Practical Text Analytics: Probabilistic Latent Semantic Analysis

Fan Dai

Iowa State University

October 29, 2019

Recall: Latent semantic analysis (LSA)

- LSA focuses on word co-occurrence and the meaning behind
- Truncated singular value decomposition (SVD)

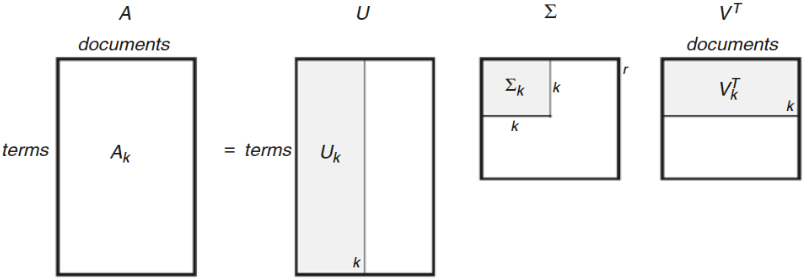


Figure: Truncated SVD process

Source: **Martin and Berry (2007)**

LSA and topics

- Truncated singular value decomposition (SVD)
  - columns (singular vectors) in value-weighted term matrix $U_k\Sigma_k$:
    - a linear combination of words

LSA and topics

- Truncated singular value decomposition (SVD)
  - columns (singular vectors) in value-weighted term matrix $U_k\Sigma_k$:
    - a linear combination of words
    - ***topics*** (latent) in the documents

      Example: weighted term matrix with $k = 2$ (column normalized)

Word	S1	S2
dog	0.54	0.02
cat	0.40	0.01
apple	0.03	0.22
blueberry	0.02	0.4
orange	0.01	0.35

* topic 1: animal
* topic 2: fruit

LSA and topics

- Truncated singular value decomposition (SVD)
  - columns (singular vectors) in value-weighted term matrix $U_k\Sigma_k$:
    - a linear combination of words
    - ***topics*** (latent) in the documents

      Example: weighted term matrix with $k = 2$ (column normalized)

Word	S1	S2
dog	0.54	0.02
cat	0.40	0.01
apple	0.03	0.22
blueberry	0.02	0.4
orange	0.01	0.35

  * topic 1: animal

  * topic 2: fruit

    - rows in U_k: words sharing with similar topical content are expected be close in the semantic term space.

LSA and topics

- Truncated singular value decomposition (SVD)
  - columns (singular vectors) in value-weighted term matrix $U_k \Sigma_k$:
    - a linear combination of words
    - ***topics*** (latent) in the documents

      Example: weighted term matrix with $k = 2$ (column normalized)

Word	S1	S2
dog	0.54	0.02
cat	0.40	0.01
apple	0.03	0.22
blueberry	0.02	0.4
orange	0.01	0.35

* topic 1: animal
* topic 2: fruit

    - rows in U_k: words sharing with similar topical content are expected be close in the semantic term space.
  - rows in weighted document matrix $V_k \Sigma_k$: documents with similar topical content will be close in the semantic document space.

LSA and topics

- Strengths of LSA
  - compress the term-document matrix ($M \rightarrow A_k$)
  - easier and faster to implement

LSA and topics

- Strengths of LSA
  - compress the term-document matrix ($M \rightarrow A_k$)
  - easier and faster to implement
- Limitations of LSA
  - lacks a generative model
  - no statistical inference for the latent variables (topics)
  - number of topics k is determined by heuristic techniques

LSA and topics

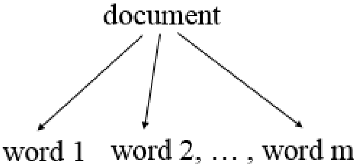
- Strengths of LSA
  - compress the term-document matrix ($M \rightarrow A_k$)
  - easier and faster to implement
- Limitations of LSA
  - lacks a generative model
  - no statistical inference for the latent variables (topics)
  - number of topics k is determined by heuristic techniques

Possible solution: **probabilistic latent semantic analysis (PLSA)**

PLSA: Generative model

What is the generation process?

Graphical model representations



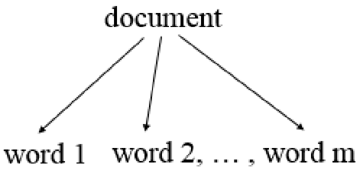
Observed variables:

- documents: $d_1, d_2, \ldots, d_n$
- words: $w_1, w_2, \ldots, w_m$

PLSA: Generative model

What is the generation process?

Graphical model representations



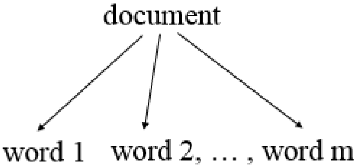
Observed variables:

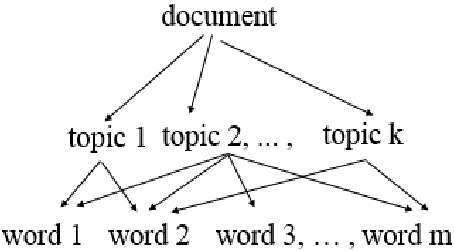
- documents: $d_1, d_2, \ldots, d_n$
- words: $w_1, w_2, \ldots, w_m$
- *some words may not appear*

PLSA: Generative model

What is the generation process?

Graphical model representations





Observed variables:

- documents: $d_1, d_2, \ldots, d_n$
- words: $w_1, w_2, \ldots, w_m$
- *some words may not appear*

Include the latent variables:

- topics: $t_1, t_2, \ldots, t_k$
- conditional independence assumption

PLSA: Generative model

The generative model can be summarized as: (Recall the formalization of topic models)

- for each document $d_i, i \in \{1, 2, \ldots, n\}$, suppose it contains $N_i <= m$ words, then for each word position $g \in \{1, 2, \ldots, N_i\}$

    1. choose a topic $t_l \sim \text{Multinomial}(\theta_{d_i})$
    2. choose a word $w_j \sim \text{Multinomial}(\phi_{t_l})$

    where,

    - $\theta_{d_i} = (p(t_1|d_i), p(t_2|d_i), \ldots, p(t_k|d_i)), \mathbf{1}'\theta_{d_i} = 1$
    - $\phi_{t_l} = (p(w_1|t_l), p(w_2|t_l), \ldots, p(w_m|t_l)), \mathbf{1}'\phi_{t_l} = 1$
    - $p(t_l|d_i)$: probability that topic t_l appears in document d_i
    - $p(w_j|t_l)$: probability that word w_j is chosen by topic t_l

PLSA: Likelihood function

- By conditional independence, $p(w_j|t_l, d_i) = p(w_j|t_l)$,
  so the probability function of a word w_j appearing at position g in document d_i is,

$$p(d_{i,g} = w_j|\Theta) = \sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i)$$

- the joint likelihood function for the whole text collection is,

$$f(\text{data}|\Theta) = \prod_{i=1}^{n} \prod_{g=1}^{N_i} \sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i)$$

PLSA: Likelihood function

- By conditional independence, $p(w_j|t_l, d_i) = p(w_j|t_l)$,
  so the probability function of a word w_j appearing at position g in document d_i is,

$$p(d_{i,g} = w_j|\Theta) = \sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i)$$

- the joint likelihood function for the whole text collection is,

$$\begin{aligned} f(\text{data}|\Theta) &= \prod_{i=1}^{n} \prod_{g=1}^{N_i} \sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m} \left[\sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i) \right]^{n(w_j,d_i)} \end{aligned}$$

  where $n(w_j, d_i)$ is the number of times term w_j appearing in document d_i (entries in the TDM/DTM).

PLSA: Prameter estimation

- Maximize $\log f(\text{data}|\Theta)$ with constraints $\mathbf{1}'\theta_{d_i} = 1$ and $\mathbf{1}'\phi_{t_l} = 1$ for $i = 1, 2, \ldots, n, l = 1, 2, \ldots, k$
  - Lagrange Multipliers

$$\arg\max \left\{ \log f(\text{data}|\Theta) + \sum_{i=1}^{n} \lambda_i [1 - \sum_{l=1}^{k} p(t_l|d_i)] + \sum_{l=1}^{k} \xi_l [1 - \sum_{j=1}^{m} p(w_j|t_l)] \right\}$$

where $\log f(\text{data}|\Theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} n(w_j, d_i) \log \left[\sum_{l=1}^{k} p(w_j|t_l) p(t_l|d_i) \right]$

  - difficult to directly optimize...

PLSA: Prameter estimation

- Maximize $\log f(\text{data}|\Theta)$ with constraints $\mathbf{1}'\theta_{d_i} = 1$ and $\mathbf{1}'\phi_{t_l} = 1$ for $i = 1, 2, \ldots, n, l = 1, 2, \ldots, k$
    - Lagrange Multipliers

$$\arg\max \left\{ \log f(\text{data}|\Theta) + \sum_{i=1}^{n} \lambda_i [1 - \sum_{l=1}^{k} p(t_l|d_i)] + \sum_{l=1}^{k} \xi_l [1 - \sum_{j=1}^{m} p(w_j|t_l)] \right\}$$

    where $\log f(\text{data}|\Theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} n(w_j, d_i) \log \left[\sum_{l=1}^{k} p(w_j|t_l) p(t_l|d_i) \right]$

        - difficult to directly optimize...

    - Expectation Maximization (EM) algorithm: *reference link*

PLSA: Prameter estimation

- Maximize $\log f(\text{data}|\Theta)$ with constraints $\mathbf{1}'\theta_{d_i} = 1$ and $\mathbf{1}'\phi_{t_l} = 1$ for $i = 1, 2, \ldots, n, l = 1, 2, \ldots, k$
  - Lagrange Multipliers

$$\arg\max\left\{\log f(\text{data}|\Theta) + \sum_{i=1}^{n} \lambda_i [1 - \sum_{l=1}^{k} p(t_l|d_i)] + \sum_{l=1}^{k} \xi_l [1 - \sum_{j=1}^{m} p(w_j|t_l)]\right\}$$

  where $\log f(\text{data}|\Theta) = \sum_{i=1}^{n} \sum_{j=1}^{m} n(w_j, d_i) \log\left[\sum_{l=1}^{k} p(w_j|t_l) p(t_l|d_i)\right]$

    - difficult to directly optimize...

  - Expectation Maximization (EM) algorithm: *reference link*
    - key idea: suppose for each word position in document d_i, the topic t_l is known
    - indicator variable (hidden) $r_{d_{i,g},l} = 1$, if topic t_l is chosen for word position g in document d_i.

PLSA: EM algorithm

- formalize the complete data log-likelihood with $\boldsymbol{r}$, the set of all the latent variables $r_{d_{i,g},l}$,

$$\log f(\text{data}|\Theta) = \log \prod_{i=1}^{n} \prod_{g=1}^{N_i} \sum_{l=1}^{k} p(w_j|t_l)\, p(t_l|d_i)$$

$$\log f(\text{data}|\boldsymbol{r},\Theta) = \log \prod_{i=1}^{n} \prod_{g=1}^{N_i} \prod_{l=1}^{k} \left[p(w_j|t_l)\, p(t_l|d_i)\right]^{r_{d_{i,g},l}}$$

$$= \sum_{i=1}^{n} \sum_{g=1}^{N_i} \sum_{l=1}^{k} r_{d_{i,g},l} \left[\log p(w_j|t_l) + \log p(t_l|d_i)\right]$$

PLSA: EM algorithm

- formalize the complete data log-likelihood with $\boldsymbol{r}$, the set of all the latent variables $r_{d_{i,g},l}$,

$$\log f(\text{data}|\Theta) = \log \prod_{i=1}^{n} \prod_{g=1}^{N_i} \sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i)$$

$$\log f(\text{data}|\boldsymbol{r},\Theta) = \log \prod_{i=1}^{n} \prod_{g=1}^{N_i} \prod_{l=1}^{k} \left[p(w_j|t_l)p(t_l|d_i)\right]^{r_{d_{i,g},l}}$$

$$= \sum_{i=1}^{n} \sum_{g=1}^{N_i} \sum_{l=1}^{k} r_{d_{i,g},l} \left[\log p(w_j|t_l) + \log p(t_l|d_i)\right]$$

- new objective function

$$\arg\max \left\{ \log f(\text{data}|\boldsymbol{r},\Theta) + \sum_{i=1}^{n} \lambda_i [1 - \sum_{l=1}^{k} p(t_l|d_i)] + \sum_{l=1}^{k} \xi_l [1 - \sum_{j=1}^{m} p(w_j|t_l)] \right\}$$

PLSA: EM algorithm

What is the value for $r_{d_{i,g},l}$?

- E-step in EM algorithm: compute the expected values of missing variables given the observed data and current parameters

$$
\begin{aligned}
\mathrm{E}[r_{d_{i,g},l} | data, \Theta] =& p(r_{d_{i,g},l} = 1 | data, \Theta) \\
=& \frac{p(r_{d_{i,g},l} = 1, data | \Theta)}{p(r_{d_{i,g},l} = 1 | \Theta)} \\
=& \cdots \\
=& \frac{p(d_{i,g} | t_l) p(t_l | d_i)}{\sum_{l=1}^{k} p(d_{i,g} | t_l) p(t_l | d_i)}
\end{aligned}
$$

  * work out $\cdots$: how do the data and parameters relate to $r_{d_{i,g},l}$? (*hint: graphical representations*)

PLSA: EM algorithm

- M-step in EM algorithm:

$$p(t_l|d_i) = \frac{\sum_{g=1}^{N_i} \mathrm{E}[r_{d_{i,g},l}|data,\Theta]}{N_i}$$

$$= \frac{\sum_{j=1}^{m} n(w_j, d_i) \frac{p(w_j|t_l)p(t_l|d_i)}{\sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i)}}{N_i}$$

$$p(w_j|t_l) = \frac{\sum_{i=1}^{n} \sum_{g=1}^{N_i} \mathrm{E}[r_{d_{i,g},l}|data,\Theta]\mathbf{I}(d_{i,g} = w_j)}{\sum_{j'=1}^{m} \sum_{i=1}^{n} \sum_{g=1}^{N_i} \mathrm{E}[r_{d_{i,g},l}|data,\Theta]\mathbf{I}(d_{i,g} = w_{j'})}$$

$$= \frac{\sum_{i=1}^{n} n(w_j, d_i) \frac{p(w_j|t_l)p(t_l|d_i)}{\sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i)}}{\sum_{j'=1}^{m} \sum_{i=1}^{n} n(w_{j'}, d_i) \frac{p(w_{j'}|t_l)p(t_l|d_i)}{\sum_{l=1}^{k} p(w_{j'}|t_l)p(t_l|d_i)}}$$

PLSA: Original formalism (Hofmann,1999)

The joint probability model of the observations (documents and words) with known topics

- parameters
  - $p(t_l)$: probability of topic t_l
  - $\boldsymbol{p(d_i|t_l)}$: probability of document d_i given topic t_l
  - $p(w_j|t_l)$: probability of word w_j given topic t_l

- by conditional probability, $p(d_i, w_j) = p(d_i)p(w_j|d_i)$ and $p(w_j, t_l|d_i) = p(w_j|t_l, d_i)p(t_l|d_i)$

PLSA: Original formalism (Hofmann,1999)

The joint probability model of the observations (documents and words) with known topics

- parameters
  - $p(t_l)$: probability of topic t_l
  - $\boldsymbol{p(d_i|t_l)}$: probability of document d_i given topic t_l
  - $p(w_j|t_l)$: probability of word w_j given topic t_l

- by conditional probability, $p(d_i, w_j) = p(d_i)p(w_j|d_i)$ and $p(w_j, t_l|d_i) = p(w_j|t_l, d_i)p(t_l|d_i)$

- by conditional independence,

$$\Rightarrow p(w_j|d_i) = \sum_{l=1}^{k} p(w_j, t_l|d_i) = \sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i)$$

PLSA: Original formalism (Hofmann,1999)

The joint probability model of the observations (documents and words) with known topics

- parameters
  - $p(t_l)$: probability of topic t_l
  - $\boldsymbol{p(d_i|t_l)}$: probability of document d_i given topic t_l
  - $p(w_j|t_l)$: probability of word w_j given topic t_l

- by conditional probability, $p(d_i, w_j) = p(d_i)p(w_j|d_i)$ and $p(w_j, t_l|d_i) = p(w_j|t_l, d_i)p(t_l|d_i)$

- by conditional independence,

$$\Rightarrow p(w_j|d_i) = \sum_{l=1}^{k} p(w_j, t_l|d_i) = \sum_{l=1}^{k} p(w_j|t_l)p(t_l|d_i)$$

- by reparameterization, $p(d_i, w_j) = \sum_{l=1}^{k} p(t_l)p(w_j|t_l)p(d_i|t_l)$

PLSA: Original formalism (Hofmann,1999)

- log-likelihood function

$$
\begin{aligned}
\ell &= \log \prod_{i=1}^{n} \prod_{j=1}^{m} p(d_i, w_j)^{n(w_j, d_i)} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} n(w_j, d_i) \log \sum_{l=1}^{k} p(t_l) p(w_j | t_l) p(d_i | t_l)
\end{aligned}
$$

PLSA: Original formalism (Hofmann,1999)

- log-likelihood function

$$\ell = \log \prod_{i=1}^{n} \prod_{j=1}^{m} p(d_i, w_j)^{n(w_j, d_i)}$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{m} n(w_j, d_i) \log \sum_{l=1}^{k} p(t_l) p(w_j|t_l) p(d_i|t_l)$$

- EM algorithm: introduce hidden variables $r(t_l, w_j, d_i) = 1$, if topic t_l is chosen to generated word w_J in document d_i

$$\ell_c = \sum_{i=1}^{n} \sum_{j=1}^{m} n(w_j, d_i) \sum_{l=1}^{k} r(t_l, w_j, d_i) \left[\log p(t_l) + \log p(w_j|t_l) + \log p(d_i|t_l)\right]$$

PLSA: Original formalism (Hofmann,1999)

- EM algorithm
  - E-step

$$p(t_l|w_j,d_i) = \frac{p(t_l)p(d_i|t_l)p(w_j|t_l)}{\sum_{l=1}^{k} p(t_{l'})p(d_i|t_{l'})p(w_j|t_{l'})}$$

  - M-step

$$p(t_l) \propto \sum_{i=1}^{n}\sum_{j=1}^{m} n(w_j,d_i)p(t_l|w_j,d_i)$$

$$p(w_j|t_l) \propto \sum_{i=1}^{n} n(w_j,d_i)p(t_l|w_j,d_i)$$

$$p(d_i|t_l) \propto \sum_{j=1}^{m} n(w_j,d_i)p(t_l|w_j,d_i)$$