Blog2

Thursday, March 23, 2017 4:02 PM

贝叶斯估计

模型已知，参数未知。这是做参数估计的前提。在最大似然估计中，我们对“未知”的参数没有做任何的限制，而是谁能使采样结果出现的机会最大，就认为谁是那个我要找的参数。然而在贝叶斯估计当中，我们将这个“未知”的参数视为随机变量，既然是随机变量，那它也是有其自身的分布的。“未知”参数自身的概率在贝叶斯估计中称为“先验概率”。将先验分布纳入到参数估计的过程中，这既是贝叶斯估计的优势，也是可能使其失灵的缺陷。

贝叶斯估计的原理
设立一个代价函数（cost function），亦称为损失函数（loss function），$\mathrm{C}(\mathrm{y},\hat{y})$，来表述估计量和实际量之间的差异，当估计量$\hat{y}$的取值使得此差异$\mathrm{C}(\mathrm{y},\hat{y})$的期望值最小时，认为此时的$\hat{y}$是最合理的。
数学表述为：

$$\mathrm{R} \equiv \mathrm{E}[\mathrm{C}(\mathrm{y},\hat{y})] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\mathrm{C}(\mathrm{y},\hat{y})\ \cdot f_{y,\boldsymbol{x}}(y,\boldsymbol{x})dyd\boldsymbol{x} = \int_{-\infty}^{\infty}\underbrace{\left[\int_{-\infty}^{\infty}\mathrm{C}(\mathrm{y},\hat{y})\ \cdot f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy\right]}_{I(y)}\cdot f_{\boldsymbol{x}}(\boldsymbol{x})d\boldsymbol{x},$$

其中，
$\boldsymbol{x}$为样本$x_1,x_2,x_3,\dots,x_n$
y为待估参数
$\hat{y}$为估计量
$f_{y|\boldsymbol{x}}(y|\boldsymbol{x})$为在样本$\boldsymbol{x}$发生的条件下y的概率密度函数

因为$f_{\boldsymbol{x}}(\boldsymbol{x}) \geq 0$，要使得R最小，只需$I(y)$最小即可。这通过寻找一阶导数为0的点可以得到。

通俗的说，就是我们先设立一个评价标准（当然，这个评价标准是由数学家们提出的）来评价估计量和实际量的差异有多大，然后找到能使这个平均差异最小的值为估计值。

常用的几种代价函数
在学术上，描述估计量和实际量的差异有多种方法，它们的出发点和适用场景不同，而在某些特殊的条件下，它们可能又是等价的。以下是常用的代价函数的定义，以及在此基础上的贝叶斯估计表达式。

1. 均方误差估计/ MSE（Mean Squared Error）
   定义 $\mathrm{C}(\mathrm{y},\hat{y}) \equiv |\mathrm{y}-\hat{y}|^2$，
   在连续的条件下，$\mathrm{E}[\mathrm{C}(\mathrm{y},\hat{y})]$的极小值在一阶导数为0处取得，整理后可得:

$$\hat{y}_{MSE} = \int_{-\infty}^{\infty} \mathrm{y} \cdot f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy = E[y|\boldsymbol{x}],$$

也就是说估计量$\hat{y}_{MSE}$是待估量y在样本x下的期望值。

2. 最大后验估计/ MAP（Maximum A Posteriori）

定义 $\mathrm{C}(\mathrm{y},\hat{y}) \equiv \begin{cases} 0, & |\mathrm{y}-\hat{y}| < \varepsilon \\ 1, & else \end{cases}$，则

$$I(y) = \int_{-\infty}^{\infty} \mathrm{C}(\mathrm{y},\hat{y}) \cdot f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy = \int_{|\mathrm{y}-\hat{y}| \geq \varepsilon} f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy = 1 - \int_{|\mathrm{y}-\hat{y}| < \varepsilon} f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy,$$

当等式右边 $\int_{|\mathrm{y}-\hat{y}| < \varepsilon} f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy$ 取最大值时，$I(y)$取最小值，所以估计量为：

$$\hat{y}_{MAP} = \underset{y}{\operatorname{argmax}} f_{y|\boldsymbol{x}}(y|\boldsymbol{x}).$$

3. 平均绝对误差估计/ MAE（Mean Absolute Error）

定义 $\mathrm{C}(\mathrm{y},\hat{y}) \equiv |\mathrm{y}-\hat{y}|$, 则$\mathrm{I}(\mathrm{y})$取最小值时有：

$\hat{y}_{MAE} = \mathrm{median}\ of\ f_{y|\boldsymbol{x}}(y|\boldsymbol{x})$，即 $\int_{-\infty}^{\hat{y}_{MAE}} f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy = \int_{\hat{y}_{MAE}}^{\infty} f_{y|\boldsymbol{x}}(y|\boldsymbol{x})dy$

MSE, MAP, MAE之间的关系

三者实际上是在不同的评价标准下对参数y进行的贝叶斯估计。其中，最大后验估计MAP和最大似然估计ML形式上极为相似，可以用来比较贝叶斯派和频率派的差异。

1. MAP与ML

$$\begin{cases} \hat{y}_{ML} = \underset{y}{\operatorname{argmax}} f_{\boldsymbol{x}|y}(\boldsymbol{x}|y), & ① \\ \hat{y}_{MAP} = \underset{y}{\operatorname{argmax}} f_{y|\boldsymbol{x}}(y|\boldsymbol{x}), & ② \end{cases}$$

根据贝叶斯公式有，

$f_{y|\boldsymbol{x}}(y|\boldsymbol{x}) = \frac{f_{\boldsymbol{x}|y}(\boldsymbol{x}|y) * f_y(y)}{f_x(x)}$，那么

$$\hat{y}_{MAP} = \underset{y}{\operatorname{argmax}} \frac{f_{\boldsymbol{x}|y}(\boldsymbol{x}|y) * f_y(y)}{f_x(x)} = \underset{y}{\operatorname{argmax}} f_{\boldsymbol{x}|y}(\boldsymbol{x}|y) * f_y(y), \qquad ③$$

比较 ①③，可以发现MAP就是在ML的基础上乘上了“未知”参数本身的先验概率密度。

2. 对称

当$f_{y|\boldsymbol{x}}(y|\boldsymbol{x})$的图像关于纵轴对称时，$\hat{y}_{MAE} = \hat{y}_{MSE}$

此时意味着取最小均方差的$\hat{y}$刚好也是中位数

3. 对称且单峰值

当$f_{y|\boldsymbol{x}}(y|\boldsymbol{x})$的图像关于纵轴对称且只有一个峰值时，$\hat{y}_{MAE} = \hat{y}_{MSE} = \hat{y}_{MAP}$

总结

贝叶斯估计的关键，在于将先验概率引入了参数估计。理论上来讲，它是比最大似然估计更好的方法，然而实际中先验概率的获取并不容易，一个错误的先验概率会导致贝叶斯估计的失灵。