贝叶斯理论介绍

苏毓淞

清华大学社会科学学院政治学系副教授

贝叶斯理论研究及临床应用培训班 @ 广西国际壮医医院

2020 年 12 月 19 日

报告大纲

1 前言

2 贝叶斯方法

3 贝叶斯先验

4 MCMC 算法

5 结论与展望

三个应用方法

- 条件概率 (Conditional Probability) 的应用八十年代后期，重新获取得青睐。
- 三个主要的应用方法：
  1. 贝叶斯方法 (Bayesian Methods)
  2. 缺失数据插补 (Missing Data Imputation)
  3. 因果推论 (Causal Inference)

三个盛行原因

- 方法和概念是旧的，但是学者们进入知行合一的阶段。
- 三个盛行原因：
  1. 计算机硬件、软件发展的成熟，使得复杂条件式概率方程式的求解不再旷日费时
  2. 受这批训练的学者逐渐投入教学工作，培养出的学生也纷纷投入这项研究
  3. 方法本身符合社会科学研究的精神，使得研究者愿意舍简就繁。

基本概念

- 统计推论 (Statistical Inference)：从已知的资料 (y) 中去预测未知的资料 ($\tilde{y}$)。以数学式表达——$\tilde{y}|y$。
- 条件概率 (Conditional Probability)：
  1. 在已知的 y 的条件下，$\tilde{y}$ 发生的概率为何。以数学式表达——$p(\tilde{y}|y)$。
  2. 从已知的资料 (y)，去推导未知的资料 ($\tilde{y}$) 发生的似然 (Likelihood)。以数学式表达——$p(y|\tilde{y})$。
- 似然 (Likelihood)：假设未知资料已知的情况下，已知资料发生的概率。

应用实例

- 股票市场的指数预测
- 选举研究中的投票预测
- $y = f(\theta x + \epsilon) \rightarrow p(\theta | y, x)$
- $p(\theta | \mathbf{X})$

条件概率求解

- 最大似然估计法 (Maximum Likelihood Estimation)
  - 似然函数：$p(\mathbf{X}|\theta)$
  - $$\widehat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \log\left[p(\mathbf{X}|\theta)\right]$$
- 贝叶斯方法 (Bayesian Method)

$$p(\theta|\mathbf{X}) = \frac{p(\theta)p(\mathbf{X}|\theta)}{p(\mathbf{X})}$$

$$p(\theta|\mathbf{X}) \propto p(\theta)p(\mathbf{X}|\theta)$$

取自然对数 $\log(p(\theta|\mathbf{X})) \propto \log(p(\theta)p(\mathbf{X}|\theta))$

$$= \log(p(\theta)) + \log(p(\mathbf{X}|\theta))$$

贝叶斯法则

- $$p(\theta|\mathbf{X}) = \frac{p(\theta, \mathbf{X})}{p(\mathbf{X})} = \frac{p(\theta)p(\mathbf{X}|\theta)}{p(\mathbf{X})}$$
- $$p(\mathbf{X}) = \sum_{\theta} p(\theta)p(\mathbf{X}|\theta)$$
  $$= \int p(\theta)p(\mathbf{X}|\theta)d\theta \qquad \text{如果 } \theta \text{ 是连续型变量}$$
- 因为省略了 $p(\mathbf{X})$，称之为非正则化 (Unnormalized) 的后验分布：
  $$p(\theta|\mathbf{X}) \propto p(\theta)p(\mathbf{X}|\theta)$$

贝叶斯法则应用案例：基因检测（θ 为离散型数据）

- 人的染色体组成男女不同，男性是 X-Y 染色体组成，女性则是 X-X 染色体组成。血友病 (Hemophilia) 是潜藏在 X 染色体的遗传性疾病，所以男性如果 X 染色体有该基因，则会发病，女性如果仅有 1 个 X 染色体有缺陷，则不会发病，但是如果两组 X 都有血有病的基因存在，对于这种情况的女性是致命的。我们来“预测”下任选一个女性 A 有血友病基因的概率 θ。
- Prior：假设女性 A 的兄长有血友病，那么他的妈妈肯定有一个 X 染色体有缺陷，我们得知他们的父亲没有血友病，因此女性 A 基因有缺陷的概率是 50%，也就是说 $p(\theta = 1) = p(\theta = 0) = 0.5$。

贝叶斯法则应用案例：基因检测（θ 为离散型数据）

- Data, Model, Likelihood：假设女性 A 有两个儿子，两人（不是双胞胎）均没有血友病，则两个儿子**不**带有血有病基因的似然为：

$$\Pr(y_1 = 0, y_2 = 0|\theta = 1) = 0.5 \times 0.5 = 0.25$$
$$\Pr(y_1 = 0, y_2 = 0|\theta = 0) = 1 \times 1 = 1$$

- Posterior Distribution：应用贝叶斯法则，则女性 A 带有血友病基因的后验概率 $p(\theta = 1|y)$ 为 (令 $y = (y_1, y_2)$)：

$$\begin{aligned} p(\theta = 1|y) &= \frac{p(y|\theta = 1)\Pr(\theta = 1)}{p(y|\theta = 1)\Pr(\theta = 1) + p(y|\theta = 0)\Pr(\theta = 0)} \\ &= \frac{0.25 \times 0.5}{0.25 \times 0.5 + 1 \times 0.5} = \frac{0.125}{0.625} = 0.2 \end{aligned}$$

贝叶斯“学习”

- 贝叶斯使用“先验”结合“似然”求得“后验”的特性，使得贝叶斯方法对于新数据的到来，富有弹性和包容的学习能力。
- 对于序贯、贯时数据贝叶斯法则可以进行便捷的“学习”。
- 贝叶斯“学习”：旧数据所得的后验可以成为新数据的先验。
- $$p(\theta|\mathbf{X}_1) \propto p(\theta)p(\mathbf{X}_1|\theta)$$
- $$p(\theta|\mathbf{X}_1, \mathbf{X}_2) \propto p(\theta|\mathbf{X}_1)p(\mathbf{X}_2|\theta)$$
- 时间序列数据：预测股票市场表现。

贝叶斯“学习”：再访血友病案例

- 假设女性 A 第三个儿子出生后，经检测也没有血友病，则则女性 A 带有血友病基因的新后验概率 θ 为：

$$
\begin{aligned}
\Pr(\theta=1|y_1,y_2,y_3) &= \frac{p(\theta=1|y_1,y_2)p(y_3|\theta=1)}{p(y_3|\theta=1)p(\theta=1|y_1,y_2)+p(y_3|\theta=0)p(\theta=0|y_1,y_2)} \\
&= \frac{0.20\times 0.50}{0.50\times 0.20+1\times(1-0.20)} = 0.111
\end{aligned}
$$

贝叶方法的一些基本逻辑

- 聚焦在以分布为主的推论
  - 参数的先验分布
  - 参数的后验分布
- 从观察到的数据更新先验（贝叶斯学习）
- 参数的先验分布大多来自于先前的知识
- 不依赖数据的来源是来自相同条件下无穷试验结果的假设

条件概率求解：先验的使用

- 当使用无信息先验 (Noninformative Prior)——$p(\theta) \propto 1$，两个解法答案是一样的。
-
$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto p(\theta)p(\mathbf{X}|\theta) \\
&= p(\mathbf{X}|\theta)
\end{aligned}
$$
- 共轭先验 (conjugate prior)：能产生闭合形式解的先验，多数情况下与似然同族的先验
- 弱信息先验 (weakly informative prior)：不全然是无信息的先验，但是又与强先验有别，一般用来约束参数边界
- 超参数先验 (hyper-parameter prior)：较为“科学”的贝叶斯介入手段，针对超参数给定先验，而给定非主要参数先验。

贝叶斯方法求解实例

- 在已知数据 **X** 为常态分布，标准差为 σ 的条件下，我们求解 **X** 未知的均值为 θ。假设先验为 $p(\theta) \sim \mathcal{N}(\mu, \tau^2)$，$\theta|\mathbf{X}$ 的后验分布为：
-

$$
\begin{aligned}
p(\theta|\mathbf{X}) &\propto p(\mathbf{X}|\theta)p(\theta) \\
&\propto \underbrace{\prod_{i=1}^{n} \exp\left[-\frac{1}{2\sigma^2}(x_i-\theta)^2\right]}_{\text{似然函数}} \underbrace{\exp\left[-\frac{1}{2\tau^2}(\theta-\mu)^2\right]}_{\text{先验}} \\
&\propto \underbrace{\exp\left[-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{n}{\sigma^2}\right)\left(\theta-\frac{\frac{\mu}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right)^2\right]}_{\text{似然函数与先验的乘积}}
\end{aligned}
$$

贝叶斯求解实例

- 已知数据 X 的条件下，θ 的后验分布是
  $$\hat{\theta} \sim \mathcal{N}\left(\frac{\frac{\mu}{\tau^2}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}, \frac{1}{\frac{1}{\tau^2}+\frac{n}{\sigma^2}}\right);$$
- 有两个重要的特点：
  1. θ 是先验的均值 μ) 和已知数据 X 的均值 ($\bar{x}$) 的加权平均 (Weighted Average)。
  2. $\lim\limits_{n\to\infty} \widehat{\mu} = \bar{x}$, $\lim\limits_{n\to\infty} = \sigma^2/n$。当已知数据数量够大 ($n \to \infty$) 时，对于 θ 的后验分布来说，先验的影响力就相对不那么重要，已知数据的特性决定了 θ 的后验；相反，当数据数量不够多时，先验的选择则会大大地影响后验。

先验的功能：当 $K > N$

- 使用先验可以解决演算过程中，遭遇未知参数无法识别 (unidentifiable) 的问题。
  - 例子：我们想要知道中国 4 个直辖市在 2010 年对于中国整体生产毛额总值的影响。
  - 收集这 4 个直辖市的数据进行分析。可能的变数有：人口数、可耕地面积、工业占所有行业比重、有无自然资源、人均教育水平、性别比等等。
  - 使用简单的回归分析 (如方程式 (1))，我们立即发现在只有 4 个直辖市的已知数据 ($n = 4$)，6 个变数 ($K = 6$) 的情况下，未知数大于已知数 ($n < K + 1$，1 为截距 β_0)，方程式 (1) 中的 $\beta's$ 是无法求解的。

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_{2i} + \ldots + \beta_{ki} + \epsilon_i, \qquad i = 1, \ldots, 4 \quad k = 1, \ldots, 6 \tag{1}$$

$K > N$ 的回归：数据

```
> pop <- rescale(c(5.4, 3.2, 6.5, 9.0))
> land <- rescale(c(2.3, 1.2, 0.5, 0.3))
> indus <- rescale(c(0.4, 0.3, 0.2, 0.15))
> resource <- c(0,1,1,0)
> eduyrs <- rescale(c(8.3, 9.4, 9.8, 13))
> gender <- c(0.3, 0.5, 0.4, 0.48)
> gdp <- rescale(c(5430, 6780, 5980, 10200))
> dat <- cbind.data.frame(gdp, pop, land, indus, resource, eduyrs, gender)
```

$K > N$ 的回归：OLS 回归

```
> M0 <- lm(gdp ~ pop + land + indus + resource + eduyrs + gender, data = dat)
> summary(M0)
...
Residuals:
ALL 4 residuals are 0: no residual degrees of freedom!

Coefficients: (3 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.706e-15         NA      NA       NA
pop         -5.040e+00         NA      NA       NA
land         2.369e+01         NA      NA       NA
indus       -2.759e+01         NA      NA       NA
resource            NA         NA      NA       NA
eduyrs              NA         NA      NA       NA
gender              NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
F-statistic:   NaN on 3 and 0 DF,  p-value: NA
```

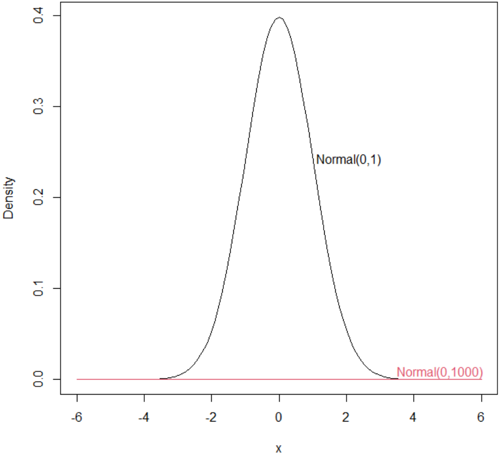
$K > N$ 的回归：贝叶斯回归（强信息先验）

```
> M1 <- stan_glm(gdp ~ pop + land + indus + resource + eduyrs + gender,
+     prior_intercept = normal(0,10), prior = normal(0,1), data = dat)
...

Estimates:
              mean   sd    10%   50%   90%
(Intercept)  0.03   0.51 -0.61  0.01  0.68
pop         -0.25   0.67 -1.03 -0.30  0.63
land        -0.21   0.80 -1.30 -0.17  0.80
indus       -0.08   0.85 -1.14 -0.03  0.96
resource    -0.33   0.56 -1.02 -0.34  0.40
eduyrs       0.68   0.71 -0.23  0.71  1.57
gender       0.31   0.97 -0.95  0.32  1.47
sigma        0.33   0.29  0.06  0.25  0.72

Fit Diagnostics:
           mean   sd    10%   50%   90%
mean_PPD -0.01   0.31 -0.29  0.00  0.27


...
```

$K > N$ 的回归：贝叶斯回归（弱信息先验）

```
> M2 <- stan_glm(gdp ~ pop + land + indus + resource + eduyrs + gender,
+     prior_intercept = normal(0,1000), prior = normal(0,1000), data = dat)
...

Estimates:
              mean    sd      10%     50%     90%
(Intercept)    -1.6   464.6  -593.2   -16.6   587.8
pop           -18.7   474.0  -635.7   -14.0   595.9
land          -19.1   831.4 -1111.9   -13.1  1026.0
indus         -29.1   842.3 -1118.6   -26.0  1040.1
resource      -30.1   485.4  -636.9   -45.3   594.1
eduyrs        -39.5   550.4  -744.8   -40.8   654.6
gender         39.6   946.4 -1152.1    33.7  1259.0
sigma           0.7     0.5     0.2     0.5     1.3

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD  0.0    0.6 -0.6   0.0   0.6

...
```

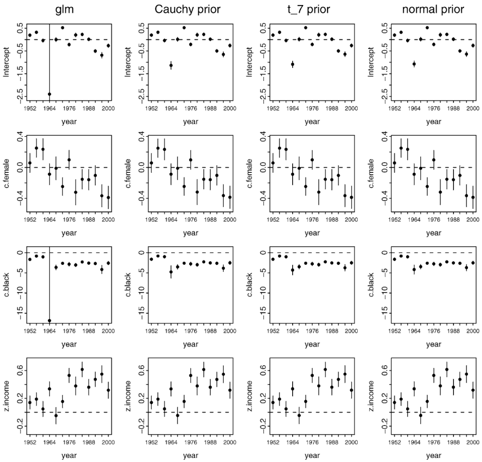
$K > N$ 的回归：强信息先验 vs. 弱信息先验



先验的功能：解决完美分离的问题

- 使用先验可以解决演算过程中，运算不稳定 (computation instability) 的问题。
- 例子：二元因变量在罗吉斯回归中，遭遇完美分离的自变量而运算崩解。

解决完美分离的问题


glm
Cauchy prior
t_7 prior
normal prior
Intercept
c.female
c.black
z.income
year

解决完美分离的问题：数据

```
> n <- 100
> x1 <- rnorm (n)
> x2 <- rbinom (n, 1, .5)
> b0 <- 1
> b1 <- 1.5
> b2 <- 2
> y <- rbinom (n, 1, invlogit(b0+b1*x1+b2*x2))
> y <- ifelse (x2==1, 1, y)
> dat <- cbind.data.frame(y, x1, x2)
```

解决完美分离的问题：logistic 回归

```
> M1 <- glm (y ~ x1 + x2, family=binomial(link="logit"), data = dat)
...
glm(formula = y ~ x1 + x2, family = binomial(link = "logit"),
    data = dat)
            coef.est coef.se
(Intercept)    0.63     0.31
x1             0.73     0.36
x2            19.15  1463.99
---
  n = 100, k = 3
  residual deviance = 60.1, null deviance = 94.3 (difference = 34.1)
```

解决完美分离的问题：logistic 回归

```
> M4 <- stan_glm(y ~ x1 + x2, family=binomial(link="logit"),
+     prior = student_t(7,0,2.5), prior_intercept = student_t(7,0,10), data = dat)
...
Estimates:
              mean   sd   10%   50%   90%
(Intercept) 0.70   0.33 0.28  0.69  1.11
x1          0.74   0.34 0.33  0.72  1.18
x2          5.10   2.03 3.03  4.71  7.64

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 0.8    0.0  0.8   0.8   0.9
...
```

先验的功能：P 值的误区

- P 值在贝叶斯方法中是没有意义的。
- 使用强先验，统计显著性是可以唾手可得的。

P 值的误区：数据

```
n <- 100
a <- 1
b <- 2
x1 <- rnorm(n)
x2 <- rnorm(n)
y <- a + b*x1 + rnorm(100)
dat <- cbind.data.frame(y, x1, x2)
```

P 值的误区：OLS 回归

```
> M6 <- lm(y ~ x1 + x2, data = dat)
> display(M6)
lm(formula = y ~ x1 + x2, data = dat)
            coef.est coef.se
(Intercept) 1.06     0.11
x1          1.96     0.11
x2          0.04     0.09
---
n = 100, k = 3
residual sd = 1.05, R-Squared = 0.76
```

P 值的误区：贝叶斯回归

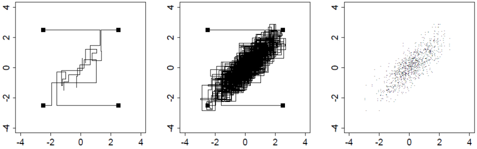
```
data {
  int<lower=0> N;
  vector[N] x1;
  vector[N] x2;
  vector[N] y;
}
parameters {
  vector<lower=0.1>[3] beta;
  real<lower=0> sigma;
}
model {
  sigma ~ normal(0, 10);
  target += normal_lpdf(y| beta[1] + beta[2] * x1 + beta[3] * x2, sigma);
}
```

P 值的误区：贝叶斯回归

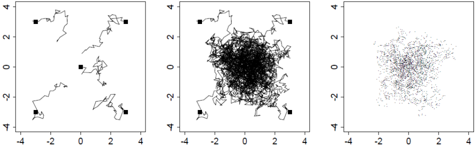
```
> dataList <- with(dat, list("N"= n, "y" = y, "x1" = x1, "x2" = x2))
> BM01 <- stan(file = 'ols.stan', data = dataList, iter = 100, chains = 1)
> BM01 <- stan(fit = BM01, data = dataList, iter = 3000, chains = 3, cores = 3)
> print(BM01)
Inference for Stan model: ols.
3 chains, each with iter=3000; warmup=1500; thin=1;
post-warmup draws per chain=1500, total post-warmup draws=4500.

           mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff Rhat
beta[1]    1.05    0.00 0.11    0.84    0.98    1.05    1.12    1.27  3613    1
beta[2]    1.96    0.00 0.11    1.74    1.89    1.96    2.04    2.18  3980    1
beta[3]    0.16    0.00 0.05    0.10    0.12    0.15    0.19    0.28  3740    1
sigma      1.07    0.00 0.08    0.92    1.01    1.06    1.12    1.24  3963    1
lp__    -150.13    0.04 1.55 -153.91 -150.92 -149.76 -148.99 -148.21  1511    1

Samples were drawn using NUTS(diag_e) at Sun Nov 08 22:43:03 2020.
For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

贝叶斯方法成功落地的原因

- 1990 年后，两大运算法被重新介绍、梳理。
  1. Gibbs Sampling



  2. Metropolis Hasting



贝叶斯方法成功落地的原因

- 大量软件问市：WinBUGS, OpenBUGS, JAGS, R2jags, rjags, MCMCpack, Stan
- Stan (Sampling through adaptive neighbors): `http://mc.stan.org`
  - No-U-Turn sampling algorithm, an extension of Hamiltonian sampling algorithm:
  - R 包：`rstan`、`rstanarm`
  - 跨系统、跨平台、跨语言的贝叶斯软件

一些贝叶斯应用上的问题

- 先验的选择、敏感性分析 (sensitivity analysis)
- 模型检验、模型比较、交叉验证 (cross-validation)
- 后验预测检验 (posterior predictive check)、可视化呈现
- 模型收敛问题、参数重新调整（re-parametrization)、变量标准化 (standardization)

小结

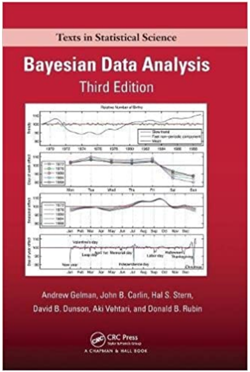
- 使用贝叶斯方法为条件概率求解提供了新的面貌。
- 先验的使用尤其解决了运算上许多难题。
- 让研究者成为更为诚实的科研人员。
- 计算机软、硬件突破性发展，让学者可以使用贝叶斯方法来研究数据上更多的问题。

展望：从贝叶斯方法的视角看待大数据

- 研究前期探索
  - large N data mining $\rightarrow$ pattern recognition $\rightarrow$ (somewhat) stronger informative prior
- 小数据方法验证大数据发现
  - 整合信息量大的先验，产出更合理精确的后验
  - 贝叶斯法则： posterior is the weighted average between the prior and the likelihood.
  - 贝叶斯学习
  - 贝叶斯后验预测
  - Ignorability

贝叶斯方法参考书

- Gelman, Andrew et al, 2014, *Bayesian Data Analysis*, 3rd Edition, Chapman and Hall/CRC.



谢谢！欢迎提问！