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

PLSAにおける同時分布

- PLSAでは、文書 $\boldsymbol{x}_d$ の i 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 $p(x_{d,i} = w, z_{d,i} = k)$ は

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

- $p(z_{d,i} = k)$ は、文書 $\boldsymbol{x}_d$ の i 番目のトークンがトピック k を表現する確率
- $p(x_{d,i} = w | z_{d,i} = k)$ は、文書 $\boldsymbol{x}_d$ の i 番目のトークンがトピック k を表現するとき、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 i, i' についても $p(z_{d,i} = k) = p(z_{d,i'} = k)$ と仮定
  - そこで、$p(z_{d,\cdot} = k) = \theta_{d,k}$ とおく
  - 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）
- どの d, d' や i, i' についても、
  $p(x_{d,i} = w | z_{d,i} = k) = p(x_{d',i'} = w | z_{d',i'} = k)$ と仮定
  - そこで、 $p(x_{\cdot,\cdot} = w | z_{\cdot,\cdot} = k) = \phi_{k,w}$ とおく
  - 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、<u>単語確率分布とトピックが一対一に対応</u>している

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

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

潜在変数である $z_{d,i}$ を周辺化

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

各トークンの独立性の仮定より

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

各文書の独立性の仮定より

$$p(\mathcal{D}) = \prod_{d=1}^{D} \prod_{i=1}^{N_d} \left(\sum_{k=1}^{K} \phi_{k,x_{d,i}}\theta_{d,k} \right) \tag{5}$$

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 $\boldsymbol{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,K})$ 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
  - $\boldsymbol{\theta}_d$ と $\boldsymbol{\theta}_{d'}$ 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、全文書にわたって $\boldsymbol{\theta}_d$ が同一のディリクレ事前分布 $\mathrm{Dir}(\boldsymbol{\alpha})$ から draw されると仮定する
- 他はそのまま
  - 各トピックの単語確率 ϕ_k についても別のディリクレ分布 $\mathrm{Dir}(\boldsymbol{\beta})$ を導入できるが、これはそうしなくてもよい。

Contents

PLSAとLDAの比較

PLSAにおける$\boldsymbol{x}_d$の尤度

$$p(\boldsymbol{x}_d; \boldsymbol{\theta}_d, \boldsymbol{\Phi}) = \sum_{\boldsymbol{z}_d} p(\boldsymbol{x}_d, \boldsymbol{z}_d; \boldsymbol{\theta}_d, \boldsymbol{\Phi}) = \sum_{\boldsymbol{z}_d} p(\boldsymbol{z}_d; \boldsymbol{\theta}_d) p(\boldsymbol{x}_d | \boldsymbol{z}_d; \boldsymbol{\Phi})$$

$$= \prod_{i=1}^{N_d} \left(\sum_{z_{d,i}=1}^{K} p(z_{d,i}; \boldsymbol{\theta}_d) p(x_{d,i} | z_{d,i}; \boldsymbol{\Phi}) \right) = \prod_{i=1}^{N_d} \left(\sum_{k=1}^{K} \phi_{k,x_{d,i}} \theta_{d,k} \right)$$

LDAにおける$\boldsymbol{x}_d$の尤度

$$p(\boldsymbol{x}_d; \boldsymbol{\Phi}, \boldsymbol{\alpha}) = \int p(\boldsymbol{\theta}_d; \boldsymbol{\alpha}) p(\boldsymbol{x}_d | \boldsymbol{\theta}_d; \boldsymbol{\Phi}) d\boldsymbol{\theta}_d$$

$$= \int \sum_{\boldsymbol{z}_d} p(\boldsymbol{\theta}_d; \boldsymbol{\alpha}) p(\boldsymbol{z}_d | \boldsymbol{\theta}_d) p(\boldsymbol{x}_d | \boldsymbol{z}_d; \boldsymbol{\Phi}) d\boldsymbol{\theta}_d \tag{6}$$

LDA の変分ベイズ法

Jensen の不等式を適用して ELBO を求める

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

$q(\boldsymbol{\theta}_d)$ を求める

$q(\boldsymbol{z}_d, \boldsymbol{\theta}_d)$ は $q(\boldsymbol{z}_d)q(\boldsymbol{\theta}_d)$ と factorize すると仮定する。そして、$q(\boldsymbol{z}_d)$ を固定する。

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

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

$$
\begin{aligned}
\sum_{\boldsymbol{z}_d} q(\boldsymbol{z}_d) \ln p(\boldsymbol{z}_d|\boldsymbol{\theta}_d) &= \sum_{\boldsymbol{z}_d} q(\boldsymbol{z}_d) \ln \prod_{i=1}^{N_d} \theta_{d,z_{d,i}} = \sum_{i=1}^{N_d} \sum_{\boldsymbol{z}_d} q(\boldsymbol{z}_d) \ln \theta_{d,z_{d,i}} \\
&= \sum_{i=1}^{N_d} \sum_{z_{d,i}=1}^{K} q(z_{d,i}) \ln \theta_{d,z_{d,i}} = \sum_{k=1}^{K} \Big(\sum_{i=1}^{N_d} q(z_{d,i} = k) \Big) \ln \theta_{d,k} = \sum_{k=1}^{K} N_{d,k} \ln \theta_{d,k} \qquad (9)
\end{aligned}
$$

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

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

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

$q(\boldsymbol{z}_d)$ を求める

今度は $q(\boldsymbol{\theta}_d)$ を固定する。

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

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

$q(\boldsymbol{\theta}_d)$ がパラメータ $\boldsymbol{\zeta}_d$ のディリクレ分布であることを使うと、

$$\int q(\boldsymbol{\theta}_d;\boldsymbol{\zeta}_d)\ln p(\boldsymbol{z}_d|\boldsymbol{\theta}_d)d\boldsymbol{\theta}_d = \int q(\boldsymbol{\theta}_d;\boldsymbol{\zeta}_d)\ln\prod_{i=1}^{N_d}\theta_{d,z_{d,i}}d\boldsymbol{\theta}_d = \sum_{i=1}^{N_d}\int q(\boldsymbol{\theta}_d;\boldsymbol{\zeta}_d)\ln\theta_{d,z_{d,i}}d\boldsymbol{\theta}_d$$
$$= \sum_{i=1}^{N_d}\left\{\psi(\zeta_{d,z_{d,i}}) - \psi(\sum_k \zeta_{d,k})\right\} = \sum_{i=1}^{N_d}\psi(\zeta_{d,z_{d,i}}) + const. \tag{12}$$

よって、

$$q(\boldsymbol{z}_d) \propto p(\boldsymbol{x}_d|\boldsymbol{z}_d;\boldsymbol{\Phi})\exp\left[\int q(\boldsymbol{\theta}_d)\ln p(\boldsymbol{z}_d|\boldsymbol{\theta}_d)d\boldsymbol{\theta}_d\right]$$
$$= \prod_{i=1}^{N_d}\phi_{z_{d,i},x_{d,i}}\exp\left(\sum_{i=1}^{N_d}\psi(\zeta_{d,z_{d,i}})\right) = \prod_{i=1}^{N_d}\phi_{z_{d,i},x_{d,i}}\exp\left(\psi(\zeta_{d,z_{d,i}})\right) \tag{13}$$

つまり、

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

変分事後分布を使ってELBOを書き下す

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

式 (15) の右辺の最初の項を計算してみる。

$$
\begin{aligned}
&\int \sum_{\boldsymbol{z}_d} q(\boldsymbol{z}_d) q(\boldsymbol{\theta}_d) \ln p(\boldsymbol{z}_d | \boldsymbol{\theta}_d) d\boldsymbol{\theta}_d = \sum_{i=1}^{N_d} \sum_{z_{d,i}=1}^{K} q(z_{d,i}) \int q(\boldsymbol{\theta}_d) \ln \theta_{d, z_{d,i}} d\boldsymbol{\theta}_d \\
&= \sum_{k=1}^{K} \left(\sum_{i=1}^{N_d} q(z_{d,i} = k) \right) \left(\psi(\zeta_{d,k}) - \psi(\sum_{l} \zeta_{d,l}) \right)
\end{aligned}
\tag{16}
$$

式 (15) の右辺の 2 番目の項を計算してみる。

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

トピック単語確率 $\boldsymbol{\Phi}$ は、この項の全文書についての和 $\sum_{d=1}^{D} \sum_{i=1}^{N_d} \sum_{k=1}^{K} q(z_{d,i} = k) \ln \phi_{k, x_{d,i}}$ 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、

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

$$\therefore \phi_{k,w} = \frac{\sum_{d=1}^{D} \sum_{i=1}^{N_d} q(z_{d,i} = k) \delta(x_{d,i} = w)}{\sum_{d=1}^{D} \sum_{i=1}^{N_d} q(z_{d,i} = k)} \tag{19}$$

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。

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

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

$$\phi_{k,w} \leftarrow \frac{\sum_{d=1}^{D} \sum_{i=1}^{N_d} q(z_{d,i}=k)\delta(x_{d,i}=w)}{\sum_{d=1}^{D} \sum_{i=1}^{N_d} q(z_{d,i}=k)} \tag{22}$$