

MLE

极大似然估计(Maximum Likelihood Estimation, MLE)

$X = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}$ 表示样本/观测点 (从"data"分布中抽取出来的)

$p_{\text{data}}(\mathbf{x})$ 表示服从"data"分布的pdf (概率密度函数)

$p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})$ 表示服从"model"分布的pdf, $\boldsymbol{\theta}$ 为"model"分布的参数

$\mathbf{x} \in \Omega$ 表示状态空间 (通俗点说就是p(x)定义域)

逐行解释:

$\boldsymbol{\theta}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} p_{\text{model}}(X; \boldsymbol{\theta})$ 表示希望找到合适的参数 $\boldsymbol{\theta}$ 使得 $p_{\text{model}}(X; \boldsymbol{\theta})$ 尽可能大,

通俗点说就是X 即真实观测样本在model分布下"出现概率"较大 (**idea**) 。

$= \arg \max_{\boldsymbol{\theta}} \prod_{i=1}^m p_{\text{model}}(\mathbf{x}^{(i)}; \boldsymbol{\theta})$ 写为乘积形式。这是由于X是 i.i.d (独立同分布)

更直观的说就是 $p(xy) = p(x)p(y)$ 这里 x, y 相互独立。

$\boldsymbol{\theta}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^m \log p_{\text{model}}(\mathbf{x}^{(i)}; \boldsymbol{\theta})$ 乘法不好计算 (数值上很容易太小) ,

取个log, 化成加法。 $\log(x)$ 是单调函数, 故可以这样操作。

$\boldsymbol{\theta}_{\text{ML}} = \arg \max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})$ 更加一般的可以写为这个形式。

其中期望表达式为 $\mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) = \int_{\Omega} \log p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta}) d p_{\text{data}}$

或者离散形式 $\sum_{\mathbf{x} \in \Omega} p_{\text{data}}(\mathbf{x}) \bullet \log p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})$

最大似然估计是机器学习 (统计学习) 中十分重要的理论基础。最大似然估计的思想是: 使观测数据发生概率最大的参数就是最好的参数

1.估计

概率论与数理统计 (浙大第四版) 对估计的定义如下:

设总体X的分布函数 $F(x; \theta)$ 的形式是已知的, θ 是待估计参数。 $X_1, X_2, \dots, X_n$ 是 X 的一个样本, $x_1, x_2, \dots, x_n$ 是相应的一个样本值。点估计 (估计分点估计和区间估计两大类) 就是要构造一个适当的统计量, $\hat{\theta}(X_1, X_2, \dots, X_n)$, 用它的观测值 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 作为未知参数 θ 的近似值, 称 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 为 θ 的估计量, $\hat{\theta}(x_1, x_2, \dots, x_n)$ 为 θ 的估计值。

在不混淆的情况下, 统称估计量和估计值为**估计**, 并都简记为 $\hat{\theta}$ 。由于估计量是关于样本的函数, 因此对于不同的样本值, θ 的估计值一般是不同的。

点估计的常用方法是 **最大似然估计** 和 **矩估计**, 下面讨论最大似然估计。

2.离散型随机变量的最大似然估计

若总体 X 是离散型, 其分布律为 $P\{X = x\} = p(x; \theta), \theta \in \Theta$, 则 $X_1, X_2, \dots, X_n$ 的联合分布律为(独立同分布)

$$\prod_{i=1}^n p(x_i; \theta). \quad (1.1)$$

事件 $\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}$ 发生的概率是

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n p(x_i; \theta), \theta \in \Theta \quad (1.2)$$

样本的观测值是确定的, $L(\theta)$ 是 θ 的函数, 称作样本的**似然函数**.

下面就是要在 θ 可能的取值范围 Θ 内挑选, 使似然函数 $L(x_1, x_2, \dots, x_n; \theta)$ 达到最大的参数, 即取 $\hat{\theta}$ 使

$$L(x_1, x_2, \dots, x_n; \hat{\theta}) = \arg \max_{\theta \in \Theta} L(x_1, x_2, \dots, x_n; \theta) \quad (1.3)$$

这样的 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 称为参数 θ 的 **最大似然估计值**, 相应的统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 称为参数 θ 的 **最大似然估计量**。

3.连续型随机变量的最大似然估计

若总体 X 是连续型, 其概率密度 $f(x; \theta), \theta \in \Theta$ 的形式是已知的, θ 为待估参数, Θ 是 θ 可能取值的范围. 设 $X_1, X_2, \dots, X_n$ 是来自 X 的样本, 则 $X_1, X_2, \dots, X_n$ 的联合密度为(独立同分布)

$$\prod_{i=1}^n f(x_i; \theta). \quad (1.4)$$

又设 $x_1, x_2, \dots, x_n$ 是样本 $X_1, X_2, \dots, X_n$ 的一个样本值。则随机点

$\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\}$ 落在 $x_1, x_2, \dots, x_n$ 的 **领域** (边长分别为 $dx_1, dx_2, \dots, dx_n$ 的 n 维立方体) 内的概率 **近似为**:

$$\prod_{i=1}^n f(x_i; \theta) dx_i. \quad (1.5)$$

其值随 θ 的取值变化, 与离散型的情况一样, 我们取 θ 的估计值 $\hat{\theta}$ 使式 1.5取到最大值, 因为因子 $\prod_{i=1}^n dx_i$ 不随 θ 而变化, 因此只要考虑函数

$$L(\theta) = L(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta) \quad (1.6)$$

的最大值。这里 $L(\theta)$ 成为样本的 **似然函数**。如果

$$L(x_1, x_2, \dots, x_n; \hat{\theta}) = \arg \max_{\theta \in \Theta} L(x_1, x_2, \dots, x_n; \theta) \quad (1.7)$$

则称 $\hat{\theta}(x_1, x_2, \dots, x_n)$ 称为参数 θ 的 **最大似然估计值**，相应的统计量 $\hat{\theta}(X_1, X_2, \dots, X_n)$ 称为参数 θ 的 **最大似然估计量**。

4.求解似然函数

$L(\theta)$ 如果关于 θ 连续可微，那么直接对 $L(\theta)$ 求导，并且令

$$\frac{d}{d\theta} L(\theta) = 0 \quad (1.8)$$

即可，但 $L(\theta)$ 的组成是连乘形式 $(\prod_{i=1}^n)$ 往往我们会对 1.8 式求对数，这样可以将连乘求导转化为连加求导，方便计算。于是有

$$\frac{d}{d\theta} \ln L(\theta) = 0 \quad (1.9)$$

1.9 式称为 **对数似然方程**。