

PLSA

正田 備也

masada@rikkyo.ac.jp

Contents

混合多項分布の問題点

PLSA (probabilistic latent semantic analysis)

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は多項分布の混合分布の改良版と見なすこともできる

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、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。

- ▶ 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、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

混合多項分布



PLSI



Figure: 混合多項分布と PLSA の違い

Shanghai is the largest city in China, located in its eastern coast at the outlet of the Yangtze River. Originally a fishing and textiles town, Shanghai grew in importance in the 19th century. In 2005 Shanghai became the world's busiest cargo port. The city is an emerging tourist destination renowned for its historical landmarks such as the Bund and Xintiandi, its modern and

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）
- ▶ 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
 - ▶ PLSA は最もシンプルなトピックモデル

LSA の概念図

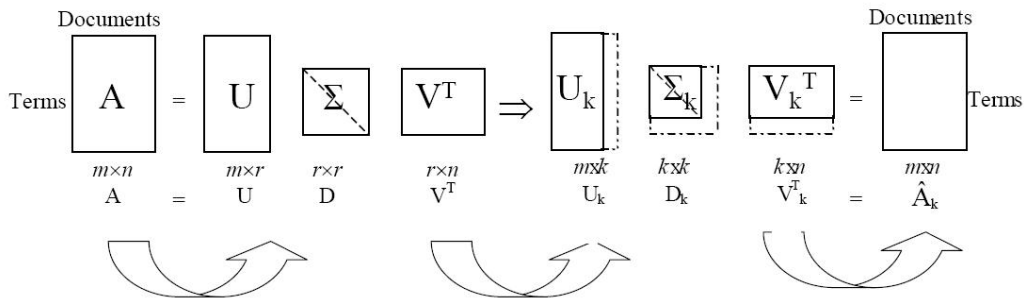


Figure: LSA の概念図

- ▶ 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 $\mathcal{V} = \{1, \dots, W\}$
- ▶ トピック集合 $\mathcal{T} = \{1, \dots, K\}$
 - ▶ 語彙やトピックをその添字と同一視している。
- ▶ 文書集合 $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_D\}$
- ▶ 文書 \mathbf{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 \mathbf{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 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

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

- ▶ $p(z_{d,i} = k)$ は、文書 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 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） 13 / 21

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、単語確率分布とトピックが一对一に対応している

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

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

潜在変数である $z_{d,i}$ を周辺化

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

各トークンの独立性の仮定より

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

各文書の独立性の仮定より

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

混合多項分布と PLSA の比較

- ▶ PLSA における \mathbf{x}_d の尤度

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

- ▶ 混合多項分布における \mathbf{x}_d の尤度

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

Jensen の不等式の適用

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

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

観測データの尤度の lower bound

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

最大化すべき目的関数は

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

PLSAのEMアルゴリズム(1/2)

M step

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

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

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

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

PLSA の EM アルゴリズム (2/2)

E step

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

$$\therefore q_{d,i,k} \propto \phi_{k,x_{d,i}} \theta_{d,k} \quad (16)$$

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

$x_{d,i} = x_{d,i'}$ 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

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

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

PLSAのEMアルゴリズムのまとめ

- ▶ 文書 x_d 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

$$q_{d,w,k} \leftarrow \frac{\phi_{k,w} \theta_{d,k}}{\sum_{k=1}^K \phi_{k,w} \theta_{d,k}} \quad (20)$$

- ▶ 文書 x_d 内の単語トークンがトピック t_k を表現する確率

$$\theta_{d,k} \leftarrow \frac{\sum_{w=1}^W n_{d,w} q_{d,w,k}}{\sum_{k=1}^K \sum_{w=1}^W n_{d,w} q_{d,w,k}} \quad (21)$$

- ▶ 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

$$\phi_{k,w} \leftarrow \frac{\sum_{d=1}^D n_{d,w} q_{d,w,k}}{\sum_{w=1}^W \sum_{d=1}^D n_{d,w} q_{d,w,k}} \quad (22)$$