

## Introduction 案能反映真实 i.d. 实际问题: 已知样本

贝叶斯决策: 已知  $p(C|W_i)$  和  $P(X|W_i)$  分类未知样本

仅应根据样本估计  $p(W_i)$  和  $p(C|W_i)$  (记为  $\hat{p}(W_i)$  和  $\hat{P}(C|W_i)$ )

$N \rightarrow \infty$  时,  $\hat{p}(W_i) \rightarrow p(W_i)$ ,  $\hat{P}(C|W_i) \rightarrow P(C|W_i)$

非参数估计  $\Rightarrow$  密度函数形式未知且不作假设 监督/非监督

参数估计  $\Rightarrow$  假设随机变量服从某种分布, 已知其形式, 估计表征 MLE / Bayes 估计

参数估计  $\Rightarrow$  假设随机变量服从某种分布, 已知其形式, 估计表征 监督  
函数参数 Parzen 窗法,  $k_n$ -近邻

## 一、MLE

### 最大似然估计和 Bayes 估计的区别

两种方法估计的参数结果接近, 但过程有区别:

• 最大似然估计将未知参数看成是确定变量, 在实际样本观察的概率最大的条件下, 获得未知参数的最好估计。

• Bayes 估计将未知参数看成是按某种分布的随机变量, 样本的观察结果由先验分布转化为后验分布, 再由后验分布修正参数的估计值。

## 思路

样本集可按类别分开, 不同类别密度函数的参数分别用各类样本来估计 确定而未知

概率密度函数形式已知, 参数未知, 我们用  $P(X|W_i, \theta)$  描述  $P(X|W_i)$  与参数  $\theta$  的依赖关系  $\Rightarrow$  按  $P(X|W_i)$  由取样本

为目标是用样本集  $X = \{x_1, x_2, \dots, x_N\}$  估计  $\theta$

## 原理

似然函数: 是关于统计模型参数  $\theta$  的函数, 写做  $L(\theta)$ 。观测结果  $X$  在参数集合  $\theta$  上的似然函数就是在给定参数  $\theta$  的基础上观察到结果  $X$  的概率, 故有  $L(\theta) = P(X|\theta)$ 。

$$\text{似然函数: } l(\theta) = p(X|\theta) = p(x_1, x_2, \dots, x_N|\theta) = \prod_{i=1}^N p(x_i|\theta)$$

假设条件:

① 参数  $\theta$  是确定的未知量 (不是随机量)

② 样本集  $X_i$  ( $i = 1, \dots, c$ ) 中的样本都是从密度为  $p(x|\theta_i)$  的总体中独立抽出来的(独立同分布, i.i.d.)

③  $p(x|\theta_i)$  具有某种确定的函数形式, 只具参数  $\theta$  未知

④ 各类样本只包含本类分布的信息

注意: 参数  $\theta$  通常向量, 比如一维正态分布  $N(\mu, \sigma^2)$  未知参数可能是  $\theta = [\mu, \sigma^2]$ , 此时,  $p(x|\theta)$  可写成  $p(x|\theta)$  或  $p(x|\theta_1, \theta_2)$ 。

鉴于上述假设, 我们可以只考虑一类样本, 记已知样本为  $X = \{x_1, x_2, \dots, x_N\}$

$$\text{似然函数: } l(\theta) = p(X|\theta) = p(x_1, x_2, \dots, x_N|\theta) = \prod_{i=1}^N p(x_i|\theta)$$

——在参数  $\theta$  下观测到样本集  $X$  的概率 (联合分布) 密度函数

基本思想:

如果在参数  $\theta = \hat{\theta}$  下  $l(\theta)$  最大, 则  $\hat{\theta}$  应是“最可能”的参数值, 称作最大似然估计量。

\* 为了便于分析, 还可以定义对数似然函数  $H(\theta) = \ln(l(\theta))$ 。

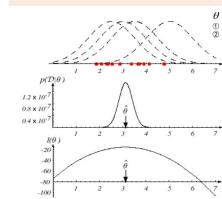
## 二、求解

求解:

• 若似然函数满足连续、可微的条件, 则最大似然估计量就是方程  $d(l(\theta))/d\theta = 0$  或  $dH(\theta)/d\theta = 0$  的解 (必要条件)。

• 若未知参数不止一个, 即  $\theta = [\theta_1, \theta_2, \dots, \theta_d]^T$ , 记梯度算子  $\nabla_\theta = \left[ \frac{\partial}{\partial \theta_1}, \frac{\partial}{\partial \theta_2}, \dots, \frac{\partial}{\partial \theta_d} \right]^T$ , 则最大似然估计量的必要条件由  $S$  个方程组成:

$$\nabla_\theta l(\theta) = 0 \quad \nabla_\theta H(\theta) = 0$$



讨论:

- 如果  $l(\theta)$  或  $H(\theta)$  连续、可微, 存在最大值, 且上述必要条件方程组有唯一解, 则其解就是最大似然估计量。(比如多元正态分布)。
- 如果必要条件有多解, 则需从中求似然函数最大者。
- 若不满足条件, 则无一般性方法, 用其它方法求最大。

### 均匀分布参数估计

$$\text{例1 随机变量 } x \text{ 服从均匀分布, 但参数 } \theta_1 \text{ 和 } \theta_2 \text{ 未知: } p(x|\theta) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \theta_1 < x < \theta_2 \\ 0 & \text{其它} \end{cases}$$

计算: 均匀分布的对数似然函数为  $\ln l(\theta|\theta_1, \theta_2) = \ln \left( \frac{1}{\theta_2 - \theta_1} \right)^N = -N \ln(\theta_2 - \theta_1)$  (例1.2.18为第二类 (0.1.1) 问题, 那从 Remez 方法  $\ln l(\theta|\theta_1, \theta_2) = \prod_{i=1}^N \ln p(x_i|\theta)$  为对数似然函数,  $\theta_1$  为  $x_i = 1$  的概率, 证明其是最大似然估计为  $\hat{\theta} = \frac{1}{N} \sum_{i=1}^N x_i$ 。

解: 图解从 Bernoulli 分布, 得

$$p(x^1, \dots, x^N | \theta) = \prod_{i=1}^N \theta^{x_i} (1-\theta)^{1-x_i}$$

$$\ln p(x^1, \dots, x^N | \theta) = \sum_{i=1}^N x_i \ln \theta + (1-x_i) \ln(1-\theta)$$

$$\nabla_\theta \ln p(x^1, \dots, x^N | \theta) = \sum_{i=1}^N \left( \frac{x_i}{\theta} - \frac{1-x_i}{1-\theta} \right) = 0$$

$$\hat{\theta} = \frac{1}{N} \sum_{i=1}^N x_i$$

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

$$\nabla_\theta H(\theta) = \sum_{k=1}^N \nabla_\theta \ln p(x_k | \theta) = 0$$

$$\ln p(x_k | \theta) = -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2\theta_2} (x_k - \theta_1)^2$$

$$\nabla_\theta \ln p(x_k | \theta) = \begin{bmatrix} \frac{1}{\theta_2} (x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2} (x_k - \theta_1)^2 \end{bmatrix}$$

$$\hat{\theta}_1 = \frac{1}{N} \sum_{k=1}^N x_k \quad \text{均值}$$

$$\hat{\theta}_2 = \frac{1}{N} \sum_{k=1}^N (x_k - \hat{\theta}_1)^2 \quad \text{方差均值}$$

## 三、正态分布下的最大似然估计

### 1. 单变量正态分布

$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[ -\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right] \quad \theta = [\mu, \sigma^2] \quad \mu = \theta_1 \quad \sigma^2 = \theta_2 \quad \text{样本集 } X = \{x_1, x_2, \dots, x_N\}$$

$$\text{似然函数: } l(\theta) = p(X|\theta) = \prod_{k=1}^N p(x_k|\theta)$$

$$\text{对数似然函数: } H(\theta) = \ln l(\theta) = \sum_{k=1}^N \ln p(x_k|\theta)$$

$$\text{最大似然估计量 } \hat{\theta} \text{ 满足方程: } \nabla_\theta H(\theta) = \sum_{k=1}^N \nabla_\theta \ln p(x_k|\theta) = 0$$

$$\text{而: } \ln p(x_k|\theta) = -\frac{1}{2} \ln 2\pi\theta_2 - \frac{1}{2\theta_2} (x_k - \theta_1)^2$$

$$\nabla_\theta \ln p(x_k|\theta) = \begin{bmatrix} \frac{1}{\theta_2} (x_k - \theta_1) \\ -\frac{1}{2\theta_2} + \frac{1}{2\theta_2^2} (x_k - \theta_1)^2 \end{bmatrix}$$

$$\hat{\theta}_1 = \frac{1}{N} \sum_{k=1}^N x_k \quad \text{均值}$$

$$\hat{\theta}_2 = \frac{1}{N} \sum_{k=1}^N (x_k - \hat{\theta}_1)^2 \quad \text{方差均值}$$

### 2. 多变量正态分布

对于一般的多元正态分布, 计算方法完全类似, 且有

$$\begin{cases} \hat{\mu} = \frac{1}{N} \sum_{k=1}^N x_k \\ \hat{\Sigma} = \frac{1}{N-1} \sum_{k=1}^N (x_k - \hat{\mu})(x_k - \hat{\mu})^T \end{cases}$$

• 均值估计是无偏的, 协方差矩阵估计是有偏的。

• 协方差矩阵的无偏估计是:

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

## 四、Bayes 估计

### 与 MLE 区别

- 最大似然估计将未知参数看成是确定变量, 在实际样本观察的概率最大的条件下, 获得未知参数的最好的估计;
- Bayes 估计将未知参数看成是按某种分布的随机变量, 样本的观察结果由先验分布转化为后验分布, 再由后验分布修正参数的估计值。

思路与贝叶斯决策类似, 而离散决策状态变为连续估计

基本思想: 是变量就有分布  
把待估计参数  $\theta$  看作具有先验概率密度函数  $p(\theta)$  的随机变量, 其取值与样本集  $X$  有关, 根据样本集  $X = \{x_1, x_2, \dots, x_N\}$  估计  $\theta$ 。

损失函数: 把  $\theta$  估计为  $\hat{\theta}$  所造成的损失, 记为  $\lambda(\hat{\theta}, \theta)$

$$x \in E^d, \theta \in \Theta$$

期望风险:  $R = \int_{E^d} \int_{\Theta} \lambda(\hat{\theta}, \theta) p(x, \theta) d\theta dx$

$P(AB) = P(A|B)P(B)$

$R(\alpha) = \int_{E^d} R(\alpha(x)|x) p(x) dx$

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

$R(\alpha_i | x) = E[\lambda(\hat{\theta}, \theta) | x] = \sum_{j=1}^k \lambda(\alpha_j, \theta_j) P(\theta_j | x), i = 1, \dots, k$

条件风险:  $R(\hat{\theta} | x) = \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta | x) d\theta$

目标: 最小化期望风险

• 期望风险是在所有可能的情况下条件风险的积分

• 条件风险是非负

最小化期望风险  $\Rightarrow$  最小化条件风险 (对所有的  $x$ )

Bayes 估计: (在样本集  $X$  下) 使条件风险  $R(\hat{\theta} | X) = \int_{\Theta} \lambda(\hat{\theta}, \theta) p(\theta | X) d\theta$

最小的估计量

常用损失函数  $\lambda(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2 = (\theta - \hat{\theta})^T (\theta - \hat{\theta})$

定理 3.1 如果采用平方误差损失函数, 则  $\theta$  的贝叶斯估计量  $\hat{\theta}$  是在给定  $x$  时  $\theta$  的条件期望,

$$\text{即: } \hat{\theta} = E[\theta | X] = \int_{\Theta} \theta p(\theta | X) d\theta.$$

$$\Rightarrow \text{在给定样本集 } X \text{ 下, } \theta \text{ 的贝叶斯估计是 } \hat{\theta} = E[\theta | X] = \int_{\Theta} \theta p(\theta | X) d\theta \quad \text{最小风险贝叶斯估计}$$

贝叶斯估计方法 (平方误差损失  $\lambda(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$  条件下)

- ① 确定  $\theta$  的先验分布概率  $p(\theta)$  0是变量
- ② 求样本集的联合分布概率  $p(X | \theta) = \prod_{i=1}^N p(x_i | \theta)$  带着的密度函数
- ③ 求  $\theta$  的后验概率分布密度  $p(\theta | X) = \frac{p(X | \theta) p(\theta)}{\int_{\Theta} p(X | \theta) p(\theta) d\theta}$   $P(x_i)$  连乘形式
- ④ 求  $\theta$  的贝叶斯估计量  $\hat{\theta} = \int_{\Theta} \theta p(\theta | X) d\theta$  全概率密度  $P(x)$  难以计算

最小风险贝叶斯估计

贝叶斯估计的另一种方法 (平方误差损失  $\lambda(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$  条件下)

对于一个训练集合  $Z$ , 贝叶斯公式就变成了:

$$p(\theta | Z) = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} = \frac{p(x | \theta) p(\theta)}{\sum_i p(x_i | \theta) p(\theta)}$$

可以假定先验概率满足:  $p(\theta_1 | Z) = p(\theta_2 | Z)$ , 同时, 假定函数独立:

$$p(\theta_1 | Z) = \frac{p(x_1 | \theta_1) p(\theta_1)}{\sum_i p(x_i | \theta_i) p(\theta_i)}, Z \text{ 是属于 } k \text{ 类的样本}$$

因此, 我们处理的核心问题, 实际上是根据一组训练样本  $Z$ , 估计分布  $p(x_i)$ , 简单记  $Z_i$  为  $Z$ ,  $p(x_i)$  为  $p(Z_i)$ 。

贝叶斯估计的另一种方法 (平方误差损失  $\lambda(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$  条件下)

直接估计条件概率密度函数:  $p(x | \theta) = \int_{\Theta} p(x, \theta) p(\theta) d\theta$

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

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

观测样本特征为由  $p(x | \theta)$  和  $p(\theta | X)$ , 希望其尖峰值逼近真值

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

思路: 如果条件概率密度形式已知, 则利用已有训练样本, 可通过  $p(\theta | Z)$  对  $p(x_i)$  进行估计。

### Bayes 估计与 Bayes 决策

#### Bayes 决策

确定  $x$  的真实状态  $\omega_i$  (模式类)

#### Bayes 估计

根据一样本集  $Z = \{x_1, x_2, \dots, x_N\}$

找出估计量  $\hat{\theta}$ , 估计  $Z$  所属总体分布的某个真实参数  $\theta$ , 使带来的 Bayes 风险最小。

状态空间  $A$  是离散空间

参数空间  $\Omega$  是连续空间

先验概率  $p(\theta)$

后验概率  $p(\theta | X)$

似然函数  $p(x_i | \theta)$

损失函数  $\lambda(\hat{\theta}, \theta)$

贝叶斯估计

最大似然估计