多項分布から Dirichlet 過程まで
Version 0.03

村脇 有吾

最終更新: 2012 年 10 月 29 日

今日の内容

- 正規分布などは無視して多項分布一本に絞り込む
- かなり基本的なところから始める
- 図をほとんど書いていないので、適宜板書するかも
- Bayesian で、主に教師なし学習で使う話をやる
- Dirichlet 過程の説明は端折る
- inference の説明は省略するので、実装できるまでは至らない
- 実装したい人は Graham さんのノンパラメトリックベイズ入門を参照のこと

多項分布 (Multinomial distribution)

確率変数 x は K 個のうちいずれかの値をとる。その確率は、例えば以下に従う。

$$\boldsymbol{\theta} = \quad (0.1, 0.6, 0.3)$$
$$\text{赤 黄 緑}$$

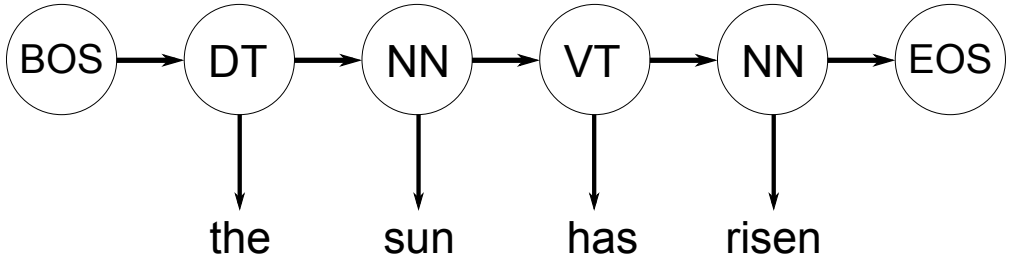
多くのタスク設定では、真の $\boldsymbol{\theta}$ は不明であり、観測済みのデータ $\boldsymbol{x} = (x_1, \cdots, x_N)$ から推定する。推定された $\boldsymbol{\theta}$ の主な用途は、新たな x' の確率を予測することである。

Bayes 推定では、$\boldsymbol{\theta}$ を積分消去してしまい、$\boldsymbol{x}$ から直接 x' の確率を予測する。

自然言語処理は多項分布の天下

- N-gram language model (V 次元の語彙から一つの単語を生成)
- HMM POS tagging
- Probabilistic context-free grammar (PCFG)
- Topic model

例: Hidden Markov Model による POS tagging



品詞数を K、単語数を V とする (簡単のため BOS と EOS は無視)。

遷移確率 (transition probability) は K 個の多項分布 $(\boldsymbol{\theta}_{\text{DT}}^{\text{TRANS}}, \boldsymbol{\theta}_{\text{NN}}^{\text{TRANS}}, \cdots)$ からなり、各多項分布は K 値。

$$\boldsymbol{\theta}_{\text{DT}}^{\text{TRANS}} = (0.21, 0.27, 0.09, \cdots)$$

NN NNS ADJ

出力確率 (emission probability) は K 個の多項分布 $(\boldsymbol{\theta}_{\text{DT}}^{\text{EMIT}}, \boldsymbol{\theta}_{\text{NN}}^{\text{EMIT}}, \cdots)$ からなり、各多項分布は V 値。

$$\boldsymbol{\theta}_{\text{NN}}^{\text{EMIT}} = (0.012, 0.002, 0.005, \cdots)$$

sun rose cat

Outline

- 多項分布を導入
- パラメータの最尤推定
- 事前確率の導入と MAP 推定
- Bayes 推定 (← 結局使うのはこれ!)

多項分布: 1 回の観測の場合

$\boldsymbol{\theta}$ が既知とする。

確率変数 $x \in \{1, \cdots, K\}$ は以下に従う。

$$p(x = k|\boldsymbol{\theta}) = \theta_k$$

ここで $0 \leq \theta_k \leq 1$, $\sum_{k=1}^{K} \theta_k = 1$。

例えば、$\boldsymbol{\theta} = (0.1, 0.6, 0.3)$ について、$p(x = 2|\boldsymbol{\theta}) = \theta_2 = 0.6$。

多項分布: N 回の観測の場合

N 回の観測 $\boldsymbol{x} = (x_1, \cdots x_N)$ を得る確率は以下に従う。

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \theta_{x_i} = \prod_{k=1}^{K} \theta_k^{n_k}$$

ここで n_k は $\boldsymbol{x}$ における値 k の観測数 ($\sum_{k=1}^{K} n_k = N$)。

$p(\boldsymbol{x}|\boldsymbol{\theta})$ は $\boldsymbol{x}$ 内の順序に依存しない。N 個の観測の総和のみに依存している (十分統計量)。

多項分布: 観測数の確率

$\boldsymbol{x} = (x_1, \cdots, x_N)$ の代わりに、観測数 $\boldsymbol{n} = (n_1, \cdots, n_K)$ を得る確率を考えると、

$$
\begin{aligned}
\mathrm{Multi}(\boldsymbol{n}|\boldsymbol{\theta}) &= \binom{N}{n_1 \cdots n_k} \prod_{k=1}^{K} \theta_k^{n_k} \\
&= \frac{N!}{n_1! \cdots n_k!} \prod_{k=1}^{K} \theta_k^{n_k} \\
&= \frac{\Gamma(N+1)}{\prod_{k=1}^{K} \Gamma(n_k+1)} \prod_{k=1}^{K} \theta_k^{n_k}
\end{aligned}
$$

となる。ここで $\Gamma(n) = (n-1)!$。$\boldsymbol{n}$ となるような $\boldsymbol{x}$ の組み合わせの数を $p(\boldsymbol{x}|\boldsymbol{\theta})$ にかけている。

自然言語処理の文献では $\boldsymbol{n}$ ではなく $\boldsymbol{x}$ の分布をを多項分布とよぶ場合が多く、注意が必要 (Minka, 2003)。$\boldsymbol{x}$ の分布を Categorical distribution とよぶ流儀もある。

N 回の観測から θ を推定

$\boldsymbol{\theta}$ が未知のとき、観測データ $\boldsymbol{n}$ から推定する。

まず尤度 (likelihood) 関数を以下のように定義。

$$L(\boldsymbol{\theta}; \boldsymbol{n}) = \mathrm{Multi}(\boldsymbol{n}|\boldsymbol{\theta})$$

以下を満たす $\boldsymbol{\theta}^{\mathrm{ML}}$ を求める。これを最尤推定 (maximum likelihood estimation) とよぶ。

$$\boldsymbol{\theta}^{\mathrm{ML}} = \operatorname*{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \boldsymbol{n}) = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathrm{Multi}(\boldsymbol{n}|\boldsymbol{\theta})$$

これを解くと以下が得られる。

$$\theta_k^{\mathrm{ML}} = \frac{n_k}{N}$$

これは Lagrange の未定乗数法を用いれば求まる (高村本 1.5.2 参照)。

最尤推定の欠点

低頻度のパラメータが信用出来ない。特に、観測回数が 0 のパラメータが出現すると、予測確率が 0 になってしまい都合が悪い。

観測データを大きくすれば大丈夫か? No. 自然言語処理で用いられる多項分布は一般に sparse (参考: Zipf の法則)。常に低頻度のパラメータが存在する (Johnson, 2007)。

Bayes の定理

事前分布 $p(\theta)$ を導入する。これはデータを観測する前に (恣意的に) 仮定する θ の分布。

Bayes の定理により

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta)p(\theta)}{p(x)} \\ &\propto p(x|\theta)p(\theta) \end{aligned}$$

$p(\theta|x)$ を事後分布 (posterior distribution)、$p(x|\theta)$ を尤度、$p(\theta)$ を事前分布 (prior distribution) とよぶ。$p(\theta|x)$ は、データを観測した後の θ の分布と解釈できる。

多項分布の事前分布

多項分布に対しては、事前分布として Dirichlet 分布 $\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ を仮定すると計算上都合がよい。

$$p(\boldsymbol{\theta}|\boldsymbol{n},\boldsymbol{\alpha}) \propto \mathrm{Multi}(\boldsymbol{n}|\boldsymbol{\theta})\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$$

ここで

$$\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(A)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)}\prod_{k=1}^{K}\theta_k^{\alpha_k-1}$$

ここで $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_K)$, $\alpha_k > 0$, $A = \sum_{k=1}^{K}\alpha_k$。特に理由がなければ $\alpha_i = \alpha_j$ とする。$\boldsymbol{\alpha}$ は事前に与えるパラメータであり、パラメータのパラメータであるためハイパーパラメータとよばれる。

n_k が値 k の観測上のカウントなのに対して、α_k は事前に仮定されたカウント (擬似カウント) と解釈できる。ただし、α_k は整数でなくてもよい。

事後分布

Dirichlet 分布と多項分布 (再掲) の式の類似に注意。

$$\mathrm{Multi}(\boldsymbol{n}|\boldsymbol{\theta}) = \frac{\Gamma(N+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)}\prod_{k=1}^{K}\theta_k^{n_k}$$

事後分布 $p(\boldsymbol{\theta}|\boldsymbol{n},\boldsymbol{\alpha}) \propto \mathrm{Multi}(\boldsymbol{n}|\boldsymbol{\theta})\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ を求める (計算省略)。

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{n},\boldsymbol{\alpha}) &= \frac{\Gamma(N+A)}{\Gamma(n_1+\alpha_1)\cdots\Gamma(n_K+\alpha_K)}\prod_{k=1}^{K}\theta_k^{n_k+\alpha_k-1} \\ &= \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{n}+\boldsymbol{\alpha}) \end{aligned}$$

事後分布は事前分布と同じ形になる。この性質を共役性 (conjugacy) とよぶ。

MAP (最大事後確率) 推定

事後確率を最大化するような $\boldsymbol{\theta}$ の推定を最大事後確率 (maximum a posteriori) 推定とよぶ。

$$\begin{aligned}\boldsymbol{\theta}^{\mathrm{MAP}} &= \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{n}) \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} p(\boldsymbol{n}|\boldsymbol{\theta})p(\boldsymbol{\theta})\end{aligned}$$

Dirichlet-多項分布の場合は以下のようになる。

$$\boldsymbol{\theta}^{\mathrm{MAP}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{n}+\boldsymbol{\alpha})$$

$$\theta_k^{\mathrm{MAP}} = \frac{n_k + \alpha_k - 1}{\sum_{k=1}^{K}(n_k + \alpha_k - 1)} = \frac{n_k + \alpha_k - 1}{N + A - K}$$

最尤推定 $\theta_k^{\mathrm{ML}} = \frac{n_k}{N}$ と比較すると、加算スムージングが行われていることが確認できる。

Bayes 推定

Bayes 推定では、MAP 推定のように θ を一つに決めるのではなく、θ の分布を保持したまま、積分消去して利用する。

$$p(\boldsymbol{x}) = \int_{\theta} p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

新たな x' の生成確率 (予測確率) を以下で求める。

$$\begin{aligned} p(x'|\boldsymbol{x}) &= \int_{\theta} p(x'|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta} \\ &= \int_{\theta} p(x'|\boldsymbol{\theta}) \frac{p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{x})}\, d\boldsymbol{\theta} \end{aligned}$$

一般に積分は困難だが、Dirichlet-多項分布の場合は解析的に解ける。

Dirichlet-多項分布の積分消去

$$
\begin{aligned}
p(\boldsymbol{n}|\boldsymbol{\alpha}) &= \int_{\boldsymbol{\theta}} \mathrm{Multi}(\boldsymbol{n}|\boldsymbol{\theta})\mathrm{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})\, d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \left(\frac{\Gamma(N+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)} \prod_{k=1}^{K} \theta_k^{n_k} \right) \left(\frac{\Gamma(A)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \theta_k^{\alpha_k-1} \right) d\boldsymbol{\theta} \\
&= \frac{\Gamma(N+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)} \frac{\Gamma(A)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \int_{\boldsymbol{\theta}} \prod_{k=1}^{K} \theta_k^{n_k+\alpha_k-1}\, d\boldsymbol{\theta} \\
&= \frac{\Gamma(N+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)} \frac{\Gamma(A)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(n_k+\alpha_k)}{\Gamma(N+A)} \\
&= \frac{\Gamma(N+1)}{\prod_{k=1}^{K}\Gamma(n_k+1)} \frac{\Gamma(A)}{\Gamma(N+A)} \prod_{k=1}^{K} \frac{\Gamma(n_k+\alpha_k)}{\Gamma(\alpha_k)}
\end{aligned}
$$

($\boldsymbol{n}$ ではなく) $\boldsymbol{x}$ の生成確率

$p(\boldsymbol{n}|\boldsymbol{\alpha})$ の最初の項は組み合わせ数。これを除くと $\boldsymbol{x}$ の生成確率となる。

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{\Gamma(A)}{\Gamma(N+A)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)}$$

なお、x の生成過程を以下のように表記する。

$$\begin{aligned} x_n|\boldsymbol{\theta} &\sim \boldsymbol{\theta} \\ \boldsymbol{\theta}|\boldsymbol{\alpha} &\sim \mathrm{Dir}(\boldsymbol{\alpha}) \end{aligned}$$

$x_n|\boldsymbol{\theta} \sim \mathrm{Multi}(\boldsymbol{\theta})$ と表記することもある。

予測確率

x' の予測確率は以下で与えられる ($\boldsymbol{\theta}^{\mathrm{MAP}}$ と比較せよ)。

$$
\begin{aligned}
p(x' = k' | \boldsymbol{x}, \boldsymbol{\alpha}) &= \frac{p(x' = k', \boldsymbol{x} | \boldsymbol{\alpha})}{p(\boldsymbol{x} | \boldsymbol{\alpha})} \\
&= \frac{\frac{\Gamma(A)}{\Gamma(N+1+A)} \prod_{k=1}^{K} \frac{\Gamma(n_k + I(k=k') + \alpha_k)}{\Gamma(\alpha_k)}}{\frac{\Gamma(A)}{\Gamma(N+A)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)}} \\
&= \frac{\Gamma(N+A)}{\Gamma(N+1+A)} \frac{\Gamma(n_{k'} + 1 + \alpha_{k'})}{\Gamma(n_{k'} + \alpha_{k'})} \\
&= \frac{n_{k'} + \alpha_{k'}}{N + A}
\end{aligned}
$$

ここで I(statement) は statement が真なら 1、そうでなければ 0。

x の逐次更新

$$
\begin{aligned}
p(\boldsymbol{x}_{1:N}|\boldsymbol{\alpha}) &= p(x_N|\boldsymbol{x}_{1:N-1},\boldsymbol{\alpha})p(\boldsymbol{x}_{1:N-1}|\boldsymbol{\alpha}) \\
&= \frac{n_{x_N}(\boldsymbol{x}_{1:N-1})+\alpha_{x_N}}{N-1+A}p(\boldsymbol{x}_{1:N-1}|\boldsymbol{\alpha}) \\
&= \frac{n_{x_N}(\boldsymbol{x}_{1:N-1})+\alpha_{x_N}}{N-1+A}p(x_{N-1}|\boldsymbol{x}_{1:N-2},\boldsymbol{\alpha})p(\boldsymbol{x}_{1:N-2}|\boldsymbol{\alpha}) \\
&= \frac{n_{x_N}(\boldsymbol{x}_{1:N-1})+\alpha_{x_N}}{N-1+A}\frac{n_{x_{N-1}}(\boldsymbol{x}_{1:N-2})+\alpha_{x_{N-1}}}{N-2+A}p(\boldsymbol{x}_{1:N-2}|\boldsymbol{\alpha}) \\
&= \frac{\prod_{k=1}^{K}\prod_{i=1}^{n_k(\boldsymbol{x}_{1:N})} i-1+\alpha_k}{\prod_{i=1}^{N} i-1+A} \\
&= \frac{\Gamma(A)}{\Gamma(N+A)}\prod_{k=1}^{K}\frac{\Gamma(n_k(\boldsymbol{x}_{1:N})+\alpha_k)}{\Gamma(\alpha_k)}
\end{aligned}
$$

ここで $\boldsymbol{x}_{1:N}=(x_1,\cdots,x_N)$, $n_k(\boldsymbol{x})=$ number of times k appears in $\boldsymbol{x}$。

交換可能性 (Exchangeability)

$p(\boldsymbol{x}|\alpha)$ は $\boldsymbol{x}$ 内の順序に依存せず、N 個の観測の総和のみに依存している。つまり、$\boldsymbol{x}$ 内の任意の要素ペア x_i, x_j を置換して作った $\boldsymbol{x}'$ を考えた時、

$$p(\boldsymbol{x}'|\alpha) = p(\boldsymbol{x}|\alpha)$$

がなりたつ。この性質を交換可能性という。交換可能性のおかげで inference が簡単に実現できるのだが、それを説明するにはまずサンプリングを説明しないといけない。

x の逐次更新: 例

赤、黄、緑の 3 種類のいずれかの値を取る x を考える。いま (赤, 赤, 黄) を観測したとき、この 3 個の生成確率は以下で求まる。

$$p(\boldsymbol{x}|\alpha) = \frac{0+\alpha_{\text{赤}}}{0+A}\,\frac{1+\alpha_{\text{赤}}}{1+A}\,\frac{0+\alpha_{\text{黄}}}{2+A}$$

この結果は観測する順序を変えても変わらない。

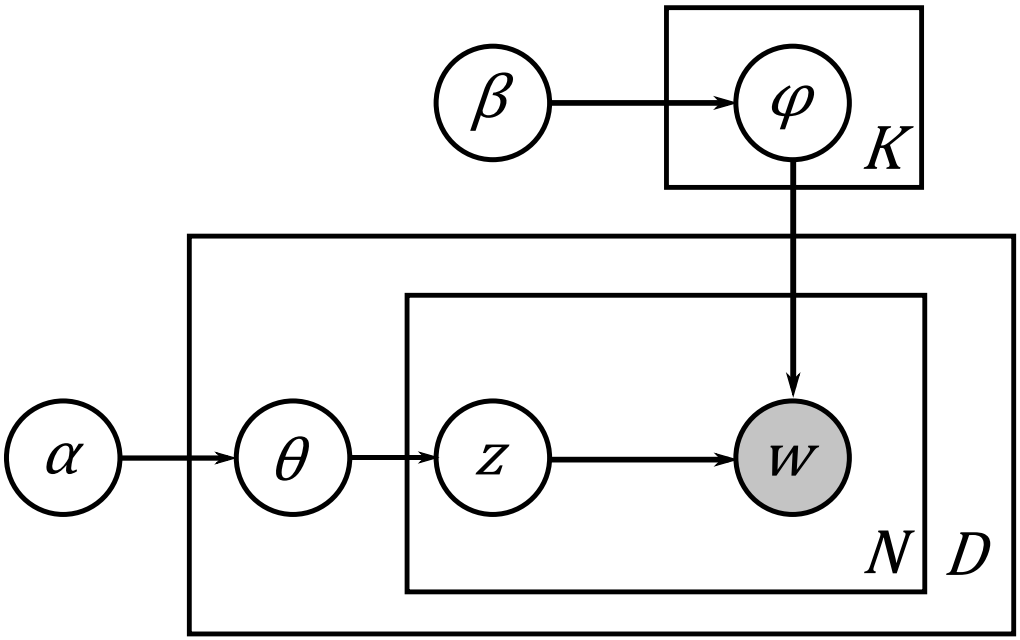
例題: LDA (Latent Dirichlet allocation)

(Blei+, 2002 のオリジナルではなく、Griffiths+, 2004 のモデル)

D 個の文書が与えられ、文書 i について N_i 個の単語が観測されるとする。LDA では、各文書は K 個のトピックの混合であると仮定する。単語 $w_{i,j}$ はトピック k に属しており、文書 i は K 個のトピックの混合分布 θ_i を持つ。トピック k は、それに対応する単語分布 φ_k を持っており、$w_{i,j}$ は単語分布 φ_k から生成されたものである。

例題: LDA (Latent Dirichlet allocation)

生成過程は有向グラフで表現できる。



プレートは繰り返し、色付きのノードは観測変数を表す。
α と β のような所与のパラメータは黒点で表す流儀もある (その方が良さそう)。

例題: LDA (Latent Dirichlet allocation)

以下の生成過程で表される。

1. 各トピック $k \in \{1, \cdots, K\}$ について、単語分布 $\boldsymbol{\varphi}_k \sim \mathrm{Dir}(\boldsymbol{\beta})$ を生成
2. 文書 $i \in \{1, \cdots, D\}$ について、トピック分布 $\boldsymbol{\theta}_i \sim \mathrm{Dir}(\boldsymbol{\alpha})$ を生成
3. 文書 i 内の単語 $j \in \{1, \cdots, N_i\}$ について
    1. トピック $z_{i,j} \sim \mathrm{Multi}(\boldsymbol{\theta}_i)$ を生成
    2. 単語 $w_{i,j} \sim \mathrm{Multi}(\boldsymbol{\varphi}_{z_{i,j}})$ を生成

以後、コーパス中の単語列を $\boldsymbol{W}$、対応するトピック列を $\boldsymbol{Z}$ で表す。いま $\boldsymbol{W}$ が観測されたとき、$\boldsymbol{Z}$ を推定したい。

例題: LDA (Latent Dirichlet allocation)

まず同時分布を考える。

$$p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \alpha, \beta) = \prod_{k=1}^{K} p(\boldsymbol{\varphi_k} | \beta) \times \prod_{i=1}^{D} \left(p(\boldsymbol{\theta_i} | \alpha) \prod_{j=1}^{N_i} p(z_{i,j} | \boldsymbol{\theta_i}) p(w_{i,j} | \boldsymbol{\varphi}_{\boldsymbol{z_{i,j}}}) \right)$$

$\boldsymbol{\theta}_i$ と $\boldsymbol{\varphi}_k$ は積分消去できる。

$$\begin{aligned} p(\boldsymbol{W}, \boldsymbol{Z} | \alpha, \beta) &= \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\varphi}} p(\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi} | \alpha, \beta) \, d\boldsymbol{\theta} d\boldsymbol{\varphi} \\ &= \prod_{i=1}^{D} \left(\frac{\Gamma(\alpha_{(\cdot)})}{\Gamma(n_{i,(\cdot)}^{(\cdot)} + \alpha_{(\cdot)})} \prod_{k=1}^{K} \frac{\Gamma(n_{i,(\cdot)}^{k} + \alpha_k)}{\Gamma(\alpha_k)} \right) \\ &\quad \times \prod_{k=1}^{K} \left(\frac{\Gamma(\beta_{(\cdot)})}{\Gamma(n_{(\cdot),(\cdot)}^{k} + \beta_{(\cdot)})} \prod_{v} \frac{\Gamma(n_{(\cdot),v}^{k} + \beta_v)}{\Gamma(\beta_v)} \right) \end{aligned}$$

ここで $n_{i,v}^{k}$ は文書 i における単語 v がトピック k を取る回数。
$D + K$ 個の Dirichlet-多項分布の組み合わせであることが確認できる。

例題: LDA (Latent Dirichlet allocation)

文書からトピックの、トピックからの単語の逐次生成を考える。
交換可能性が成り立つことも容易に確かめられる。

文書 ID	トピック	単語	生成確率
1	政治	オバマ	$\frac{0+\alpha_{\text{政治}}}{0+\alpha_{(\cdot)}} \times \frac{0+\beta_{\text{オバマ}}}{0+\beta_{(\cdot)}}$
	経済	FRB	$\frac{0+\alpha_{\text{経済}}}{1+\alpha_{(\cdot)}} \times \frac{0+\beta_{\text{FRB}}}{0+\beta_{(\cdot)}}$
	政治	選挙	$\frac{1+\alpha_{\text{政治}}}{2+\alpha_{(\cdot)}} \times \frac{0+\beta_{\text{選挙}}}{1+\beta_{(\cdot)}}$
	政治	オバマ	$\frac{2+\alpha_{\text{政治}}}{3+\alpha_{(\cdot)}} \times \frac{1+\beta_{\text{オバマ}}}{2+\beta_{(\cdot)}}$
2	経済	バーナンキ	$\frac{0+\alpha_{\text{経済}}}{0+\alpha_{(\cdot)}} \times \frac{0+\beta_{\text{バーナンキ}}}{1+\beta_{(\cdot)}}$
	経済	FRB	$\frac{1+\alpha_{\text{経済}}}{1+\alpha_{(\cdot)}} \times \frac{1+\beta_{\text{FRB}}}{2+\beta_{(\cdot)}}$

例題: LDA (Latent Dirichlet allocation)

求めたいのは

$$\operatorname*{argmax}_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{W},\boldsymbol{\alpha},\boldsymbol{\beta}) = \operatorname*{argmax}_{\boldsymbol{Z}} p(\boldsymbol{W},\boldsymbol{Z}|\boldsymbol{\alpha},\boldsymbol{\beta})$$

探索空間は $K^{|W|}$ 通り。すべての組み合わせを調べるのは現実的でない。そこで、Gibbs サンプリングを用いた random walk による探索がよく用いられる (詳細省略)。
ただし、

1. サンプリング自体は、argmax を求めるアルゴリズムではない。
2. サンプリングが簡単に実現できるのは、交換可能性がなりたつから。

Dirichlet 過程 (Dirichlet process)

LDA のトピック数 K は事前に人手で与える。しかし、最適な K は事前にはわからず、データから自動推定したい。

そこで導入されるのが Dirichlet 過程である。Dirichlet 過程も多項分布と共役であり、事後分布も Dirichlet 過程となる。

$$\begin{aligned} x|G &\sim G \\ G|\alpha, G_0 &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

ここで G_0 は基底測度。確率分布と思えばよい。Dirichlet 分布の α_k が $\alpha G_0(k)$ に相当する。ただし、Dirichlet 過程においては、K は無限である。もっとも、有限の N 個のデータが属すトピック数 K' は有限 $(K' \leq N)$ である。よって、無限のトピック数のうち、K' 個のトピックのみを明示的に扱い、残りのトピックは遅延評価すれば手続き的に扱える。

Dirichlet 過程の予測確率

G を積分消去する (Blackwell and MacQueen, 1973)。

$$p(x' = k' | \mathbf{x}, \alpha, G_0) = \int_G p(x' = k' | G) P(G | \mathbf{x}, \alpha, G_0)\, dG$$

$$\begin{aligned} x' | \mathbf{x}, \alpha, G_0 &\sim \frac{1}{N+\alpha} \sum_{j=1}^{N} \delta(x_j) + \frac{\alpha}{N+\alpha} G_0 \\ &\sim \sum_k \frac{n_k}{N+\alpha} \delta(k) + \frac{\alpha}{N+\alpha} G_0 \end{aligned}$$

ここで $\delta(k)$ は k にすべての確率質量を置く分布。

中華料理店過程 (Chinese Restaurant Process)

Dirichlet 過程は中華料理店過程により解釈できる。無限個のテーブルを持ち、各テーブルに無限の客を収容できる料理店を考える。新たに入ってきた客は、既に客がいるテーブルに客の数に比例した確率で座る。あるいは、新たなテーブルに α に比例する確率で座る。

$$p(x_i = l_k | \boldsymbol{x}, \alpha) = \begin{cases} \frac{n_k}{N+\alpha} & 1 \leq k \leq K(\boldsymbol{x}) \\ \frac{\alpha}{N+\alpha} & k = K(\boldsymbol{x}) + 1 \end{cases}$$

ここで $K(\boldsymbol{x})$ は、$\boldsymbol{x}$ が座っているテーブルの総数。なお、最初の客は必ず新たなテーブルに座る。

トピック・モデルの場合、k はトピックのインデックスを表し、単語分布 φ_k に対応付けられている。G_0 を一様な連続分布とすれば、各テーブルがトピックに対応する。ただし、このままではトピック・モデルは機能しない。各文書の G_i 間でトピックを共有するには階層 Dirichlet 過程が必要となる (Teh et al., 2006)。以下では、より簡単な unigram 単語分布を例題とする。

2 段階生成モデル

文を単語へ分割することを考える (いわゆる単語分割)。考えられる文の分割、それに応じた単語ラベルの総数は組み合わせ爆発を起こす。そこで、任意の文字列に対して 0 より大きな確率を与える単語確率が必要となる。こうしたモデルを Dirichlet 過程を用いて実現する。
以下のように、文字列から単語を生成すれば、任意の文字列からなる単語に適当な確率を割り振れる。

$$G_0(l = c_1, \cdots, c_M) = p_\#(1 - p_\#)^{M-1} \prod_{i=1}^{M} p(c_i)$$

ここで $p_\#$ は単語境界の確率。この G_0 を基底分布とする単語分布を考える (unigram モデル)。

2 段階生成モデル

CRP の各テーブルに単語ラベル l を関連付ける必要がある。これは、G_0 による単語ラベルの生成 (generator) と、CRP による客のテーブルへの割り当て (adaptor) という 2 段階の生成によって実現される。
新たな客が来たとき以下のいずれかが行われる。

- $\frac{n_k}{N+\alpha}$ の確率で既存のテーブル k に座る。このとき、割り当て済みの単語ラベル l_k が使われる。
- $\frac{\alpha}{N+\alpha}$ の確率で新たなテーブルに座る。このとき、G_0 に従って新たなラベル $l_{K(x+1)}$ を生成し、これをテーブルに割り当てる。

一般に複数のテーブルに同じラベルが割り当てられることに注意。

2 段階生成モデル

単語列 $\boldsymbol{w}_{1:N}$ およびそのテーブル割り当て $\boldsymbol{z}_{1:N}$ が与えられたとき、予測確率は以下で与えられる。

$$
\begin{aligned}
p(w_i = l | \boldsymbol{w}_{1:N}, \boldsymbol{z}_{1:N}, \alpha, G_0) &= \sum_{k=1}^{K(\boldsymbol{z}_{1:N})} I(l_k = l) \frac{n_k(\boldsymbol{z}_{1:N})}{N + \alpha} + \frac{\alpha}{\alpha + N} G_0(l) \\
&= \frac{n_l(\boldsymbol{w}_{1:N}) + \alpha G_0(l)}{N + \alpha}
\end{aligned}
$$

ここで $n_l(\boldsymbol{w})$ は、$\boldsymbol{w}$ におけるラベル l の出現総数。つまり、実は予測確率はテーブル割り当てに関係なく

$$
p(w_i = l | \boldsymbol{w}_{1:N}, \alpha, G_0) = \frac{n_l(\boldsymbol{w}_{1:N}) + \alpha G_0(l)}{N + \alpha}
$$

となる (階層 Dirichlet 過程や、(階層) Pitman-Yor 過程を用いた場合はテーブル割り当てに依存する)。

Unigram 言語モデル

分割なしのコーパス $\boldsymbol{C}$、これに対応する単語列 $\boldsymbol{W}$ の生成確率は、Dirichlet-多項分布の生成確率と同様に以下で表される。

$$p(\boldsymbol{C}, \boldsymbol{W}|\alpha, G_0) = \frac{\Gamma(\alpha)}{\Gamma(n_{(\cdot)}(\boldsymbol{W}) + \alpha)} \prod_{l \in voc(\boldsymbol{W})} \frac{\Gamma(n_l(\boldsymbol{W}) + \alpha G_0(l))}{\Gamma(\alpha G_0(l))}$$

ここで $voc(\boldsymbol{W})$ は $\boldsymbol{W}$ に含まれる単語ラベルの集合を返す関数。

$\mathrm{argmax}_{\boldsymbol{W}}\, p(\boldsymbol{W}|\boldsymbol{C}, \alpha, G_0) = \mathrm{argmax}_{\boldsymbol{W}}\, p(\boldsymbol{C}, \boldsymbol{W}|\alpha, G_0)$ を求めたい。すべての $\boldsymbol{W}$ を調べるのは不可能なので、Gibbs サンプリングによる randomized search が用いられる。

今回触れなかった話題

- Gibbs サンプリングの理論的背景、中身 & 実装まわり
- ブロック・サンプリング、特に type-based のサンプリング
- Stick-breaking 過程による Dirichlet 過程の解釈
- 変分 Bayes (variational Bayes)
- 階層 Dirichlet 過程
- Dirichlet 過程の一般化である Pitman-Yor 過程