

定理 10

固定 x 和 h, 令估计的密度是 $\hat{p}(x)$, 如果 $x \in I_j, p_j = \int_{I_j} \hat{p}(x)\mathrm{d}x$, 有

$$E\hat{p}(x) = p_j/h, \quad \operatorname{var} \hat{p}(x) = \frac{p_j\left(1-p_j\right)}{nh^2}$$

证明提示: 注意到 $E\hat{p}_j = n_j/n = \int_{I_j} \hat{p}(x)\mathrm{d}x, \operatorname{var} \hat{p}_j = p_j\left(1-p_j\right)/n.$

例 17

下面使用了鸢尾花数据集中的山鸢尾（*Setosa*）和维吉尼亚鸢尾（*Virginical*）两种花花萼长度的观测数据，共计 *150* 条．在下图中，我们从左到右，分别采用逐渐增加的带宽间隔：$h_l = 0.40$, $h_m = 0.19$, $h_r = 0.09$ 制作了 *3* 个直方图．可以发现当带宽很小的时候，个体特征比较明显，从图中可以看到多个峰值；而带宽过大的最左边的图上，很多峰都不明显了．中间的图比较合适，它有两个主要的峰，提供了最为重要的特征信息．实际上，参与直方图运算的是山鸢尾和维吉尼亚鸢尾两种花花萼长度的混合数据，经验表明，大部分山鸢尾的花萼长度与维吉尼亚鸢尾的花萼长度有一定的差别，因而两个峰是合适的．

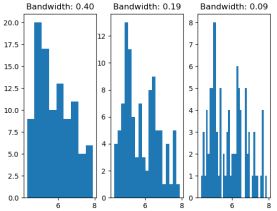


图 2: 山鸢尾和维吉尼亚鸢尾花萼长度

理论性质和最优带宽

由于带宽的不同，会得到不同的估计结果。因此选择合适的带宽，对于得到好的密度估计非常重要。在计算最优带宽前，我们先定义 $\hat{p}$ 的平方损失风险：

$$R(\hat{p}, p) = \int (\hat{p}(x) - p(x))^2 \, \mathrm{d}x.$$

定理 11

假设 $\int p'(x)\mathrm{d}x < +\infty$，则在平方损失风险下，有

$$R(\hat{p}, p) \approx \frac{h^2}{12} \int (p'(u))^2 \, \mathrm{d}u + \frac{1}{nh}.$$

极小化上式，得到理想带宽为

$$h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int p'(x)^2 \, \mathrm{d}x} \right)^{1/3}$$

于是理想的带宽为 $h = Cn^{-1/3}$.

理论性质和最优带宽

- 在大多数情况下，我们不知道密度 $p(x)$，因此也不知道 $p'(x)$. 对于理想带宽 $h^* = \frac{1}{n^{1/3}} \left(\frac{6}{\int p'(x)^2 \ \mathrm{d}x} \right)^{1/3}$ 也无法计算，在实际操作中，经常假设 $p(x)$ 为一个标准正态分布，并进而得到一个带宽 $h_0 \approx 3.5 n^{-1/3}$.
- 直方图密度估计的优势在于简单易懂，在计算过程中也不涉及复杂的模型计算，只需要计算 I_j 中样本点的个数. 另一方面，直方图密度估计只能给出一个阶梯函数，该估计不够光滑. 另外一个问题是直方图密度估计的收敛速度比较慢，也就是说，$\hat{p}(x) \longrightarrow p(x)$ 比较慢.

2. 多维直方图

下面我们扩展一维直方图的密度定义公式到任意维空间。

- 设有 n 个观测点 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$，将空间分成若干小区域 R，V 是区域 R 所包含的体积。如果有 k 个点落入 R，则可以得到如下密度公式：$p(x)$ 的估计为

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

- 如果这个体积和所有的样本体积相比很小，就会得到一个很不稳定的估计，这时，密度值局部变化很大，呈现多峰不稳定的特点；反之，如果这个体积太大，则会圈进大量样本，从而使估计过于平滑。

如何平衡不稳定与过度光滑产生两种可能的解决方法：核估计和 k 近邻估计。核估计法的总体思想：

- 固定体积 V 不变，它与样本总数呈反比关系即可。注意到，在直方图密度估计中，每一点的密度估计只与它是否属于某个 I_i 有关，而 I_i 是预先给定的与该点无关的区域。不仅如此，区域 I_i 中每个点共有相等的密度，这相当于待估点的密度取邻域 R 的平均密度。
- 现在以待估点为中心，作体积为 V 的邻域，令该点的密度估计与纳入该邻域中的样本点的多少呈正比，如果纳入的点多，则取密度大，反之亦然。
- 这一点还可以进一步扩展开去，将密度估计不再局限于 R 内的带内，而是将体积 V 合理拆分到所有样本点对待估计点贡献的加权平均，同时保证距离远的点取较小的权，距离近的点取较大的权，这样就形成了核函数密度估计法的基本思想。后面我们将看到，这些方法都可能获得较为稳健而适度光滑的估计。

k 近邻估计的总体思想：

- 固定 k 值不变，它与样本总数呈一定关系即可。
- 根据数据之间的疏密情况调整 V，这样就导致了另外一种密度估计方法—k 近邻法。

下面介绍核估计和 k 近邻估计两种非参数方法。

25.5.2 核密度估计：一维情形

直方图是不连续的。核密度估计较光滑且比直方图估计较快地收敛到真正的密度。先考虑一维的情况。

定义 18

假设数据 $x_1, x_2, \cdots, x_n$ 取自连续分布 $p(x)$，在任意点 x 处的一种核密度估计定义为

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\omega_i = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x - x_i}{h}\right) \tag{5}$$

其中 $h > 0$，称作为带宽；$K(\cdot)$ 称为核函数（*kernel function*）并且满足

$$K(x) \geqslant 0, \quad \int K(x)\mathrm{d}x = 1.$$

核函数的基本概念

定义中关于核函数 K 的分布密度的约束可以保证 $\hat{p}(x)$ 作为概率密度函数的合理性，也即其值非负并且积分结果为 1。实际上，容易验证有

$$\int \hat{p}(x)\mathrm{d}x = \int \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{x-x_i}{h}\right)\mathrm{d}x = \frac{1}{n}\sum_{i=1}^{n}\int \frac{1}{h}K\left(\frac{x-x_i}{h}\right)\mathrm{d}x$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int K(u)\mathrm{d}u = \frac{1}{n}\cdot n = 1 \quad \left(\text{其中} u = \frac{x-x_i}{h}\right).$$

因此上述定义的 $\hat{p}(x)$ 是一个合理的密度估计函数。

常用的核函数

核密度估计中，一个重要的部分就是核函数。以一维为例，常用的核函数如表所示。

核函数名称	核函数$K(u)$
Parzen 窗 (Uniform)	$\frac{1}{2}I(\|u\| \leqslant 1)$
三角 (Triangle)	$(1-\|u\|)I(\|u\| \leqslant 1)$
Epanechikov	$\frac{3}{4}\left(1-u^2\right)I(\|u\| \leqslant 1)$
四次 (Quartic)	$\frac{15}{16}\left(1-u^2\right)I(\|u\| \leqslant 1)$
三权 (Triweight)	$\frac{35}{32}\left(1-u^2\right)^3 I(\|u\| \leqslant 1)$
高斯 (Gauss)	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}u^2\right)$
余弦 (Cosinus)	$\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)I(\|u\| \leqslant 1)$
指数 (Exponent)	$\exp\{\|u\|\}$

核函数的基本概念

例 18

下图给出了各种带宽之下根据正态核函数做出的密度估计曲线. 由图可知，带宽 $h = 0.40$ 是最平滑的（左边），相反带宽 $h = 0.09$ 噪声很多，它在密度中引入了很多虚假的波形. 从图中比较，带宽 $h = 0.19$ 是较为理想的，它在不稳定和过于平滑之间作了较好的折中.

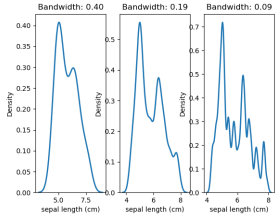


图 3: 山鸢尾和维吉尼亚鸢尾的花萼长度密度核估计

带宽的重要性

为了构造一个核密度估计，需要选择一个核函数 K 和一个带宽 h。理论和经验表明 K 的选择不是关键的，但是带宽 h 的选择非常重要。带宽对模型光滑程度的影响作用较大.

- 如果 h 非常大，将有更多的点对 x 处的密度产生影响. 由于分布是归一化的，即

$$\int \omega_i (x - x_i) \, \mathrm{d}x = \int \frac{1}{h} K\left(\frac{x - x_i}{h}\right) \mathrm{d}x = \int K(u) \mathrm{d}u = 1,$$

  因而距离 x_i 较远的点也分担了对 x 的部分权重，从而较近的点的权重 ω_i 减弱，距离远和距离近的点的权重相差不大. 在这种情况下，$\hat{p}(x)$ 是 n 个变化幅度不大的函数的叠加，因此 $\hat{p}(x)$ 非常平滑。
- 反之，如果 h 很小，则各点之间的权重由于距离的影响而出现大的落差，因而 $\hat{p}(x)$ 是 n 个以样本点为中心的尖脉冲的叠加，就好像是一个充满噪声的估计.

核估计与带宽关系定理

如何选择合适的带宽，是核函数密度估计能够成功应用的关键. 通过分析密度估计与真实密度之间的均方误差，有如下定理：

定理 12

假设 $\hat{p}(x)$ 定义如式(5)，是 $p(x)$ 的核估计，令 $\operatorname{supp}(p) = \{x : p(x) > 0\}$ 是密度 p 的支撑. 设 $x \in \operatorname{supp}(p) \subset \mathbb{R}$ 为 $\operatorname{supp}(p)$ 的内点 (非边界点)，当 $n \to +\infty$ 时，$h \to 0, nh \to +\infty$，核估计有如下性质：

$$\operatorname{Bias}(x) = \frac{h^2}{2}\mu_2(K)p^{(2)}(x) + O\left(h^2\right)$$

$$\mathrm{V}(x) = (nh)^{-1}p(x)R(K) + O\left((nh)^{-1}\right) + O\left(n^{-1}\right)$$

若 $\sqrt{(nh)}h^2 \longrightarrow 0$，则

$$\sqrt{(nh)}\left(\hat{p}_n(x) - p(x)\right) \longrightarrow N(0, p(x)R(K))$$

其中 $R(K) = \int K(x)^2 \, \mathrm{d}x$.

核估计与带宽关系定理

- 从均方误差的偏差和方差分解来看，带宽 h 越小，核估计的偏差越小，但核估计的方差越大；
- 反之，带宽 h 增大，则核估计的方差变小，但核估计偏差却增大.
- 所以，带宽 h 的变化不可能一方面使核估计的偏差减小，同时又使核估计的方差减小.
- 因而，最佳带宽选择的标准必须在核估计的偏差和方差之间作一个权衡，使积分均方误差达最小.

最优带宽

实际上，由上述定理，我们可以得到渐近积分均方误差 (AMISE)

$$\frac{h^4}{4}\mu_2^2 \int p^{(2)}(x)^2 \, \mathrm{d}x + n^{-1}h^{-1} \int K(x)^2 \, \mathrm{d}x,$$

由此可知，最优带宽为

$$h_{\text{opt}} = \mu_2(K)^{-4/5} \left\{ \int K(x)^2 \, \mathrm{d}x \right\}^{1/5} \left\{ \int p^{(2)}(x)^2 \, \mathrm{d}x \right\}^{-1/5} n^{-1/5}.$$

对于上式中的最优带宽，核函数 $K(u)$ 是已知的，但是密度函数 $p(x)$ 是未知的. 在实际操作中，我们经常把 $p(x)$ 看成正态分布去求解，即 $\int p^{(2)}(x)^2 \, \mathrm{d}x = \frac{3}{8}\pi^{-1/2}\sigma^{-5}$，这样，对于不同的核函数，我们可以得到相应的最优带宽. 例如当核函数是高斯时，可以得到 $\mu_2 = 1, \int K(u)^2 \, \mathrm{d}u = \int \frac{1}{2\pi} \exp\left(-u^2\right) \mathrm{d}u = \pi^{-1/2}$，这样，最优带宽就是 $h_{\text{opt}} = 1.06\sigma n^{-1/5}$.

最优带宽的递推方法

除了上述的方法，从实际计算的角度，Rudemo(1982) 和 Bowman(1984) 提出用交叉验证法确定最终带宽的递推方法. 具体来说，考虑积分平方误差

$$\mathrm{ISE}(h) = \int (\hat{p}(x) - p(x))^2 \,\mathrm{d}x = \int \hat{p}^2 \,\mathrm{d}x + \int p^2 \,\mathrm{d}x - 2\int \hat{p}p \,\mathrm{d}x$$

达到最小，将右边展开，因此这等价于最小化式:

$$\mathrm{ISE}(h)_{\mathrm{opt}} = \int \hat{p}^2 \,\mathrm{d}x - 2\int \hat{p}p \,\mathrm{d}x.$$

注意到等式的第二项为 $\int \hat{p}p \,\mathrm{d}x = E(\hat{p})$，因此，可以用 $\int \hat{p}p \,\mathrm{d}x$ 的一个无偏估计 $n^{-1}\sum_{i=1}^{n} \hat{p}_{-i}(X_i)$ 来估计，其中 $\hat{p}_{-i}$ 是将第 i 个观测点剔除后的概率密度估计. 下面只要估计第一项即可.

最优带宽的递推方法

将核估计定义式代入第一项，不难验证：

$$\begin{aligned}\int \hat{p}^2 \, \mathrm{d}x &= n^{-2}h^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_x K\left(\frac{X_i - x}{h}\right) K\left(\frac{X_j - x}{h}\right)\mathrm{d}x \\ &= n^{-2}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_t K\left(\frac{X_i - X_j}{h} - t\right) K(t)\mathrm{d}t\end{aligned}$$

于是，$\int \hat{p}^2 \, \mathrm{d}x$ 可用 $n^{-2}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} K \cdot K\left(\frac{X_i - X_j}{h}\right)$ 估计，其中 $K \cdot K(u) = \int_t K(u-\mathrm{t})K(t)\mathrm{d}t$ 是卷积.

最优带宽的递推方法

所以, Rudemo 和 Bowman 提出的交叉验证法 (cross validation) 实际上是选择 h 使下一步

$$\mathrm{ISE}(h)_1 = n^{-2}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n} K\cdot K\left(\frac{X_i - X_j}{h}\right) - 2n^{-1}\sum_{i=1}^{n}\hat{p}_{-i}(X_i)$$

达到最小. 当 K 是标准正态密度函数时, $K\cdot K$ 是 $N(0,2)$ 密度函数, 有

$$\begin{aligned}\mathrm{ISE}(h)_1 =& \frac{1}{2\sqrt{\pi}n^2h}\sum_i\sum_j \exp\left[-\frac{1}{4}\left(\frac{X_i - X_j}{h}\right)^2\right] \\ &- \frac{2}{\sqrt{2\pi}n(n-1)h}\sum_i\sum_{j\neq i}\exp\left[-\frac{1}{2}\left(\frac{X_i - X_j}{h}\right)^2\right].\end{aligned}$$

2. 多维密度估计

前面考虑的是一维情况下的核密度估计，下面考虑多维情形.

定义 19

假设数据 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ 是 d 维向量，并取自一个连续分布 $p(\boldsymbol{x})$，在任意点 $\boldsymbol{x}$ 处的一种核密度估计定义为

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

其中 h 是带宽，K 是定义在 d 维空间上的核函数，即 $K : \mathbb{R}^d \to \mathbb{R}$，并满足如下条件：

$$K(\boldsymbol{x}) \geqslant 0, \quad \int K(\boldsymbol{x}) \mathrm{d}\boldsymbol{u} = 1.$$

类似于一维情况，可以证明 $\int_{\mathbb{R}^d} \widehat{p}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1$，即 $\widehat{p}(\boldsymbol{x})$ 是一个密度估计.

常用的多维核函数

对于核函数的选择，我们经常选取对称的多维密度函数来作为核函数. 例如可以选取多维标准正态密度函数来作为核函数，$K_n(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}/2\right)$. 其他常用的核函数还有

- $K_2(\boldsymbol{x}) = 3\pi^{-1}\left(1 - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right)^2 I\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} < 1\right)$
- $K_3(\boldsymbol{x}) = 4\pi^{-1}\left(1 - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right)^3 I\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} < 1\right)$
- $K_e(\boldsymbol{x}) = \frac{1}{2}c_d^{-1}(d+2)\left(1 - \boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}\right) I\left(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x} < 1\right)$.

K_e 被称为多维 Epanechinikov 核函数，其中 c_d 是一个和维度有关的常数，$c_1 = 2$，$c_2 = \pi$，$c_3 = 4\pi/3$.

不同的带宽参数

上述的多维核密度估计中，我们只使用了一个带宽参数 h，这意味着在不同方向上，我们取的带宽是一样的. 事实上，我们可以对不同方向取不同的带宽参数，即

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh_1 \cdots h_d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{x}_i}{\boldsymbol{h}}\right)$$

其中，$\boldsymbol{h} = (h_1, h_2, \cdots, h_d)$ 是一个 d 维向量. 在实际数据中，有时候一个维度上的数据比另外一个维度上的数据分散得多，这个时候上述的核函数就有用了. 比如说数据在一个维度上分布在 $(0, 100)$ 区间上，而在另一个维度上仅仅分布在区间 $(0, 1)$ 上，这时候采用不同带宽的多维核函数就比较合理了.

例 19

下例是鸢尾花数据集中的数据，它包含 *150* 对数据，分别为鸢尾花数据集花萼长度和花瓣长度. 我们以此数据估计花萼长度和花瓣长度的联合密度函数.

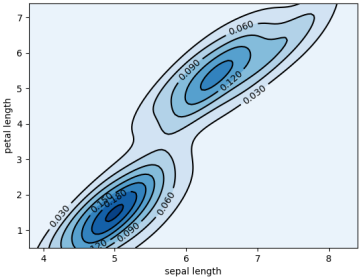


图 4: 鸢尾花数据集花萼长度和花瓣长度

最优带宽

关于最优带宽的选择，我们也有类似一维情况下的结论. 对于多维核密度估计，利用多维泰勒展开，有

$$\mathrm{Bias}(\boldsymbol{x}) \approx \frac{1}{2}h^2\alpha\nabla^2 p(\boldsymbol{x}),$$
$$V(\widehat{p}(\boldsymbol{x})) \approx n^{-1}h^{-d}\beta p(\boldsymbol{x}).$$

其中，$\alpha = \int \boldsymbol{x}^2 K(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \beta = \int K(\boldsymbol{x})^2\ \mathrm{d}\boldsymbol{x}$. 因此我们可以得到渐进积分均方误差

$$\mathrm{AMISE} = \frac{1}{4}h^4\alpha^2\int \nabla^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + n^{-1}h^{-d}\beta.$$

由此可得最优带宽为

$$h_{\mathrm{opt}} = \left\{d\beta\alpha^{-2}\left(\int \nabla^2 p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\right\}^{1/(d+4)} n^{-1/(d+4)}.$$

最优带宽

在上述的最优带宽中，真实密度 $p(\boldsymbol{x})$ 是未知的，因此我们可以采用多维正态密度 $\phi(\boldsymbol{x})$ 来代替，进而得到

$$h_{\text{opt}} = A(K)n^{-1/(d+4)},$$

其中 $A(K) = \left\{d\beta\alpha^{-2}\left(\int \nabla^2\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\right)\right\}^{1/(d+4)}$. 对于 $A(K)$，在知道估计中的核函数类型后，可以计算出来，并进而得到最优带宽 h_{opt}. 以下是不同核函数的 $A(K)$：

Kernel	Dimensionality	$A(K)$
K_n	2	1
K_n	d	$\{4/(d+2)\}^{1/(d+4)}$
K_e	2	2.40
K_e	3	2.49
K_e	d	$\left\{8c_d^{-1}(d+4)(2\sqrt{\pi})\right\}^{1/(d+4)}$
K_2	2	2.78
K_3	2	3.12

3. 贝叶斯决策和非参数估计

在机器学习领域，分类是一个基本的任务。在统计学中，分类被看成一个决策。分类决策是对一个概念的归属作决定的过程。一个分类框架一般由 4 项基本元素构成.

1. 参数集: 概念所有可能的不同自然状态. 在分类问题中, 自然参数是可数个, 用 $\theta = \{\theta_0, \theta_1, \cdots\}$ 表示.
2. 决策集: 所有可能的决策结果 $\mathcal{A} = \{a\}$. 比如: 买或卖、是否癌症、是否为垃圾邮件, 在分类问题中, 决策结果就是决策类别的归属, 所以决策集与参数集往往是一致的.
3. 决策函数集: $\Delta = \{\delta\}$, 函数 $\delta : \theta \to \mathcal{A}$.
4. 损失函数: 联系于参数和决策之间的一个损失函数. 如果概念和参数都是有限可数的, 那么所有的概念和相应的决策所对应的损失就构成了一个矩阵.

贝叶斯决策和非参数估计

例 20

两类问题中，真实的参数集为 θ_1 和 θ_0 (分别简记为 *1* 或 *0*)，可能的决策集由 *4* 个可能的决策构成 $\Delta = \{\delta_{1,1}, \delta_{0,0}, \delta_{0,1}, \delta_{1,0}\}$. 其中，$\delta_{i,j}$ 表示把 i 判为 $j, i, j = 0, 1$，相应的损失矩阵可能为

$$\boldsymbol{L} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

这表示判对没有损失，判错有损失. 真实的情况为 *1* 判为 *0*，或真实的情况为 *0* 判为 *1*，则发生损失 *1*，称为 “$0-1$” 损失.

贝叶斯决策和非参数估计

从分布的角度来看，分类问题本质上是概念属性分布的辨识问题，于是可能通过密度估计回答概念归属的问题. 以两类问题为例: 真实的参数集为 θ_1 和 θ_0，在没有观测之前，对 θ_1 和 θ_0 的决策函数可以应用先验 $p(\theta_1)$ 和 $p(\theta_0)$ 确定，即定义决策函数

$$\delta = \begin{cases} \theta_1, & p(\theta_1) > p(\theta_0), \\ \theta_0, & p(\theta_1) < p(\theta_0). \end{cases}$$

贝叶斯决策和非参数估计

很多情况下，我们对概念能够收集到更多的观测数据，于是可以建立类条件概率密度 $p(x \mid \theta_1)$, $p(x \mid \theta_0)$. 显然，两个不同的概念在一些关键属性上一定存在差异，这表现为两个类别在某些属性上面分布呈现差异. 综合先验信息，可以对类别的归属通过贝叶斯公式重新组织，即

$$p(\theta_1 \mid x) = \frac{p(x \mid \theta_1)\, p(\theta_1)}{p(x)}, \quad p(\theta_0 \mid x) = \frac{p(x \mid \theta_0)\, p(\theta_0)}{p(x)}.$$

根据贝叶斯公式，我们可以通过后验分布制定决策:

$$\delta = \begin{cases} \theta_1, & p(\theta_1 \mid x) > p(\theta_0 \mid x), \\ \theta_0, & p(\theta_1 \mid x) < p(\theta_0 \mid x). \end{cases}$$

注意到后验概率比较中，本质的部分是分子，所以上式等价于

$$\delta = \begin{cases} \theta_1, & p(x \mid \theta_1)\, p(\theta_1) > p(x \mid \theta_0)\, p(\theta_0), \\ \theta_0, & p(x \mid \theta_1)\, p(\theta_1) < p(x \mid \theta_0)\, p(\theta_0). \end{cases}$$

贝叶斯决策和非参数估计

定理 13

后验概率最大化分类决策是“$0-1$”损失下的最优风险.

证明.

注意到条件风险

$$R(\theta_1 \mid x) = p(\theta_0 \mid x) L(\theta_0, \theta_1) + p(\theta_1 \mid x) L(\theta_1, \theta_1) = 1 - p(\theta_1 \mid x).$$

□

注：上述定理很容易扩展到 $k, k \geqslant 3$ 个不同的分类。

贝叶斯决策和非参数估计

于是给出如下的非参数核密度估计分类计算步骤（后验分布构造贝叶斯分类）：

1. $\forall i = 1, 2, \cdots, k, \theta_i$ 下观测 $x_{i1}, x_{i2}, \cdots, x_{in} \sim p(x \mid \theta_i)$;
2. 估计 $p(\theta_i), i = 1, 2, \cdots, k$;
3. 估计 $p(x \mid \theta_i), i = 1, 2, \cdots, k$;
4. 对新待分类点 x, 计算 $p(x \mid \theta_i) p(\theta_i)$;
5. 计算 $\theta^* = \operatorname{argmax}\{p(x \mid \theta_i) p(\theta_i)\}$.

分类的例子

例 21

根据核密度估计贝叶斯分类对前面例题中提及的两类鸢尾花进行分类。

解

假设 θ_0 表示山鸢尾，θ_1 表示维吉尼亚鸢尾，记两类花的先验分布为

$$\text{山鸢尾：}\ \hat{p}(\theta_0) \leftrightarrow \text{维吉尼亚鸢尾：}\ \hat{p}(\theta_1).$$

用两类分别占用全部数据的频率估计先验概率。在本例中，由于山鸢尾和维吉尼亚鸢尾各为一半，两类的先验概率分别估计为 $\hat{p}(\theta_0) = \hat{p}(\theta_1) = 0.5$。然后，对每一类考虑用核概率密度估计类条件概率：　　山鸢尾：$\hat{p}(x \mid \theta_0) \leftrightarrow$ 维吉尼亚鸢尾：$\hat{p}(x \mid \theta_1)$.

最后，根据最大后验概率进行分类：

$$\forall x, \quad \delta_x \in \begin{cases} \theta_0, & \text{当} p(\theta_0 \mid x) > p(\theta_1 \mid x), \\ \theta_1, & \text{当} p(\theta_1 \mid x) > p(\theta_0 \mid x). \end{cases}$$

分类的例子

下面我们针对一组数据点，得到如表所示的分类结果：

数值	$p^*(\theta_0 \mid x)$	$p^*(\theta_1 \mid x)$	真实的类别	判断的类别
5	0.9916	0.0358	0	0
7.1	0.0000	0.3140	1	1
4.4	0.3535	0.0018	0	0
4.9	0.9489	0.0400	1	0
6.5	0.0003	0.6855	1	1
5.1	0.9554	0.0267	0	0
7.2	0.0000	0.28563	1	1
5.0	0.9916	0.0358	0	0

注：p^* 表示没有归一化的分布密度

非参数估计的优缺点

- 上述的概率密度估计和分类的例子已经较好地说明了非参数密度估计的优点. 如果能采集足够多的训练样本，无论实际采取哪一种核函数形式，从理论上最终可以得到一个可靠的收敛于密度的估计结果.
- 概率密度估计和分类例子的主要缺点是为了获得满意的密度估计，实际需要的样本量却是非常惊人的. 非参数估计要求的样本量远超过在已知分布参数形式下估计所需要的样本量. 这种方法对时间和内存空间的消耗都是巨大的，人们也正在努力寻找有效降低估计样本量的方法.

非参数估计的维数灾难问题

然而，非参数密度估计最严重的问题是高维应用问题. 一般在高维空间上，会考虑定义一个 d 维核函数为一维核函数的乘积，每个核函数有自己的带宽，记为 $h_1, h_2, \cdots, h_d$，参数数量与空间维数呈线性关系. 然而在高维空间中，任何一个点的邻域里没有数据点是很正常的，因而出现了体积很小的邻域中的任意两个点之间的距离却很远，比如 10 维空间上位于一个体积为 0.001 的小邻域内的两个点的距离可以允许高到 0.5，这样基于体积概念定义的核函数没有样本点估计. 这种现象被称为“维数灾难”问题（curse of dimensionality）. 为了使核估计能够应用，则需要更多的样本作为代价. 因此这也严重限制了非参数密度估计在高维空间上的应用.

25.5.3 k 近邻估计：引入

- Parzen 窗估计一个潜在的问题是每个点都选用固定的体积.
- 如果 h_n 定的过大，则那些分布较密的点由于受到过多点的支持，使得本应突出的尖峰变得扁平；而对于另一些相对稀疏的位置或离群点，则可能因为体积设定过小，而没有样本点纳入邻域，从而使密度估计为零.
- 虽然可能选择像正态密度等一些连续核函数，能够在一定程度上弱化该问题，但很多情况下并不具有实质性的突破，仍然没有一个标准指明应该按照哪些数据的分布情况制定带宽.

k 近邻估计：样本点与体积

一种可行的解决方法就是让体积成为样本的函数，不硬性规定窗函数为全体样本个数的某个函数，而是固定贡献的样本点数，以点 $\boldsymbol{x}$ 为中心，令体积扩张，直到包含进 k_n 个样本为止，其中的 k_n 是关于 n 的某一个特定函数. 被吸收到邻域中的样本就称为点 $\boldsymbol{x}$ 的 k_n 个最近邻. 用停止时的体积定义估计点的密度如下:

$$\tilde{p}_n(\mathbf{x}) = \frac{k_n/n}{V_n}.$$

如果在点 $\boldsymbol{x}$ 附近有很多样本点，那么这个体积就相对较小，得到很大的概率密度；而如果在点 $\boldsymbol{x}$ 附近很稀疏，那么这个体积就会变大，直到进入某个概率密度很高的区域，这个体积就会停止生长，从而概率密度比较小.

k 近邻估计：样本点与体积

如果样本点增多，则 k_n 也相应增大，以防止 V_n 快速增大导致密度趋于无穷.
另一方面，我们还希望 k_n 的增加能够足够慢，使得为了包含进 k_n 个样本的体积能够逐渐地趋于零. 在选择 k_n 方面，Fukunaga 和 Hosterler(1973) 给出了一个计算 k_n 的公式，对于正态分布而言:

$$k = k_0 n^{4/(d+4)}$$

式中，k_0 是常数，与样本量 n 和空间维数 d 无关.

k 近邻估计：样本点与体积

如果取 $k_n = \sqrt{n}$, 并且假设 $\tilde{p}_n(\mathbf{x})$ 是 $p(\mathbf{x})$ 的一个较准确的估计，那么根据上式，有

$$V_n \approx 1/(\sqrt{n}p(x)).$$

这与核函数中的情况是一样的. 但是这里的初始体积是根据样本数据的具体情况确定的，而不是事先选定的. 而且不连续梯度的点常常并不出现在样本点处，见下图.

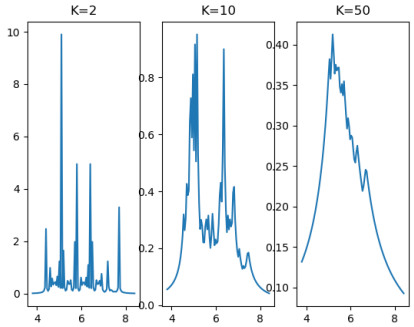


图 5: 鸢尾花数据集花萼长度 k_n 近邻估计图

k 近邻估计：存在的问题

与核函数一样，k_n 近邻估计也同样存在维度问题. 除此之外，虽然 $\tilde{p}_n(\boldsymbol{x})$ 是连续的，但 k 近邻密度估计的梯度却不一定连续。k_n 近邻估计需要的计算量相当大，同时还要防止 k_n 增加过慢导致密度估计扩散到无穷. 这些缺点使得用 k_n 近邻法产生密度并不多见，k_n 近邻法更常用于分类问题.

本讲小结

概率密度估计方法：

参数估计

- 极大似然估计
- 贝叶斯估计（最大后验概率估计）
- 最小最大估计
- 非监督参数估计

非参估计

- 直方图
- 核密度估计
- k 近邻估计

除此之外，还有半参的方法，用于混合密度估计。对于含有隐变量的密度函数的估计，可以采用 EM 算法。注意：概率密度函数包含随机变量的全部信息，是导致估计困难的重要原因。高维概率分布的估计无论在理论上还是实际操作中都是十分困难的问题。