

Bayesian Analysis with Python

一

1. Bayesian model 构建步骤：
2. 给定一些数据以及这些数据是如何生成的假设，
3. 利用Bayesian 理论将数据和模型结合起来，根据数据和假设推导出逻辑结论。
4. 根据多种标准，包括真实数据和对研究问题的先验知识，判断模型拟合是否合理
5. p(H)：先验概率

p(D|H)：似然

p(H|D)：后验概率

p(D)：证据

1. 通用模型：p(y|θ) ∝ p(θ|y) p(θ)
2. 似然的选择：

例如抛硬币的例子，因为多次抛硬币之间的结果没有影响，故可选择二项分布为p(θ|y)

1. 先验的选择：

在抛硬币的例子中可以选用Beta分布

二．

1. 推断引擎：

非Markov方法：网格计算（暴力计算，但不适用于多参数）

二次近似（正态近似）

变分方法

Markov方法：蒙特卡洛

Markov链：Metropolis—Hasting算法：

1. 随机or根据经验初始化参数xi
2. 从某个简单的分布Q（xi+1|xi）中选取一个新的xi+1。（如高斯分布，均匀分布）
3. 根据Metropolis—Hasting准则计算接受一个新参数的概率
4. 从（0,1）上的均匀分布中随机抽取一个值与（3）的计算结果相比较，若（3）的值大则接受新值，否则拒绝。
5. 继续代送

汉密尔顿蒙特卡洛方法/不掉向采样：

函数梯度的计算

其他MCMC方法：

副本交换（并行退火）

注：pm.sample函数调用时默认调用CPU和GPU，直接运行可能报错，添加参数cores=1

2.pymc3库简要使用

计算法解决抛硬币问题：

模型描述，

推断，

诊断采样过程：增加样本次数，老化（去除一定的样本），重新参数化，转化数据

收敛性：traceplot函数可以返回核密度估计图（KDE）和采样值，

自相关：

有效采样大小

3.总结后验

基于后验的决策

ROPE：实用等价区间，比较ROPE与HPD区间关系

损失函数：

三：多参和分层模型

冗余参数和边缘分布：

两参数的贝叶斯模型：p(θ1,θ2|y) ∝ p(y|θ1,θ2) p(θ1,θ2)

边缘分布：p(θ1,y)=∫p(θ1,θ2|y)dθ2

正态模型

存在异常值时的t分布：

分层模型和收缩：

分层：在先验之前一个“超先验”，即不再固定参数，而是从数据中将其估计出来

收缩：所有组通过超先验共享了部分信息

介于对数据分组建模与将数据作为一个大组建模之间

组间比较：使用seaborn中的‘tips’数据集

Cohen’s d：（u1-u2）/ √[(a12+a22)/2]

概率优势：ps=φ（δ/√2）

φ:标准正态分布分布函数

δ:Cohen’d

四．线性回归

一元线性回归：

核心： yi = a + bxi

从概率角度：y ～ N（u = a+b, σ= ）

先验设置： a ～ N（u1，σ1）#较为平坦的正态分布

b ～ N（u2，σ2 ）#

ε～ N（0， h）#也可使用半正态和半高斯

从自相关图中可看出，a,b具有很强的自相关性

拟合直线穿过中心点只有在最小二乘才会实现

贝叶斯方法下限制被放松

解决高自相关性问题：数据中心化，更换采样方法

相关系数，根据多元正态分布计算相关系数

稳健线性回归：当正态假设并不成立或者近似程度不佳时，可以采用t分布

分层线性回归：相关性和因果性不能混为一谈

多项式回归：使用x的n次方加和去拟合曲线（实际中超过二阶和三阶的多项式回归意义不大），同时要注意回归系数的实际意义解释

多元线性回归：模型中包含多个自变量；

存在多余和混淆变量；

存在多重共线性和自相关过高问题；

可以通过增加相互作用（如变量之间的乘积来修正）

五．利用逻辑回归对结果进行分类

多元逻辑回归：逻辑回归模型主要用于分类

逻辑函数： logistic（z）= 

线性回归模型转化为分类模型：



~Bern()

决策边界：



解释逻辑回归的系数：

逻辑函数的逆函数：



因此





即



此时模型中的是y=1时的概率



其中称为发生比

Softmax回归或多项逻辑回归：



该模型与逻辑回归的另一个区别是：Bernoulli 分布换成了类别分布

六．模型比较：

模型简约性与准确性：

参数太多导致的过拟合和参数过少导致的欠拟合

高方差与高偏差之间的平衡

正则先验：

拉普拉斯模型，对称分布，其比正态分布更趋近于0，在0附近导数不存在

衡量预测准确性：

方法：交叉检验，信息准则

交叉验证：将数据分为多个子集，轮流将其中一个子集作为测试集

适用于数据量大的情况

信息量准则：

log似然与偏差

赤池信息量准则

偏差信息量准则

通用信息量准则

帕累托平滑重要性采样与留一交叉验证

贝叶斯信息准则