ノート：ディリクレ分布の $\boldsymbol{\alpha}$ のサンプリング *

LDA のハイパーパラメータはトピック分布およびトピック–単語分布を生成するディリクレ分布のハイパーパラメータ $\boldsymbol{\alpha}$, η ですが, これらはどうやって決めたらいいのでしょうか.

一般的にはトピック数を K として, $\boldsymbol{\alpha}=(50/K,\cdots,50/K)$, $\eta=0.01$ のようにヒューリスティックに決められることが多いのですが, これらのハイパーパラメータの影響を詳しく調べた研究[145]によると, これらもデータから学習した方がよいモデルとなることが示されています. とくに, 非対称な $\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_K)$ を推定することで, α_k が大きいトピックに「が」「の」, 英語なら "of", "the" といった機能語 *28 が集まり, 他のトピックにこれらが混ざることが少なくなります. また, トピック数を増やしても不要なトピックは α_k が小さく, 自然に使われなくなることも報告されています.

$\boldsymbol{\alpha}$ と η についての尤度関数は式(5.78)ですから, 数学的にはこれは, このポリア分布のハイパーパラメータを推定することと等価です. ただし, このカウント $n(i,k)$, $m(k,v)$ はあくまで学習中の $\mathbf{z}_1^N$ から計算される値ですから, 式(3.61)で学習中に $\boldsymbol{\alpha}$ を最適化してしまうと, 局所解に陥る危険があります.*29 よって, きちんと $\boldsymbol{\alpha}$ のベイズ推定を行うのが正しい方法でしょう.

ポリア分布の式

$$p(\mathbf{n}|\boldsymbol{\alpha}) = \underbrace{\frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N+\sum_k \alpha_k)}}_{\text{(A)}} \underbrace{\prod_{k=1}^{K} \frac{\Gamma(\alpha_k+n_k)}{\Gamma(\alpha_k)}}_{\text{(B)}} \tag{5.79}$$

は α_k がガンマ関数 $\Gamma()$ の中に入っているため, そのままではサンプリングできる形になりません. しかし, 巧妙な補助変数 θ, x を導入すると, ガンマ事後分布からサンプリングすることができます[60, Appendix C].*30

*28 もしこうした機能語を「ストップワード」として除いても, 236 ページで説明したように, 同様に意味の薄い語が必ず出現し, 機能語とそれ以外は簡単に二分することができません.

*29 すなわち, 学習途中で $\boldsymbol{\alpha}$ を最適化してしまうと, 学習アルゴリズム全体が Gibbs サンプリングではなく, 245 ページのモンテカルロ EM アルゴリズムを行うことになってしまいます.

(A) ディリクレ分布で 2 次元の場合を考えると, **ベータ関数**

$$B(a,b)=\int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{5.80}$$

が得られます. よって,

$$B(\textstyle\sum_k \alpha_k, N)=\dfrac{\Gamma(\sum_k \alpha_k)\Gamma(N)}{\Gamma(\sum_k \alpha_k+N)} \tag{5.81}$$

ですから, (A) は補助変数 θ を積分消去した形で,

(5.82)
$$\frac{\Gamma(\sum_k \alpha_k)}{\Gamma(N+\sum_k \alpha_k)}=\frac{B(\sum_k \alpha_k, N)}{\Gamma(N)}=\frac{1}{\Gamma(N)}\int_0^1 \theta^{\sum_k \alpha_k-1}(1-\theta)^{N-1}d\theta$$

と表すことができます.

(B) 110 ページのコラムでみたように, Pochhammer 関数 $\Gamma(\alpha+n)/\Gamma(\alpha)$ は

$$\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}=\alpha(\alpha+1)\cdots(\alpha+n-1)=\prod_{i=0}^{n-1}(\alpha+i) \tag{5.83}$$

と表すことができますが, これも $\{0,1\}$ の補助変数 x_{ki} を周辺化した形で

$$\frac{\Gamma(\alpha_k+n_k)}{\Gamma(\alpha_k)}=\prod_{i=0}^{n_k-1}(\alpha_k+i)=\prod_{i=0}^{n_k-1}\sum_{x_{ki}\in\{0,1\}}\alpha_k^{x_{ki}}i^{1-x_{ki}} \tag{5.84}$$

と書くことができることに注意しましょう.
よって, α_k がガンマ事前分布

$$p(\alpha_k)=\mathrm{Ga}(a,b)=\frac{b^a}{\Gamma(a)}\alpha_k^{a-1}e^{-b\alpha_k} \tag{5.85}$$

に従っているとすると, 式(5.82), 式(5.84)より, α_k の事後分布は, ベイズの定理から

∗30 これは MCMC 法の基本テクニックの一つで, **補助変数法**とよばれています[146].

$$(5.86)\quad p(\boldsymbol{\alpha}|\mathbf{n}) \propto p(\mathbf{n}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}) = \prod_{k=1}^{K} \frac{b^a}{\Gamma(a)} \alpha_k^{a-1} e^{-b\alpha_k} \times \frac{1}{\Gamma(N)} \int_0^1 \theta^{\sum_k \alpha_k - 1} (1-\theta)^{N-1} d\theta \times \prod_{k=1}^{K} \left(\prod_{i=0}^{n_k-1} \sum_{x_{ki} \in \{0,1\}} \alpha_k^{x_{ki}} i^{1-x_{ki}} \right)$$

となります. これは θ と x_{ki} を周辺化した形ですから, 同時確率は

$$(5.87)\quad p(\boldsymbol{\alpha}, \theta, x|\mathbf{n}) \propto \prod_{k=1}^{K} \frac{b^a}{\Gamma(a)} \alpha_k^{a-1} e^{-b\alpha_k} \times \frac{1}{\Gamma(N)} \theta^{\sum_k \alpha_k - 1} (1-\theta)^{N-1} \times \prod_{k=1}^{K} \left(\prod_{i=0}^{n_k-1} \alpha_k^{x_{ki}} i^{1-x_{ki}} \right)$$

です. よって各 α_k の事後分布は, 式(5.87)から α_k に関する項を抜き出すと, 補助変数 θ, x_{ki} が与えられた下では

$$(5.88)\quad \begin{aligned} p(\alpha_k|\mathbf{n}, \theta, x) &\propto p(\alpha_k, \theta, x|\mathbf{n}) = \alpha_k^{a-1} \cdot e^{-b\alpha_k} \cdot \theta^{\alpha_k} \cdot \prod_{i=0}^{n_k-1} \alpha_k^{x_{ki}} \\ &= \alpha_k^{a+\sum_{i=0}^{n_k-1} x_{ki} - 1} e^{-(b-\log\theta)\alpha_k} \\ &\sim \mathrm{Ga}(a + \textstyle\sum_{i=0}^{n_k-1} x_{ki},\ b - \log\theta) \end{aligned}$$

となることがわかります. 同様にして補助変数 θ, x_{ki} についても対応する項を抜き出すと, その分布は

$$(5.89)\quad p(\theta|\mathbf{n}, \boldsymbol{\alpha}, y) \sim \mathrm{Be}(\textstyle\sum_k \alpha_k, N)$$

$$(5.90)\quad \begin{aligned} p(x_{ki}|\mathbf{n}, \boldsymbol{\alpha}, \theta) &\propto \alpha_k^{x_{ki}}\, i^{1-x_{ki}} \propto \left(\frac{\alpha_k}{\alpha_k + i}\right)^{x_{ki}} \left(\frac{i}{\alpha_k + i}\right)^{1-x_{ki}} \\ &\sim \text{Bernoulli}\left(\frac{\alpha_k}{\alpha_k + i}\right) \end{aligned}$$

となります. この分布から補助変数 θ, x_{ki} をサンプリングしてから, 式(5.88)を使って α_k をサンプリングします. 実際には, 式(5.79)の $p(\mathbf{n}|\boldsymbol{\alpha})$ は N 個の文書について存在しますから, 同様に計算すると, N_i を文書 i の長さとして

$$(5.91)\quad \begin{cases} p(\theta_i|n,\boldsymbol{\alpha},x) & \sim \mathrm{Be}\left(\sum_k \alpha_k,\, N_i\right) \\ p(x_{ikj}|n,\boldsymbol{\alpha},\theta) & \sim \mathrm{Bernoulli}\left(\dfrac{\alpha_k}{\alpha_k+j}\right) \\ p(\alpha_k|n,\theta,x) & \sim \mathrm{Ga}\left(a+\sum_{i=1}^{N}\sum_{j=0}^{n(i,k)-1} x_{ikj},\, b-\sum_{i=1}^{N}\log\theta_i\right) \end{cases}$$

で, $\boldsymbol{\alpha}$ をガンマ事後分布からサンプリングすることができます. □

η についても, 式(5.78)から同様にして計算すれば, 補助変数 ϕ_k, y_{kvj} を導入することで

$$(5.92)\quad \begin{cases} p(\phi_k|m,\eta,y) & \sim \mathrm{Be}\left(V\eta,\, \sum_{v=1}^{V} m(k,v)\right) \\ p(y_{kvj}|m,\eta,\phi) & \sim \mathrm{Bernoulli}\left(\dfrac{\eta}{\eta+j}\right) \\ p(\eta|m,\phi,y) & \sim \mathrm{Ga}\left(a+\sum_{k=1}^{K}\sum_{v=1}^{V}\sum_{j=0}^{m(k,v)-1} y_{kvj},\right. \\ & \qquad \left. b-\sum_{k=1}^{K}\log\phi_k\right) \end{cases}$$

で, ガンマ事後分布からサンプリングを行えることがわかります. □