Dirichlet 过程及非参数 Bayes 模型

献给严加安院士 80 华诞

张钧曦, 胡耀忠*

Department of Mathematical and Statistical Sciences, University of Alberta at Edmonton, Alberta T6G 2G1, Canada
E-mail: junxi3@ualberta.ca, yaozhong@ualberta.ca

收稿日期: 2020-03-30; 接受日期: 2020-06-15; 网络出版日期: 2021-12-08; * 通信作者
Natural Sciences and Engineering Research Council of Canada (Grant No. RES0038963) 资助项目

摘要 自从 Ferguson 的里程碑式的工作以来, 非参数 Bayes 模型在统计和机器学习等领域中有着广泛的应用, 近年来得到了蓬勃的发展. 它的一个重要的理论基础是一个特殊的随机概率测度族, 即 Dirichlet 过程. 本文介绍 Dirichlet 过程的构造、性质、推广以及它在非参数 Bayes 估计问题中的应用. 另外, 本文也提到双参数 Poisson-Dirichlet 过程、Beta 过程和更一般的断棍 (stick-breaking) 过程以及相关性质.

关键词 Dirichlet 过程 断棍过程 双参数 Poisson-Dirichlet 过程 Poisson 过程 中国餐馆模型 印度自助餐模型 非参数 Bayes 模型 U- 统计量

MSC (2020) 主题分类 60G57, 62F15, 62G05

1 引言

在统计学中, 参数统计首先得到了大量深刻的理论研究和实际应用. 由于在机器学习和大数据处理等相关领域中原始数据量的急剧膨胀, 参数统计中的数据分布拟合工具受到了不同程度的制约. 为了在统计估计、数据拟合和聚类分析等问题中尽可能多地保留和利用原始数据中所包含的大量信息, 近年来, 研究人员广泛地发展和应用非参数统计理论.

像参数统计一样, 非参数统计也可以分为非 Bayes 非参数统计和 Bayes 非参数统计. 前者也有大量的研究结果, 如核密度估计、概率密度估计和 Wilcoxon 符号秩检验等. 在参数 Bayes 统计模型中, 假定样本是从 $F(x, \theta)$ 中抽取的, 其中对每个给定的参数 θ , $F(\cdot, \theta)$ 是一个给定的分布函数, 而参数 θ 本身又被假定是一个给定的随机变量 (先验分布). 有了样本之后, 就可以对 θ 进行估计 (通过 θ 的后验分布), 从而知道这个分布 $F(x, \theta)$. 为了便于计算 θ 的后验分布, 通常假定 θ 本身服从一些特殊的分布, 如 Gauss 分布和 Beta 分布等. 那么非参数 Bayes 模型应该是什么样子呢? 一个办法就是将传统

英文引用格式: Zhang J X, Hu Y Z. Dirichlet process and Bayesian nonparametric models (in Chinese). Sci Sin Math, 2021, 51: 1895–1932, doi: 10.1360/SSM-2020-0097

Bayes 理论中的参数空间, 即参数 θ 所取值的空间, 变成一个无穷维空间. 去掉 $F(\cdot, \theta)$ 中的 θ , 把整个 $F(\cdot)$ 看成是一个参数 ($F(\cdot)$ 本身是未知的, 即非参数统计), 同时再假定 $F(\cdot)$ 是随机的. 换句话说, 假定样本数据所服从的分布 (先验分布) 是一个随机概率测度^[1]. 有了样本观察之后, 再计算这个随机概率测度关于样本观察的条件分布 (后验分布). 当然, 这个先验随机概率测度也不能过于广泛, 我们还是希望它属于某一类型. 这样, 问题就转化成了如何选择这些先验随机概率测度, 即如何选择非参数 Bayes 模型的先验 (随机) 分布. 关于这点, 很自然地应该遵守以下两个原则:

- (1) 先验分布的随机概率族空间需要足够大 (使得它能够包含数据所服从的分布);
- (2) 有了样本观察值之后, 想要得到的后验分布要便于计算和分析.

假定 $(\Omega, \mathcal{F}, P)$ 是一个概率空间, $(\mathbb{X}, \mathcal{X})$ 是一个可测空间. Dirichlet 过程 $\{P(A, \omega), A \in \mathcal{X}, \omega \in \Omega\}$ 是一族特殊的随机概率测度: 对几乎所有的 $\omega \in \Omega$, $P(\cdot, \omega)$ 是 $(\mathbb{X}, \mathcal{X})$ 上的概率测度, 而固定 $A \in \mathcal{X}$, $P(A, \cdot)$ 是 $(\Omega, \mathcal{F}, P)$ 上的随机变量; 若对 $\mathbb{X}$ 的任意有限可测分割 $(B_1, \dots, B_k)$, 都有 $(P(B_1), \dots, P(B_k))$ 是 Dirichlet 分布, 则把 P 叫作 Dirichlet 过程 (参见下一节的定义 2.4). Dirichlet 过程可以看成是 Dirichlet 分布的无穷维推广, 是一个无尾的分布 (见定义 2.7, 以后会更详细地讨论). 自从 Ferguson^[2] 于 1973 年奠定了里程碑式的基础以来, Dirichlet 过程在非参数 Bayes 模型中一直占据着非常重要的位置, 非参数 Bayes 统计也得到了长足的发展. Dirichlet 过程之所以能在非参数 Bayes 模型中有着这样一个重要的作用, 主要在于它具有下面这个特性: 如果先验分布是一个 Dirichlet 过程, 那么所得到的后验分布也是一个 Dirichlet 过程. 而且 Dirichlet 过程所确定的随机概率族也包含足够多的分布. 这样, 它满足上面所提到的两个原则, 构成了非参数 Bayes 模型中一类非常理想的随机概率族. 它在很多领域, 包括在机器学习中, 有着广泛的应用, 很多学者对此做了大量的工作. 以下简单提及一些结果.

(1) 假定样本观察值 $X_1, \dots, X_n, \dots$ 是从先验随机概率测度 P 得到的. 即假定在已知 P 的条件下, $X_1, \dots, X_n, \dots$ 是独立同分布的, $X_1, \dots, X_n, \dots$ 本身一般并不是独立的. 但在很多问题中, 例如, 在对一些统计量的分析中, 需要知道 $X_1, \dots, X_n, \dots$ 本身的有限维分布. 如何确定这些有限维分布呢? Blackwell 和 MacQueen^[3] 提出了 Dirichlet 过程的 Pólya 罐子模型. 该模型给出了 $X_{n+1} | X_1, \dots, X_n$ 的分布的显式表达式. 通过递推的办法就可以用它来找到随机变量 $X_1, \dots, X_n$ 的联合分布. 在应用中也可以通过 Pólya 罐子模型结合 Markov 链 Monte Carlo 方法作一些具体的计算 (参见文献 [4-7]).

(2) Sethuraman^[8] 提出了用断棍 (stick-breaking) 构造的方法来构造 Dirichlet 过程. 这一构造使得我们可以把 Dirichlet 过程看作是一个离散的随机概率测度族. 另外, 断棍过程是用一个随机级数来表达的, 所以, 在实际的计算中, 在允许的误差范围内, 人们可以利用这一随机级数的有限项截断来逼近这一 Dirichlet 过程. 除了 Dirichlet 过程外, 一个很重要的随机概率测度是由 Hjort^[9] 首先提出的 Beta 过程. 这一特殊的随机概率测度的有限维分布也有显式的表达式. 与 Dirichlet 过程不同的是, Beta 过程还是一个 Lévy 过程, 可以应用一整套关于 Lévy 过程的理论 (参见文献 [10, 11]) 分析 Beta 过程, 它在生存分析中有重要应用. 自从有了 Beta 过程的断棍构造的显式表达式后, 它在机器学习问题中的隐条件模型中得到了重要应用 (参见文献 [12, 13]). 另外, 如果把 Dirichlet 过程的断棍过程显式表达式中的某些限制 (独立性、权重的构造方式等) 加以变化, 可以得到更一般的断棍过程. 在这些推广的断棍过程中, 一个非常重要的随机概率测度是, 由 Pitman 和 Yor 于 1997 年提出的双参数 Poisson-Dirichlet 过程 (参见文献 [14, 15]). 该过程在生物种群信息、基因表达聚类问题等有广泛的应用. 不同于 Dirichlet 过程和 Beta 过程的是, 该过程的有限维联合分布没有简单的显式表达式, 因此, 断棍过程的表达形式是用来研究双参数 Poisson-Dirichlet 过程的较为有效的方法. 在其他的一些具体问题的应用中, 如金融数学和环境科学, 人们经常需要构造多元随机波动模型或者空间过程, Rodríguez 和 Dunson^[16] 引进了 Probit 断棍过程, 作为一个特殊的断棍过程, 它在金融和环境问题中有较为广泛

的应用.

根据在不同领域里的应用中的需要, 人们已经对 Dirichlet 过程进行了多种多样的推广. 其中有代表性的有混合 Dirichlet 过程^[17]、Dirichlet 混合模型^[17]、层次 Dirichlet 过程^[18]、嵌入 Dirichlet 过程^[19] 和关联 Dirichlet 过程^[20,21] 等. 本文也会对这些推广进行一个简单的介绍.

(3) 正如上述所提到的, Dirichlet 过程这个随机概率测度族在某种意义上是离散型的. 正是由于它的这种离散性, 使得 Dirichlet 过程可以应用到无监督学习中的聚类问题上. 这方面的应用主要是通过中国餐馆模型来实现. 以后将会看到, 中国餐馆模型从聚类的角度描述了 Pólya 罐子模型, 两者是等价的. 因此, 中国餐馆模型也是 Dirichlet 过程的又一个表示方式, 该模型也给出了解决聚类问题的算法的理论基础. 与此有关的另一个模型是用来表示 Beta 过程的印度自助餐模型.

本文余下内容的结构如下: 第 2 节首先从随机测度的角度给出 Dirichlet 过程的定义, 并给出 Dirichlet 过程的一些主要性质和其他构造方法; 第 3 节介绍断棍过程、双参数 Poisson-Dirichlet (Pitman-Yor) 过程、Beta 过程以及这些过程的一些主要推广; 第 4 节给出如何在非参数 Bayes 模型中使用 Dirichlet 过程对不同的统计量进行估计.

2 Dirichlet 过程

2.1 随机测度

随机 (概率) 测度是概率理论中一个重要的概念. Dirichlet 过程是一个特殊的随机概率测度. 本节简要介绍随机测度的概念, 关于随机测度的更多内容可参见文献 [1].

首先, 定义核函数.

定义 2.1 (核函数) 假定 $(S, \mathcal{S})$ 和 $(T, \mathcal{T})$ 为两个可测空间. 从 $S \times \mathcal{T}$ 到 $[0, +\infty]$ 的函数

$$\mu(\cdot, \cdot) : S \times \mathcal{T} \rightarrow [0, +\infty]$$

被称为从 S 到 T 的核函数, 如果这个函数 μ 满足如下的两个条件:

- (1) 对于任何给定的 $B \in \mathcal{T}$, $\mu(\cdot, B)$ 是 $\mathcal{S}$ -可测的;
- (2) 对于任何给定的 $s \in S$, $\mu(s, \cdot)$ 是 $(T, \mathcal{T})$ 上的测度.

核函数的一个简单例子是 $\mu(s, A) = \delta_s(A) = \mathbb{1}_A(s)$, 其中 $s \in S$, $A \subseteq S$, 而

$$\mathbb{1}_A(s) = \begin{cases} 1, & \text{当 } s \in A \text{ 时,} \\ 0, & \text{当 } s \notin A \text{ 时} \end{cases}$$

为示性函数. 如果把上述定义核函数的第二个条件换成“对于任何给定的 $s \in S$, $\mu(s, \cdot)$ 是 $(T, \mathcal{T})$ 上的一个概率测度”, 则 μ 称为概率核函数 (probability kernel function). 通过核函数的定义可以给出以下随机测度的定义.

定义 2.2 (随机测度) 假定 $(\Omega, \mathcal{F}, P)$ 是一个概率空间, $\mathbb{X}$ 是度量空间, $\mathcal{X} = \mathcal{B}(\mathbb{X})$ 是 $\mathbb{X}$ 的 Borel σ -代数. 称从 Ω 到 $\mathbb{X}$ 的局部有限核函数 ζ 为在 $(\mathbb{X}, \mathcal{X})$ 上的随机测度 (也叫作 $(\Omega, \mathcal{F}, P) \times (\mathbb{X}, \mathcal{X})$ 上, 或简单一点, $\Omega \times \mathbb{X}$ 上的随机测度). 这里“局部有限”是指, 对于所有有界可测集合 $\tilde{B} \in \mathcal{X}$ 和所有 $\omega \in \Omega$ (除去一个 P -零测子集), 都有

$$\zeta(\omega, \tilde{B}) < \infty.$$

如果随机测度 ζ 是一个概率核函数, 则这个随机测度叫作随机概率测度. 如果对于 $\mathbb{X}$ 中任意有限多个互不相交的集合 $B_1, B_2, \dots, B_n$, 相对应的随机变量 (概率核) $\zeta(B_1), \zeta(B_2), \dots, \zeta(B_n)$ 是互相独立的, 则该概率核函数被称为完全随机的 (completely random).

有关随机测度的研究, 尤其是关于随机概率测度的研究, 在非参数 Bayes 模型中有着重要的应用. 这种现象的产生主要是因为非参数 Bayes 模型中的先验分布不再是基于某个带 (随机) 参数的特定分布, 而是将整个先验分布作为一个 “参数”. 因此, 在非参数 Bayes 模型中, 先验分布需要用随机概率测度表示. 而一系列特殊的随机概率测度, 如 Dirichlet 过程和双参数 Poisson-Dirichlet 过程等, 在机器学习、生物统计等领域受到广泛的欢迎. 要强调的是, 在许多关于非参数 Bayes 统计模型的文献中, 人们往往把随机概率测度叫作随机过程, 本文也就沿用这个名称.

注 2.1 本文除非明确标出, 所有的随机变量都是定义在概率空间 $(\Omega, \mathcal{F}, P)$ 上的, 随机概率测度都像上面一样定义在 $(\Omega, \mathcal{F}, P) \times (\mathbb{X}, \mathcal{X})$ 上. 在不引起混淆的情形下, 也将省略概率空间 $(\Omega, \mathcal{F}, P)$ 和测度空间 $(\mathbb{X}, \mathcal{X})$. 当 $\mathbb{X}$ 是 Polish 空间时, 将 $\mathbb{X}$ 和它的 Borel σ -代数统称为 Polish 空间 $(\mathbb{X}, \mathcal{X})$.

2.2 Dirichlet 过程

Dirichlet 过程和非参数 Bayes 模型是由 Ferguson 首次在他 1973 年所发表的奠基性的文献 [2] 中提出的并对其进行了深入的讨论. 之后, Dirichlet 过程在非参数 Bayes 模型中得到了进一步的应用, 成为了一个最基本的工具. 本小节给出 Dirichlet 过程的定义、一些基本的性质和几个用处广泛的表示形式, 以及其在非参数 Bayes 模型的应用基础.

定义 2.3 (Dirichlet 分布) 令 $\alpha_1, \dots, \alpha_n \in (0, \infty)$ 为 n 个给定的正数. 称随机向量 $(X_1, \dots, X_n)$ 服从参数为 $(\alpha_1, \dots, \alpha_n)$ 的 Dirichlet 分布, 记作 $(X_1, \dots, X_n) \sim \text{Dir}(\alpha_1, \dots, \alpha_n)$, 如果以下两个条件得到满足:

- (1) 对所有的 $i \in \{1, \dots, n\}$, $X_i \geq 0$ 且 $\sum_{i=1}^n X_i = 1$;
- (2) $(X_1, \dots, X_{n-1})$ 的联合密度函数可表示为

$$f(x_1, \dots, x_{n-1}) = \frac{\Gamma(|\alpha|)}{\prod_{i=1}^n \Gamma(\alpha_i)} \prod_{i=1}^{n-1} x_i^{\alpha_i-1} \left(1 - \sum_{i=1}^{n-1} x_i\right)^{\alpha_n-1},$$

$$0 \leq x_1 + \dots + x_{n-1} \leq 1,$$

其中 $|\alpha| = \sum_{i=1}^n \alpha_i$, $\Gamma(a) = \int_0^\infty x^{a-1} dx$ ($a > 0$) 是 Gamma 函数.

Dirichlet 分布有许多重要的性质, 这些性质是研究 Dirichlet 过程的工具, 在此进行简要介绍并给出证明概要.

命题 2.1 假定 $Y_1, \dots, Y_n$ 是一组随机变量, 记 $Y = \sum_{i=1}^n Y_i$.

- (1) 如果 $Y_i \stackrel{\text{ind}}{\sim} \text{Gamma}(\alpha_i, 1)$, $i = 1, \dots, n$, 则

$$\left(\frac{Y_1}{Y}, \dots, \frac{Y_n}{Y}\right) \sim \text{Dir}(\alpha_1, \dots, \alpha_n),$$

而且 $(\frac{Y_1}{Y}, \dots, \frac{Y_n}{Y})$ 与 Y 是相互独立的.

- (2) 如果 $Y_i \stackrel{\text{ind}}{\sim} \text{Beta}(\alpha_i, 1)$, $i = 1, \dots, n$, 则

$$(Y_1, \dots, Y_n) | Y = 1 \sim \text{Dir}(\alpha_1, \dots, \alpha_n).$$

(3) 如果 $Y_i \stackrel{\text{ind}}{\sim} \text{Exp}(\alpha_i)$, $i = 1, \dots, n$, 则

$$(e^{-Y_1}, \dots, e^{-Y_n}) \mid \sum_{i=1}^n e^{-Y_i} = 1 \sim \text{Dir}(\alpha_1, \dots, \alpha_n).$$

证明 (1) 令 $X_i = \frac{Y_i}{Y}$, 构造如下一对一的变换 T :

$$(Y_1, \dots, Y_n) = T(Y, X_1, \dots, X_{n-1}) = \left(YX_1, \dots, YX_{n-1}, Y \left(1 - \sum_{i=1}^{n-1} X_i \right) \right),$$

通过变换 T 和 $(Y_1, \dots, Y_n)$ 的联合分布可以得到 $(Y, X_1, \dots, X_{n-1})$ 的联合分布, 对 Y 进行积分可得 $(X_1, \dots, X_{n-1})$ 的联合分布.

(2) 构造如下一对一的变换 S :

$$(Y_1, \dots, Y_n) = S(Y, X_1, \dots, X_{n-1}) = \left(X_1, \dots, X_{n-1}, Y - \sum_{i=1}^{n-1} X_i \right),$$

通过变换 S 和 $(Y_1, \dots, Y_n)$ 的联合分布可以得到 $(Y, X_1, \dots, X_{n-1})$ 的联合分布, 由此可以看出 $X_1, \dots, X_{n-1}$ 和 Y 是相互独立的. 把所得到的联合分布对 Y 进行积分可得 $(X_1, \dots, X_{n-1})$ 的联合分布, 进而得到 $(X_1, \dots, X_{n-1}) \mid Y$ 的分布.

(3) 如果 $Y_i \stackrel{\text{ind}}{\sim} \text{Exp}(\alpha_i)$, 那么 $e^{-Y_i} \stackrel{\text{ind}}{\sim} \text{Beta}(\alpha_i, 1)$, 由结论 (2) 可证 (3). $\square$

命题 2.2 假定 $(X_1, \dots, X_n) \sim \text{Dir}(\alpha_1, \dots, \alpha_n)$, $\{I_1, \dots, I_m\}$ 是 $\{1, \dots, n\}$ 的任意一个分割, 记 $Z_j = \sum_{i \in I_j} X_i$, 则

(1) $(Z_1, \dots, Z_m) \sim \text{Dir}(\beta_1, \dots, \beta_m)$, 其中 $\beta_j = \sum_{i \in I_j} \alpha_i$;

(2) 对于任意的 $j \in \{1, \dots, m\}$, 有 $(\frac{X_i}{Z_j} : i \in I_j) \stackrel{\text{ind}}{\sim} \text{Dir}(\alpha_i : i \in I_j)$;

(3) 对于任意的 $j \in \{1, \dots, m\}$, $(Z_1, \dots, Z_m)$ 和 $(\frac{X_i}{Z_j} : i \in I_j)$ 是相互独立的.

反过来, 给定 $\{1, \dots, n\}$ 的一个分割 $\{I_1, \dots, I_m\}$ 和 $Z_j = \sum_{i \in I_j} X_i$, 如果随机向量 $(X_1, \dots, X_n)$ 满足上述的 (1)–(3), 则有 $(X_1, \dots, X_n) \sim \text{Dir}(\alpha_1, \dots, \alpha_n)$.

证明 根据命题 2.1 可知, 对任意的 $i \in \{1, \dots, n\}$, 有 $X_i = \frac{Y_i}{Y}$, 其中 $Y_i \stackrel{\text{ind}}{\sim} \text{Gamma}(\alpha_i, 1)$. 结论 (1)–(3) 和最后的逆命题都可以通过这样的表示方式证明. $\square$

命题 2.3 假定 $p = (p_1, \dots, p_n) \sim \text{Dir}(\alpha_1, \dots, \alpha_n)$, 随机向量 $N = (N_1, \dots, N_n)$ 满足 $N \mid p \sim \text{MN}_n(p_1, \dots, p_n)$, 其中 $\text{MN}_n(p_1, \dots, p_n)$ 是参数为 p 的 n 维多项式分布, 则 $p \mid N \sim \text{Dir}(\alpha_1 + N_1, \dots, \alpha_n + N_n)$.

证明 由于 p 和 $N \mid p$ 的分布都可以显式地给出, 所以可以通过 p 和 $N \mid p$ 的密度函数得到 (p, N) 的联合分布和 $p \mid N$ 的条件分布. $\square$

定义 2.4 (Dirichlet 过程) 假定 $(\mathbb{X}, \mathcal{X})$ 是 Polish 空间, α 是 $(\mathbb{X}, \mathcal{X})$ 上的非负局部有限测度, 称 $(\mathbb{X}, \mathcal{X})$ 上的随机概率测度 P 为 $(\mathbb{X}, \mathcal{X})$ 上的参数为 α 的 Dirichlet 过程, 如果对任意的 $k \in \mathbb{N}$ 和 $\mathbb{X}$ 的任意可测分割 $(B_1, \dots, B_k)$ 都有

$$(P(B_1), \dots, P(B_k)) \sim \text{Dir}(\alpha(B_1), \dots, \alpha(B_k)). \quad (2.1)$$

参数 α 被称为基测度 (base measure).

通常称 $|\alpha| = \alpha(\mathbb{X})$ 为集中参数 (concentration parameter), 而如果对测度 α 进行标准化, $\bar{\alpha} = \frac{\alpha}{|\alpha|}$ 则是一个概率测度, 被称作中心测度 (center parameter). 用 $P \sim \text{DP}(\alpha)$ 或者 $P \sim \text{DP}(|\alpha|, \bar{\alpha})$ 表示 P 是一个参数为 α 的 Dirichlet 过程.

假定 $(\mathbb{Y}, \mathcal{Y})$ 为标准的 Borel 空间, $\psi: \mathbb{X} \rightarrow \mathbb{Y}$ 为任意可测函数, P 为 $(\mathbb{X}, \mathcal{X})$ 上参数为 α 的 Dirichlet 过程, 则 $P \circ \psi^{-1} \sim \text{DP}(\alpha \circ \psi^{-1})$ 是 $(\mathbb{Y}, \mathcal{Y})$ 上的 Dirichlet 过程.

注 2.2 (1) 记 $\mathbb{M}$ 为 $(\mathbb{X}, \mathcal{X})$ 上所有概率测度的全体, 在 $\mathbb{M}$ 上可以引进全变差度量, 或其他度量 (这在本文中并不重要), 这样, $\mathbb{M}$ 就成了一个度量空间. 记 $\mathcal{M}$ 为 $\mathbb{M}$ 的相对应的 Borel σ -代数. 这样, $(\mathbb{M}, \mathcal{M})$ 就是一个可测空间, 是概率论中一个重要概念 (参见文献 [22]). 显然, 从定义 2.4 中可知, 对于任何 $\omega \in \Omega$, 有 $P(\omega) \in \mathbb{M}$, 即可以把 Dirichlet 过程看成是一个取值于 $\mathbb{M}$ 上的随机变量, 或取值于 $(\mathbb{M}, \mathcal{M})$ 上的分布.

(2) 一个取值于度量空间 $(\mathbb{M}, \mathcal{M})$ (也即 $(\mathbb{X}, \mathcal{X})$ 上所有概率测度的全体) 中的随机过程被称为测度值过程. 这样的过程在分支过程、超过程和随机偏微分方程等领域有重要应用 (参见文献 [23–26]).

2.3 Dirichlet 过程的其他构造和表示方法

多种 Dirichlet 过程的构造方式给出了 Dirichlet 过程多个不同但等价的定义方法. 本文涉及其中主要的几种, 即 Gamma 过程和断棍过程的等价表达方法.

定义 2.5 假定 α 是 $(\mathbb{X}, \mathcal{X})$ 上的非负局部有限测度. $(\mathbb{X}, \mathcal{X})$ 上的随机概率测度 U 被称为参数为 α 的 Gamma 过程, 如果对于 $\mathbb{X}$ 的任意可测的划分 $(A_1, \dots, A_n)$, 有 $U(A_i) \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha(A_i), 1)$, $i = 1, \dots, n$, 也即 $U(A_i)$ 的分布为

$$f(x) = \frac{x^{\alpha(A_i)-1} e^{-x}}{\Gamma(\alpha(A_i))}, \quad x \geq 0.$$

命题 2.4 假定 U 是 $(\mathbb{X}, \mathcal{X})$ 上的参数为 α 的 Gamma 过程, 则 $P(\cdot) = \frac{U(\cdot)}{U(\mathbb{X})}$ 是 $(\mathbb{X}, \mathcal{X})$ 上的参数为 α 的 Dirichlet 过程, 即 $P \sim \text{DP}(\alpha)$.

证明 只要证明对任何一组随机变量 $Y_1, \dots, Y_k$, 如果 $Y_i \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha_i)$, 则

$$\left(\frac{Y_1}{Y}, \dots, \frac{Y_k}{Y} \right) \sim \text{Dir}(\alpha_1, \dots, \alpha_k), \quad \text{其中 } Y = \sum_{i=1}^k Y_i.$$

而上述结论是显然的. □

定理 2.1 (断棍过程的构造 [8]) 假定 $\theta_1, \theta_2, \dots \stackrel{\text{iid}}{\sim} \bar{\alpha}$, $v_1, v_2, \dots \stackrel{\text{iid}}{\sim} \text{Beta}(1, |\alpha|)$, 并且 $(\theta_1, \theta_2, \dots)$ 和 $(v_1, v_2, \dots)$ 是独立的. 令

$$w_j = v_j \prod_{l=1}^{j-1} (1 - v_l), \quad j = 1, 2, \dots, \quad (2.2)$$

其中, 当 $j = 1$ 时, 约定 $\prod_{l=1}^{j-1} (1 - v_l) = 1$, 则

$$P = \sum_{j=1}^{\infty} w_j \delta_{\theta_j} \sim \text{DP}(\alpha). \quad (2.3)$$

证明 首先,

$$\lim_{j \rightarrow \infty} \mathbb{E} \left[1 - \sum_{l=1}^j w_l \right] = \lim_{j \rightarrow \infty} \mathbb{E} \left[\prod_{l=1}^{j-1} (1 - v_l) \right] = \lim_{j \rightarrow \infty} \left(\frac{|\alpha|}{1 + |\alpha|} \right)^j = 0.$$

由于 $1 - \sum_{l=1}^j w_l$ 对于 j 是几乎处处递减、非负且有界的, 因此, 上述结果推出几乎处处有 $\sum_{l=1}^{\infty} w_l = 1$. 由此可知 P 定义了随机概率测度.

对于所有的 $j \geq 1$, 定义 $\tilde{w}_j = v_{j+1} \prod_{l=2}^j (1 - v_l)$, $\tilde{\theta}_j = \theta_{j+1}$, 则有 $w_{j+1} = (1 - v_1) \tilde{w}_j$, 因此可得等式

$$P = w_1 \delta_{\theta_1} + \sum_{j=2}^{\infty} w_j \delta_{\theta_j} = v_1 \delta_{\theta_1} + (1 - v_1) \sum_{j=1}^{\infty} \tilde{w}_j \delta_{\tilde{\theta}_j}.$$

但根据上述 $\tilde{w}_j$ 和 $\tilde{\theta}_j$ 的构造可知, $\tilde{P} := \sum_{j=1}^{\infty} \tilde{w}_j \delta_{\tilde{\theta}_j}$ 与 P 有相同的结构, 因此, 它们也有相同的分布, 而且 $\tilde{P}$ 与 (v_1, θ_1) 相互独立, 这样, P 满足等式

$$P \stackrel{d}{=} v \delta_{\theta} + (1 - v) P, \quad (2.4)$$

其中 $P, v \sim \text{Beta}(1, |\alpha|)$, $\theta \sim \bar{\alpha}$ 是相互独立的.

对于任意给定的整数 $k \geq 2$, 假定 $p \sim \text{Dir}(\alpha_1, \dots, \alpha_k)$, $v \sim \text{Beta}(1, |\alpha|)$, $N \sim \text{MN}_k(\alpha_1, \dots, \alpha_k)$ 相互独立, 其中 $\text{MN}_k(\alpha_1, \dots, \alpha_k)$ 为 k 维多项式分布, 则容易证明 $vN + (1 - v)p \sim \text{Dir}(\alpha_1, \dots, \alpha_k)$. 因为 P 满足 (2.4), 而 δ_{θ} 是一个特殊的多项式分布, 可知 $P \sim \text{DP}(\alpha)$.

还可以证明等式 (2.4) 的解在分布意义下是唯一的. 假设 P 和 P' 同时满足等式 (2.4) 却不同分布, 令 $(u_n, \vartheta_n)_{n=1}^{\infty}$ 为 (v, θ) 的相互独立的复制, 而且对任意的 $n \in \mathbb{N}^+$, (u_n, ϑ_n) 与 P 和 P' 互相独立. 定义两组如下的随机变量 $\{P_n\}_{n=1}^{\infty}$ 和 $\{P'_n\}_{n=1}^{\infty}$:

$$\begin{aligned} P_1 &= P, & P_{n+1} &= u_n \delta_{\vartheta_n} + (1 - u_n) P_n, \\ P'_1 &= P', & P'_{n+1} &= u_n \delta_{\vartheta_n} + (1 - u_n) P'_n. \end{aligned}$$

根据以上对等式 (2.4) 的分析可知, 对任意的 $n \in \mathbb{N}^+$, 可推出 P_n 与 P 同分布, P'_n 与 P' 同分布. 然而, 因为 $\{u_n\}_{n=1}^{\infty}$ 是独立同分布的随机变量, 服从 $\text{Beta}(1, |\alpha|)$ 分布, 所以, $P(|1 - u_1| \leq 1) = 1$, 而且 $P(|1 - u_1| = 1) = 0$, 因此, $P(|1 - u_1| < 1) = 1$. 从而, 依概率 1 有

$$\begin{aligned} |P_{n+1}(A) - P'_{n+1}(A)| &= |1 - u_n| |P_n(A) - P'_n(A)| \\ &= \prod_{i=1}^n |1 - u_i| |P_1(A) - P'_1(A)| \\ &\xrightarrow{n \rightarrow \infty} 0, \quad \forall A \in \mathcal{X}. \end{aligned}$$

这与假设 P 和 P' 同时满足等式 (2.4) 却不同分布的命题互相矛盾. 因此, 等式 (2.4) 的解唯一. $\square$

注 2.3 由 (2.3) 定义的过程被称为断棍 (stick-breaking) 过程. 该表达方法由 Sethuraman^[8] 提出并给予证明. 作为 Dirichlet 过程的一种表示方法, 这种断棍的构造允许在实际应用中用有限项截断的断棍过程^[27, 28]逼近 Dirichlet 过程. 另外, 从 Dirichlet 过程的断棍表示方法可以看出 P 是离散测度. 如果对定理 2.1 中关于 θ_i 和 v_i 所加的限制加以变化将会得到更一般的断棍过程, 这在第 3 节将会有更详细的讨论.

2.4 一些常用性质

本小节给出 Dirichlet 过程常用的一些性质.

命题 2.5 (Dirichlet 过程的矩) 假定 P 为在 $(\mathbb{X}, \mathcal{X})$ 上参数为 α 的 Dirichlet 过程. 对于任何的 $A, B \in \mathcal{X}$, 有

$$E[P(A)] = \bar{\alpha}(A), \quad (2.5)$$

$$\text{Var}[P(A)] = \frac{\bar{\alpha}(A)\bar{\alpha}(A^c)}{|\alpha| + 1}, \quad (2.6)$$

$$\text{Cov}[P(A), P(B)] = \frac{\bar{\alpha}(A \cap B) - \bar{\alpha}(A)\bar{\alpha}(B)}{|\alpha| + 1}. \quad (2.7)$$

证明 (2.5) 和 (2.6) 可以利用 $\text{Dir}(\alpha(A), \alpha(A^c)) = \text{Beta}(\alpha(A), \alpha(A^c))$ 的分布的显式表达式通过直接计算得到. 而 (2.7) 可以分以下两种情形来讨论.

情形 1 当 $A \cap B = \emptyset$ 时, $(P(A), P(B), P(A^c \cap B^c)) \sim \text{Dir}(\alpha(A), \alpha(B), \alpha(A^c \cap B^c))$, 协方差可以从相对应的 $\text{Dir}(\alpha(A), \alpha(B), \alpha(A^c \cap B^c))$ 分布的协方差得到.

情形 2 当 $A \cap B \neq \emptyset$ 时, $A = (A \cap B) \cup (A \cap B^c)$, $B = (B \cap A) \cup (B \cap A^c)$. 相对应地, 有

$$\begin{aligned} \text{Cov}[P(A), P(B)] &= \text{Cov}[P((A \cap B) \cup (A \cap B^c)), P((B \cap A) \cup (B \cap A^c))] \\ &= \text{Cov}[P(A \cap B) + P(A \cap B^c), P(B \cap A) + P(B \cap A^c)]. \end{aligned} \quad (2.8)$$

展开上述协方差后, 可以利用情形 1 得到的结果 (即 (2.6)) 来得到 (2.7). $\square$

命题 2.6 ^[2] 假定 P 为 $(\mathbb{X}, \mathcal{X})$ 上参数为 α 的 Dirichlet 过程. 对于 $B \in \mathcal{X}$, 如果 $\alpha(B) = 0$, 则依概率 1 有 $P(B) = 0$. 如果 $\alpha(B) > 0$, 则依概率 1 有 $P(B) > 0$.

证明 考虑一个对 $\mathbb{X}$ 的分割 (A, A^c) , 根据定义 2.4 可知 $P(A) \sim \text{Beta}(\alpha(A), \alpha(A^c))$. 而上述命题可由 Beta 分布的性质直接得到. $\square$

命题 2.6 给人的感觉是说 α 和 P 是相互绝对连续的, 可事实并非如此, 这主要是因为 P 是离散的测度, 而 α 可以是连续的.

命题 2.7 ^[2] 如果 $P \sim \text{DP}(\alpha)$, 则对一序列给定的递减的可测集 $A_n \searrow \emptyset$, 依概率 1 有 $P(A_n) \rightarrow 0$.

证明 根据 α 的 σ -可加性可知依概率 1 有 $\alpha(A_n) \rightarrow 0$. 因此, 存在 $\{A_n\}$ 的一个子列 $\{A_{n_j}\}$ 使得 $\sum_{j=1}^{\infty} \alpha(A_{n_j}) < \infty$. 对于固定的 $\epsilon > 0$, 有

$$\sum_{j=1}^{\infty} P[P(A_{n_j}) < \epsilon] \leq \sum_{j=1}^{\infty} \frac{E[P(A_{n_j})]}{\epsilon} = \frac{\sum_{j=1}^{\infty} \alpha(A_{n_j})}{\epsilon|\alpha|} < \infty.$$

根据 Borel-Cantelli 定理可知 $P[P(A_{n_j}) > \epsilon, \text{i.o.}] = 0$. 这样依概率 1 有 $P(A_{n_j}) \rightarrow 0$. 因为依概率 1 有 $P(A_1) > P(A_2) > \dots$, 所以依概率 1 有 $P(A_n) \rightarrow 0$. $\square$

命题 2.8 (支撑 ^[2]) 令 $P \sim \text{DP}(\alpha)$ 而且 $Q \ll \alpha$ 是 $(\mathbb{X}, \mathcal{X})$ 上的概率测度. 对于任何正整数 m 和可测集合 $A_1, \dots, A_m$ 以及任意的 $\epsilon > 0$, 有

$$P(|P(A_i) - Q(A_i)| < \epsilon, i = 1, \dots, m) > 0.$$

证明 对任意的 $j \in \{1, \dots, m\}$, 假定 $v_j \in \{0, 1\}$, 定义 $B_{v_1, \dots, v_m}$ 如下:

$$B_{v_1, \dots, v_m} = \bigcap_{j=1}^m A_j^{v_j},$$

其中 $A_j^1 = A_j$, $A_j^0 = A_j^c$ 是 A_j 的补集. 通过这样的定义可知, $\{B_{v_1, \dots, v_m}\}$ 是 $\mathbb{X}$ 的一个分割, 而且 $A_j = \bigcup_{(v_1, \dots, v_m): v_j=1} B_{v_1, \dots, v_m}$, 则有

$$\begin{aligned} & \mathbb{P}(|P(A_i) - Q(A_i)| < \epsilon, i = 1, \dots, m) \\ & \geq \mathbb{P}\left(\sum_{(v_1, \dots, v_m): v_i=1} |P(B_{v_1, \dots, v_m}) - Q(B_{v_1, \dots, v_m})| < \epsilon, i = 1, \dots, m\right). \end{aligned}$$

因此只需证明

$$\mathbb{P}\left(|P(B_{v_1, \dots, v_m}) - Q(B_{v_1, \dots, v_m})| < \frac{\epsilon}{2^m}, \forall (v_1, \dots, v_m)\right) > 0. \quad (2.9)$$

分两种情形来证.

(1) 对 $(v_1, \dots, v_m)$ 使得 $\alpha(B_{v_1, \dots, v_m}) = 0$, 有 $Q(B_{v_1, \dots, v_m}) = 0$, 而且根据命题 2.6, 依概率 1 有 $P(B_{v_1, \dots, v_m}) = 0$, 因此依概率 1 有

$$|P(B_{v_1, \dots, v_m}) - Q(B_{v_1, \dots, v_m})| = 0.$$

(2) 对 $(v_1, \dots, v_m)$ 使得 $\alpha(B_{v_1, \dots, v_m}) > 0$, 有 $P(B_{v_1, \dots, v_m}) \sim \text{Beta}(\alpha(B_{v_1, \dots, v_m}), \alpha(B_{v_1, \dots, v_m}^c))$, 根据 Beta 分布的定义有

$$\begin{aligned} & \mathbb{P}\left(|P(B_{v_1, \dots, v_m}) - Q(B_{v_1, \dots, v_m})| < \frac{\epsilon}{2^m}\right) \\ & = \mathbb{P}\left(Q(B_{v_1, \dots, v_m}) - \frac{\epsilon}{2^m} < P(B_{v_1, \dots, v_m}) < Q(B_{v_1, \dots, v_m}) + \frac{\epsilon}{2^m}\right) \\ & = \kappa \int_{Q(B_{v_1, \dots, v_m}) - \frac{\epsilon}{2^m}}^{Q(B_{v_1, \dots, v_m}) + \frac{\epsilon}{2^m}} x^{\alpha(B_{v_1, \dots, v_m})-1} (1-x)^{\alpha(B_{v_1, \dots, v_m}^c)-1} dx > 0, \end{aligned}$$

其中

$$\kappa = \frac{\Gamma(\alpha(\mathbb{X}))}{\Gamma(\alpha(B_{v_1, \dots, v_m}))\Gamma(\alpha(B_{v_1, \dots, v_m}^c))}.$$

综合上述情形 (1) 和 (2) 可知不等式 (2.9) 成立. $\square$

上述命题推出了 $\text{DP}(\alpha)$ 的概率族空间为 $\{P \in \mathbb{M} : \text{supp}(P) \subset \text{supp}(\alpha)\}$.

对 $(\mathbb{X}, \mathcal{X})$ 上的任意测度 μ 和可测集 $A, B \in \mathcal{X}$, 定义

$$\mu|_B(A) = \mu(A \cap B), \quad \mu_B(A) = \mu(A|B) = \frac{\mu(A \cap B)}{\mu(B)}.$$

这样, $P_B(\omega, A) = P(\omega, A|B) = \frac{P(\omega, A \cap B)}{P(\omega, B)}$ 仍然是一个 (条件) 随机概率测度.

定理 2.2 (自相似性) 假定 $P \sim \text{DP}(\alpha)$. 如果 $B \in \mathcal{X}$ 使得 $\alpha(B) > 0$, 则 $P_B \sim \text{DP}(\alpha|_B)$. 另外, $P(B)$, $(P_B(A) : A \in \mathcal{X})$, $(P_{B^c}(A) : A \in \mathcal{X})$ 是互相独立的.

证明 首先, 通过 Dirichlet 过程的定义可知 $P(B) \sim \text{Beta}(\alpha(B), \alpha(B^c))$, 根据命题 2.6 和 $\alpha(B) > 0$ 可推出 $P(B) > 0$ 几乎处处成立. 因此, P_B 是定义好的. 对于给定的 B 的分割 $(A_1, \dots, A_r)$ 和 B^c 的分割 $(C_1, \dots, C_s)$, 定义向量

$$Y := (Y_1, \dots, Y_{r+s}) = (P(A_1), \dots, P(A_r), P(C_1), \dots, P(C_s)),$$

则 $Y \sim \text{Dir}(\alpha(A_1), \dots, \alpha(A_r), \alpha(C_1), \dots, \alpha(C_s))$. 根据命题 2.2 可知以下随机变量和随机向量是互相独立的:

$$Z_1 := \sum_{i=1}^r Y_i, \quad Z_2 := \sum_{i=r+1}^{r+s} Y_i, \quad \left(\frac{Y_1}{Z_1}, \dots, \frac{Y_r}{Z_1} \right), \quad \left(\frac{Y_{r+1}}{Z_2}, \dots, \frac{Y_{r+s}}{Z_2} \right),$$

并且

$$\begin{aligned} \left(\frac{Y_1}{Z_1}, \dots, \frac{Y_r}{Z_1} \right) &\sim \text{Dir}(\alpha(A_1), \dots, \alpha(A_r)), \\ \left(\frac{Y_{r+1}}{Z_2}, \dots, \frac{Y_{r+s}}{Z_2} \right) &\sim \text{Dir}(\alpha(C_1), \dots, \alpha(C_s)). \end{aligned}$$

通过上述构造可知,

$$P(B) = Z_1, \quad (P_B(A) : A \in \mathcal{X}) = \left(\frac{Y_1}{Z_1}, \dots, \frac{Y_r}{Z_1} \right), \quad (P_{B^c}(A) : A \in \mathcal{X}) = \left(\frac{Y_{r+1}}{Z_2}, \dots, \frac{Y_{r+s}}{Z_2} \right).$$

这样, 就证明了定理. $\square$

2.5 Dirichlet 过程的抽样分析

给定了一个带参数的分布函数 $F(x, \theta)$ (如果给定一个参数 θ , 则该分布 $F(x, \theta)$ 已知), 对分布 $F(x, \theta)$ 进行抽样就是给定有限多个独立同分布的随机变量 $X_1, \dots, X_n$. 在 Bayes 参数统计中, 还假定 θ 是一个随机变量, 服从某类已知的分布 (先验分布), 然后计算统计量 $E[g(\theta) | X_1, \dots, X_n]$. 在非参数 Bayes 模型中, 因为没有参数, 无法假定参数的随机性. 下面先给出抽样的定义, 然后再用一个注记来作一些解释.

定义 2.6 (随机概率测度的抽样) 假定 P 为在 $(\mathbb{X}, \mathcal{X})$ 上的随机概率测度, $X_1, \dots, X_n$ 是取值在 $\mathbb{X}$ 上的一组随机变量, 称 $X_1, \dots, X_n$ 为服从 P 的 (相互独立的) 一组样本 (从 P 抽出的相互独立的样本), 如果对于任何的 $m, n \in \mathbb{N}$ 和任何的可测集合 $A_1, \dots, A_m, C_1, \dots, C_n \in \mathcal{X}$, 有

$$P(X_1 \in C_1, \dots, X_n \in C_n | P(A_1), \dots, P(A_m), P(C_1), \dots, P(C_n)) = \prod_{i=1}^n P(C_i) \quad \text{a.s.}$$

注 2.4 为了理解上述定义, 引进另一个概率空间 $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$, 测度空间 $(\mathbb{X}, \mathcal{X})$ 上的一个概率测度可以等同于一个定义在 $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ 上取值在 $\mathbb{X}$ 中的随机变量. 对于固定的 $\omega_1 \in \Omega$, 概率测度 $P(\omega_1, \cdot)$ 相应于概率空间 $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ 上的一个取值在 $(\mathbb{X}, \mathcal{X})$ 中的随机变量 $(X(\omega_1, \omega_2), \omega_2 \in \tilde{\Omega})$ (固定 $\omega_1 \in \Omega$). 定义在乘积空间 $\Omega \times \tilde{\Omega}$ 上的随机变量 X 可以写成 $X(\omega_1, \omega_2)$, $\omega_1 \in \Omega$, $\omega_2 \in \tilde{\Omega}$. 上述关于随机概率测度的抽样的定义则可如下地理解: 从 P 抽出的独立的样本可以看成是定义在乘积空间 $(\Omega, \mathcal{F}, P) \times (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ 上取值在 $\mathbb{X}$ 中的随机变量族 $X_1(\omega_1, \omega_2), \dots, X_n(\omega_1, \omega_2)$, 使得对于任何给定的 $\omega_1 \in \Omega$, $X_1(\omega_1, \cdot), \dots, X_n(\omega_1, \cdot)$ 是服从 $P(\omega_1, \cdot)$ 的定义在 $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$ 上取值在 $\mathbb{X}$ 中的独立同分布随机变量族.

在上面提到的参数 Bayes 统计中, 如果要具体地计算统计量 $E[g(\theta) | X_1, \dots, X_n]$, 则需要知道后验分布 $\theta | X_1, \dots, X_n$ 的显式表达式. 当 θ 服从正态分布或者 Beta 分布等一些特殊分布时, 这样的计算可以显式给出. 回到非参数 Bayes 统计, 对于一般的先验随机概率测度, 后验分布非常难以计算出来. 但当先验随机概率测度是 Dirichlet 过程时, $P | X_1, \dots, X_n$ 可以显式地给出. 这一结果由下述定理给出. 该定理是说在非参数 Bayes 模型中, 如果以 Dirichlet 过程为先验, 则其后验依然是一个 Dirichlet 过程. 而且其中的参数的计算也非常简单: 只需进行适当的“更新”.

定理 2.3 (Dirichlet 过程的共轭性^[2]) 如果 $X_1, \dots, X_n$ 为一组 $P \sim \text{DP}(\alpha)$ 的互相独立的样本, 那么,

$$P | X_1, \dots, X_n \sim \text{DP}\left(\alpha + \sum_{i=1}^n \delta_{X_i}\right), \quad (2.10)$$

其中 δ_X 是一个 Dirac 测度, 具体的定义是

$$\delta_X(A) = \begin{cases} 1, & \text{当 } X \in A \text{ 时,} \\ 0, & \text{当 } X \notin A \text{ 时.} \end{cases}$$

在给出证明之前, 首先考虑对 $\mathbb{X}$ 的一系列可测划分:

$$\mathcal{T}_0 = \{\mathbb{X}\}, \quad \mathcal{T}_1 = \{A_0, A_1\}, \quad \mathcal{T}_2 = \{A_{00}, A_{01}, A_{10}, A_{11}\},$$

其中 $\mathcal{T}_2$ 中的 4 个元素为 $\mathcal{T}_1$ 中两个元素的分割, 即 $A_0 = A_{00} \cup A_{01}$, $A_1 = A_{10} \cup A_{11}$. 可以类似地依次定义 $\mathcal{T}_3, \mathcal{T}_4, \dots$. 分割 $\mathcal{T}_m$ 包含了 2^m 个元素, 它们的下标为 $\lambda \in \Lambda^m$, $\Lambda = \{0, 1\}$. 记 $\Lambda^* = \bigcup_{m=0}^{\infty} \Lambda^m$, 则有 $A_\lambda = A_{\lambda_0} \cup A_{\lambda_1}$, $\forall \lambda \in \Lambda^*$. 如此的划分序列被称为划分树, 假设这样的划分生成了 $\mathbb{X}$ 的 Borel σ -集, 定义在 $\mathbb{X}$ 中的概率测度 P 满足 $P(A_\lambda) = P(A_{\lambda_0}) + P(A_{\lambda_1})$. 通过这样的划分树, 可定义 $V_{\lambda_0} = P(A_{\lambda_0} | A_\lambda)$, $V_{\lambda_1} = P(A_{\lambda_1} | A_\lambda)$.

定义 2.7 假定 P 是 $\mathbb{X}$ 上的一个随机测度. 如果对于 $\mathbb{X}$ 的任意一个划分树 $\mathcal{T}$, 有

$$\{V_0\} \perp\!\!\!\perp \{V_{00}, V_{01}\} \perp\!\!\!\perp \{V_{000}, V_{001}, V_{010}, V_{011}\} \perp\!\!\!\perp \dots,$$

则称随机测度 P 是一个无尾 (tailfree) 过程, 其中 $\{V_0\} \perp\!\!\!\perp \{V_{00}, V_{01}\}$ 表示 $\{V_0\}$ 与 $\{V_{00}, V_{01}\}$ 相互独立.

通过定理 2.3 可知, Dirichlet 过程是无尾过程, 而无尾过程的后验分布有着如下命题所述的表示方式.

命题 2.9 $(\mathbb{X}, \mathcal{X})$ 上的随机概率测度 P 是无尾的当且仅当对于任意的正整数 m 和 n , 对于 $\mathbb{X}$ 的任何一系列可生成 Borel 集的划分 $\mathcal{T}_m = \{A_\lambda : \lambda \in \Lambda^m\}$ 和任何样本 $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} P$, $(P(A_\lambda) : \lambda \in \Lambda^m) | X_1, \dots, X_n$ 的分布与 $(P(A_\lambda) : \lambda \in \Lambda^m) | (N_\lambda : \lambda \in \Lambda^m)$ 的分布几乎处处相等, 其中 $N_\lambda := \#\{1 \leq i \leq n : X_i \in A_\lambda\}$.

这一命题的证明有些长, 参见文献 [29, 引理 3.13].

通过上述命题, 给出定理 2.3 的证明 (参见文献 [29]).

定理 2.3 的证明 假定 $X_1, \dots, X_n$ 为一组 $\text{DP}(\alpha)$ 的互相独立的样本. $A_1, \dots, A_k$ 是 $\mathbb{X}$ 的任意一个可测划分. 记 $N_j = \#\{1 \leq i \leq n : X_i \in A_j\}$, $j = 1, \dots, k$. 由命题 2.9 可推出

$$(P(A_1), \dots, P(A_k)) | (X_1, \dots, X_n)$$

的条件分布与

$$(P(A_1), \dots, P(A_k)) \mid (N_1, \dots, N_k)$$

的条件分布相同. 根据 $(N_1, \dots, N_k)$ 的定义可知, $(N_1, \dots, N_k)$ 服从参数为 $(P(A_1), \dots, P(A_k))$ 的多项式分布, 并且这些参数本身服从

$$(P(A_1), \dots, P(A_k)) \sim \text{Dir}(\alpha(A_1), \dots, \alpha(A_k)).$$

根据命题 2.3 可知,

$$(P(A_1), \dots, P(A_k)) \mid N_1, \dots, N_k \sim \text{Dir}(\alpha(A_1) + N_1, \dots, \alpha(A_k) + N_k),$$

即

$$(P(A_1), \dots, P(A_k)) \mid X_1, \dots, X_n \sim \text{Dir}\left(\alpha(A_1) + \sum_{i=1}^n \delta_{X_i}(A_1), \dots, \alpha(A_k) + \sum_{i=1}^n \delta_{X_i}(A_k)\right).$$

命题因此得证. $\square$

命题 2.10 $P \sim \text{DP}(\alpha)$, 则 P 是离散测度 a.s., 即 $\text{DP}_\alpha(P \in \mathbb{M} : P \text{ 是一个离散概率}) = 1$.

该命题首次由 Blackwell^[30] 证明, 而且 P 的离散性也可以通过 Dirichlet 过程的断棍构造看出. 这里给出另一个简单的证明.

命题 2.10 的证明 P 是离散测度当且仅当 $P(x : P(\{x\}) > 0) = 1$. 如果令 $P \sim \text{DP}(\alpha)$ 且 $X \mid P \sim P$, 则

$$P(P(\{X\}) > 0) = \int P(x : P(\{x\}) > 0) d\text{DP}_\alpha(P).$$

需要证明对于每个 x , 有

$$P(P(\{X_0\}) > 0 \mid X_0 = x) = 1.$$

根据定理 2.3 可知 $P \mid X_0 = x \sim \text{DP}(\alpha + \delta_x)$, 那么给定 $X_0 = x$, $P(\{X_0\}) \sim \text{Beta}(\alpha(\{x\}) + 1, \alpha(\{x\}^c))$, 所以 $P(P(\{X_0\}) > 0 \mid X_0 = x) = 1$ a.s. $\square$

定义 2.8 (可交换随机变量) 称一组有限多个随机变量 $X_1, \dots, X_n$ 为可交换的, 如果该组随机变量的任何重新排列都有同样的联合分布. 类似地, 一组无限多个随机变量是可交换的当且仅当其任意有限多个随机变量子集是可交换的.

定理 2.4 (Pólya 罐子模型^[29]) 假定 $X_1, X_2, \dots$ 为 $\text{DP}(\alpha)$ 的一组样本, 则 X_1 的边缘分布是 $\bar{\alpha}$. 当 $n \geq 1$ 时,

$$X_{n+1} \mid X_1, \dots, X_n \sim \frac{1}{|\alpha| + n} \left(\alpha + \sum_{i=1}^n \delta_{X_i} \right), \quad (2.11)$$

其中 $|\alpha| = \alpha(\mathbb{X})$. 而且, $X_1, X_2, \dots$ 的联合分布是可交换的.

证明 对于任何一个可测集 $A \in \mathcal{X}$, 有

$$P(X_1 \in A) = E[P(X_1 \in A) \mid P] = E[P(A)] = \bar{\alpha}(A).$$

因此, X_1 的边缘分布是 $\bar{\alpha}$.

由于 $X_2 | (P, X_1) \sim P$ 而且 $P | X_1 \sim \text{DP}(\alpha + \delta_{X_1})$, 根据定理 2.3 可推出 $X_2 | X_1$ 的条件分布是

$$P(X_2 \in A | X_1) = E[P(A) | X_1] = \frac{\alpha(A) + \delta_{X_1}(A)}{\alpha(\mathbb{X}) + \delta_{X_1}(\mathbb{X})} = \frac{1}{|\alpha| + 1} \alpha(A) + \frac{1}{|\alpha| + 1} \delta_{X_1}(A).$$

重复上述过程以及已有结论 $P | X_1, \dots, X_n \sim \text{DP}(\alpha + \sum_{l=1}^n \delta_{X_l})$ 可得到定理中的结论. $\square$

定理 2.4 之所以被称为 Pólya 罐子模型是因为该模型可以表述如下: 假设在一个罐子中有很多球, $\mathbb{X}$ 是这些球的可能的颜色 (假设颜色的种类是连续的) 的集合, 一开始球的个数是 $|\alpha|$, 这些球的颜色分布是 $\bar{\alpha}$. 首先从罐子里抽取第一个球, 标记其颜色是 X_1 , 然后把该球放回并再放进去一个同样颜色的球. 这时, 罐子里的球有 $|\alpha| + 1$ 个, 它们颜色的分布是 $\frac{|\alpha|}{|\alpha|+1} \bar{\alpha} + \frac{1}{|\alpha|+1} \delta_{X_1}$. 重复该过程, 则得到上述命题中的结果. 上述命题中的条件分布给出了 Dirichlet 过程样本的一些特性, 即新生成的样本有多大的概率等于已生成的样本, 又有多大的概率不等于任何已生成的样本. 当集中参数 $|\alpha|$ 较大时, 新生成的样本大概率不等于已生成的样本. 形象地说, 可以用“富者更富”来表示这种特性.

如果 $X_1, X_2, \dots, X_n$ 是一组 $\text{DP}(\alpha)$ 的样本, 将观测值相等的样本放在同一个组里. 根据定理 2.4 中的条件分布可知, 如果有 m 组这样的分类, 则称样本 $X_1, X_2, \dots, X_n$ 有 m 个聚类. 这些聚类的特征和分析则引出了中国餐馆模型 (Chinese restaurant process, CRP)^[31], 下一节对该过程进行介绍.

另外, 定理 2.4 也给出了一种构造 Dirichlet 过程的方法, 可以通过样本的分布结构表示 Dirichlet 过程. 首先定义随机变量的可交换性, 并给出著名的 De Finetti 表示定理.

定理 2.5 (De Finetti 表示定理^[32]) 假定 $(\Omega, \mathcal{F}, P)$ 是一个概率空间, $(Y, \mathcal{Y})$ 是一个 Borel 空间, $X_n : \Omega \rightarrow Y$ ($n \in \mathbb{N}$) 是序列可测映射, 则序列 $\{X_n\}_{n=1}^\infty$ 是可交换的当且仅当在 $(Y, \mathcal{Y})$ 上存在一个随机概率测度 μ 使得 $\{X_n\}_{n=1}^\infty | (\mu = P) \stackrel{\text{iid}}{\sim} P$. 更进一步地, 如果序列 $\{X_n\}_{n=1}^\infty$ 是可交换的, 则 μ 的分布是唯一的, 并且对每个 $A \in \mathcal{Y}$ 有 $\mu_n(A) \xrightarrow{\text{a.s.}} \mu(A)$, 其中 $\mu_n(B) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_B(X_i)$ 是 $\{X_i\}_{i=1}^n$ 的经验概率测度.

通过上述定理, 可以给出以下命题的证明.

命题 2.11 假定 $X_1, X_2, \dots$ 是一族参数为 α 的 Pólya 随机变量序列, 即它们满足如下条件:

$$\begin{aligned} X_1 &\sim \bar{\alpha}, \\ X_{n+1} | X_1, \dots, X_n &\sim \frac{\alpha + \sum_{i=1}^n \delta_{X_i}}{|\alpha| + n}, \quad n \geq 1, \end{aligned}$$

则存在 $P \sim \text{DP}(\alpha)$ 使得 $X_i | P \stackrel{\text{iid}}{\sim} P, i = 1, 2, \dots$

证明 首先 $X_1, X_2, \dots$ 是可交换的. 根据定理 2.5 可知, 存在随机概率测度 P 使得 $X_i | P \stackrel{\text{iid}}{\sim} P$ ($i = 1, 2, \dots$), 而且 P 的分布是唯一的. 通过 Pólya 罐子模型 (定理 2.4), 以及 P 的分布的唯一性可知, 当 $P \sim \text{DP}(\alpha)$ 时, 样本 $X_1, X_2, \dots$ 是一组参数为 α 的 Pólya 数列. $\square$

令 $\mathcal{L}_n$ 表示对 $\{1, \dots, n\}$ 的所有分割的全体, 用 $\mathfrak{S} = \{\mathfrak{S}_1, \dots, \mathfrak{S}_N\} \in \mathcal{L}_n$ 表示 $\{1, \dots, n\}$ 的一个分割, 其中 $N = |\mathfrak{S}|$ 表示分割 $\mathfrak{S}$ 中元素的个数; 用 $(s)_n = s(s+1) \cdots (s+n-1)$ 表示 s 的上升阶乘.

命题 2.12^[29] 如果 $X_1, X_2, \dots, X_n$ 是一组 $\text{DP}(\alpha)$ 的样本, 则对于任何可测函数 $g_i : \mathbb{X} \rightarrow \mathbb{R}$ 都有

$$E \left[\prod_{i=1}^n g_i(X_i) \right] = \frac{1}{(|\alpha|)_n} \sum_{\mathfrak{S} \in \mathcal{L}_n} \prod_{j=1}^{|\mathfrak{S}|} \left[(|\mathfrak{S}_j| - 1)! \int_{\mathbb{X}} \prod_{l \in \mathfrak{S}_j} g_l(x) d\alpha(x) \right]. \quad (2.12)$$

令 $X_1, X_2, \dots, X_n$ 为一组 $\text{DP}(\alpha)$ 的样本, 这一组样本中不同样本值的个数, 即对 $\{1, \dots, n\}$ 的分割的个数, 也有一系列特性. 在此, 假设 α 满足对于任何一个 $x \in \mathbb{X}$ 有 $\alpha(\{x\}) = 0$, 这样可以保证, 如果 X_i 是从 $\bar{\alpha}$ 中取得的样本, 那么 X_i 与 $X_1, \dots, X_{i-1}$ 都不相同. 对于 $i \in \mathbb{N}$, 如果 $X_i \notin \{X_1, \dots, X_{i-1}\}$, 则定义 $D_i = 1$, 否则 $D_i = 0$. 那么 $K_n := \sum_{i=1}^n D_i$ 代表了这组样本中样本值不相等的样本的数量. 用 $\mathfrak{L}(K_n)$ 表示 K_n 的分布.

命题 2.13 ^[29] 令 $X_1, X_2, \dots, X_n$ 为一组 $\text{DP}(\alpha)$ 的样本, 则随机变量 $D_1, D_2, \dots$ 是互相独立并且服从参数为 $p = \text{P}(D_i = 1) = \frac{|\alpha|}{|\alpha|+i-1}$ 的 Bernoulli 分布, 而且对于固定的 $|\alpha|$, 当 $n \rightarrow \infty$ 时,

- (1) $\text{E}[K_n] \sim |\alpha| \log n \sim \text{Var}(K_n)$;
- (2) $\frac{K_n}{\log n} \rightarrow |\alpha|, \text{a.s.}$;
- (3) $\frac{K_n - \text{E}[K_n]}{\sqrt{\text{Var}(K_n)}} \xrightarrow{d} N(0, 1)$;
- (4) $d_{\text{TV}}(\mathfrak{L}(K_n), \text{Poisson}(\text{E}[K_n])) = O(\frac{1}{\log n})$;
- (5) 对任意的 $k \in \mathbb{N}$, 有

$$\text{P}(K_n = k) = C_n(k) n! |\alpha|^k \frac{\Gamma(|\alpha|)}{|\alpha| + n}, \quad (2.13)$$

其中

$$C_n(k) = \frac{1}{n!} \sum_{\mathfrak{S} \in \{1, \dots, n-1\}: |\mathfrak{S}| = n-k} \prod_{j \in \mathfrak{S}} j. \quad (2.14)$$

证明 (1) 根据定理 2.4 可知,

$$\text{P}(D_i = 1) = \text{P}(X_i \notin \{X_1, \dots, X_{i-1}\}) = \text{P}(X_i \sim \bar{\alpha}) = \frac{|\alpha|}{|\alpha| + i - 1}.$$

因此, 可以得到

$$\begin{aligned} \text{E}[K_n] &= \sum_{i=1}^n \frac{|\alpha|}{|\alpha| + i - 1}, \\ \text{Var}(K_n) &= \sum_{i=1}^n \frac{|\alpha|(i-1)}{(|\alpha| + i - 1)^2}. \end{aligned}$$

通过简单的计算可知, 对于任意 $|\alpha|$ 和正整数 n , 有

$$1 \vee |\alpha| \log \left(1 + \frac{n}{|\alpha|} \right) \leq \text{E}[K_n] \leq 1 + |\alpha| \log \left(a + \frac{n}{|\alpha|} \right).$$

显然地有 $\lim_{n \rightarrow \infty} \frac{\text{E}[K_n]}{|\alpha| \log n} = 1$.

类似地, 根据

$$\text{Var}(K_n) = \sum_{i=1}^n \frac{|\alpha|}{|\alpha| + i - 1} - \sum_{i=1}^n \frac{|\alpha|^2}{(|\alpha| + i - 1)^2},$$

可以证明 $\lim_{n \rightarrow \infty} \frac{\text{Var}[K_n]}{|\alpha| \log n} = 1$, 从而证明了 (1).

(2) 首先有

$$\sum_{i=1}^{\infty} \text{Var} \left(\frac{D_i}{\log i} \right) = \sum_{i=1}^{\infty} \frac{|\alpha|(i-1)}{(|\alpha| + i - 1)^2 (\log i)^2}$$

$$= |\alpha| \left[\sum_{i=1}^{\infty} \frac{1}{(|\alpha| + i - 1)^2 (\log i)^2} - |\alpha| \sum_{i=1}^{\infty} \frac{1}{(|\alpha| + i - 1)^2 (\log i)^2} \right] \\ < \infty.$$

由 Kolmogorov 强大数定律可知, 当 $n \rightarrow \infty$ 时,

$$\frac{K_n}{\log n} = \frac{\sum_{i=1}^n D_i}{\log n} \xrightarrow{\text{a.s.}} \frac{E[K_n]}{\log n} \sim |\alpha|.$$

(3) 可以由 Lindeberg 中心极限定理直接得到.

(4) 对独立同分布的 Bernoulli 随机变量的分布应用 Chen-Stein 估计^[33], 可得到

$$d_{\text{TV}}(\mathfrak{L}(K_n), \text{Poisson}(E[K_n])) \leq 2 \frac{1 - e^{-E[K_n]}}{E[K_n]} \sum_{i=1}^n (E[D_i])^2.$$

再根据 $\sum_{i=1}^n (E[D_i])^2 = |\alpha|^2 \sum_{i=1}^n \frac{|\alpha|^2}{(|\alpha| + i - 1)^2} \leq n$ 和 $E[K_n] \sim |\alpha| \log n$ 可得到结论 (4).

(5) 从 K_n 和 D_i 的定义可知 $K_n = k$ 可以通过两种可能的方式实现, 第一种是 $K_{n-1} = k - 1$ 并且 $D_n = 1$; 第二种是 $K_{n-1} = k$ 并且 $D_n = 0$. 因此, 如果对 $1 \leq k \leq n$, 定义

$$p_n(k, |\alpha|) = P(K_n = k \mid \alpha(\mathbb{X}) = |\alpha|),$$

即当 $\alpha(\mathbb{X}) = |\alpha|$ 时 $K_n = k$ 的条件概率, 则可得到如下递归关系:

$$p_n(k, |\alpha|) = \frac{|\alpha|}{|\alpha| + n - 1} p_{n-1}(k - 1, |\alpha|) + \frac{n - 1}{|\alpha| + n - 1} p_{n-1}(k, |\alpha|). \quad (2.15)$$

当 $|\alpha| = 1$ 时, $C_n(k) = p_n(k, 1)$, 则上述关系可推出

$$C_n(k) = \frac{1}{n} C_{n-1}(k - 1) + \frac{n - 1}{n} C_{n-1}(k). \quad (2.16)$$

由此, 将用数学归纳法证明 (2.13). 当 $n = 1$ 时, k 只能取 1, 显然, $P(K_1) = C_1(1)$. 假设 (2.13) 在 $n - 1$ 时成立, 则

$$\begin{aligned} p_n(k, |\alpha|) &= \frac{|\alpha|}{|\alpha| + n - 1} C_{n-1}(k - 1) (n - 1)! |\alpha|^{k-1} \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + n - 1)} \\ &\quad + \frac{n - 1}{|\alpha| + n - 1} C_{n-1}(k) (n - 1)! |\alpha|^k \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + n - 1)} \\ &= \frac{|\alpha|^k}{|\alpha| + n - 1} \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + n - 1)} (n - 1)! [C_{n-1}(k - 1) + (n - 1) C_{n-1}(k)] \\ &= C_n(k) n! |\alpha|^k \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + n)}. \end{aligned}$$

这就证明了 (2.13). 通过母函数来证明 (2.14). $C_n(k)$ 的母函数可以表示为 $A_n(s) = \sum_{k=1}^{\infty} C_n(k) s^k$, 根据 (2.16) 可知,

$$A_n(s) = \frac{1}{n} [s A_{n-1}(s) + (n - 1) A_{n-1}(s)] = \frac{s + n - 1}{n} A_{n-1}(s),$$

上述递归关系结合 $A_1(s) = s$ 可以得到

$$A_n(s) = A_1(s) \prod_{j=1}^n \frac{s + j - 1}{j} = \frac{1}{n!} \prod_{j=0}^{n-1} (s + j).$$

因此有

$$\sum_{k=1}^{\infty} C_n(k) s^k = \frac{1}{n!} \prod_{j=0}^{n-1} (s+j),$$

对比 s^k 的系数可以得到 (2.14). □

在许多聚类问题中, 一个主要的问题是, 如何决定一组数据应该分多少类, 上述命题给出了, 如果在聚类问题中使用 Dirichlet 过程, 则当样本大小为 n 时, 应该分成大约 $|\alpha| \log n$ 个类. 这就是说, 类别的个数随着样本大小的增大而增大, 而且以 $\log n$ 的速度增长, 这样的结果符合统计中的惯用的做法, 这也是 Dirichlet 过程在机器学习的无监督学习中有广泛应用的一个主要原因.

称样本 $X_1, \dots, X_n$ 服从多重类别, 记作 $C(m_1, \dots, m_n)$, 如果样本值的集合 $\{X_1, \dots, X_n\}$ 中有 m_1 个数值只出现过一次, m_2 个数值 (每个) 出现两次, $\dots$, 这样 $(m_1, \dots, m_k)$ 是一组非负整数, 而且 $m_1 + 2m_2 + \dots + nm_n = n$. 在统计学领域中, 一个重要的公式是 Ewens 样本公式 (Ewens's sampling formula), 该公式给出了一组样本服从某个给定的多重类别的概率, 这在随机分割的研究中有重要的应用.

命题 2.14 (Ewens 样本公式^[34]) 令 $X_1, X_2, \dots, X_n$ 为一组 $DP(\alpha)$ 的样本, 则这组样本服从多重类别 $C(m_1, \dots, m_n)$ 的概率是

$$P(\text{样本中有 } m_1 \text{ 个数值出现过一次, } \dots, m_n \text{ 个数值出现 } n \text{ 次}) = \frac{n!}{(|\alpha|)_n} \prod_{i=1}^n \frac{|\alpha|^{m_i}}{i^{m_i} m_i!}. \quad (2.17)$$

Dirichlet 过程还有很多其他重要性质, 例如, Sethuraman 和 Tiwari^[35] 研究了由 Dirichlet 过程组成的序列, 给出了当集中参数 $|\alpha|$ 趋于 0 、 ∞ 或 $0 < M < \infty$ 时, 该序列的收敛结果. Korwar 和 Hollander^[36] 证明了, 如果两个 Dirichlet 过程的基测度没有共同的原子 (atom), 则这两个 Dirichlet 过程是相互奇异的. Doss 和 Sellke^[37] 描述了 Dirichlet 过程的尾部性质并给出了具体的表示形式.

3 其他有关的一些过程

3.1 断棍过程

很多随机过程, 如 Dirichlet 过程、双参数 Poisson-Dirichlet 过程和 Beta 过程以及更广泛的带独立增量的齐次标准化的随机测度 (homogeneous normalized random measure with independent increments, HNRMI)^[38, 39] 等, 都可以用一类特殊的方式表示和构造. 而对这种模型的推广可以定义一种更一般的随机过程, 统称为断棍过程. 下面给出断棍过程的定义, 仍然假定 $(\mathbb{X}, \mathcal{X})$ 为一个 Polish 空间, H 是 $(\mathbb{X}, \mathcal{X})$ 上的一个概率分布, 满足 $H(\{x\}) = 0, \forall x \in \mathbb{X}$.

定义 3.1 (断棍过程) 假定 $v_1, v_2, \dots$ 是独立的取值在 $[0, 1]$ 上的随机变量族, $\theta_1, \theta_2, \dots$ 是独立同分布取值在 $\mathbb{X}$ 上服从分布 H 的随机变量 (这里的 H 被称为基础测度). 定义如下一组新的随机变量 $\{w_i\}_{i=1}^{\infty}$:

$$w_1 = v_1, \quad w_i = v_i \prod_{l=1}^{i-1} (1 - v_l), \quad i \geq 2.$$

那么, 在 $(\mathbb{X}, \mathcal{X})$ 上的随机测度

$$P = \sum_{i=1}^L w_i \delta_{\theta_i} \quad (3.1)$$

被称为断棍测度.

其中 L 可以是有限的也可以是 ∞ . 如果 $L < \infty$, 则需设定 $v_L = 1$, 这样可以保证以概率 1 有 $\sum_{i=1}^L w_i = 1$. 如果 $L = \infty$, 则保证 $\sum_{i=1}^L w_i = 1$ a.s. 的充要条件 (参见文献 [27]) 是

$$\sum_{i=1}^{\infty} \mathbb{E}[\log(1 - v_i)] = -\infty.$$

断棍过程可以形象地描述如下: 有一根长度为 1 的棍子, 第一次折取 v_1 的长度 (记这一次折取掉长度为 w_1), 然后在剩下的 $1 - v_1$ 长度中折取比例为 v_2 的长度 (记这一次折取掉长度为 w_2), 按照如此的方式一直折取下去, 这些 w_i 就成为了 P 中对于 θ_i 的权重. 这有点像《庄子》中的“一尺之棰, 日取其半, 万世不竭”的思想. 在实际应用中, v_i 的分布可以取为 $\text{Beta}(\alpha_i, \beta_i)$ 分布或其他在 $[0, 1]$ 上的分布, 而 α_i 和 β_i 的取值则决定了相对应的断棍过程的分布. 通过简单的计算可以得知, 对于任何一个 $A \in \mathcal{X}$, 如果 $v_i \sim \text{Beta}(\alpha_i, \beta_i)$ ($i = 1, 2, \dots$), 则

$$\mathbb{E}[P(A)] = H(A),$$

$$\text{Var}[P(A)] = H(A)(1 - H(A)) \sum_{i=1}^L \left[\frac{\alpha_i(\alpha_i + 1)(\alpha_i + \beta_i)}{(\alpha_i + \beta_i)^2(\alpha_i + \beta_i + 1)} \prod_{l=1}^{i-1} \frac{\beta_l(\beta_l + 1)(\alpha_l + \beta_l)}{(\alpha_l + \beta_l)^2(\alpha_l + \beta_l + 1)} \right].$$

从上面的等式可以看出, 基础测度 H 决定了断棍过程的期望, 而方差则主要由 $(\alpha_i, \beta_i)_{i=1}^{\infty}$ 决定. 当 $v_i \sim \text{Beta}(\alpha_i, \beta_i)$ 时, 断棍过程的一些特例目前在各个领域中有广泛的应用. 这些特例主要包括 ($L = \infty$): (1) 当 $\alpha_i = 1, \beta_i = a, a > 0$ 时, 上述断棍过程构造了 Dirichlet 过程; (2) 当 $\alpha_i = 1 - b, \beta_i = a + ib$ (其中 $b \in [0, 1), a > -b$) 时, 上述断棍过程为双参数的 Poisson-Dirichlet (Pitman-Yor) 过程.

3.2 Probit 断棍过程

假定 P 是由 (3.1) 定义的 $(\mathbb{X}, \mathcal{X})$ 上的断棍过程, 为了在应用问题中便于对先验分布进行计算, 人们通常要求 v_i 服从 Beta 分布, 这限制了 P 的灵活性. 但是, 在解决具体问题时, 人们需要 P 能满足更多方面的要求, 这时需要对 v_i 进行更灵活的选择, 但同时又要便于计算. 在众多 v_i 的好的选择中, Probit 断棍^[16] 是其中一个.

定义 3.2 如果 $v_i = \Phi(\alpha_i)$, 其中 $\alpha_i \sim N(\mu, \sigma^2)$, 而 $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{u^2}{2}} du$ 是正态分布函数, 则相应地在 (3.1) 中定义的 P 被称为 Probit 断棍过程.

通过简单的计算可以得到 v_i 的 n 次矩为

$$e_n := \mathbb{E}[v_i^n] = P(T_1 > 0, \dots, T_n > 0),$$

其中 $(T_1, \dots, T_n)$ 服从满足如下特征的 n 维正态分布: $\mathbb{E}[T_l] = \mu, \text{Var}(T_l) = \sigma^2, \text{Cov}(T_l, T_k) = 1 + \sigma^2, k, l \in \{1, \dots, n\}$. 因此, 可得到 Probit 断棍过程 P 的期望和方差,

$$\mathbb{E}[P(A)] = H(A),$$

$$\text{Var}(P(A)) = H(A)(1 - H(A))e_2 \frac{1 - (1 - 2e_1 + e_2)^L}{2e_1 - e_2}, \quad \forall A \in \mathcal{X}.$$

当 $L = \infty$ 时, $\text{Var}(P(A)) = H(A)(1 - H(A)) \frac{e_2}{2e_1 - e_2}$.

在使用 Probit 断棍过程处理聚类问题时, 有时需要假定分布有特定的渐近性. 这个可以通过改变 α_i 所服从的正态分布的期望实现, 例如, 假定 $\alpha_i \sim N(\mu + i\gamma, \sigma^2)$ 时对 Probit 断棍过程的 n 个样本进

行分类: 当 $\gamma = 0$ 时, 所得到聚类个数大约为 $\log n$; 当 $\gamma < 0$ 时, 聚类个数大于 $\log n$; 当 $\gamma > 0$ 时, 聚类个数小于 $\log n$. Probit 断棍过程的另一个灵活性体现在: 它可以做类似于 Dirichlet 过程的推广, 例如, 类似于相关 Dirichlet 过程, 可以定义相关 Probit 断棍过程.

除了 Probit 断棍过程以外, 在机器学习和生物统计等领域, 其数据结构往往是空间数据, 数据依赖于其空间位置, 而有时不同位置的数据有着非独立的关系, 因此需要对一般的断棍过程做更多的推广, 如核断棍过程^[40], 假定

$$\{\psi_i, v_i, G_i^* : i \in \{1, 2, \dots\}\}$$

是相互独立的随机变量, 其中 $\psi_i \stackrel{\text{iid}}{\sim} G_0$ 代表空间位置, $v_i \stackrel{\text{iid}}{\sim} \text{Beta}(a_i, b_i)$ 是概率权重, $\{G_i^*\}_{i=1}^\infty$ 是一族在 $(\mathbb{X}, \mathcal{X})$ 上的随机概率测度 (这里可以假设 $G_i^* \sim \text{DP}(\alpha_i)$), G_0 是定义在 $(\mathbb{R}^m, \mathcal{B}(\mathbb{R}^m))$ 上的随机概率测度. 定义随机测度 P_s 如下:

$$\begin{cases} P_s = \sum_{i=1}^{\infty} \pi_i(s; v_i, \psi_i) G_i^*, \\ \pi_i(s; v_i, \psi_i) = v_i K(s, \psi_i) \prod_{l=1}^{i-1} [1 - v_l K(s, \psi_l)], \end{cases}$$

其中 $s \in \mathbb{R}^m$, $K(s, \psi_i) : \mathbb{R}^m \times \mathbb{R}^m \rightarrow [0, 1]$ 是有界的核函数, 代表了空间坐标 s 与空间坐标 ψ_i 的距离, 称随机测度 P_s 服从核断棍过程. 核断棍过程主要用于对空间数据的分类, 这样的分类依赖于数据的空间位置信息, 具体来讲, P_s 赋予第一个基测度 G_1^* 的权重为 $v_1 K(s, \psi_1)$, 随着 s 与 ψ_1 之间的距离的增大, $K(s, \psi_1)$ 的取值从 0 递增到 1 (其递增的方式取决于 K 的定义), 相对应的权重 $v_1 K(s, \psi_1)$ 也从 0 变化到 v_1 ; 类似地, P_s 赋予第二个基测度 G_2^* 的权重为 $v_2 K(s, \psi_2)(1 - v_1 K(s, \psi_1))$, 该权重的变化取决于 s 与 ψ_2 的距离.

在解决空间数据的分类问题中, 另一个重要的工具是 logistic 断棍过程^[41]. 假定 $\{\theta_i\}_{i=1}^\infty$ 为参数空间 Θ 中的一组参数, 给定 n 个样本点 $\{D_k\}_{k=1}^n$, $D_k = (y_k, s_k)$, 其中 y_k 为服从参数为 $\theta_k \in \{\theta_i\}_{i=1}^\infty$ 的某个分布的随机变量, 代表第 k 个数据的特征向量, $s_k \in \mathbb{R}^m$ 为相对应的空间位置, 我们希望得到与空间位置距离相关的分类. 定义如下随机测度 P_{s_k} :

$$P_{s_k} = \sum_{i=1}^{\infty} \pi_i(s_k) \delta_{\theta_i}, \quad \pi_i(s_k) = v_i(s_k) \prod_{l=1}^{i-1} [1 - v_l(s_k)],$$

其中 $v_i(s_k)$ 由 logistic 方程定义, 具体地, 考虑 $\{\hat{s}_j\}_{j=1}^N$ 为 N 个定义域内的位置 (例如, 选取一些样本 D_k 的位置), 通过如下 logistic 回归估计一系列 $v_i(s_k)$:

$$\log \left(\frac{v_i(s_k)}{1 - v_i(s_k)} \right) = \sum_{j=1}^N \beta_{kj} K(s_k, \hat{s}_j; \phi_k) + \beta_{k0},$$

其中 ϕ_k 为参数, $K(s_k, \hat{s}_j; \phi_k) : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ 为自定义的函数用来衡量 s_n 与 $\hat{s}_j$ 距离, 例如, $K(s_k, \hat{s}_j; \phi_k) = \exp(-\frac{\|s_n - \hat{s}_j\|^2}{\phi_k})$ 为表示 s_n 与 $\hat{s}_j$ 距离的 Gauss 核函数. 称随机测度 P_{s_k} 服从 logistic 断棍过程.

另一种推广被称作隐断棍过程^[42], 该过程适用于非独立的空间数据. 假定 $(\mathbb{X}, \mathcal{X})$ 是一个 Polish 空间, H 是 $(\mathbb{X}, \mathcal{X})$ 上的一个概率测度, $s \in D \subset \mathbb{R}^m$ 代表空间坐标. 给定如下相互独立的随机变量:

$$\{z(s) : s \in D\}, \quad \{v_i\}_{i=1}^\infty, \quad \{\theta_i^*\}_{i=1}^\infty,$$

$$z(s) \sim \mathcal{N}(0, 1), \quad \text{Cov}(z(s), z(s')) = \gamma(s, s'), \quad v_1, v_2, \dots \stackrel{\text{ind}}{\sim} \text{Beta}(\alpha_i, \beta_i),$$

其中 $\gamma(s, s') : D \times D \rightarrow \mathbb{R}$ 为自定义的协方差函数. 定义随机过程 $U = \{U(s) = \Phi(z(s)) : s \in D\}$, 其中 Φ 是标准正态分布函数, 则 U 的边缘分布是均匀分布. 定义 $(\mathbb{X}, \mathcal{X})$ 上的随机测度

$$P(\cdot) = \sum_{i=1}^{\infty} v_i \prod_{l=1}^{i-1} (1 - v_l) \delta_{\theta_i^*}(\cdot),$$

其中 $\theta_1^* \sim H$; 当 $i > 1$ 时, $\theta_i^* \sim H_i$, H_i 满足 $H_i(B) = \frac{H(B \cap S_i)}{H(S_i)}$, $B \in \mathcal{X}$, $S_i = \{\theta : \theta > \theta_{i-1}^*\}$. 定义一组与空间坐标 s 相关的随机样本 $\theta = \{\theta(s) = P^{-1}(U(s)) : s \in D\}$, 其中 $P^{-1}(x) = \inf\{\theta : P(\{y : y \leq \theta\}) > x\}$. 通过上述定义有

$$\theta(s) \mid P \sim P, \quad \text{Cov}(\theta(s), \theta(s') \mid P) \neq 0.$$

而且, 对于有限多个空间位置 $s_1, \dots, s_n \in D$, 有

$$\begin{aligned} P(\theta(s_1) = \theta_{i_1}^*, \dots, \theta(s_n) = \theta_{i_n}^* \mid \{\theta_i^*\}_{i=1}^L, \{V_i\}_{i=1}^L) \\ = P(z(s_1) \in [\Phi^{-1}(\pi_{i_1-1}), \Phi^{-1}(\pi_{i_1})], \dots, z(s_n) \in [\Phi^{-1}(\pi_{i_n-1}), \Phi^{-1}(\pi_{i_n})]), \end{aligned}$$

其中 $\pi_i = 1 - \prod_{l=1}^i (1 - v_l)$. 称上述定义中的随机测度 P 服从隐断棍过程.

除了核断棍过程、logistic 断棍过程和隐断棍过程之外, 基于实际应用中的各种数据结构, 断棍过程还有其他形式的推广, 在此不一一介绍, 请读者自行阅读相关文献.

3.3 混合 Dirichlet 过程、Dirichlet 混合模型和层次 Dirichlet 过程

混合 Dirichlet 过程^[17]、Dirichlet 混合模型^[17] 和层次 Dirichlet 过程^[18] 是 Dirichlet 过程的重要推广, 是各领域应用中的主要工具.

混合 Dirichlet 过程是一种参数基于另一个或多个参数的 Dirichlet 过程, 因此, 混合 Dirichlet 过程也拥有 Dirichlet 过程的所有性质.

定义 3.3 (混合 Dirichlet 过程) 如果 Dirichlet 过程 $P \sim \text{DP}(\alpha_\xi)$ 中的基测度 α_ξ 依赖于另外一个参数 ξ , 而其中的参数 ξ 是服从分布 π 的一个随机变量, 则称 P 为混合 Dirichlet 过程 (mixed Dirichlet process, MDP), 记作 $P \sim \text{MDP}(\alpha_\xi, \xi \sim \pi)$.

假定 $P \sim \text{MDP}(\alpha_\xi, \xi \sim \pi)$, 相对应的 Bayes 模型是

$$X_1, \dots, X_n \mid (P, \xi) \stackrel{\text{iid}}{\sim} P, \quad P \mid \xi \sim \text{DP}(\alpha_\xi), \quad \xi \sim \pi. \quad (3.2)$$

一个用处广泛的特例是, 当参数 ξ 服从的分布是 Gauss 分布时, 这个混合 Dirichlet 过程被称为 Dirichlet Gauss 混合模型. Dirichlet 混合模型是实际应用中常见的模型之一. Dirichlet 混合模型结合了 Dirichlet 过程与一系列随机样本, 其展示的关系可以让我们对样本进行分类或拟合.

定义 3.4 (Dirichlet 混合模型) 假定 $(\mathbb{X}, \mathcal{X})$ 和 $(\mathbb{Y}, \mathcal{Y})$ 是两个 Polish 空间, 对于任意确定的 $\theta \in \mathbb{Y}$, 函数 $\psi(x; \theta) : \mathbb{X} \rightarrow \mathbb{R}$ 是一个密度函数, 即 $\int_{\mathbb{X}} \psi(x; \theta) dx = 1$; 如果 $P \sim \text{DP}(\alpha)$ 是定义在 $(\mathbb{Y}, \mathcal{Y})$ 上的 Dirichlet 过程, 则称 $f_P(x) = \int_{\mathbb{Y}} \psi(x; \theta) P(d\theta)$ 为 Dirichlet 混合模型.

假如 $X_1, \dots, X_n$ 是 Dirichlet 混合模型的样本, 则 $X_i \in \mathbb{X}$ ($i \in \{1, \dots, n\}$) 服从如下密度函数:

$$X_i \stackrel{\text{iid}}{\sim} f_{i,P}(x) = \int_{\mathbb{Y}} \psi_i(x; \theta) P(d\theta), \quad i = 1, 2, \dots, n. \quad (3.3)$$

Dirichlet 混合模型的 Bayes 模型为

$$X_i \mid (\theta_i, P) \stackrel{\text{ind}}{\sim} \psi_i(\cdot; \theta_i), \quad \theta_i \mid P \stackrel{\text{iid}}{\sim} P, \quad P \sim \text{DP}(\alpha), \quad i = 1, \dots, n. \quad (3.4)$$

注 3.1 Dirichlet 混合模型的一组样本 $X_1, \dots, X_n$ 是取值在 $(\mathbb{X}, \mathcal{X})$ 上的相互独立的随机变量, X_i 服从一个参数为 θ_i 的密度函数, 其中 $\{\theta_i\}_{i=1}^n$ 是取值在 $(\mathbb{Y}, \mathcal{Y})$ 上的 $\text{DP}(\alpha)$ 的一组样本, 参见定义 2.6.

在这个 Bayes 模型中, 人们也需要得到后验分布 $P \mid X_1, \dots, X_n$. 假定 $F(\theta)$ 是概率测度 P 所对应的分布函数, 根据 Dirichlet 过程的共轭性, 即定理 2.3, 可以推出 $P \mid \theta_1, \dots, \theta_n \sim \text{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$. 这样, 对于任何可测函数 ψ , 有

$$\mathbb{E} \left[\int_{\mathbb{Y}} \psi dF \mid \theta_1, \dots, \theta_n, X_1, \dots, X_n \right] = \frac{1}{|\alpha| + n} \left[\int_{\mathbb{Y}} \psi d\alpha + \sum_{j=1}^n \psi(\theta_j) \right]. \quad (3.5)$$

这样的结果虽然并不能使我们得到希望得到的后验分布 $P \mid X_1, \dots, X_n$, 但可以通过它得到 $\mathbb{E}[\int_{\mathbb{Y}} \psi dF \mid X_1, \dots, X_n]$, 这样的后验期望足够我们去了解与后验分布 $P \mid X_1, \dots, X_n$ 相关的统计量.

给出以下两种得到 $\mathbb{E}[\int_{\mathbb{Y}} \psi dF \mid X_1, \dots, X_n]$ 的方法, 第一种方法通过理论推导得到 $\mathbb{E}[\int_{\mathbb{Y}} \psi dF \mid X_1, \dots, X_n]$ 的数学表达式; 第二种方法通过 Gibbs 抽样^[27] 估计 $\mathbb{E}[\int_{\mathbb{Y}} \psi dF \mid X_1, \dots, X_n]$. 两种方法的证明过程可参见文献 [29, 第 5.1 小节].

方法一: 用 $\mathcal{L}_n$ 表示 $\{1, \dots, n\}$ 的所有分割的全体, 令 $\mathfrak{S} = \{\mathfrak{S}_1, \dots, \mathfrak{S}_N\}$ 表示一个分割, 其中 $N = |\mathfrak{S}|$ 是分割的个数, $N_j = |\mathfrak{S}_j|$ 是 $\mathfrak{S}_j$ 中包含元素的个数. 假定 W 为 $\mathcal{L}_n$ 上的概率测度, 满足

$$W(\mathfrak{S}) \propto \prod_{j=1}^N (N_j - 1)! \int_{\mathbb{Y}} \prod_{l \in \mathfrak{S}_j} \psi_l(X_l; \theta) d\alpha(\theta).$$

命题 3.1^[29] 在模型 (3.4) 中, 对于任何非负或可积函数 ψ , 有

$$\begin{aligned} & \mathbb{E} \left[\int_{\mathbb{Y}} \psi dF \mid X_1, \dots, X_n \right] \\ &= \frac{1}{|\alpha| + n} \left[\int_{\mathbb{Y}} \psi d\alpha + \sum_{\mathfrak{S} \in \mathcal{L}_n} W(\mathfrak{S}) \sum_{j=1}^N N_j \frac{\int_{\mathbb{Y}} \psi(\theta) \prod_{l \in \mathfrak{S}_j} \psi_l(X_l; \theta) d\alpha(\theta)}{\int_{\mathbb{Y}} \prod_{l \in \mathfrak{S}_j} \psi_l(X_l; \theta) d\alpha(\theta)} \right]. \end{aligned}$$

方法二: 首先通过 Gibbs 抽样得到 Dirichlet 混合模型中 θ_i 的后验分布, 即 $\theta \mid \hat{\theta}, X_1, \dots, X_n$ 的分布. 假定一组参数 $\theta = (\theta_1, \dots, \theta_n)$, 记 $\hat{\theta}_i = (\theta_j : j \neq i)$.

命题 3.2 (Gibbs 抽样^[29]) 在模型 (3.4) 中, θ_i 的后验分布为

$$\theta_i \mid \hat{\theta}_i, X_1, \dots, X_n \sim \sum_{j \neq i} q_{i,j} \delta_{\theta_j} + q_{i,0} G_{b,i}, \quad i, j \in \{1, \dots, n\}, \quad (3.6)$$

其中 $\{q_{i,j} : j \in \{0, 1, \dots, n\} \setminus \{i\}\}$ 满足如下条件:

$$q_{i,j} \propto \begin{cases} \psi_i(X_i; \theta_j), & j \neq i, \quad j \geq 1, \\ \int_{\mathbb{Y}} \psi_i(X_i; \theta) d\alpha(\theta), & j = 0. \end{cases}$$

而且 $G_{b,i}$ 是一个后验基测度, 满足如下条件:

$$dG_{b,i}(\theta \mid X_i) \propto \psi_i(X_i; \theta) d\alpha(\theta). \quad (3.7)$$

通过上述命题可以取得样本 $\theta_1, \dots, \theta_n$, 当 n 足够大时, θ_i 的分布可以通过样本 $\theta_1, \dots, \theta_n$ 的经验分布 $\tilde{F}(\theta)$ 估计. 这样一来, $P \mid X_1, \dots, X_n$ 就可以通过如下混合 Dirichlet 过程 (定义见 3.3) 得到:

$$P \mid X_1, \dots, X_n \sim \text{MDP}\left(\alpha + \sum_{i=1}^n \delta_{\theta_i}, \theta_i \sim \tilde{F}(\theta)\right).$$

可以看到从方法一中推导的结果包含一些复杂的积分运算, 在实际问题中很难应用. 方法二利用 Gibbs 抽样的方法绕过了大量的积分运算. 另外, Gibbs 抽样的方法有很多推广, 可以通过一些已知程序快速得到结果, 适合在实际问题中应用.

层次 Dirichlet 过程在社会学、信息检索和文本分类等领域有广泛的应用. 当处理数据时, 有时需要将数据分为很多组, 我们希望在分析每组数据时考虑各组之间的联系, 在这种情形下可以用到层次 Dirichlet 过程.

定义 3.5 (层次 Dirichlet 过程) 假定 $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$, $\bar{\alpha}$ 和 G 为 $(\mathbb{X}, \mathcal{X})$ 上的概率测度, $|\alpha| \in (0, \infty)$ 和 $\gamma \in (0, \infty)$ 为中心参数, $\{P_j\}_{j=1}^m$ 是 $(\mathbb{X}, \mathcal{X})$ 上的一组随机概率测度, $X_{j,i} : \Omega \rightarrow \mathbb{X}$ 是随机变量, 则层次 Dirichlet 过程可用如下模型来给出:

$$\begin{cases} X_{j,1}, \dots, X_{j,n} \mid P_j \stackrel{\text{iid}}{\sim} P_j, & j = 1, \dots, m, \\ P_j \mid (\gamma, G) \sim \text{DP}(\gamma, G), & j = 1, \dots, m, \\ G \mid (|\alpha|, \bar{\alpha}) \sim \text{DP}(|\alpha|, \bar{\alpha}). \end{cases}$$

根据以上定义可知, 层次 Dirichlet 过程赋予每个组一个 Dirichlet 过程, 而所有组的基测度都服从一个 Dirichlet 过程. 更进一步地, 可以对每个组赋予一个 Dirichlet 混合模型, 这种情形下该模型可以被称为层次 Dirichlet 混合模型.

定义 3.6 (层次 Dirichlet 混合模型) 假定 $(\mathbb{X}, \mathcal{X})$ 和 $(\mathbb{Y}, \mathcal{Y})$ 是两个 Polish 空间, $\bar{\alpha}$ 和 G 为 $(\mathbb{X}, \mathcal{X})$ 上的概率测度, $|\alpha| \in (0, \infty)$ 和 $\gamma \in (0, \infty)$ 为中心参数, $\{P_j\}_{j=1}^m$ 是 $(\mathbb{X}, \mathcal{X})$ 上的一组随机概率测度; $i \in \{1, \dots, n\}$, $j \in \{1, \dots, m\}$, $\theta_{j,i}$ 是取值在 $(\mathbb{X}, \mathcal{X})$ 上的随机变量, $X_{j,i}$ 是取值在 $(\mathbb{Y}, \mathcal{Y})$ 上的随机变量; 给定 $\theta_{j,i}$, $f_{\theta_{j,i}}(x) : \mathbb{Y} \rightarrow \mathbb{R}$ 是 $X_{j,i}$ 的密度函数. 通过以下 Bayes 模型描述层次 Dirichlet 混合模型:

$$\begin{cases} X_{j,i} \mid \theta_{j,i} \sim f_{\theta_{j,i}}(x), & i \in \{1, \dots, n\}, \quad j \in \{1, \dots, m\}, \\ \theta_{j,i} \mid P_j \sim P_j, \\ P_j \mid (\gamma, G) \sim \text{DP}(\gamma, G), \\ G \mid (|\alpha|, \bar{\alpha}) \sim \text{DP}(|\alpha|, \bar{\alpha}). \end{cases}$$

通过上述定义中的构造可知, 层次 Dirichlet 混合模型是在层次 Dirichlet 过程中每一组都赋予了一个 Dirichlet 混合模型. 层次 Dirichlet 混合模型在实际问题中应用的一个典型例子是处理文本数据的聚类. 假设一篇文本分为 m 个章节, 每个章节里包含了 n 个话题, 想要对这篇文本中所有的话题进行聚类, 那么就可以应用层次 Dirichlet 混合模型. m 个章节就是 m 个小组, 每一个小组都赋予一个 Dirichlet 混合模型, 每个小组中的 n 个样本就是每个章节中的 n 个话题. 这样的应用可以让 m 个章节相互有联系, 想要对这种联系做清晰的描述, 需要通过以下层次 Dirichlet 过程的断棍构造来分析.

首先通过 Dirichlet 过程的断棍构造可知,

$$G = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}, \quad (3.8)$$

其中 w_k 定义见定理 2.1, $\theta_k \sim \bar{\alpha}$. 也就是说, G 定义在 $\Theta = \{\theta_k\}_{k=1}^\infty$ 上, 那么每个 P_j 也必须定义在 $\{\theta_k\}_{k=1}^\infty$ 上, 因此可以得到对 P_j 的断棍表示方法:

$$P_j = \sum_{k=1}^{\infty} \pi_{j,k} \delta_{\theta_k}, \quad j \in \{1, \dots, m\}, \quad (3.9)$$

为了刻画层次 Dirichlet 过程中的层次关系^[18], 需要找到 $W = \{w_k\}_{k=1}^\infty$ 与 $\Pi_j = \{\pi_{j,k}\}_{k=1}^\infty$ 之间的关系. 令 $(A_1, \dots, A_r)$ 为 $\mathbb{X}$ 的一个可测分割, 对 $l \in \{1, \dots, r\}$, 令 $K_l = \{k : \theta_k \in A_l\}$, 这里可以把 $(K_1, \dots, K_r)$ 看作对自然数的一个分割. 根据 Dirichlet 过程的定义可知,

$$(P_j(A_1), \dots, P_j(A_r)) \sim \text{Dir}(|\alpha| \bar{\alpha}(A_1), \dots, |\alpha| \bar{\alpha}(A_r)),$$

即

$$\left(\sum_{k \in K_1} \pi_{j,k}, \dots, \sum_{k \in K_r} \pi_{j,k} \right) \sim \text{Dir} \left(|\alpha| \sum_{k \in K_1} w_k, \dots, |\alpha| \sum_{k \in K_r} w_k \right). \quad (3.10)$$

如果将 W 和 Π 看作是自然数空间上的概率测度, 则通过 (3.10) 可知 $\Pi \sim \text{DP}(|\alpha|, W)$, 这表示出 $W = \{w_k\}_{k=1}^\infty$ 与 $\Pi_j = \{\pi_{j,k}\}_{k=1}^\infty$ 之间的关系. 更进一步地, 可以通过断棍模型给出这个关系的具体表达:

$$v_k \stackrel{\text{iid}}{\sim} \text{Beta}(1, |\alpha|), \quad w_k = v_k \prod_{l=1}^{k-1} (1 - v_l), \quad k \in \mathbb{N}^+, \quad (3.11)$$

$$e_{j,k} \stackrel{\text{iid}}{\sim} \text{Beta} \left(|\alpha| w_k, |\alpha| \left(1 - \sum_{l=1}^k w_l \right) \right), \quad \pi_{j,k} = e_{j,k} \prod_{l=1}^{k-1} (1 - e_{j,l}). \quad (3.12)$$

3.4 双参数 Poisson-Dirichlet (Pitman-Yor) 过程

双参数 Poisson-Dirichlet 过程由 Pitman 和 Yor^[14] 于 1997 年提出并进行了深入的研究, 因此该过程也被称为 Pitman-Yor 过程. 要研究双参数 Poisson-Dirichlet 过程首先需要研究双参数 Poisson-Dirichlet 分布 (参见文献 [15]).

定义 3.7 假定 $b \in [0, 1)$, $a > -b$, $v_i \stackrel{\text{iid}}{\sim} \text{Beta}(1 - b, a + ib)$, $\forall i \in \mathbb{N}$. 定义

$$w_1 = v_1, \quad w_i = v_i \prod_{l=1}^{i-1} (1 - v_l), \quad i \geq 2.$$

将 $w_1, w_2, \dots$ 从大到小进行重新排序得到 $\tilde{w}_1 \geq \tilde{w}_2 \geq \dots$, 称 $(w_1, w_2, \dots)$ 的分布为 GEM (Griffiths-Engen-McCloskey) 分布, 记作 $\text{GEM}(b, a)$. 称 $(\tilde{w}_1, \tilde{w}_2, \dots)$ 的分布为双参数 Poisson-Dirichlet (Pitman-Yor) 分布, 记作 $\text{PD}(b, a)$.

双参数 Poisson-Dirichlet 过程可以如下给定.

定义 3.8 假定 $\theta_i \stackrel{\text{iid}}{\sim} \bar{\alpha}$, 其中 $\bar{\alpha}$ 是 $(X, \mathcal{X})$ 上的概率测度, 满足条件 $\bar{\alpha}(\{x\}) = 0$, $\forall x \in X$. 假定 $\tilde{w}_1 \geq \tilde{w}_2 \geq \dots$ 为与 θ_i ($i = 1, 2, \dots$) 独立的双参数 Poisson-Dirichlet (Pitman-Yor) 分布. 如下定义的随机测度:

$$P_{b,a} = \sum_{i=1}^{\infty} \tilde{w}_i \delta_{\theta_i} \quad (3.13)$$

被称为双参数 Poisson-Dirichlet (或 Pitman-Yor) 过程, 记作 $P_{b,a} \sim \text{PDP}(b, a, \bar{\alpha})$.

显然, 如果 $b = 0$, 则 $v_1, v_2, \dots \stackrel{\text{iid}}{\sim} \text{Beta}(1, a)$, 这时双参数 Poisson-Dirichlet 过程就是 Dirichlet 过程, 即 $\text{PDP}(0, a, \bar{\alpha}) = \text{DP}(a, \bar{\alpha})$.

命题 3.3 假定 $Q_{b,a} = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$, 则有

$$P_{b,a} \stackrel{d}{=} Q_{b,a}. \quad (3.14)$$

证明 令 E^θ 表示对 $\{\theta_k\}_{k=1}^{\infty}$ 的期望, 相对应的概率测度为 $\bar{\alpha}$, 同样地, 令 E^ω 表示对 $\{v_k\}_{k=1}^{\infty}$ 的期望. 考虑 $P_{b,a}$ 和 $Q_{b,a}$ 的特征函数, 对 $t \in \mathbb{R}$, 有

$$C_{P_{b,a}}(t) = E[e^{it \sum_{k=1}^{\infty} \tilde{w}_k \delta_{\theta_k}}] = E^\omega \left[\prod_{k=1}^{\infty} E^\theta [e^{it \tilde{w}_k \delta_{\theta_k}}] \right] = E^\omega \left[\prod_{k=1}^{\infty} f(t \tilde{w}_k) \right],$$

其中 $f(t \tilde{w}_k) = E^\theta [e^{it \tilde{w}_k \delta_{\theta_k}}]$. 由于 $(\tilde{w}_1, \tilde{w}_2, \dots)$ 是 $(w_1, w_2, \dots)$ 的重新排序, 因此有

$$\prod_{k=1}^{\infty} f(t \tilde{w}_k) = \prod_{k=1}^{\infty} f(t w_k).$$

可由此推出 $C_{P_{b,a}}(t) = C_{Q_{b,a}}(t)$, $\forall t \in \mathbb{R}$. □

3.5 双参数 Poisson-Dirichlet 过程的其他表示方法和一些推广

类似于 Dirichlet 过程, 也可以用 Pólya 罐子模型^[43]表示双参数 Poisson-Dirichlet 过程.

令 $X_1, X_2, \dots$ 为 $\text{PDP}(b, a, \bar{\alpha})$ 的样本, 即

$$X_1, X_2, \dots \mid P_{b,a} \stackrel{\text{iid}}{\sim} P_{b,a}, \quad P_{b,a} \sim \text{PDP}(b, a, \bar{\alpha}). \quad (3.15)$$

对于任何一个 $i \in \mathbb{N}$, 将 $\{X_1, \dots, X_i\}$ 中相等的观测值放进一个组, 记这些组的个数为 K_i , 也记 $\{X_1^*, \dots, X_{K_i}^*\}$ 为 $\{X_1, \dots, X_i\}$ 的观测值互不相等的值的全体, 记 $\mathcal{N}_l^i = \#\{1 \leq j \leq i : X_j = X_l^*\}$, 即第 l 组里元素的个数, 其中 $l \in \{1, \dots, K_i\}$, 则 Pólya 罐子模型^[44]可表示为

$$X_1 \sim \bar{\alpha}, \quad X_{i+1} \mid X_1, \dots, X_i \sim \frac{a + bK_i}{a + i} \bar{\alpha} + \sum_{l=1}^{K_i} \frac{\mathcal{N}_l^i - b}{a + i} \delta_{X_l^*}.$$

通过上述 Pólya 罐子模型可以得到如下 $E[K_{i+1}]$ 与 $E[K_i]$ 的迭代关系:

$$E[K_{i+1}] = \sum_{l=1}^{E[K_i]} \frac{\mathcal{N}_l^i - b}{a + i} E[K_i] + \frac{a + bE[K_i]}{a + i} E[K_i + 1] = E[K_i] + \frac{a + bE[K_i]}{a + i}.$$

这样就有

$$E[K_i] = \frac{a}{b} \left[\prod_{j=1}^i \frac{a + b + j - 1}{a + j - 1} - 1 \right]. \quad (3.16)$$

用 Stirling 逼近公式对 (3.16) 进行估计, 可得

$$E[K_i] \approx \frac{\Gamma(a+1)}{b\Gamma(a+b)} n^b. \quad (3.17)$$

(3.17) 推出, 在聚类问题中, 如果用双参数 Poisson-Dirichlet 过程, 则当样本足够多时, 可以得到大约 $\frac{\Gamma(a+1)}{b\Gamma(a+b)}n^b$ 个聚类, 分类的个数随着样本数量的增大呈现出比 Dirichlet 过程更快的增速. 假定 $\mathcal{N}_{(1)}^i \geq \mathcal{N}_2^i \geq \dots \geq \mathcal{N}_{(K_i)}^i$ 是 $(\mathcal{N}_1^i, \dots, \mathcal{N}_{K_i}^i)$ 的从大到小的重新排列, 对任意的一个 $k \in \{1, \dots, K_i\}$, 记 $V_k = \lim_{i \rightarrow \infty} \frac{\mathcal{N}_{(k)}^i}{i}$, 则有 (参见文献 [14])

$$E[V_k] \approx D_{a,b} k^{-\frac{1}{a}}, \quad (3.18)$$

其中 $D_{a,b}$ 是一个常数. 这种现象被称作幂尾部性质 (power-law tail behavior), 该特性符合了自然界中大部分的数据, 例如, 自然语言和图像分割等中的数据所呈现的结构. 这也是双参数 Poisson-Dirichlet 过程被广泛应用的原因之一.

同样地, 类似于混合 Dirichlet 过程、Dirichlet 混合模型和层次 Dirichlet 过程等, 双参数 Poisson-Dirichlet 过程也可以推广到混合双参数 Poisson-Dirichlet 过程、双参数 Poisson-Dirichlet 混合模型和层次双参数 Poisson-Dirichlet 过程等. 我们将作一个简明的介绍.

在以下的几个双参数 Poisson-Dirichlet 过程的推广模型中, 假定 $(\mathbb{X}, \mathcal{X})$ 和 $(\mathbb{Y}, \mathcal{Y})$ 是两个 Polish 空间, $b, \gamma \in [0, 1)$, $a > -b$, $\beta > -\gamma$, $P_{b,a}$ 、 $G_{b,a}$ 和 $\{P_{\beta,\gamma}^j\}_{j=1}^m$ 是 $(\mathbb{X}, \mathcal{X})$ 上的随机测度, $\bar{\alpha}_\xi$ 和 $\bar{\alpha}$ 是 $(\mathbb{X}, \mathcal{X})$ 上的概率测度.

定义 3.9 (混合双参数 Poisson-Dirichlet 过程) 如果 $P_{b,a} \sim \text{PDP}(b, a, \bar{\alpha}_\xi)$, 其中 ξ 是服从分布 π 的随机变量, 称 $P_{b,a}$ 服从混合双参数 Poisson-Dirichlet 过程, 记作 $\text{MPDP}(b, a, \bar{\alpha}_\xi, \xi \sim \pi)$. 假定 $X_1, X_2, \dots$ 是取值在 $\mathbb{X}$ 上的一组 $\text{MPDP}(b, a, \bar{\alpha}_\xi, \xi \sim \pi)$ 的样本, 则该过程由如下模型表示:

$$X_1, X_2, \dots \mid P_{b,a} \stackrel{\text{iid}}{\sim} P_{b,a}, \quad P_{b,a} \mid (b, a, \xi) \sim \text{PDP}(b, a, \bar{\alpha}_\xi), \quad \xi \sim \pi.$$

定义 3.10 (双参数 Poisson-Dirichlet 混合模型) 假定 $i \in \{1, 2, \dots, n\}$, $\theta_i \in \mathbb{X}$ 和 $X_i \in \mathbb{Y}$ 是随机变量, 给定一个 θ_i , $\psi(\cdot; \theta_i) : \mathbb{Y} \rightarrow \mathbb{R}$ 是非负可测函数使得 $\int_{\mathbb{Y}} \psi(x; \theta_i) dx = 1$, 称

$$\int_{\mathbb{X}} \psi(x; \theta_i) P_{b,a}(d\theta_i)$$

是双参数 Poisson-Dirichlet 混合模型, 记作 $\text{PDPM}(b, a, \bar{\alpha}, \psi)$. 双参数 Poisson-Dirichlet 过程有如下表示:

$$X_i \mid \theta_i \stackrel{\text{iid}}{\sim} \psi(\cdot; \theta_i), \quad \theta_i \mid P_{b,a} \sim P_{b,a}, \quad P_{b,a} \sim \text{PDP}(b, a, \bar{\alpha}), \quad i \in \{1, \dots, n\}.$$

定义 3.11 (层次双参数 Poisson-Dirichlet 过程) 假定 $i \in \{1, 2, \dots, n\}$, $j \in \{1, 2, \dots, m\}$, $X_{j,i} \in \mathbb{Y}$ 是随机变量, 则层次双参数 Poisson-Dirichlet 过程可以由如下模型表示:

$$\begin{cases} X_{j,i} \mid P_{\beta,\gamma}^j \sim P_{\beta,\gamma}^j, & i \in \{1, 2, \dots, n\}, \quad j \in \{1, 2, \dots, m\}, \\ P_{\beta,\gamma}^j \mid (\beta, \gamma, G) \sim \text{PDP}(\beta, \gamma, G_{b,a}), \\ G_{b,a} \mid (b, a, \bar{\alpha}) \sim \text{PDP}(b, a, \bar{\alpha}). \end{cases}$$

定义 3.12 (层次双参数 Poisson-Dirichlet 混合模型) 假定 $i \in \{1, 2, \dots, n\}$, $j \in \{1, 2, \dots, m\}$, $X_{j,i} \in \mathbb{Y}$ 是随机变量, $\theta_{j,i}$ 是定义在 $(\mathbb{X}, \mathcal{X})$ 上的随机变量, 则层次双参数 Poisson-Dirichlet 混合模型可

由如下模型表示:

$$\begin{cases} X_{j,i} \mid \theta_{j,i} \sim \psi(\cdot; \theta_{j,i}), & i \in \{1, 2, \dots, n\}, \quad j \in \{1, 2, \dots, m\}, \\ \theta_{j,i} \mid P_{\beta,\gamma}^j \sim P_{\beta,\gamma}^j, \\ P_{\beta,\gamma}^j \mid (\beta, \gamma, G) \sim \text{PDP}(\beta, \gamma, G_{b,a}), \\ G_{b,a} \mid (b, a, \bar{\alpha}) \sim \text{PDP}(b, a, \bar{\alpha}). \end{cases}$$

3.6 Beta 过程

Beta 过程最早由 Hjort^[9] 提出和研究. 它在生存分析、统计学习和机器学习, 特别是在非参数的隐条件模型 (nonparametric latent factor model) 中有重要的应用 (参见文献 [12, 13]).

定义 3.13 (Beta 过程) 假定 $(\mathbb{X}, \mathcal{X})$ 是一个可测空间, N 为定义在 $(\mathbb{X} \times [0, 1], \mathcal{X} \times \mathcal{B})$ 上的 Poisson 随机测度, 其期望测度为 $\nu(du, d\pi) = \alpha(\theta)\pi^{-1}(1-\pi)^{\alpha(\theta)-1}d\pi\mu(d\theta)$, 其中 μ 是在 $(\mathbb{X}, \mathcal{X})$ 上的一个 σ -有限的测度, 函数 $\alpha(\theta) \in (0, \infty)$. 对任意一个紧致集 $A \in \mathcal{X}$, 定义以下完全随机的概率测度:

$$B(A) = \int_{A \times [0, 1]} \pi N(d\theta, d\pi). \quad (3.19)$$

那么, B 被称为 Beta 过程, 记作 $B \sim \text{BP}(\alpha, \mu)$, 其中 $\alpha(\cdot)$ 被称为集中函数, μ 被称为基测度.

通过上述定义可看出 Beta 过程是一个正的 Lévy 过程, 其 Lévy 测度为

$$\nu(d\theta, d\pi) = \alpha(\theta)\pi^{-1}(1-\pi)^{\alpha(\theta)-1}d\pi\mu(d\theta).$$

命题 3.4^[45] 假定 $t < 0$, 对任意的 $A \in \mathcal{X}$, $B(A)$ 的 Laplace 变换为

$$\mathbb{E}[e^{tB(A)}] = \exp \left\{ - \int_{A \times [0, 1]} (1 - e^{t\pi}) \nu(d\theta, d\pi) \right\}, \quad (3.20)$$

其中期望测度为 $\nu(d\theta, d\pi) = \alpha(\theta)\pi^{-1}(1-\pi)^{\alpha(\theta)-1}d\pi\mu(d\theta)$.

除了基于 Poisson 过程的定义之外, Beta 过程还有其他的构造方式.

命题 3.5 (有限逼近^[46]) 令 $\alpha(\theta) = \alpha$, 定义 μ 如同定义 3.13. 对于整数 $K > \mu(\mathbb{X})$, 定义如下过程:

$$B_K = \sum_{k=1}^K \pi_k \delta_{\theta_k}, \quad (3.21)$$

其中 $\pi_k \stackrel{\text{iid}}{\sim} \text{Beta}(\frac{\alpha\mu(\mathbb{X})}{K}, \frac{\alpha(K-\mu(\mathbb{X}))}{K})$ 与 $\theta_k \sim \frac{\mu}{\mu(\mathbb{X})}$ 互相独立, 则有

$$\lim_{K \rightarrow \infty} B_K = B \sim \text{BP}(\alpha, \mu).$$

上述命题给出了 Beta 过程的一个类似于断棍模型的逼近构造, 更进一步地可以给出一个非逼近形式的断棍模型.

命题 3.6 (Beta 过程的断棍构造^[46, 47]) 假定 μ 和 α 如同定义 3.13. 定义 $(\mathbb{X}, \mathcal{X})$ 上的随机测度 B 如下:

$$B = \sum_{i=1}^{\infty} \sum_{j=1}^{C_i} v_{ij}^{(i)} \prod_{l=1}^{i-1} (1 - v_{ij}^{(l)}) \delta_{\theta_{ij}}, \quad (3.22)$$

其中

$$C_i \stackrel{\text{iid}}{\sim} \text{Pois}(\mu(\mathbb{X})), \quad v_{ij}^{(l)} | \theta_{ij} \stackrel{\text{ind}}{\sim} \text{Beta}(1, \alpha(\theta_{ij})), \quad \theta_{ij} \stackrel{\text{iid}}{\sim} \frac{\mu}{\mu(U)},$$

则有 $B \sim \text{BP}(\alpha, \mu)$.

Beta 过程的重要性还在于它是一类 Lévy 过程, 即 Bernoulli 过程的共轭过程.

定义 3.14 (Bernoulli 过程) 假定 ν 是 $(\mathbb{X}, \mathcal{X})$ 上的一个测度, 如果 X 是一个 Lévy 测度为

$$\mathfrak{L}(d\pi, d\theta) = \delta_1(d\pi)\nu(d\theta)$$

的 $(\mathbb{X}, \mathcal{X})$ 上的 Lévy 过程, 称 X 为危险测度 (hazard measure) 为 ν 的 Bernoulli 过程, 记作 $X \sim \text{BeP}(\nu)$.

命题 3.7 (Beta 过程对 Bernoulli 过程的共轭性^[9]) 令 $B \sim \text{BP}(\alpha, \mu)$ (可以假设 $\alpha \in (0, \infty)$ 为定值), 而且 $X_1, \dots, X_n | B \stackrel{\text{iid}}{\sim} \text{BeP}(B)$, 即给定测度 B , $\{X_i\}_{i=1}^n$ 是序列危险测度为 B 的 Bernoulli 过程, 则后验分布满足

$$B | X_1, \dots, X_n \sim \text{BP}\left(\alpha + n, \frac{\alpha}{\alpha + n}\mu + \frac{1}{\alpha + n} \sum_{i=1}^n X_i\right). \quad (3.23)$$

Beta 过程也可以相应地推广到 Beta 混合过程和层次 Beta 过程 (参见文献 [48]), 其具体表示形式可由 Dirichlet 过程的推广类似得到.

3.7 中国餐馆模型

在介绍 Dirichlet 过程的性质和构造时介绍了 Pólya 罐子模型, 该模型可以用来计算 Dirichlet 过程的样本的联合分布, 也可以通过 Pólya 数列构造 Dirichlet 过程, 由定理 2.4 可知 Pólya 罐子模型主要描述样本 $X_1, X_2, \dots$, 而通过 Pólya 罐子模型可知样本 $X_1, X_2, \dots, X_n$ 可以做聚类划分, 即数值相同的样本被划分为一类, 如果对这些类别的分布感兴趣, 则需要一种新的模型, 该模型被称为中国餐馆模型. 定义 $|\alpha| = \alpha(\mathbb{X})$ 与定理 2.4 中一致. 假设有一个餐馆, 餐馆中的餐桌数量是无限的, 而每张餐桌都可以容纳无限多个消费者, 假设一开始餐馆里没有消费者, 消费者依序一个一个地进入餐馆, 并依下列规则选择餐桌就餐:

(1) 第一个消费者进入餐馆选择第一张餐桌坐下就餐.

(2) 假定餐馆已经有了 i 个消费者. 第 $i+1$ 个消费者进入餐馆后有如下两个选择.

选择一: 选择一个空餐桌, 假定发生的概率为 $\frac{|\alpha|}{|\alpha|+i}$.

选择二: 选择第 k 个已经有人的餐桌, 假定发生的概率为 $\frac{c_k}{|\alpha|+i}$, 其中 $k = 1, \dots, i$, c_k 表示第 k 个已经有人的餐桌的人数.

在该模型中, 餐桌可以用来代表样本被划分的类, 消费者代表了样本 $X_1, X_2, \dots$, 可以将 Pólya 罐子模型中 $X_i = X_j$ ($i \neq j$), 理解为消费者 X_i 和 X_j 选择了同一个桌子, 显然可知中国餐馆模型与 Pólya 罐子模型是互相吻合的. 在简单的聚类问题中, 为了将样本 $X_1, \dots, X_n$ 进行分类, 可以通过上述的中国餐馆模型中消费者选择餐桌的策略生成算法, 用计算机模拟对 $X_1, \dots, X_n$ 的分类. 也就是说, 该模型为解决聚类问题的算法提供了直观的解释.

3.8 印度自助餐模型

印度自助餐模型^[12, 49] (Indian buffet process) 的提出主要是为了解决, 在因子模型 (factorial model) 中, 如果每个数据都与一系列隐 Bernoulli 型的特征 (latent Bernoulli feature) 相关, 那么在非参数 Bayes

的意义下, 如何选取先验随机概率测度的问题. 类似于中国餐馆模型, 印度自助餐模型是对 Beta 过程从样本的角度进行描述. 假设在一个自助餐厅, N 个人排队品尝已经排好顺序的菜品, 假设菜品的数量是无限多个, 记 $Z_{i,k} \in \{0, 1\}$ 为一个 Bernoulli 样本, 如果 $Z_{i,k} = 1$, 则表示第 i 个消费者品尝了第 k 道菜. 沿用定义 3.13 中的符号和意义, 并记 $\nu(\mathbb{X}) = \gamma$. 假定消费者进入自助餐厅后按照下列原则品尝菜品:

(1) 第一个消费者进入自助餐厅品尝的菜品的数量服从 $\text{Poisson}(\gamma)$;

(2) 第 i 个消费者品尝了第 k 道菜, 即 $Z_{i,k} = 1$ 的概率为 $\frac{m_k}{i}$, 也即 $Z_{i,k} \sim \text{Bernoulli}(\frac{m_k}{i})$, 其中 m_k 表示品尝过第 k 道菜的消费者个数;

(3) 第 i 个消费者品尝新菜品 (未被品尝过的菜品) 的数量服从 $\text{Poisson}(\frac{\gamma}{i})$. 可以用一个矩阵 $\mathcal{Z}$ 来表述这 N 个消费者的选择, 矩阵的每个元素的取值范围为 $\{0, 1\}$, 即 $\mathcal{Z} = [Z_{i,k}]_{1 \leq i \leq N, 1 \leq k < \infty}$, 该矩阵有 N 行、无穷多列. 记 $\mathbb{M}_N$ 为这样的矩阵的全体, 即

$$\mathbb{M}_N = \{z = [z_{i,k}]_{1 \leq i \leq N, 1 \leq k < \infty}, z_{ik} = 0, 1\}.$$

记

$$\begin{cases} K_1^{(i)}(z) = \text{第 } i \text{ 个消费者品尝的新菜品的数量} = \sum_{k=1}^{\infty} z_{ik}, \\ H_N(\mathcal{Z}) = \sum_{j=1}^N \frac{1}{j}, \\ K_+(z) = \sum_{h=1}^{2^N-1} K_h(z), \quad \text{其中 } K_h(z) = \#\left\{k : m_k(z) := \sum_{i=1}^N z_{ik} = h\right\}, \end{cases}$$

则在印度自助餐模型中产生得到某个特定的 $\mathcal{Z} = z$ 的概率^[12, 49] 可表示为

$$P(\mathcal{Z} = z) = \frac{\gamma^{K_+(z)}}{\prod_{i=1}^N K_1^{(i)}(z)!} e^{-\gamma H_N} \prod_{k=1}^{K_+(z)} \frac{(N - m_k(z))! (m_k(z) - 1)!}{N!}, \quad \forall z \in \mathbb{M}_N. \quad (3.24)$$

可以通过 Beta 过程对 Bernoulli 过程的共轭性 (即命题 3.7) 来刻画 Beta 过程与印度自助餐模型的关系. 在印度自助餐模型中, $\mathbb{X}$ 表示了餐馆里菜品的全集, 假定 $i \in \mathbb{N}^+$, 用 X_i 表示第 i 个消费者选择的菜品, $X_i = X_j$ ($i \neq j$) 表示第 i 和 j 个消费者选取相同的菜品. 首先 X_1 的分布可以由

$$E[X_1] = E[E(X_1 | B)] = \mu$$

得到. 通过 (3.23) 和等式 $P(X_{n+1} | X_1, \dots, X_n) = E_{B | X_1, \dots, X_n}[P(X_{n+1} | B)]$ 得到

$$X_{n+1} | X_1, \dots, X_n \sim \text{BeP}\left(\frac{\alpha}{\alpha+n}\mu + \frac{1}{\alpha+n} \sum_{i=1}^n X_i\right) = \text{BeP}\left(\frac{\alpha}{\alpha+n}\mu + \sum_j \frac{m_{n,j}}{\alpha+n} \delta_{\omega_j}\right), \quad (3.25)$$

其中 $m_{n,j}$ 表示消费者 $X_1, \dots, X_n$ 中品尝菜品 ω_j 的个数, 为方便起见, 记 α 为常数 (其实 α 可以是一个 $\mathbb{X}$ 上的测度), $\mu(\mathbb{X}) = \gamma$. 通过上述结果可知, $X_1 \sim \text{BeP}(\mu)$, 也就是说 $X_1(\mathbb{X}) \sim \text{Poisson}(\gamma)$, 即第一个消费者品尝的菜品数量服从参数为 γ 的 Poisson 分布. 而 X_{n+1} 的条件分布可以写成两部分: $X_{n+1} = Y + Z$, $Y \sim \text{BeP}(\sum_j \frac{m_{n,j}}{\alpha+n} \delta_{\omega_j})$ 表示以概率 $\frac{m_{n,j}}{\alpha+n}$ 品尝菜品 ω_j (第 j 道菜), 而 $Z \sim \text{BeP}(\frac{\alpha}{\alpha+n}\mu)$ 则是一个强度 (intensity) 为 $\frac{\alpha}{\alpha+n}\mu$ 的 Poisson 过程, 也就是说品尝新菜品的数量服从 $\text{Poisson}(\frac{\alpha\gamma}{\alpha+n})$ 分布.

4 非参数 Bayes 估计

4.1 一般的思想

一般的参数 Bayes 模型可以表示如下:

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x, \theta),$$

其中对于任何给定的 θ , $f(\cdot, \theta)$ 是一个分布函数, 再假定 θ 本身也是一个分布为 π (先验分布) 的随机变量. 根据 Bayes 定理, 后验分布为

$$\theta | X_1, \dots, X_n \sim \frac{L_n(\theta)\pi(\theta)}{\int L_n(\theta)\pi(\theta)d\theta},$$

其中 $L_n(\theta) = \prod_{i=1}^n f(X_i, \theta)$ 为似然函数. 可以通过后验分布估计某些统计量, 例如, 用后验分布的期望估计 θ 本身.

在非参数 Bayes 模型中, 样本 $X_1, \dots, X_n$ 是从 $(\mathbb{X}, \mathcal{X})$ 上的一个随机概率测度 P 中抽样的. 把这个 P 本身看成是一个参数. 这样, 就需要通过计算 $P | X_1, \dots, X_n$ 得到对 $(\mathbb{X}, \mathcal{X})$ 上随机概率测度 P 的估计. 这就是说, 有了样本 $X_1, \dots, X_n$ 的具体观察值之后, 就可以 (在理论上) 得到 $(\mathbb{X}, \mathcal{X})$ 上的一个确定的概率测度 $P | X_1, \dots, X_n$ (后验分布). 有了这个后验分布, 就可以计算任何一个统计量了.

参数 Bayes 模型也可以看成上述思想的一个特例. 如果参数 θ 是随机的, 则 $f(x, \theta)$ 可以看成是一个随机概率测度. 在进行了抽样 $X_1, \dots, X_n$ 之后对 θ 进行估计, 并把 θ 的估计值代入 $f(x, \theta)$. 这样得到的分布 $f(x, \theta | X_1, \dots, X_n)$ 就相当于 $P | X_1, \dots, X_n$.

对于一般的先验随机概率测度 P , 后验概率测度 $P | X_1, \dots, X_n$ 是非常难以计算的. 这一困难即使是在参数 Bayes 模型中也是存在的. 所以, 怎样选择一个恰当的先验随机概率测度是非参数 Bayes 模型的一个最重要的任务: 人们既要先验随机概率测度包括尽量多的概率测度使得它们确实能够用来拟合需要处理的数据, 又要使得后验概率测度 $P | X_1, \dots, X_n$ 的计算不是太难. 这是一个很艰难的任务. Dirichlet 过程的出现为解决这一困难提出了一个方案. 这就是为什么 Dirichlet 过程是非参数 Bayes 模型的核心. 首先给出两个有关 Dirichlet 过程的性质, 并在随后的几小节中对期望和方差等统计量进行估计.

假定 $(\mathbb{X}, \mathcal{X})$ 为一个 Polish 空间, α 是 $(\mathbb{X}, \mathcal{X})$ 上的一个非负局部有限测度 (非负而且 σ -可加), 本节仍然假定 $P \sim \text{DP}(\alpha)$ 是一个 Dirichlet 过程.

性质 4.1 ^[2] 假定 ϕ 为 $(\mathbb{X}, \mathcal{X})$ 上的可测函数, 如果 $\int_{\mathbb{X}} |\phi| d\alpha < \infty$, 则依概率 1 有 $\int_{\mathbb{X}} |\phi| dP < \infty$, 而且

$$\mathbb{E} \left[\int_{\mathbb{X}} \phi dP \right] = \int_{\mathbb{X}} \phi d\mathbb{E}[P] = \frac{\int_{\mathbb{X}} \phi d\alpha}{|\alpha|}.$$

证明 根据 Dirichlet 过程的断棍构造 (2.3) 可知,

$$\int_{\mathbb{X}} |\phi| dP = \sum_{j=1}^{\infty} w_j \int_{\mathbb{X}} |\phi(x)| \delta_{\theta_j}(dx) = \sum_{j=1}^{\infty} |\phi(\theta_j)| w_j. \quad (4.1)$$

根据单调收敛定理可知,

$$\mathbb{E} \left[\int_{\mathbb{X}} |\phi| dP \right] = \sum_{j=1}^{\infty} \mathbb{E}[|\phi(\theta_j)|] \mathbb{E}[w_j] = \frac{1}{|\alpha|} \int_{\mathbb{X}} |\phi| d\alpha \sum_{j=1}^{\infty} \mathbb{E}[w_j] = \frac{1}{|\alpha|} \int_{\mathbb{X}} |\phi| d\alpha < \infty.$$

因此, 依概率 1 有 $\int_{\mathbb{X}} |\phi| dP < \infty$, 而且 $\int_{\mathbb{X}} \phi dP = \sum_{j=1}^{\infty} \phi(\theta_j) w_j$ 依概率 1 绝对收敛, 另外由 (4.1) 可知其有界. 所以可根据有界收敛定理得

$$\mathbb{E} \left[\int_{\mathbb{X}} \phi dP \right] = \sum_{j=1}^{\infty} \mathbb{E}[\phi(\theta_j)] \mathbb{E}[w_j] = \frac{\int_{\mathbb{X}} \phi d\alpha}{|\alpha|}.$$

证毕. $\square$

该性质更直观地给出了 P 与 α 之间的关系, 如果 $(\mathbb{X}, \mathcal{X}) = (\mathbb{R}, \mathcal{B})$, 那么可知, 如果 α 存在 n 阶矩, 则 P 也存在 n 阶矩.

性质 4.2 ^[2] 令 ϕ_1 和 ϕ_2 为定义在 $(\mathbb{X}, \mathcal{X})$ 上的可测函数. 如果 $\int_{\mathbb{X}} |\phi_1| d\alpha < \infty$, $\int_{\mathbb{X}} |\phi_2| d\alpha < \infty$, $\int_{\mathbb{X}} |\phi_1 \phi_2| d\alpha < \infty$, 则

$$\mathbb{E} \left[\int_{\mathbb{X}} \phi_1 dP \int_{\mathbb{X}} \phi_2 dP \right] = \frac{\sigma_{12}}{|\alpha| + 1} + \mu_1 \mu_2, \quad (4.2)$$

其中

$$\begin{cases} \mu_i = \frac{\int_{\mathbb{X}} \phi_i d\alpha}{|\alpha|}, & i \in \{1, 2\}, \\ \sigma_{12} = \frac{\int_{\mathbb{X}} \phi_1 \phi_2 d\alpha}{|\alpha|} - \mu_1 \mu_2. \end{cases}$$

证明 类似于性质 4.1 的证明, 我们知道 $\sum_{i=1}^{\infty} \phi_1(\theta_i) w_i$ 和 $\sum_{j=1}^{\infty} \phi_2(\theta_j) w_j$ 依概率 1 绝对收敛, 因此,

$$\int_{\mathbb{X}} \phi_1 dP \int_{\mathbb{X}} \phi_2 dP = \sum_{i=1}^{\infty} \phi_1(\theta_i) w_i \sum_{j=1}^{\infty} \phi_2(\theta_j) w_j = \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \phi_1(\theta_i) w_i \phi_2(\theta_j) w_j.$$

而且, 上述级数的上界为 $\sum_{i=1}^{\infty} \sum_{j=1}^{\infty} |\phi_1(\theta_i) \phi_2(\theta_j)| w_i w_j$. 因此,

$$\begin{aligned} \mathbb{E} \left[\int_{\mathbb{X}} \phi_1 dP \int_{\mathbb{X}} \phi_2 dP \right] &= \sum_{i=1}^{\infty} \sum_{j=1}^{\infty} \mathbb{E}[\phi_1(\theta_i) \phi_2(\theta_j)] \mathbb{E}[w_i w_j] \\ &= \sum_{i \neq j} \mathbb{E}[\phi_1(\theta_i)] \mathbb{E}[\phi_2(\theta_j)] \mathbb{E}[w_i w_j] + \sum_{i=1}^{\infty} \mathbb{E}[\phi_1(\theta_i) \phi_2(\theta_i)] \mathbb{E}[w_i^2] \\ &= \mu_1 \mu_2 \sum_{i \neq j} \mathbb{E}[w_i w_j] + (\sigma_{12} + \mu_1 \mu_2) \sum_{i=1}^{\infty} \mathbb{E}[w_i^2] \\ &= \mu_1 \mu_2 + \sigma_{12} \mathbb{E} \left[\sum_{i=1}^{\infty} w_i^2 \right]. \end{aligned} \quad (4.3)$$

现在可以通过以下方法来计算上面的 $\mathbb{E}[\sum_{i=1}^{\infty} w_i^2]$:

$$\begin{aligned} \mathbb{E} \left[\sum_{i=1}^{\infty} w_i^2 \right] &= \mathbb{E} \left[\sum_{i=1}^{\infty} v_i^2 \prod_{l=1}^{i-1} (1 - v_l)^2 \right] = \sum_{i=1}^{\infty} \mathbb{E}[v_i^2] \prod_{l=1}^{i-1} \mathbb{E}[(1 - v_l)^2] \\ &= \sum_{i=1}^{\infty} \frac{2}{(|\alpha| + 1)(|\alpha| + 2)} \left(\frac{|\alpha|}{|\alpha| + 2} \right)^{i-1} \\ &= \frac{1}{|\alpha| + 1}. \end{aligned}$$

将上述结果代入 (4.3), 即得到 (4.2). $\square$

4.2 分布函数的非参数 Bayes 估计

以后的几个小节中假定 $(\mathbb{X}, \mathcal{X}) = (\mathbb{R}, \mathcal{B})$, $P \sim \text{DP}(\alpha)$ 是 $(\mathbb{R}, \mathcal{B})$ 上的一个 Dirichlet 过程. 我们将给出 Dirichlet 过程在解决一些统计问题中的应用.

首先估计分布函数 $F_0(t) = P(X \leq t)$. 因为 F_0 的值不知道, 假定它是 $F(t) = P((-\infty, t])$. 现在, $F(t)$ 是一个随机变量, 在对 P 抽样之后, 要对它进行估计. 在非参数 Bayes 模型下, 用如下损失函数 (loss function) 估计分布函数 $F(t)$:

$$L(P, \hat{F}) = \int_{\mathbb{R}} (F(t) - \hat{F}(t))^2 dW(t), \quad (4.4)$$

其中 W 是一个在 $(\mathbb{R}, \mathcal{B})$ 上的有限测度, 被称为权重函数 (weight function). 在无样本的情形下,

$$E[L(P, \hat{F})] = \int_{\mathbb{R}} E[(F(t) - \hat{F}(t))^2] dW(t)$$

在 $\hat{F}_0(t) = E[F(t)]$ 时取最小值.

因此, 在无样本的情形下对分布函数 $F(t)$ 的估计为

$$\hat{F}_0(t) = E[F(t)] = E[P((-\infty, t])] = \frac{\alpha((-\infty, t])}{|\alpha|}. \quad (4.5)$$

现在假定已经有了样本 $X_1, \dots, X_n$. 将通过解决

$$E[L(P, \hat{F}) | X_1, \dots, X_n] = \int_{\mathbb{R}} E[(F(t) - \hat{F}(t))^2 | X_1, \dots, X_n] dW(t)$$

的极小化问题来找到分布函数 $F(t)$ 的估计. 很显然, 这一极小化问题的解是 $\hat{F}_n = E[F(t) | X_1, \dots, X_n]$. 由定理 2.3 知 $P | X_1, \dots, X_n$ 由 (2.10) 给出. 再利用性质 4.1, 有

$$\begin{aligned} \hat{F}_n &= E[F(t) | X_1, \dots, X_n] \\ &= \frac{\alpha((-\infty, t]) + \sum_{i=1}^n \delta_{X_i}((-\infty, t])}{|\alpha| + n} \\ &= r_n \hat{F}_0(t) + (1 - r_n) F_n(t | X_1, \dots, X_n), \end{aligned} \quad (4.6)$$

其中

$$r_n = \frac{|\alpha|}{|\alpha| + n}, \quad F_n(t | X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}((-\infty, t]). \quad (4.7)$$

上面的公式可以解释为在有了样本的观测值之后, 对分布函数的估计是 $F_0(t)$ 与 $F_n(t | X_1, \dots, X_n)$ 的线性组合, 相对应的权重分别为 r_n 和 $1 - r_n$. 当集中参数 $|\alpha|$ 相对于 n 足够大时, 观测值对于分布函数的估计没有太大的影响, 而当样本个数 n 趋于无穷时, $\hat{F}(t)$ 则变成了样本的经验分布 $F_n(t | X_1, \dots, X_n)$, 它将会几乎处处地收敛到真实的分布.

4.3 期望、中值和 q 分位值的非参数 Bayes 估计

假设 α 的期望有限, 用如下损失函数来估计期望 μ :

$$L(P, \hat{\mu}) = (\mu - \hat{\mu})^2, \quad (4.8)$$

其中 μ 可表示为 $\mu = \int_{\mathbb{R}} x dP(x)$.

假设 $P \sim \text{DP}(\alpha)$, 根据性质 4.1 可知 $\mu = \int_{\mathbb{R}} x dP(x)$ 存在, 而且

$$\mathbb{E}[\mu] = \mathbb{E}\left[\int_{\mathbb{R}} x dP(x)\right] = \frac{\int_{\mathbb{R}} x d\alpha(x)}{|\alpha|}.$$

在无样本的情形下, 当 $\mu = \frac{\int_{\mathbb{R}} x d\alpha(x)}{|\alpha|}$ 时, $\mathbb{E}[L(P, \hat{\mu})]$ 取得最小值. 这样, 在无样本的情形下, μ 的估计为 $\hat{\mu}_0 = \frac{\int_{\mathbb{R}} x d\alpha(x)}{|\alpha|}$. 现在假定有了样本 $X_1, \dots, X_n$, 根据定理 2.3, 有

$$\hat{\mu}(X_1, \dots, X_n) = \frac{\int_{\mathbb{R}} x d(\alpha(x) + \sum_{i=1}^n \delta_{X_i}(x))}{|\alpha| + n} = r_n \hat{\mu}_0 + (1 - r_n) \bar{X}_n,$$

其中 $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$, r_n 与 (4.7) 一致. 容易看出, 当 $n \rightarrow \infty$ 时, $r_n \rightarrow 0$, 此时对期望的估计与样本均值一致, 它收敛到真实期望.

下面要估计一个未知分布 P 的中位数 $m = \text{med}(P)$. 假定 $P \sim \text{DP}(\alpha)$, $F(t) = P((-\infty, t])$, 首先需要证明依概率 1 中位数 m 是唯一的. 假设存在两个不同的中值 m 和 m' , 不妨令 $m > m'$, 根据中值的定义可知,

$$\begin{aligned} \int_{-\infty}^m dF(t) &\geq \frac{1}{2}, & \int_m^{\infty} dF(t) &\geq \frac{1}{2}, \\ \int_{-\infty}^{m'} dF(t) &\geq \frac{1}{2}, & \int_{m'}^{\infty} dF(t) &\geq \frac{1}{2}. \end{aligned}$$

那么, $\int_{-\infty}^{m'} dF(t) + \int_m^{\infty} dF(t) \geq 1$, 因此, 依概率 1 有 $\int_{m'}^m dF(t) = 0$. 但是, 根据 Dirichlet 过程的性质可知, $F(t) = P((-\infty, t]) \sim \text{Beta}(\alpha((-\infty, t]), \alpha((t, \infty)))$, 依概率 1 有 $\int_{m'}^m dF(t) = P((m', m)) > 0$, 这与我们的假设互相矛盾, 由此可推出 m 是唯一的. 而且对于 $P \sim \text{DP}(\alpha)$, m 是一个随机变量.

为了要估计 P 的中位数, 我们所用损失函数为

$$L(P, \hat{m}) = |m - \hat{m}|.$$

根据性质 4.1 可知, 随机变量 m 的分布的中值就是 m 的 Bayes 估计量 (参见文献 [50]), 也即当 $\hat{m} = \text{med}(\text{dist. med}(P))$ 时, $\mathbb{E}_m[L(P, \hat{m})]$ 取得最小值. 假设 t 是 m 的分布的中值, 则有

$$\mathbb{P}[m < t] \leq \frac{1}{2} \leq \mathbb{P}[m \leq t]. \quad (4.9)$$

根据对 m 的定义可知 $\mathbb{P}[m \leq t] = \mathbb{P}[F(t) > \frac{1}{2}]$ 且 $\mathbb{P}[m < t] = \mathbb{P}[F(t) > \frac{1}{2}]$, 可推出 $\mathbb{P}[F(t) > \frac{1}{2}] = \frac{1}{2}$. 另外, 有 $F(t) \sim \text{Beta}(\alpha((-\infty, t]), \alpha((t, \infty)))$. 根据 Beta 分布的性质可知, $\mathbb{P}[F(t) > \frac{1}{2}] = \frac{1}{2}$ 当且仅当 Beta 分布的两个参数相等, 即 t 满足如下不等式:

$$\frac{\alpha((-\infty, t))}{|\alpha|} \leq \frac{1}{2} \leq \frac{\alpha((-\infty, t))}{|\alpha|}.$$

也就是说 t 正是分布 $\alpha = P$ 的中值.

那么结合问题 4.2, 在无样本的情形下, $\hat{m} = \text{med}(\hat{F}(t)) = \text{med}(\mathbb{E}[F(t)]) = \text{med}(\hat{F}_0(t))$, 其中 $\hat{F}_0(t)$ 与 (4.5) 相同.

同理, 在给定样本 $X_1, \dots, X_n$ 的情形下, $\hat{m}(X_1, \dots, X_n) = \text{med}(\hat{F}(t | X_1, \dots, X_n))$, 其中 $\hat{F}(t | X_1, \dots, X_n)$ 由问题 4.2 中的 (4.6) 定义.

对 q 分位数的估计类似于对中值的估计, 记 t_q 为分布 P 的 q 分位数, 即

$$P((-\infty, t_q)) \leq q \leq P((-\infty, t_q]).$$

在非参数 Bayes 模型下, 如果 $P \sim \text{DP}(\alpha)$, 类似于中值估计的证明可知, 对于 $q \in (0, 1)$, 依概率 1 有 t_q 是唯一的, 而且 t_q 是一个随机变量. 用以下损失函数来估计 t_q , 令 $p \in (0, 1)$,

$$L(P, \hat{t}_q) = \begin{cases} p(t_q - \hat{t}_q), & \text{如果 } t_q \geq \hat{t}_q, \\ (1-p)(t_q - \hat{t}_q), & \text{如果 } t_q < \hat{t}_q. \end{cases}$$

根据 Bayes 定理可知 t_q 的 Bayes 估计量为 t_q 的分布的 p 分位值 (参见文献 [50]). 根据对 t_q 的定义可得 t_q 的分布可以由如下形式给出:

$$\begin{aligned} P[t_q \leq t] &= P[F(t) > q] \\ &= \int_q^1 \frac{\Gamma(|\alpha|)}{\Gamma(\alpha((-\infty, t]))\Gamma(\alpha((t, \infty)))} z^{\Gamma(\alpha((-\infty, t]))-1} (1-z)^{\Gamma(\alpha((t, \infty)))-1} dz. \end{aligned} \quad (4.10)$$

那么, t_q 的分布的 p 分位值可通过以下等式来计算:

$$\int_q^1 \frac{\Gamma(|\alpha|)}{\Gamma(\alpha((-\infty, t]))\Gamma(\alpha((t, \infty)))} z^{\Gamma(\alpha((-\infty, t]))-1} (1-z)^{\Gamma(\alpha((t, \infty)))-1} dz = p. \quad (4.11)$$

固定 p, q 和 $|\alpha|$, 假设由等式 (4.11) 解得 $t = u(p, q, |\alpha|)$, 那么根据 Beta 分布的性质, $t = u(p, q, |\alpha|)$ 就是分布 $\alpha = E[P]$ 的 p 分位值.

在无样本的情形下, $\hat{t}_q = \hat{F}(t)$ 的第 $u(p, q, |\alpha|)$ 分位值等于 $\hat{F}_0(t)$ 的第 $u(p, q, |\alpha|)$ 分位值, 其中 $\hat{F}_0(t)$ 与 (4.5) 相同.

在给定样本 $X_1, \dots, X_n$ 的情形下, $\hat{t}_q(X_1, \dots, X_n)$ 等于 $\hat{F}(t | X_1, \dots, X_n)$ 的第 $u(p, q, |\alpha|)$ 分位值, 其中 $\hat{F}(t | X_1, \dots, X_n)$ 与问题 4.2 中的 (4.6) 定义相同.

4.4 方差、协方差的非参数 Bayes 估计

首先使用如下损失函数估计未知分布 P 的方差:

$$L(P, \hat{\sigma}^2) = (\text{Var}(P) - \hat{\sigma}^2)^2.$$

假设 α 的二阶矩有限, 继续假设 $P \sim \text{DP}(\alpha)$ 是一个 Dirichlet 过程. 未知分布 P 的方差

$$\text{Var}(P) = \int_{\mathbb{R}} x^2 dP(x) - \left(\int_{\mathbb{R}} x dP(x) \right)^2$$

是一个随机变量, 其期望如下:

$$E[\text{Var}(P)] = E \left[\int_{\mathbb{R}} x^2 dP(x) \right] - E \left[\left(\int_{\mathbb{R}} x dP(x) \right)^2 \right] = (\sigma_0^2 + \mu_0^2) - \left(\frac{\sigma_0^2}{|\alpha| + 1} + \mu_0^2 \right) = \frac{|\alpha|}{|\alpha| + 1} \sigma_0^2,$$

其中 $\mu_0 = \int_{\mathbb{R}} x d\frac{\alpha(x)}{|\alpha|}$, $\sigma_0^2 = \frac{\int_{\mathbb{R}} x^2 d\alpha(x)}{|\alpha|} - \mu_0^2$.

这样, 在无样本的情形下,

$$\hat{\sigma}_0^2 = E[\text{Var}(P)] = \frac{|\alpha|}{|\alpha| + 1} \sigma_0^2.$$

在有了样本 $X_1, \dots, X_n$ 的情形下, 根据定理 2.3 可知, $P \mid X_1, \dots, X_n$ 仍然是一个 Dirichlet 过程. 再根据性质 4.2, 有

$$\begin{aligned}\hat{\sigma}_0^2(X_1, \dots, X_n) &= \mathbb{E}[\text{Var}(P \mid X_1, \dots, X_n) \mid X_1, \dots, X_n] \\ &= \frac{|\alpha| + n}{|\alpha| + n + 1} \text{Var}(\hat{F}(t \mid X_1, \dots, X_n)) \\ &= \frac{|\alpha| + n}{|\alpha| + n + 1} (r_n^2 \sigma_0^2 + (1 - r_n)^2 s_n^2 + r_n(1 - r_n)(\mu_0 - \bar{X}_n)^2),\end{aligned}\quad (4.12)$$

其中 $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$, r_n 与 (4.7) 一致, $s_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ 是样本方差.

假定 $\mathbb{X} = \mathbb{R}^2$, 通过如下损失函数对未知分布 P 的协方差进行估计:

$$L(P, \hat{\sigma}_{12}) = (\text{Cov}(P) - \hat{\sigma}_{12})^2.$$

在非参数 Bayes 模型下, 令 $P \sim \text{DP}(\alpha)$, 在无样本的情形下,

$$\hat{\sigma}_{12} = \mathbb{E}[\text{Cov}(P)] = \frac{|\alpha|}{|\alpha| + 1} \sigma_{12},$$

其中

$$\sigma_{12} = \frac{\int_{\mathbb{R}} xy d\alpha(x, y)}{|\alpha|} - \mu_1 \mu_2, \quad \mu_1 = \frac{\int_{\mathbb{R}} x d\alpha(x, y)}{|\alpha|}, \quad \mu_2 = \frac{\int_{\mathbb{R}} y d\alpha(x, y)}{|\alpha|}.$$

在给定样本 $(X_1, Y_1), \dots, (X_n, Y_n)$ 的情形下, 类似于上面的计算, 可以得到

$$\hat{\sigma}_{12}((X_1, Y_1), \dots, (X_n, Y_n)) = \frac{|\alpha| + n}{|\alpha| + n + 1} (r_n^2 \sigma_{12} + (1 - r_n)^2 s_{12} + r_n(1 - r_n)(\mu_1 - \bar{X}_n)(\mu_2 - \bar{Y}_n)),$$

其中 $s_{12} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)$.

4.5 $\int FdG$ 的非参数 Bayes 估计

假设 X 和 Y 是两个独立的实值随机变量. 它们的分布函数 F 和 G 是未知的. 想要估计 $P(X \leq Y) = \int FdG$. 如果用非参数 Bayes 来估计, 首先假定 $P_1 \sim \text{DP}(\alpha_1)$ 和 $P_2 \sim \text{DP}(\alpha_2)$ 是 $\mathbb{R}$ 上的互相独立的 Dirichlet 过程. 令 $X_1, \dots, X_m$ 为从 P_1 取得的样本, $Y_1, \dots, Y_n$ 为从 P_2 取得的样本. 通过如下损失函数估计 $P(X \leq Y)$:

$$L(P_1, P_2; \hat{\eta}) = \left(\int_{\mathbb{R}} FdG - \hat{\eta} \right)^2.$$

在无样本的情形下, 有如下的关于 $P(X \leq Y)$ 的 Bayes 估计:

$$\hat{\eta} = \mathbb{E} \left[\int_{\mathbb{R}} FdG \right] = \int_{\mathbb{R}} \mu_F d\mu_G,$$

其中 $\mu_F = \mathbb{E}[F]$, $\mu_G = \mathbb{E}[G]$.

在有了样本 $X_1, \dots, X_m$ 和 $Y_1, \dots, Y_n$ 的情形下, 仍然可以利用定理 2.3 和性质 4.2 来得到

$$\hat{\eta}(X_1, \dots, X_m, Y_1, \dots, Y_n) = \mathbb{E} \left[\int_{\mathbb{R}} \hat{F}_m d\hat{G}_n \mid X_1, \dots, X_m, Y_1, \dots, Y_n \right]$$

$$\begin{aligned}
 &= r_{1,m} p_{2,n} \hat{\eta} + r_{1,m} (1 - r_{2,n}) \frac{\sum_{j=1}^n \mu_F(Y_j)}{\mu_F(Y_j)} + (1 - r_{1,m}) r_{2,n} \frac{\sum_{i=1}^m (1 - \mu_G(X_i))}{m} \\
 &\quad + (1 - r_{1,m}) (1 - r_{2,n}) \frac{U}{mn},
 \end{aligned}$$

其中

$$\begin{cases} \hat{F}_m = P_1 | X_1, \dots, X_m, \\ \hat{G}_n = P_2 | Y_1, \dots, Y_n, \\ U = \sum_{i=1}^m \sum_{j=1}^n \mathbb{I}_{(-\infty, Y_j]}(X_i), \end{cases} \quad \text{这里 } \mathbb{I}_A(x) \text{ 为示性函数.}$$

4.6 “容忍”区间的非参数 Bayes 估计

在统计学中, 容忍区间 (tolerance region)^[51] 是一个随机的区间 (T_L, T_U) , 使得在一定的显著水平 ϵ 下有至少百分之 $\beta \times 100$ 的样本将会落在这个区间内, 即

$$P[F(T_U) - F(T_L) \geq \beta] \geq 1 - \epsilon.$$

类似于单边置信区间, 定义显著水平为 ϵ 的单边容忍区间 $(-\infty, a]$ 满足

$$P[F(a) \geq \beta] = 1 - \epsilon,$$

其中 a 被称为容忍上界.

本节中所用的“容忍”上界 a 并非上述定义的容忍上界, 而是利用了容忍区间定义的思想. 我们的问题是通过如下损失函数来估计未知分布 P 的 q 分位值 t_q :

$$L(P, \hat{a}) = pP((-\infty, \hat{a}]) + q\mathbb{I}_{(\hat{a}, \infty)}(t_q),$$

其中 $p \in (0, 1)$ 为给定的常数. 想要找到最小的区间 $(-\infty, a]$, 使得上界 a 能保证 $E[L(P, a)]$ 取得最小值, 在这种意义下, 把 a 叫作“容忍”上界.

假如在给定 t_q 时, 损失函数在 $\hat{a} = t_q$ 时取得最小值; 假如 t_q 是随机的, 则需要保证在不使 $E[pP((-\infty, a])]$ 过大时尽量过高地预测 t_q , 这样一来可以保证 $E[q\mathbb{I}_{(a, \infty)}(t_q)]$ 尽可能地小. 假定 $P \sim \text{DP}(\alpha)$, 则 t_q 为随机变量. 在无样本的情形下, $E[L(P, a)]$ 可由以下步骤得到:

$$\begin{aligned}
 E[L(P, a)] &= pE[P((-\infty, \hat{a}))] + qP(t_q > a) = p\hat{F}_0(a) + qP(F(a) \leq q) \\
 &= pu + q \int_0^q \frac{\Gamma(|\alpha|)}{\Gamma(u|\alpha|)\Gamma((1-u)|\alpha|)} z^{u|\alpha|-1} (1-z)^{(1-u)|\alpha|-1} dz,
 \end{aligned} \tag{4.13}$$

其中 $u = \hat{F}_0(a)$ 与 (4.5) 中一致. 给定 p, q 和 $|\alpha|$ 时, 存在一个 $u \in [0, 1]$ 使得 (4.13) 取得最小值, 假设这样的 $u = f(p, q, |\alpha|)$, 即 a 是 $\hat{F}_0$ 的 $f(p, q, |\alpha|)$ 分位数.

因此, 对 a 的 Bayes 估计值为

$$\hat{a} = F_0 \text{ 的第 } f(p, q, |\alpha|) \text{ 分位值.}$$

同理, 在给定样本 $X_1, \dots, X_n$ 的情形下,

$$\hat{a}(X_1, \dots, X_n) = \hat{F}_n(t | X_1, \dots, X_n) \text{ 的第 } f(p, q, |\alpha| + n) \text{ 分位值,}$$

其中 $\hat{F}_n(t | X_1, \dots, X_n)$ 的定义与 (4.6) 相同.

4.7 对分位值的假设检验

假定 t_q 为未知分布 P 的 q 分位值, 对假设 $t_q \leq 0$ 进行检验, a_0 和 a_1 为可选的两种策略, 损失函数为

$$\begin{aligned} L(P, a_0) &= w_0 \mathbb{I}_{(0, \infty)}(t_q), \\ L(P, a_1) &= w_1 \mathbb{I}_{(-\infty, 0]}(t_q), \end{aligned}$$

其中 $w_0 > 0$ 和 $w_1 > 0$ 为常数. 在非参数 Bayes 模型中, 假设 $P \sim \text{DP}(\alpha)$, 根据 Bayes 定理选择使得 L 期望最小的策略. 也就是说, 如果

$$P(F(0) > q) > \frac{w_0}{w_0 + w_1},$$

则选择策略 a_0 , 否则就选择策略 a_1 .

根据等式 (4.11) 可知, 在无样本的情形下, 如果

$$u\left(\frac{w_0}{w_0 + w_1}, q, |\alpha|\right) < F_0(0),$$

则选择策略 a_0 .

在给定样本 $X_1, \dots, X_n$ 的情形下, 如果

$$u\left(\frac{w_0}{w_0 + w_1}, q, |\alpha| + n\right) < r_n F_0(0) + (1 - r_n) \frac{W_n}{n},$$

则选择策略 a_0 , 其中 $W_n = \sum_{i=1}^n \mathbb{I}_{(-\infty, 0]}(X_i)$.

假定 $w_0 = w_1$, $q = \frac{1}{2}$, 则有 $u(\frac{1}{2}, \frac{1}{2}, |\alpha|) = \frac{1}{2}$. 如果

$$W_n > \frac{1}{2}n + |\alpha|\left(\frac{1}{2} - F_0(0)\right),$$

则不拒绝假设 $\text{med}(P) \leq 0$.

4.8 U- 统计量

在统计学中, U- 统计量最早由 Hoeffding^[52] 提出, 它可以用来估计很多统计参数. $X_1, X_2, \dots, X_n$ 为取值在 $\mathbb{X}$ 上服从真实概率测度为 P_0 的独立同分布的随机变量. 假定函数 $h: \mathbb{X}^m \rightarrow \mathbb{R}$ 是可测的. 定义“参数” $\theta = \theta(P_0): \mathbb{M} \rightarrow \mathbb{R}$ 如下:

$$\theta(P_0) = E_{P_0}[h(X_1, \dots, X_m)] = \int \dots \int h(x_1, \dots, x_m) P_0(dx_1) \cdots P_0(dx_m),$$

其中 $h = h(x_1, \dots, x_m)$ 是对称的, 也就是说, 假定 π 是任意的一个排列, 则有

$$h(x_1, \dots, x_m) = h(x_{\pi(1)}, \dots, h(x_{\pi(m)})).$$

总是可以假设 h 是对称的, 因为如果存在一个 h^* 不是对称的, 可以选择

$$h(x_1, \dots, x_m) = \frac{1}{m!} \sum_{\text{all } \pi} h^*(x_{\pi(1)}, \dots, h(x_{\pi(m)})).$$

例 4.1 如果 $m = 1$, $h(x_1) = x_1$, 则 $\theta_1 = \theta(P_0)$ 是 P_0 的均值. 如果 $m = 1$, $h(x_1) = x_1^2$, 则 $\theta_2 := \theta(P_0)$ 是 P_0 的二阶矩. 方差是 $\theta_3 = \theta_2 - \theta_1^2$. 真实分布 P_0 的很多参数都可以通过某个 h 来表达.

现在因为真实分布 P_0 不知道, 假定它是一个 Dirichlet 过程 $P \sim \text{Dir}(\alpha)$. 这样, 要计算的参数就成了

$$\theta(P) = \int \cdots \int h(x_1, \dots, x_m) P(dx_1) \cdots P(dx_m).$$

在未给定样本时, 对 θ 的估计为

$$\begin{aligned} \hat{\theta} &= \int \cdots \int h(x_1, \dots, x_m) \mathbb{E}[P(dx_1)] \cdots \mathbb{E}[P(dx_m)] \\ &= \int \cdots \int h(x_1, \dots, x_m) d\hat{F}_0(x_1) \cdots d\hat{F}_0(x_m), \end{aligned}$$

其中 $\hat{F}_0$ 的定义与 (4.6) 中一致.

在给定样本 $X_1, \dots, X_n$ 时, 对 θ 的估计为

$$\begin{aligned} \hat{\theta}(X_1, \dots, X_n) &= \int \cdots \int h(x_1, \dots, x_m) \mathbb{E}[P(dx_1) | X_1, \dots, X_n] \cdots \mathbb{E}[P(dx_m) | X_1, \dots, X_n] \\ &= \int \cdots \int h(x_1, \dots, x_m) d\hat{F}_n(x_1 | X_1, \dots, X_n) \cdots d\hat{F}_n(x_m | X_1, \dots, X_n), \end{aligned}$$

其中 $\hat{F}_n$ 的定义与 (4.6) 中一致. 如果将 (4.6) 代入上述表达式, 则可以得到 U- 统计量.

参考文献

- 1 Kallenberg O. Random Measures, Theory and Applications. Cham: Springer, 2017
- 2 Ferguson T S. A Bayesian analysis of some nonparametric problems. Ann Statist, 1973, 1: 209–230
- 3 Blackwell D, MacQueen J B. Ferguson distributions via Pólya schemes. Ann Statist, 1973, 1: 353–355
- 4 Blei D M, Jordan M I. Variational inference for Dirichlet process mixtures. Bayesian Anal, 2006, 1: 121–143
- 5 Wang C, Blei D M. A split-merge MCMC algorithm for the hierarchical Dirichlet process. arXiv:1201.1657, 2012
- 6 Daume H. Fast search for Dirichlet process mixture models. In: Artificial Intelligence and Statistics. 2007, 83–90, <http://users.umi.acs.umd.edu/~hal/docs/daume07astar-dp.pdf>
- 7 Wang L, Dunson D B. Fast Bayesian inference in Dirichlet process mixture models. J Comput Graph Statist, 2011, 20: 196–216
- 8 Sethuraman J. A constructive definition of Dirichlet priors. Statist Sinica, 1994, 4: 639–650
- 9 Hjort N L. Nonparametric Bayes estimators based on Beta processes in models for life history data. Ann Statist, 1990, 18: 1259–1294
- 10 Sato K. Lévy processes and infinitely divisible distributions. Translated from the 1990 Japanese original. Revised edition of the 1999 English translation. Cambridge Studies in Advanced Mathematics, vol. 68. Cambridge: Cambridge University Press, 2013
- 11 Applebaum D. Lévy Processes and Stochastic Calculus. Cambridge Studies in Advanced Mathematics, vol. 93. Cambridge: Cambridge University Press, 2004
- 12 Ghahramani Z, Griffiths T L. Infinite latent feature models and the Indian buffet process. In: Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2006, 475–482
- 13 Bernardo J M, Bayarri M J, Berger J O, et al. Bayesian nonparametric latent feature models. Bayesian Statist, 2007, 8: 1–25
- 14 Pitman J, Yor M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. Ann Probab, 1997, 25: 855–900
- 15 Feng S. The Poisson-Dirichlet Distribution and Related Topics: Models and Asymptotic Behaviors. Heidelberg: Springer, 2010

- 16 Rodríguez A, Dunson D B. Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Anal*, 2011, 6: 145–177
- 17 Antoniak C E. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann Statist*, 1974, 2: 1152–1174
- 18 Teh Y W, Jordan M I, Beal M J, et al. Hierarchical Dirichlet processes. *J Amer Statist Assoc*, 2006, 101: 1566–1581
- 19 Rodríguez A, Dunson D B, Gelfand A E. The nested Dirichlet process. *J Amer Statist Assoc*, 2008, 103: 1131–1154
- 20 MacEachern S N. Dependent Dirichlet processes. Unpublished manuscript, Department of Statistics, Ohio State University, 2000, 8: 1–40
- 21 MacEachern S N. Dependent nonparametric processes. In: *ASA Proceedings of the Section on Bayesian Statistical Science*, 1999, 1: 50–55
- 22 Bogachev V I. *Weak Convergence of Measures. Mathematical Surveys and Monographs*, vol. 234. Providence: Amer Math Soc, 2018
- 23 Li Z. *Measure-Valued Branching Markov Processes. Probability and Its Applications*. Berlin-Heidelberg: Springer-Verlag, 2011
- 24 Perkins E. Dawson-Watanabe superprocesses and measure-valued diffusions. In: *Lectures on Probability Theory and Statistics (Saint-Flour, 1999)*. *Lecture Notes in Mathematics*, vol. 1781. Berlin: Springer, 2002, 125–324
- 25 Xiong J. *Three Classes of Nonlinear Stochastic Partial Differential Equations*. Hackensack: World Scientific, 2013
- 26 Feng S. Reversible measure-valued processes associated with the Poisson-Dirichlet distribution (in Chinese). *Sci Sin Math*, 2019, 49: 377–388 [冯水. Poisson-Dirichlet 分布及可逆测度值过程. *中国科学: 数学*, 2019, 49: 377–388]
- 27 Ishwaran H, James L F. Gibbs sampling methods for stick-breaking priors. *J Amer Statist Assoc*, 2001, 96: 161–173
- 28 Muliere P, Tardella L. Approximating distributions of random functionals of Ferguson-Dirichlet priors. *Canad J Statist*, 1998, 26: 283–297
- 29 Ghosal S, Vaart V A. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge: Cambridge University Press, 2017
- 30 Blackwell D. Discreteness of Ferguson selections. *Ann Statist*, 1973, 1: 356–358
- 31 Aldous D. *Exchangeability and Related Topics*. Berlin: Springer, 1985
- 32 Schervish M J. *Theory of Statistics*. Pittsburgh: Springer, 1995
- 33 Chen S N, Lichtin N N, Stein G. Fuel formation from aqueous ferric bromide by photolysis in the visible. *Science*, 1975, 190: 879–880
- 34 Ewens W J. The sampling theory of selectively neutral alleles. *Theor Popul Biol*, 1972, 3: 87–112
- 35 Sethuraman J, Tiwari R C. Convergence of Dirichlet measures and the interpretation of their parameter. In: *Statistical Decision Theory and Related Topics III*. New York: Academic Press, 1982, 305–315
- 36 Korwar R M, Hollander M. Contributions to the theory of Dirichlet processes. *Ann Probab*, 1973, 1: 705–711
- 37 Doss H, Sellke T. The tails of probabilities chosen from a Dirichlet prior. *Ann Statist*, 1982, 10: 1302–1305
- 38 Prünster I, Lijoi A, Regazzini E. Distributional results for means of normalized random measures with independent increments. *Ann Statist*, 2003, 31: 560–585
- 39 James L F, Lijoi A, Prünster I. Posterior analysis for normalized random measures with independent increments. *Scand J Statist*, 2009, 36: 76–97
- 40 Dunson D B, Park J H. Kernel stick-breaking processes. *Biometrika*, 2008, 95: 307–323
- 41 Ren L, Du L, Carin L, et al. Logistic stick-breaking process. *J Mach Learn Res*, 2011, 12: 203–239
- 42 Rodríguez A, Dunson D B, Gelfand A E. Latent stick-breaking processes. *J Amer Statist Assoc*, 2010, 105: 647–659
- 43 Pitman J. Some developments of the Blackwell-MacQueen urn scheme. *Statistics, Probability, and Game Theory: Papers in Honor of David Blackwell*, vol. 30. Hayward: Institute of Mathematical Statistics, 1996, 245–267
- 44 Pitman J. Exchangeable and partially exchangeable random partitions. *Probab Theory Related Fields*, 1995, 102: 145–158
- 45 Çinlar E. *Probability and Stochastics. Graduate Texts in Mathematics*, vol. 261. New York: Springer-Verlag, 2011
- 46 Paisley J W, Zaas A K, Woods C W, et al. A stick-breaking construction of the Beta process. In: *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*. 2010, 847–854, <http://www.columbia.edu/~jwp2128/Papers/Paisleyetal2010.pdf>
- 47 Broderick T, Jordan M I, Pitman J. Beta processes, stick-breaking and power laws. *Bayesian Anal*, 2012, 7: 439–476
- 48 Thibaux R, Jordan M I. Hierarchical Beta processes and the Indian buffet process. In: *Artificial Intelligence and Statistics*. 2007, 564–571, <https://people.eecs.berkeley.edu/~jordan/papers/thibaux-jordan-aistats07.pdf>
- 49 Griffiths T L, Ghahramani Z. The Indian buffet process: An introduction and review. *J Mach Learn Res*, 2011, 12: 1185–1224

- 50 Brown G W. On small-sample estimation. Ann of Math Stud, 1947, 18: 582–585
- 51 Fraser D A S, Guttman I. Tolerance regions. Ann of Math Stat, 1956, 27: 162–179
- 52 Hoeffding W. A class of statistics with asymptotically normal distribution. Ann Math Statist, 1948, 19: 293–325

Dirichlet process and Bayesian nonparametric models

Junxi Zhang & Yaozhong Hu

Abstract Bayesian nonparametric models have been extensively developed and widely used in statistics, machine learning and other areas since the ground breaking work of Ferguson. The fundamental of Bayesian nonparametric models is a special class of random probability measures: Dirichlet processes. This paper introduces the constructions, properties and some recent developments of the Dirichlet processes as well as their applications to Bayesian nonparametric estimation problems. We are also concerned with two-parameter Poisson-Dirichlet processes, Beta processes and more general stick-breaking processes and their properties.

Keywords Dirichlet process, stick-breaking process, two-parameter Poisson-Dirichlet process, Poisson process, Chinese restaurant process, Indian buffet process, Bayesian nonparametric model, U-statistics

MSC(2020) 60G57, 62F15, 62G05

doi: 10.1360/SSM-2020-0097