PLSA

正田 備也

masada@rikkyo.ac.jp

Contents

混合多項分布

- 混合多項分布モデルでは、例えば文書クラスタリングへの応用の場合、一つの文書がそれ全体で意味的なまとまりを持つ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。
- つまり、一つの文書内は意味的に均一だと、仮定している
- しかし、この仮定は文書の実態に合わない
- 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
  - 政治について書かれたテキストと、スポーツについて書いたテキストとでは、どの単語がどのくらいの確率で出現するかが異なる、という考え方。
- 混合多項分布とは異なり、一つの文書に含まれる単語トークン群が、唯一の単語多項分布からではなく、複数の単語多項分布から生成されると、仮定する→PLSA モデル
  - 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。
  - 同じ文書が複数の「トピック」を含みうる、という考え方。

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どのくらい現れやすいか）から、単語トークン毎にトピックを draw
  - 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どのくらい使われやすいか）から単語を draw

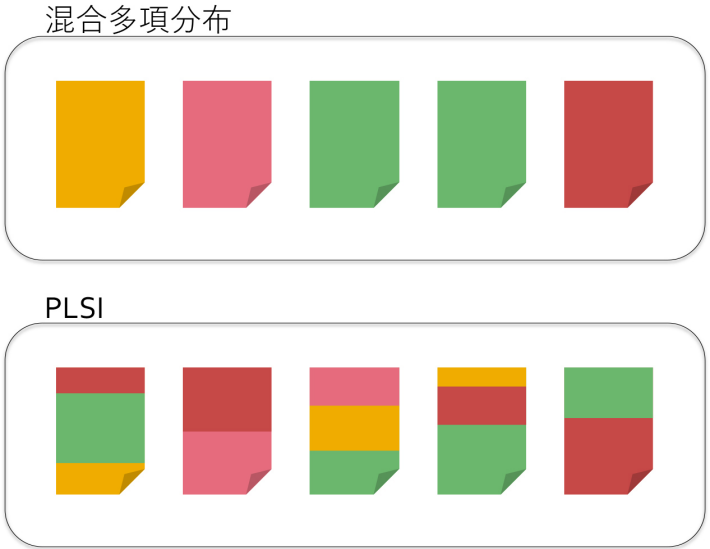

Figure: 混合多項分布と PLSA の違い

Figure: PLSA では同じ文書の単語トークンが複数の単語多項分布に由来しうる

Contents

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）
- どのトピックがどのくらいの確率で使われるかが、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
  - 混合多項分布では、複数ある単語多項分布を、クラスタやコンポーネントと呼んでいた
  - PLSA は最もシンプルなトピックモデル

LSA の概念図

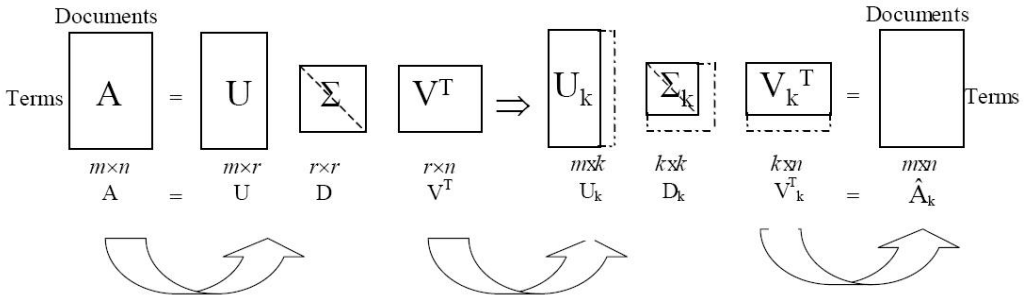

Figure: LSA の概念図

- 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 $\mathcal{V} = \{1, \ldots, W\}$ （単語とその index を同一視することとする）
- トピック集合 $\mathcal{T} = \{1, \ldots, K\}$ （トピックとその index を同一視することとする）
- 文書集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D\}$
- 文書 $\boldsymbol{x}_d$ の i 番目に出現する単語を $x_{d,i}$ という確率変数で表す
- 文書 $\boldsymbol{x}_d$ の i 番目に出現する単語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、$x_{d,i}$ がトピック t_k を表現し、かつそのトピックを表現するために現にその位置にある単語（v_w とする）が使われる同時確率、つまり $p(x_{d,i} = w, z_{d,i} = k)$ は

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

- $p(z_{d,i} = k)$ は、文書 $\boldsymbol{x}_d$ の i 番目の単語が、他のトピックではなく、第 k トピックを表現する確率
- $p(x_{d,i} = w | z_{d,i} = k)$ は、第 k トピックを表現するときに、他の単語ではなく、第 w 単語が使われる確率
- 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 $p(z_{d,i} = k) = p(z_{d,i'} = k)$ と仮定する
  - 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、第 k トピックを表現する確率は、同じ（場所によってトピック確率が違ったりしない）
- どの d, d' や i, i' についても、
  $p(x_{d,i} = w|z_{d,i} = k) = p(x_{d',i'} = w|z_{d',i'} = k)$ と仮定する
  - 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、それが第 k トピック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）、どの単語がどの確率で第 k トピックを表現するかは、同じ（単語確率分布とトピックが一対一に対応している）

PLSA における観測データの尤度

同時分布は

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

潜在変数を周辺化

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

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

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

混合多項分布とPLSAの比較

- PLSAにおける$\boldsymbol{x}_d$の尤度

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

- 混合多項分布における$\boldsymbol{x}_d$の尤度

$$p(\boldsymbol{x}_d) = \sum_{k=1}^{K} \theta_k \prod_{i=1}^{N_d} \phi_{k,x_{d,i}} \tag{7}$$

Jensen の不等式の適用

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

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

周辺尤度の lower bound

$$\ln p(\mathcal{D}) \geq \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} q_{d,i,k} \ln \phi_{k,x_{d,i}} + \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} q_{d,i,k} \ln \theta_{d,k} - \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} q_{d,i,k} \ln q_{d,i,k} \tag{9}$$

最大化すべき目的関数は

$$\begin{aligned}
&\mathcal{L}(\{\boldsymbol{\theta}_d\}, \{\boldsymbol{\phi}_k\}, \{\boldsymbol{q}_{d,i}\}) \\
&= \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} q_{d,i,k} \ln \phi_{k,x_{d,i}} + \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} q_{d,i,k} \ln \theta_{d,k} - \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} q_{d,i,k} \ln q_{d,i,k} \\
&\quad + \sum_{d=1}^{D} \lambda_d \left(1 - \sum_{k=1}^{K} \theta_{d,k}\right) + \sum_{k=1}^{K} \mu_k \left(1 - \sum_{w=1}^{W} \phi_{k,w}\right) + \sum_{d=1}^{D}\sum_{i=1}^{N_d} \nu_{d,i} \left(1 - \sum_{k=1}^{K} q_{d,i,k}\right)
\end{aligned} \tag{10}$$

PLSA の EM アルゴリズム (1/2)

M step

$$\frac{\partial \mathcal{L}}{\partial \theta_{d,k}} = \sum_{i=1}^{N_d} \frac{q_{d,i,k}}{\theta_{d,k}} - \lambda_d \tag{11}$$

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

$$\frac{\partial \mathcal{L}}{\partial \phi_{k,w}} = \frac{\sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{1}(x_{d,i} = w) q_{d,i,k}}{\phi_{k,w}} - \mu_k \tag{13}$$

$$\begin{aligned} \therefore \phi_{k,w} &= \frac{\sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{1}(x_{d,i} = w) q_{d,i,k}}{\sum_{w=1}^{W} \sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{1}(x_{d,i} = w) q_{d,i,k}} \\ &= \frac{\sum_{d=1}^{D} \sum_{i=1}^{N_d} \mathbb{1}(x_{d,i} = w) q_{d,i,k}}{\sum_{d=1}^{D} \sum_{i=1}^{N_d} q_{d,i,k}} \end{aligned} \tag{14}$$

PLSA の EM アルゴリズム (1/2)

E step

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

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

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

$x_{d,i} = x_{d,i'}$ ならば $q_{d,i,k} = q_{d,i',k}$ 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 d 文書での第 w 単語の TF を $n_{d,w}$ とすると

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

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