组会

2024 年 3 月 12 日

目录

1 有关 Bayesian analysis 的基础

1.1 Bayes' theorem

从概率论的基本公式出发

$$P(X|I) + P(\bar{X}|I) = 1 \tag{1}$$

和

$$P(X, Y|I) = P(X|Y, I) \times P(Y|I) \tag{2}$$

基本意思: 事件 X 和事件 Y 同时发生的概率是事件 Y 发生的概率乘以事件 Y 发生前提下事件 X 发生的概率

因此重要推论:(Bayes' theorem)

$$P(X|Y) = \frac{P(Y|X) \times P(X)}{P(Y)} \tag{3}$$

证明比较简单, 一步就可以说明

$$P(Y,X) = P(X,Y) = P(Y|X) \times P(X) \tag{4}$$

若将 X 和 Y 分别用 hypothesis 和 data 代替可以得到

$$P(hypothesis|data) = \frac{P(data|hypothesis) \times P(hypothesis)}{P(data)} \propto P(data|hypothesis) \times P(hypothesis) \tag{5}$$

(其中忽略了分母, 本质上分母是一个归一化常数, 这个事情只用一行也能说明)

$$P(Y) = \sum_X P(XY) = \sum_X P(Y|X) \times P(X) \tag{6}$$

然后是正式概念:

- $P(hypothesis)$: prior probability, 先验概率, 指的是分析现有数据之前对假设 hypothesis 的了解程度
- $P(data|hypothesis)$: 似然函数 likelyhood, 用于修正 prior, 以构成 posterior 后验, 可随实验数据积累而优化修正
- $P(hypothesis|data)$: posterior probability, 后验概率, 反映了在已知数据的情况下, 模型假设的真实性

1.2 posterior 定量描述

posterior 可靠性的定量描述: 最概然估计 (best estimates), 误差棒 (error-bars), 置信区间 (confidence-intervals)

最概然估计 (best estimates): 我们相信该参数落在该值附近区间的程度, 因此取 posterior 分布的最大值

然后用公式描述, 假设我们研究的参数是 X, 以及其对应的后验 posterior 分布

$$posterior = P(X|data) \tag{7}$$

那么它的最概然估计值 X_0 的位置应当满足

$$\frac{dP}{dX}|_{X_0} = 0 \tag{8}$$

再严格一点, 我们还需 check 它的二阶导数的正负号,

$$\frac{d^2P}{dX^2}|_{X_0} < 0 \tag{9}$$

它的二阶导数应该为负, 才能确认这是一个极大值, 因为二阶导数为负表现了上凸函数的性质。
由于函数的对数比函数本身更慢变, 因此采用对数分析

$$L = ln[P(X|data)] \tag{10}$$

并展开

$$L = L(X_0) + \frac{1}{2}\frac{d^2L}{dX^2}|_{X_0}(X - X_0)^2 + \cdots \tag{11}$$

具有零阶和二阶项
因此将原本 posterior 近似写作指数形式

$$P(X|data) \simeq Aexp[\frac{1}{2}\frac{d^2L}{dX^2}|_{X_0}(X - X_0)^2] \tag{12}$$

将上式和 gauss 分布做比较

$$P(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp - \left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{13}$$

可以得到

$$\sigma = \left(-\frac{d^2L}{dX^2}|_{X_0}\right)^{-1/2} \tag{14}$$

这就是标准差, 也即误差棒 error-bar
从而置信概率的概念被自然引出

$$P(X_0 - \sigma \leq X < X_0 + \sigma|data) = \int_{X_0-\sigma}^{X_0+\sigma} P(X|data)dX \simeq 0.67 \tag{15}$$

意思是感兴趣量 X 的值位于一倍的 error-bar(σ) 的范围内的置信概率约是 0.67

2 示例

以掷硬币为例, 在掷硬币之前, 我们认为硬币均匀, 这就对应了 prior 部分, 即实验之间我们对系统了解的信息, 因此

$$P(H) = \begin{cases} 1 & 0 \leq H \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

(在掷硬币类型的实验里面,H 代表了单次实验正面 (特征面) 向上的概率)
这里 prior 选取了常数 1, 常数形式的 prior 叫做均匀先验, 也可以叫做' 无信息先验'(non-informative prior)
意思是: 硬币以任何可能的特定概率出现正面的概率是相同的
下一步是 likelyhood 的选取, 对于这个例子 (N 次投掷,R 次朝上),likelyhood 的形式应满足二项分布

$$P(data|H) \propto H^R(1-H)^{N-R} \tag{17}$$

因此根据 Bayes' theorem, posterior 后验分布由上面两式相乘得到
比如, 在这里进行了多次实验, 有不同的实验结果, 当实验次数比较大的时候, 一般情况下 likelyhood 会向一个更符合实际的好的方向变化, 从而调制 posterior

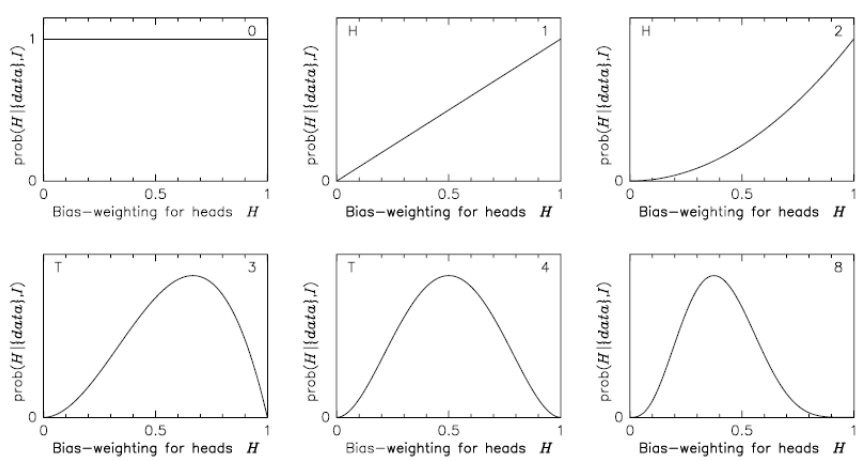

图 1: 实验次数 N 比较小的 posterior 分布

最后随着实验次数的增大, 参数 H 的 posterior 分布的 peak 收敛到了一个

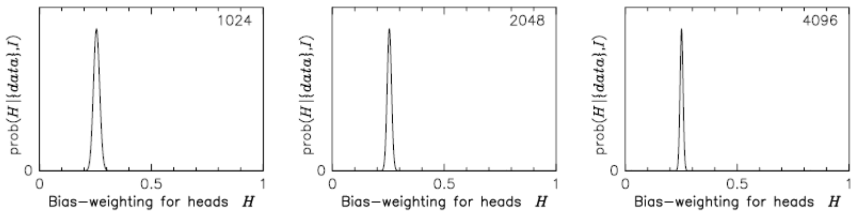

图 2: 实验次数 N 比较大的 posterior 分布

约为 0.25 的位置, 这说明硬币不均匀
因此对于该种不均匀硬币投掷实验, 我们可以预测单次实验特征面向上的概率 H 为 0.25, 从而可以根据 H 的值预测 N 次实验所有可能结果的概率
上述过程就是一个预测过程, 这里总结一下相对简单例子的流程

1. 根据我们所知的内容给定 prior 先验分布, 而由于 prior 是不完备的, 我们需在第二步做修正

2. 通过查询或者别的方式 (如何给？) 给出 likelyhood 的表达式

3. 通过一定量 (多一点好) 的实验数据去给出上述两者组合成的 posterior 分布图, 找到它的 peak 位置, 这个位置就对应了理论的各个参数应该取的具体值

4. 这就能够给出一类实验的理论模型 (包括各个参数的具体大小), 从而我们可以用该模型去预测或解释所有同类型的实验和现象。

3 BAND framework

- 首先包括先验 (Bayesian prior): 对外部信息, 专家意见定量编码得到 prior, 然后是 likelyhood 似然进行编码, 这是表示出数据 (涉及到, 被考虑的) 限制参数范围的方式 (公式)。因此在开始前要: 开发先验 prior 和可能性 likelyhood 的合适的公式, 将它们作为 input 放入 BAND 工具框架中。(计算工具 A: emulation)

- 第一步是在有了模型或模型集 f 之后, 对于很大计算量的情况, 通过 emulator 模拟降低计算量, 因此 emulator 是一个计算成本低的模型插值器。

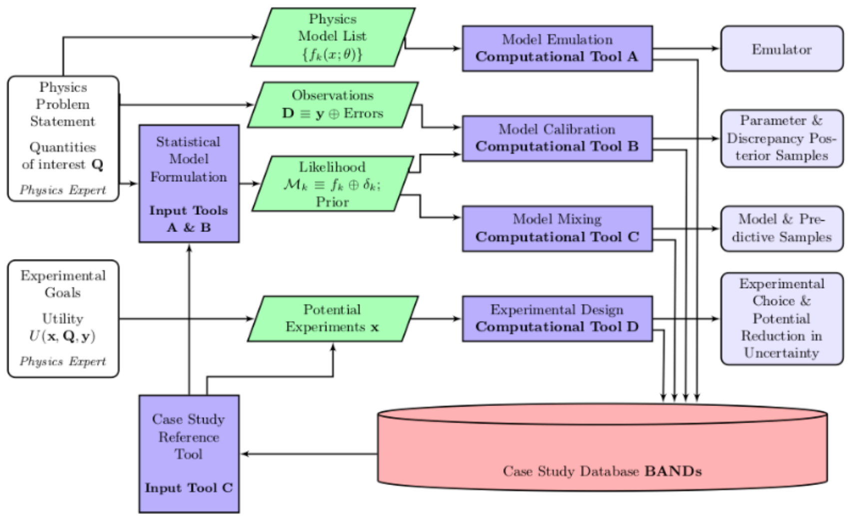

图 3

- 在指定了观测数据, 似然函数和先验后, 使用 emulator 样本获得模型校准块中每个模型参数的后验的分布函数 (计算工具 B: calibration)
- A 和 B 工具 (校准和仿真) 只能完成单个模型的信息, 而模型混合: 模块 C 完成。
- 计算工具 D 为反应模型不确定度真实程度的实验设计提供指导, 因此 D 部分完成了预测的工作

3.1 prior 和 likelyhood 的选取

- 首先: constant 形式的 prior 不够 informative
- 因此希望 prior 更 informative: 使 prior 含参 (但应当和我们理论所关注的参数 Q 要区分开), Hyperparameter
- Hyperparameter 由其他信息 I 来引入

最后通过重复使用 Bayes' theorem 写出 posterior

$$P(Q|D) \propto P(D|Q,I)P(Q|I)P(I) \propto P(D|Q)P(Q|I)P(I) \tag{18}$$

(上式被称为 hierarchical Bayesian model, 其中关键用到 $P(D|Q,I) = P(D|Q)$, 意思是实验测量的数据 D 和这里的 prior 的额外信息 I 相互独立没有影响)

3.2 举例

拟合 N 阶多项式

$$f(x,\theta) = a_0 + a_1 x + \cdots + a_M x^M \tag{19}$$

因此在这个模型中, 我们感兴趣的量 (参数集)Q 设为集合 θ,$\theta = \{a_0, a_1, \cdots, a_M\}$, 即所有的系数

而 bayesian 的关键步骤在于建立一个自然的模型 (model naturalness), 意思就是假设所有的参数 (这里指所有的系数 a_i) 都从一个普遍一般性的总体中选取

然后加入超参数, 并假定参数具有 0 的均值以及相同的标准差 σ_a, 引出了正态分布形式的 prior

$$p(a_0, a_1, \cdots, a_M|\sigma_a) \propto exp\left(-\frac{a_0^2 + a_1^2 + \cdots + a_M^2}{2\sigma_a^2}\right) \tag{20}$$

prior 选取定性说明: prior 的性质可以用 informative 这个词来描述, 包含信息更多的称为更 informative, 结论是: 越 informative 的 prior 会导致我们研究的参数的置信区间越小

可以看到 prior 选取的示例是正态分布

然后查看 likelyhood 的选取示例

- 首先将 x 和 y(自变量和因变量) 各自赋予其实际意义: x 作为输入 input 代表了类似质子的动能或者是中子数之类的量,output y 就代表了观测量 observations, 如截面, 质量等
- x 和 y 之间应该用某理论建立联系:(理想情况)

$$y = f(x,\theta) \tag{21}$$

- 实际: 测量到的东西跟建立的理论之间应该有一定区别, 不完全相等 (统计相关 error)

$$y = f(x,\theta) + error \tag{22}$$

- 因此我们熟悉的所谓 χ^2 分布就描述了：当每个测量的实验点 i 相互独立并具有均值 0 以及方差 σ^2 的时候, likelyhood 也有正态分布

$$P(D|\theta,\{\sigma_i^2\}) \propto exp\left(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-f(x_i,\theta))^2}{\sigma_i^2}\right) \tag{23}$$

- D 代表了实验测量的结果数据集 (input 和 observation)

$$D=\{x=(x_1,\cdots,x_n),y=(y_1,\cdots,y_n)\} \tag{24}$$

- 由于 emulation 基于统计模型, 所以常选取正态分布,Possion 分布等统计行为

3.3 Bayesian prediction

所谓预测就是外推实验上无法获得的数据结果, 即'experimentally inaccessible conditions' $\tilde{x}$ 对应的 experimentally inaccessible values $\tilde{y}$

$$p(\tilde{y}|\tilde{x},D)=\int_\theta p(\tilde{y}|\tilde{x},\theta,D)p(\theta|\tilde{x},D)d\theta \tag{25}$$

简化记号得到

$$P(\tilde{D}|D)=\int_\theta P(\tilde{D}|\theta,D)P(\theta|D)d\theta \tag{26}$$

证明上式只需要使用 bayesian 定理的全概率分解形式 (前提是事件组合 C_i 互斥且并集是必然事件 S)

$$\begin{aligned}
P(A|B) &= \sum_i P(AC_i|B) \\
&= \sum_i \frac{P(ABC_i)}{P(B)} \\
&= \sum_i \frac{P(ABC_i)}{P(C_i|B)P(B)}P(C_i|B) \\
&= \sum_i \frac{P(ABC_i)}{P(BC_i)}P(C_i|B) \\
&= \sum_i P(A|C_i,B)P(C_i|B)
\end{aligned} \tag{27}$$

前式积分有两项:

- 第一个乘积因子项代表了 likelyhood 似然函数, 具体来说就是这里对要预测的实验数据集 $\tilde{D}$ 应用了先前用在原来的实验数据集 D 上的相同的 likelyhood 函数形式
- 第二个乘积因子项就是我们在原本数据集 D 上得到的 posterior 后验函数

4 code

- Software: Surmise (for calibration, uncertainty quantification, and other tools)
- Github: https://github.com/bandframework