LDA
(latent Dirichlet allocation)

正田 備也

masada@rikkyo.ac.jp

Contents

PLSA (probabilistic latent semantic analysis)

- 同じ文書内でも、異なる単語トークンは異なる単語多項分布から生成されうる（＝異なるトピックを表現しうる）
- 文書によって、各トピックの出現確率が異なる

- PLSA では、単語多項分布をトピック (topic) と呼ぶ
- PLSA は最もシンプルなトピックモデル
  - トピックモデルは、単語トークンの “クラスタリング”
  - 同一文書内の同一単語の異なるトークンは区別されない (bag-of-words)

Notations

- 語彙集合 $\{1, \ldots, W\}$
- トピック集合 $\{1, \ldots, K\}$
  - 語彙やトピックをその添字と同一視している。
- 文書集合 $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$
- 文書 $\boldsymbol{x}_i$ の j 番目のトークンとして現れる単語を、$x_{i,j}$ という確率変数で表す
- 文書 $\boldsymbol{x}_i$ の j 番目の単語 $x_{i,j}$ が表現するトピックを、$z_{i,j}$ という確率変数で表す
- $x_{i,j}$ の値は観測されているが、$z_{i,j}$ の値は観測されていない
  - つまり、$z_{i,j}$ は潜在変数。

PLSAにおける同時分布

- PLSAでは、文書$\boldsymbol{x}_i$のj番目のトークンがトピックkを表現し、かつそのトピックを表現するために単語wが使われる同時確率、つまり$p(x_{i,j} = w, z_{i,j} = k)$は

$$p(x_{i,j} = w, z_{i,j} = k) = p(z_{i,j} = k)p(x_{i,j} = w|z_{i,j} = k) \quad (1)$$

  - $p(z_{i,j} = k)$は、文書$\boldsymbol{x}_i$のj番目のトークンが（他のトピックでなく）トピックkを表現する確率
  - $p(x_{i,j} = w|z_{i,j} = k)$は、文書$\boldsymbol{x}_i$の$j$番目のトークンがトピック$k$を表現するとき（他の単語でなく）単語$w$が使われる確率
- さらに、PLSAでは以下のように仮定する（次スライド）

PLSA において仮定すること

- どの j, j' についても $p(z_{i,j} = k) = p(z_{i,j'} = k)$ と仮定
  - 同じ文書内なら、どの単語トークンであれ、トピック k を表現する確率は、同じ（場所によってトピックの確率が違ったりしない）
  - そこで、$p(z_{i,\cdot} = k) = \theta_{i,k}$ とおく
- どの i, i' と j, j' についても、$p(x_{i,j} = w | z_{i,j} = k) = p(x_{i',j'} = w | z_{i',j'} = k)$ と仮定
  - 同じコーパス内なら、どの文書のどの単語トークンであれ、それがトピック k を表現するために使われるならば（条件付き確率の条件の部分）、k を表現するためにどの単語が使われるかの確率は、同じ
  - つまり、単語確率分布とトピックが一対一に対応している
  - そこで、 $p(x_{\cdot,\cdot} = w | z_{\cdot,\cdot} = k) = \phi_{k,w}$ とおく

PLSAにおける観測データの尤度

個々の単語トークンにおけるトピックと単語の同時分布は

$$p(x_{i,j} = w, z_{i,j} = k) = p(z_{i,j} = k)p(x_{i,j} = w | z_{i,j} = k) = \phi_{k,x_{i,j}} \theta_{i,k} \tag{2}$$

潜在変数である $z_{i,j}$ を周辺化

$$p(x_{i,j} = w) = \sum_{z_i=1}^{K} p(x_{i,j} = w, z_{i,j} = k) = \sum_{k=1}^{K} \phi_{k,x_{i,j}} \theta_{i,k} \tag{3}$$

各トークンの独立性の仮定より

$$p(\boldsymbol{x}_i) = \prod_{j=1}^{n_i} p(x_{i,j}) = \prod_{j=1}^{n_i} \left(\sum_{k=1}^{K} \phi_{k,x_{i,j}} \theta_{i,k} \right) \tag{4}$$

各文書の独立性の仮定より

$$p(\mathcal{X}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} \left(\sum_{k=1}^{K} \phi_{k,x_{i,j}} \theta_{i,k} \right) \tag{5}$$

Contents

PLSA の問題点とベイズ化による改良

- 各文書におけるトピック確率 $\boldsymbol{\theta}_i = (\theta_{i,1}, \ldots, \theta_{i,K})$ に関して、異なる文書の間で何の関係性も仮定されていない
  - $\boldsymbol{\theta}_i$ と $\boldsymbol{\theta}_{i'}$ の間に何の関係もない。
- このことが過学習をもたらすかもしれない
- そこで、コーパスに属する全文書の $\boldsymbol{\theta}_i$ が、同一のディリクレ事前分布 Dir($\boldsymbol{\alpha}$) から draw されると仮定する
- 他は PLSA のまま
  - 各トピックの単語確率 $\boldsymbol{\phi}_k$ についても別のディリクレ分布 Dir($\boldsymbol{\beta}$) を導入できるが、そうしなくてもよい

Contents

PLSA と LDA の比較

PLSA における $\boldsymbol{x}_i$ の尤度

$$\begin{aligned} p(\boldsymbol{x}_i; \boldsymbol{\theta}_i, \boldsymbol{\Phi}) &= \sum_{\boldsymbol{z}_i} p(\boldsymbol{x}_i, \boldsymbol{z}_i; \boldsymbol{\theta}_i, \boldsymbol{\Phi}) = \sum_{\boldsymbol{z}_i} p(\boldsymbol{z}_i; \boldsymbol{\theta}_i) p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi}) \\ &= \prod_{j=1}^{n_i} \left(\sum_{z_{i,j}=1}^{K} p(z_{i,j}; \boldsymbol{\theta}_i) p(x_{i,j} | z_{i,j}; \boldsymbol{\Phi}) \right) = \prod_{j=1}^{n_i} \left(\sum_{k=1}^{K} \phi_{k, x_{i,j}} \theta_{i,k} \right) \end{aligned}$$

LDA における $\boldsymbol{x}_i$ の尤度（$p(\boldsymbol{x}_i; \boldsymbol{\theta}_i, \boldsymbol{\Phi})$ は $p(\boldsymbol{x}_i | \boldsymbol{\theta}_i; \boldsymbol{\Phi})$ に変わる）

$$\begin{aligned} p(\boldsymbol{x}_i; \boldsymbol{\Phi}, \boldsymbol{\alpha}) &= \int p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{x}_i | \boldsymbol{\theta}_i; \boldsymbol{\Phi}) d\boldsymbol{\theta}_i \\ &= \int \sum_{\boldsymbol{z}_i} p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi}) d\boldsymbol{\theta}_i \\ &= \int \left(\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \right) \prod_{j=1}^{n_i} \left(\sum_{k=1}^{K} \phi_{k, x_{i,j}} \theta_{i,k} \right) d\boldsymbol{\theta}_i \qquad (6) \end{aligned}$$

LDAの変分ベイズ法

Jensenの不等式を適用してELBOを求める

$$
\begin{aligned}
\ln p(\boldsymbol{x}_i; \boldsymbol{\Phi}, \boldsymbol{\alpha}) &= \ln \int \sum_{\boldsymbol{z}_i} p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi}) d\boldsymbol{\theta}_i \\
&= \ln \int \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i, \boldsymbol{\theta}_i) \frac{p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi})}{q(\boldsymbol{z}_i, \boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i \\
&\geq \int \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i, \boldsymbol{\theta}_i) \ln \frac{p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi})}{q(\boldsymbol{z}_i, \boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i
\end{aligned}
\tag{7}
$$

以下、$q(\boldsymbol{z}_i, \boldsymbol{\theta}_i) = q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i)$ と factorize すると仮定する。

$q(\boldsymbol{\theta}_i)$ を求める

$q(\boldsymbol{z}_i)$ を固定する。

$$
\begin{aligned}
\ln p(\boldsymbol{x}_i; \boldsymbol{\Phi}, \boldsymbol{\alpha}) &\geq \int \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i) \ln \frac{p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi})}{q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i \\
&= \int q(\boldsymbol{\theta}_i) \Big[\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) \Big] d\boldsymbol{\theta}_i - \int q(\boldsymbol{\theta}_i) \ln q(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + const. \\
&= -D_{\mathsf{KL}}\Big(q(\boldsymbol{\theta}_i) \parallel \frac{1}{Z} \exp \Big[\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) \Big] \Big) + const.
\end{aligned}
\tag{8}
$$

以上より、$q(\boldsymbol{\theta}_i) \propto \exp \left[\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) \right]$ のとき、ELBO は最大。つまり、
$q(\boldsymbol{\theta}_i) \propto p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) \exp \left[\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) \right]$ のとき、ELBO は最大。

$$
\begin{aligned}
\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln p(\boldsymbol{z}_i|\boldsymbol{\theta}_i) &= \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln \prod_{j=1}^{n_i} \theta_{i,z_{i,j}} = \sum_{j=1}^{n_i} \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln \theta_{i,z_{i,j}} \\
&= \sum_{j=1}^{n_i} \sum_{z_{i,j}=1}^{K} q(z_{i,j}) \ln \theta_{i,z_{i,j}} = \sum_{k=1}^{K} \Big(\sum_{j=1}^{n_i} q(z_{i,j} = k) \Big) \ln \theta_{i,k} = \sum_{k=1}^{K} n_{i,k} \ln \theta_{i,k} \qquad (9)
\end{aligned}
$$

ただし、$n_{i,k} \equiv \sum_{j=1}^{n_i} q(z_{i,j} = k)$ と定義した。よって

$$
\begin{aligned}
q(\boldsymbol{\theta}_i) &\propto \prod_{k=1}^{K} \theta_{i,k}^{\alpha_k - 1} \times \exp\Big[\sum_{k=1}^{K} n_{i,k} \ln \theta_{i,k} \Big] \\
&= \prod_{k=1}^{K} \theta_{i,k}^{\alpha_k + n_{i,k} - 1} \qquad (10)
\end{aligned}
$$

これは、変分事後分布 $q(\boldsymbol{\theta}_i)$ がディリクレ分布であることを意味する。
変分ディリクレ事後分布 $q(\boldsymbol{\theta}_i)$ のパラメータを $\boldsymbol{\zeta}_i$ とすると、$\zeta_{i,k} = \alpha_k + n_{i,k}$ が成り立つ。

$q(\boldsymbol{z}_i)$ を求める

今度は $q(\boldsymbol{\theta}_i)$ を固定する。

$$
\begin{aligned}
\ln p(\boldsymbol{x}_i; \boldsymbol{\Phi}, \boldsymbol{\alpha}) &\geq \int \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i) \ln \frac{p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi})}{q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i \\
&= \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \left[\ln p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi}) + \int q(\boldsymbol{\theta}_i) \ln p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \right] - \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln q(\boldsymbol{z}_i) + const. \\
&= -D_{\mathsf{KL}}(q(\boldsymbol{z}_i) \parallel \frac{1}{Z} \exp \left[\ln p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi}) + \int q(\boldsymbol{\theta}_i) \ln p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \right]) + const. \qquad (11)
\end{aligned}
$$

以上より、$q(\boldsymbol{z}_i) \propto p(\boldsymbol{x}_i | \boldsymbol{z}_i; \boldsymbol{\Phi}) \exp \left[\int q(\boldsymbol{\theta}_i) \ln p(\boldsymbol{z}_i | \boldsymbol{\theta}_i) d\boldsymbol{\theta}_i \right]$ のとき、ELBO は最大。

変分ディリクレ事後分布 $q(\boldsymbol{\theta}_i)$ のパラメータが $\boldsymbol{\zeta}_i$ であることを使うと、

$$\int q(\boldsymbol{\theta}_i;\boldsymbol{\zeta}_i)\ln p(\boldsymbol{z}_i|\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i = \int q(\boldsymbol{\theta}_i;\boldsymbol{\zeta}_i)\ln\prod_{j=1}^{n_i}\theta_{i,z_{i,j}}d\boldsymbol{\theta}_i = \sum_{j=1}^{n_i}\int q(\boldsymbol{\theta}_i;\boldsymbol{\zeta}_i)\ln\theta_{i,z_{i,j}}d\boldsymbol{\theta}_i$$

$$= \sum_{j=1}^{n_i}\left\{\psi(\zeta_{i,z_{i,j}}) - \psi(\sum_k \zeta_{i,k})\right\} = \sum_{j=1}^{n_i}\psi(\zeta_{i,z_{i,j}}) + const. \tag{12}$$

よって、

$$q(\boldsymbol{z}_i) \propto p(\boldsymbol{x}_i|\boldsymbol{z}_i;\boldsymbol{\Phi})\exp\left[\int q(\boldsymbol{\theta}_i)\ln p(\boldsymbol{z}_i|\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i\right] = \prod_{j=1}^{n_i}\phi_{z_{i,j},x_{i,j}} \times \exp\Big(\sum_{j=1}^{n_i}\psi(\zeta_{i,z_{i,j}})\Big)$$

$$= \prod_{j=1}^{n_i}\phi_{z_{i,j},x_{i,j}} \times \prod_{j=1}^{n_i}\exp(\psi(\zeta_{i,z_{i,j}})) = \prod_{j=1}^{n_i}\phi_{z_{i,j},x_{i,j}}\exp\big(\psi(\zeta_{i,z_{i,j}})\big) \tag{13}$$

つまり、

$$q(z_{i,j}=k) = \frac{\phi_{k,x_{i,j}}\exp\big(\psi(\zeta_{i,k})\big)}{\sum_{l=1}^{K}\phi_{l,x_{i,j}}\exp\big(\psi(\zeta_{i,l})\big)} \tag{14}$$

変分事後分布を使ってELBOを書き下す

$$
\begin{aligned}
\ln p(\boldsymbol{x}_i; \boldsymbol{\Phi}, \boldsymbol{\alpha}) &\geq \int \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i) \ln \frac{p(\boldsymbol{\theta}_i; \boldsymbol{\alpha}) p(\boldsymbol{z}_i|\boldsymbol{\theta}_i) p(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\Phi})}{q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i)} d\boldsymbol{\theta}_i \\
&= \int \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i) \ln p(\boldsymbol{z}_i|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i + \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\Phi}) \\
&\quad - D_{\mathsf{KL}}(q(\boldsymbol{\theta}_i) \parallel p(\boldsymbol{\theta}_i; \boldsymbol{\alpha})) - \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln q(\boldsymbol{z}_i) \qquad (15)
\end{aligned}
$$

式 (15) の右辺の最初の項を計算してみる。

$$
\begin{aligned}
\int \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) q(\boldsymbol{\theta}_i) \ln p(\boldsymbol{z}_i|\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i &= \sum_{j=1}^{n_i} \sum_{z_{i,j}=1}^{K} q(z_{i,j}) \int q(\boldsymbol{\theta}_i) \ln \theta_{i, z_{i,j}} d\boldsymbol{\theta}_i \\
&= \sum_{k=1}^{K} \Big(\sum_{j=1}^{n_i} q(z_{i,j} = k) \Big) \Big(\psi(\zeta_{i,k}) - \psi(\sum_{l} \zeta_{i,l}) \Big) \qquad (16)
\end{aligned}
$$

式 (15) の右辺の 2 番目の項を計算してみる。

$$\sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i) \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i; \boldsymbol{\Phi}) = \sum_{j=1}^{n_i} \sum_{z_{i,j}=1}^{K} q(z_{i,j}) \ln \phi_{z_{i,j},x_{i,j}} = \sum_{j=1}^{n_i} \sum_{k=1}^{K} q(z_{i,j} = k) \ln \phi_{k,x_{i,j}} \tag{17}$$

トピック単語確率 $\boldsymbol{\Phi}$ は、この項の全文書についての和 $\sum_{i=1}^{N} \sum_{j=1}^{n_i} \sum_{k=1}^{K} q(z_{i,j} = k) \ln \phi_{k,x_{i,j}}$ を最大化することで求めることができる。(ELBO の中で $\boldsymbol{\Phi}$ を含むのはこの項だけだから。)

全文書の ELBO の和を $\mathcal{L}$ と書くことにする。

$\sum_{w=1}^{W} \phi_{k,w} = 1$ が満たされなければならないので、ラグランジュの未定乗数法を使えば、

$$\frac{\partial \mathcal{L}}{\partial \phi_{k,w}} + \frac{\partial}{\partial \phi_{k,w}} \lambda_k \left(1 - \sum_{w=1}^{W} \phi_{k,w}\right) = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} q(z_{i,j} = k) \delta(x_{i,j} = w)}{\phi_{k,w}} - \lambda_k \tag{18}$$

$$\therefore \phi_{k,w} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} q(z_{i,j} = k) \delta(x_{i,j} = w)}{\sum_{i=1}^{N} \sum_{j=1}^{n_i} q(z_{i,j} = k)} \tag{19}$$

LDAの変分ベイズ法のまとめ

以下の更新を繰り返し実行する。

$$q(z_{i,j}=k) \leftarrow \frac{\phi_{k,x_{i,j}} \exp\left(\psi(\zeta_{i,k})\right)}{\sum_{l=1}^{K} \phi_{l,x_{i,j}} \exp\left(\psi(\zeta_{i,l})\right)} \tag{20}$$

$$\zeta_{i,k} \leftarrow \alpha_k + \sum_{j=1}^{n_i} q(z_{i,j}=k) \tag{21}$$

$$\phi_{k,w} \leftarrow \frac{\sum_{i=1}^{N} \sum_{j=1}^{n_i} q(z_{i,j}=k)\delta(x_{i,j}=w)}{\sum_{i=1}^{N} \sum_{j=1}^{n_i} q(z_{i,j}=k)} \tag{22}$$

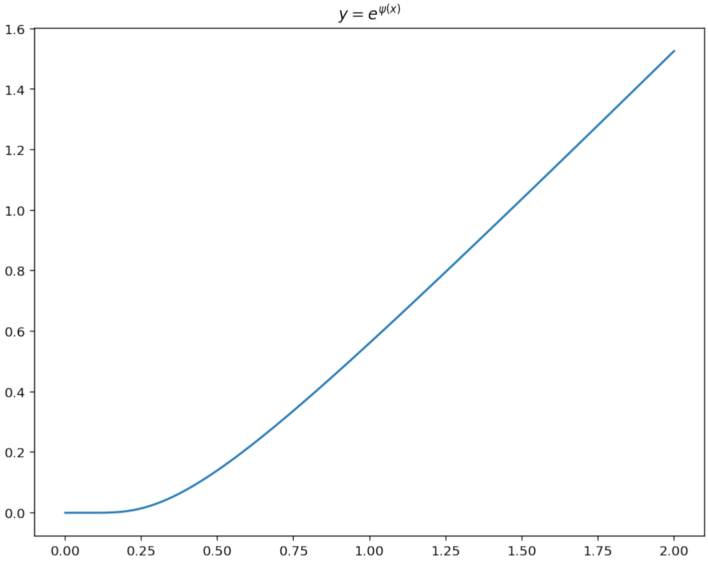

Figure: $y = e^{\psi(x)}$ のグラフ