2023 Spring《数理统计》课程大作业

2023 年 5 月 6 日

问题：已知利用加噪的正弦函数 $y_i = f(x_i) + \epsilon_i = 10\sin(0.3x_i) + \epsilon_i$ 在区间 $[0, 20]$ 内生成了一组由 100 个数据对 (x_i, y_i) 构成的数据，其中 ϵ_i 表示数据噪声. 现在我们利用多项式回归模型

$$y_i = \boldsymbol{w}^\top \boldsymbol{x}_i + \epsilon_i, \quad \boldsymbol{x}_i = [1, x_i, \cdots, x_i^m]^\top$$

来建模该组数据，其中 $\boldsymbol{w}$ 表示模型参数，$\boldsymbol{x}_i$ 表示模型的第 i 个输入，y_i 表示对应的观测值，而噪声 ϵ_i 服从高斯分布 $\mathcal{N}(\epsilon_i; 0, \sigma_0^2)$ $(\sigma_0^2 = 5.0)$，m 表示多项式的次数. 若假设模型参数 $\boldsymbol{w}$ 的先验分布为

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \mathbf{0}, \sigma_w^2 \mathbf{I}),$$

试求模型参数 $\boldsymbol{w}$ 的最大似然估计和贝叶斯估计.

数据下载地址：

github.com/Mephestopheles/Mathematical-Statistics-2023Spring/tree/main/Assignment1

训练数据由 data.csv 给出，其中数据形式如下表所示：

No	x	y
1	1.6302148253728963	6.241160657344439
2	0.8351081552813885	4.0151310636573365
⋮	⋮	⋮
99	17.60037082165521	-9.361072876930848
100	18.97287441658695	-8.612303468512675

作业要求：

1) 分别给出模型参数 $\boldsymbol{w}$ 的最大似然估计和贝叶斯估计的理论结果；

2) 实现最大似然估计和贝叶斯估计的代码，要求只调用 Numpy 软件包中的基本运算以及必要的数据读取工具（如 pandas 库）和绘图工具（如 matplotlib 库）；

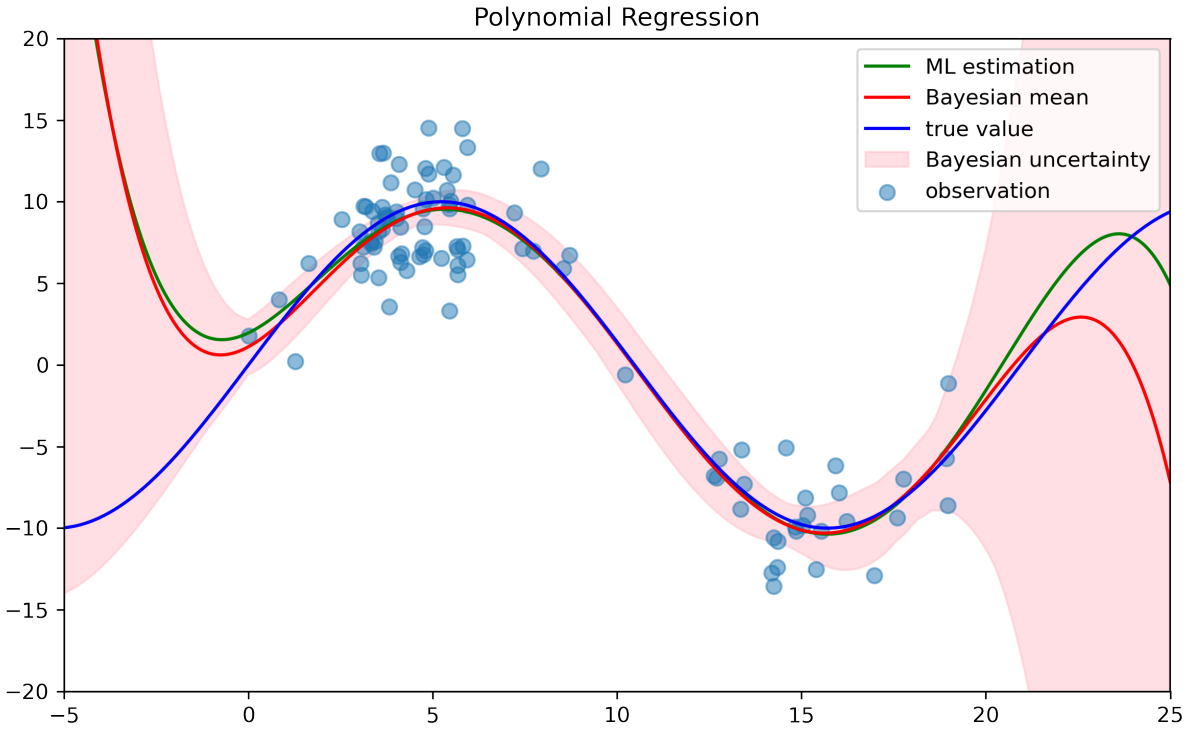
3) 探索多项式次数 m 对最大似然估计和贝叶斯估计的影响(可分别取 $m = 1, 2, 3, 4, 5$ 等)；

4) 在贝叶斯估计的实验中，探索采用不同的 σ_w 取值对实验结果的影响.

实验报告需包含（但不限于）:

1) 分别对单数据点和多数据点的情形，给出模型参数 $\boldsymbol{w}$ 最大似然估计和贝叶斯估计的理论推导；

2) 分别为最大似然估计和贝叶斯估计绘制如下所示的曲线图（包括 y_i 的观测值、真实值、最大似然估计值、贝叶斯估计均值和贝叶斯估计不确定性）:



- **最大似然估计：** 先利用上述观测数据计算出模型参数 $\boldsymbol{w}$ 的最大似然估计值 $\boldsymbol{w}^*$，再使用 Numpy 中的 linspace 函数均匀采样测试数据 x_i，并计算去噪后的对应真实值 $f(x_i)$ 和最大似然估计值 $y_i = (\boldsymbol{w}^*)^\top \boldsymbol{x}_i$
- **贝叶斯估计：** 先利用上述观测数据计算出模型参数 $\boldsymbol{w}$ 的后验分布 $p(\boldsymbol{w}|\mathbf{X}, \boldsymbol{y})$ 并得到其均值 $\boldsymbol{\mu}$ 和协方差矩阵 $\boldsymbol{\Sigma}$，并利用 $\boldsymbol{\mu}$ 来绘制测试数据 x_i 所对应的贝叶斯均值 $y_i = \boldsymbol{\mu}^\top \boldsymbol{x}_i$ 曲线；另外，可以通过 Numpy 库中的 random.multivariate_normal 函数对该后验分布进行采样得到一系列模型参数 $\{\boldsymbol{w}_j\}$，

再利用测试数据 x_i 和采样得到的模型参数来计算相应的贝叶斯估计值 $y_i = \boldsymbol{w}_j^{\top}\boldsymbol{x}_i$，并根据 y_i 的范围（如 75% 或 100% 范围）来绘制贝叶斯不确定性范围（可以使用 matplotlib.pyplot 库中的 fill_between 函数进行绘图）；

3) 通过观察实验结果，结合理论知识，说明最大似然估计与贝叶斯估计之间的关系；

4) 结合实验结果，分析并比较最大似然估计值、贝叶斯均值和贝叶斯不确定性之间的优劣（可从计算复杂度和模型不确定性等方面进行讨论）.

作业提交：

将实验报告（.doc 或.pdf）和代码打包成 zip 文件，文件包的命名规则为：学号 + 姓名.zip，并提交到助教邮箱：xjxtech@126.com

提交截止时间：

2023 年 5 月 28 日 0:00