PLSA

正田 備也

masada@rikkyo.ac.jp

Contents

なぜPLSAの話をするか

- 次回LDA(latent Dirichlet allocation)の話をするため
- LDAはPLSA(probabilistic latent semantic analysis)をベイズ化したモデル
- PLSAは混合多項分布の改良版と見なすこともできる

混合多項分布によるデータの生成

- 混合多項分布を使ったモデリングでは、N 件の文書 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ が以下のように生成されると仮定する

1. カテゴリカル分布 Cat$(\boldsymbol{\theta})$ から、確率変数 z_i の値を draw
2. z_i 番目のコンポーネントを表す多項分布 Multi$(\boldsymbol{\phi}_{z_i})$ から、$x_{i,1}, \ldots, x_{i,n_i}$ それぞれの値を draw

   例: $x_{i,j} =$ "apple" は、i 番目の文書の j 番目の単語トークンとして、"apple" が出現していることを示す

- $\boldsymbol{x}_i$ は単語の並びを表すが、実質的には各単語の出現頻度だけをモデリングしている（bag-of-words モデルだから）

混合多項分布の問題点

- 混合多項分布モデルでは、一つの文書がそれ全体で意味的に均一だと仮定することになる
  - ニュース記事であれば、一つの記事まるごとが、特定のカテゴリ（ex. 政治、経済、スポーツ、etc）に割り振られる。
- しかし、文書内は意味的に均一という仮定は実態に合わない
- というのも、一つの文書は複数の話題を含みうるからである

混合多項分布の改良としてのPLSA

- 混合多項分布と同様、カテゴリの違いは、語彙集合上に定義された多項分布の違いとして表す

  例: 政治について書かれたテキストと経済について書かれたテキストとでは、どの単語がいくらの確率で出現するかが、異なる

- 同じ文書に含まれる単語トークン群が複数の単語多項分布から生成されると仮定する
  - 同じ文書内に、異なる単語多項分布に由来する単語トークンが混ざっていてもよい、という考え方
  - つまり、同じ文書が複数の「トピック」を含みうる、という考え方
  - トピック＝単語多項分布 Multinomial(ϕ_k)

PLSA (probabilistic latent semantic analysis)

- PLSI (probabilistic latent semantic indexing) とも呼ばれる
- LSA の確率モデル版、ということ
  - LSA は、単語-文書行列の特異値分解で次元圧縮する手法（後述）
- 生成モデルとして記述すると…
  - 文書に固有のトピック多項分布（その文書でどのトピックがいくらの確率で現れるか）から、単語トークン毎にトピックを draw
  - 各単語トークンについて、そのトピックに対応する単語多項分布（そのトピックについて書くときどの単語がいくらの確率で使われるか）から単語を draw

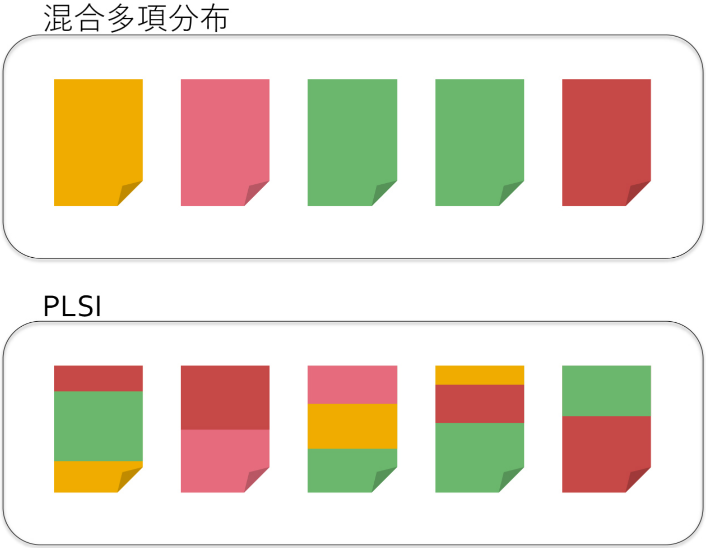

Figure: 混合多項分布と PLSA の違いのイメージ

Figure: PLSA では同一文書内の単語トークンが複数の単語多項分布に由来

Contents

混合多項分布の問題点

PLSA (probabilistic latent semantic analysis)

PLSA (probabilistic latent semantic analysis)

- LSA(latent semantic analysis) を probabilistic にしたモデル
  - LSA については次スライドの図を参照（実態は単なる SVD）

- 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）
- また、どのトピックがいくらの確率で現れるかが、文書によって異なる

- PLSA における単語多項分布を、トピック (topic) と呼ぶ
  - PLSA は最もシンプルなトピックモデル

LSA の概念図

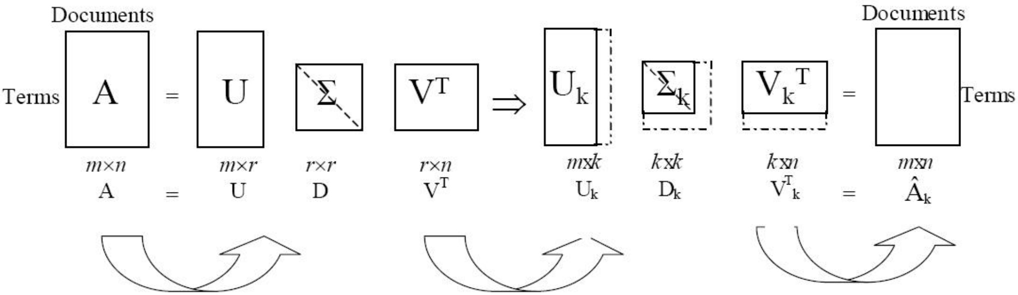

Figure: LSA の概念図

- 左から順に、データ行列の特異値分解、低ランク近似、元のデータ行列の再現
- m が語彙サイズ、n が文書数、k がトピック数（r は元のデータ行列のランク）

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（例：$x_{412,27} = 8203$）
- 文書 $\boldsymbol{x}_i$ の j 番目のトークンが表現するトピックを $z_{i,j}$ という確率変数で表す（例：$z_{412,27} = 69$）
- $x_{i,j}$ 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

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

  - $p(z_{i,j} = k)$ 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、<u>単語確率分布とトピックが一対一に対応</u>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

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

混合多項分布とPLSAの比較

- PLSAにおける$\boldsymbol{x}_i$の尤度

$$p(\boldsymbol{x}_i) = \prod_{j=1}^{n_i} \left(\sum_{k=1}^{K} \phi_{k,x_{i,j}} \theta_{i,k} \right) \tag{6}$$

- 混合多項分布における$\boldsymbol{x}_i$の尤度

$$p(\boldsymbol{x}_i) = \sum_{k=1}^{K} \theta_k \prod_{j=1}^{n_i} \phi_{k,x_{i,j}} \tag{7}$$

対数尤度最大化のため Jensen の不等式を適用

$$
\begin{aligned}
\ln p(\boldsymbol{x}_i) &= \ln \prod_{j=1}^{n_i} \left(\sum_{k=1}^{K} \phi_{k,x_{i,j}} \theta_{i,k} \right) \\
&= \sum_{j=1}^{n_i} \ln \left(\sum_{k=1}^{K} q_{i,j,k} \frac{\phi_{k,x_{i,j}} \theta_{i,k}}{q_{i,j,k}} \right) \\
&\geq \sum_{i=1}^{n_i} \left(\sum_{k=1}^{K} q_{i,j,k} \ln \frac{\phi_{k,x_{i,j}} \theta_{i,k}}{q_{i,j,k}} \right) \\
&= \sum_{i=1}^{n_i} \sum_{k=1}^{K} q_{i,j,k} \ln(\phi_{k,x_{i,j}} \theta_{i,k}) - \sum_{i=1}^{n_i} \sum_{k=1}^{K} q_{i,j,k} \ln q_{i,j,k} \\
&= \sum_{i=1}^{n_i} \sum_{k=1}^{K} q_{i,j,k} \ln \phi_{k,x_{i,j}} + \sum_{i=1}^{n_i} \sum_{k=1}^{K} q_{i,j,k} \ln \theta_{i,k} - \sum_{i=1}^{n_i} \sum_{k=1}^{K} q_{i,j,k} \ln q_{i,j,k} \qquad (8)
\end{aligned}
$$

where $\sum_{k=1}^{K} q_{i,j,k} = 1$ holds for all i, j.

対数尤度の lower bound

$$\ln p(\mathcal{X}) \geq \sum_{i=1}^{N}\sum_{i=1}^{n_i}\sum_{k=1}^{K} q_{i,j,k} \ln \phi_{k,x_{i,j}} + \sum_{i=1}^{N}\sum_{i=1}^{n_i}\sum_{k=1}^{K} q_{i,j,k} \ln \theta_{i,k} - \sum_{i=1}^{N}\sum_{i=1}^{n_i}\sum_{k=1}^{K} q_{i,j,k} \ln q_{i,j,k} \quad (9)$$

最大化すべき目的関数は

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\Phi}, \boldsymbol{Q}) \\
&= \sum_{i=1}^{N}\sum_{i=1}^{n_i}\sum_{k=1}^{K} q_{i,j,k} \ln \phi_{k,x_{i,j}} + \sum_{i=1}^{N}\sum_{i=1}^{n_i}\sum_{k=1}^{K} q_{i,j,k} \ln \theta_{i,k} - \sum_{i=1}^{N}\sum_{i=1}^{n_i}\sum_{k=1}^{K} q_{i,j,k} \ln q_{i,j,k} \\
&\quad + \sum_{i=1}^{N} \lambda_i \left(1 - \sum_{k=1}^{K} \theta_{i,k}\right) + \sum_{k=1}^{K} \mu_k \left(1 - \sum_{w=1}^{W} \phi_{k,w}\right) + \sum_{i=1}^{N}\sum_{i=1}^{n_i} \nu_{i,j} \left(1 - \sum_{k=1}^{K} q_{i,j,k}\right) \quad (10)
\end{aligned}$$

PLSA の EM アルゴリズム (1/2)

M step

$$\frac{\partial \mathcal{L}}{\partial \theta_{i,k}} = \sum_{i=1}^{n_i} \frac{q_{i,j,k}}{\theta_{i,k}} - \lambda_i \tag{11}$$

$$\therefore \theta_{i,k} = \frac{\sum_{i=1}^{n_i} q_{i,j,k}}{\sum_{k=1}^{K} \sum_{i=1}^{n_i} q_{i,j,k}} \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \phi_{k,w}} = \frac{\sum_{i=1}^{N} \sum_{i=1}^{n_i} \delta(x_{i,j} = v_w) q_{i,j,k}}{\phi_{k,w}} - \mu_k \tag{13}$$

$$\begin{aligned} \therefore \phi_{k,w} &= \frac{\sum_{i=1}^{N} \sum_{i=1}^{n_i} \delta(x_{i,j} = v_w) q_{i,j,k}}{\sum_{w=1}^{W} \sum_{i=1}^{N} \sum_{i=1}^{n_i} \delta(x_{i,j} = v_w) q_{i,j,k}} \\ &= \frac{\sum_{i=1}^{N} \sum_{i=1}^{n_i} \delta(x_{i,j} = v_w) q_{i,j,k}}{\sum_{i=1}^{N} \sum_{i=1}^{n_i} q_{i,j,k}} \end{aligned} \tag{14}$$

PLSA の EM アルゴリズム (2/2)

E step

$$\frac{\partial \mathcal{L}}{\partial q_{i,j,k}} = \ln \phi_{k,x_{i,j}} + \ln \theta_{i,k} - \ln q_{i,j,k} - 1 - \nu_{i,j} \tag{15}$$

$$\therefore q_{i,j,k} \propto \phi_{k,x_{i,j}} \theta_{i,k} \tag{16}$$

$$\therefore q_{i,j,k} = \frac{\phi_{k,x_{i,j}} \theta_{i,k}}{\sum_{k=1}^{K} \phi_{k,x_{i,j}} \theta_{i,k}} \tag{17}$$

よって、$x_{i,j} = x_{i,j'}$ ならば $q_{i,j,k} = q_{i,j',k}$ となる。つまり、PLSA では同一文書内で別の場所に現れる同じ単語を区別できない。よって、第 i 文書での単語 w の TF を $n_{i,w}$ とすると

$$q_{i,w,k} = \frac{\phi_{k,w} \theta_{i,k}}{\sum_{k=1}^{K} \phi_{k,w} \theta_{i,k}} \tag{18}$$

$$\theta_{i,k} = \frac{\sum_{w=1}^{W} n_{i,w} q_{i,w,k}}{\sum_{k=1}^{K} \sum_{w=1}^{W} n_{i,w} q_{i,w,k}}, \quad \phi_{k,w} = \frac{\sum_{i=1}^{N} n_{i,w} q_{i,w,k}}{\sum_{w=1}^{W} \sum_{i=1}^{N} n_{i,w} q_{i,w,k}} \tag{19}$$

PLSA の EM アルゴリズムのまとめ

- 文書 $\boldsymbol{x}_i$ 内の単語 v_w のトークンがトピック t_k を表現する確率

$$q_{i,w,k} \leftarrow \frac{\phi_{k,w}\theta_{i,k}}{\sum_{k=1}^{K} \phi_{k,w}\theta_{i,k}} \tag{20}$$

- 文書 $\boldsymbol{x}_i$ 内の単語トークンがトピック t_k を表現する確率

$$\theta_{i,k} \leftarrow \frac{\sum_{w=1}^{W} n_{i,w} q_{i,w,k}}{\sum_{k=1}^{K} \sum_{w=1}^{W} n_{i,w} q_{i,w,k}} \tag{21}$$

- トピック t_k が単語 v_w のトークンによって表現される確率

$$\phi_{k,w} \leftarrow \frac{\sum_{i=1}^{N} n_{i,w} q_{i,w,k}}{\sum_{w=1}^{W} \sum_{i=1}^{N} n_{i,w} q_{i,w,k}} \tag{22}$$

直感的な意味

- $q_{i,w,k}$ は、i 番目の文書に現れる単語 w が、他のトピックではなくトピック k を表現する確率
- $\sum_{w=1}^{W} n_{i,w} q_{i,w,k}$ は、i 番目の文書のなかで、トピック k を表現している単語トークンの個数
- $\sum_{i=1}^{N} n_{i,w} q_{i,w,k}$ は、コーパス全体のなかで、単語 w のトークン群のうち、トピック k を表現しているトークンの個数