R 语言编程：基于 tidyverse

第 20 讲（选讲） 贝叶斯学派与最大后验估计

张敬信

2022 年 3 月 4 日

哈尔滨商业大学

贝叶斯学派

先来介绍一下频率学派与贝叶斯学派[1]，频率学派和贝叶斯学派对世界的认知有本质不同：

- 频率学派认为世界是确定的，有一个本体，这个本体的真值是不变的，我们的目标是要找到这个真值或真值所在的范围；
- 而贝叶斯学派认为世界是不确定的，人们对世界先有一个预判，而后通过观测数据对这个预判做调整，我们的目标是要找到最优的描述这个世界的概率分布。

[1]引用自夏飞知乎：聊一聊机器学习的 MLE 和 MAP：最大似然估计和最大后验估计.

在对数据建模时，用 θ 表示模型的参数，解决问题的本质就是求 θ:

- **频率学派认为：存在唯一真值** θ

比如抛硬币抛一枚硬币 `100` 次，有 `20` 次正面朝上，要估计抛硬币正面朝上的概率，即伯努利分布的参数 $p = P(\text{正})$.

在频率学派看来，$p = 20/100 = 0.2$, 简单直观。当抛硬币次数趋于无穷时，该方法能给出精确的估计；然而次数不够大时，可能会产生严重的偏差。

- **贝叶斯学派认为：θ 是一个随机变量，符合一定的概率分布**

在贝叶斯学派里有两大输入和一大输出，输入是先验 (prior) 和似然 (likelihood)，输出是后验 (posterior)。

先验，即 $P(\theta)$ ，是指在没有观测到任何数据时对 θ 的预先判断，比如抛一枚硬币，一种可行的先验是认为该硬币有较大概率是均匀的；似然，即 $P(X|\theta)$, 是假设 θ 已知后观察到的数据应该是什么样子的；

后验，即 $P(\theta|X)$, 是最终的参数分布。

贝叶斯估计的基础是贝叶斯公式：

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

同样是抛硬币，对一枚均匀硬币抛 5 次得到 5 次正面，如果先验认为该硬币大概率是均匀的，那么参数 p，即 $P(\theta|X)$，是一个概率分布，最大值会介于 `0.5~1` 之间，而不是武断地认为 $p = 1$.

随着数据量的增加，参数分布会越来越向数据靠拢，先验的影响力会越来越小；如果先验是均匀分布（本质上表示对事物没有任何预判），则贝叶斯方法等价于频率方法。

贝叶斯统计也越来越兴起，特别是出现了专门用于贝叶斯推断的 `Stan` 语言，其与 R 语言的接口是 `rstan` 包，以及更方便的统计建模包：`rstanarm`，`brms`，`tidybayes` 等。

最大后验估计

贝叶斯学派常用的估计方法，同样假设数据 $x_1, x_2, \cdots, x_n$ 是独立同分布的一组抽样，记 $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, 则最大后验估计参数 θ, 其推导基于贝叶斯公式：

$$
\begin{aligned}
\hat{\theta}_{\mathrm{MAP}} &= \arg\max P(\theta \mid \mathbf{x}) \\
&= \arg\max \ln P(\theta \mid \mathbf{x}) \\
&= \arg\max \ln P(\mathbf{x} \mid \theta) + \ln P(\theta) - \ln P(\mathbf{x}) \\
&= \arg\max \ln P(\mathbf{x} \mid \theta) + \ln P(\theta)
\end{aligned}
$$

可见，与最大似然估计的不同在于相差一个先验 $\ln(P(\theta))$. 有趣的是，若该先验是正态分布，`MAP` 等价于 `MLE+L2` 正则。

本篇主要参阅 (张敬信, 2022), 知乎文章。

参考文献

张敬信 (2022). *R 语言编程：基于 tidyverse*. 人民邮电出版社, 北京.