Statistical Learning for Text Data Analytics
Text Clustering and Topic Models

Yangqiu Song

Hong Kong University of Science and Technology

yqsong@cse.ust.hk

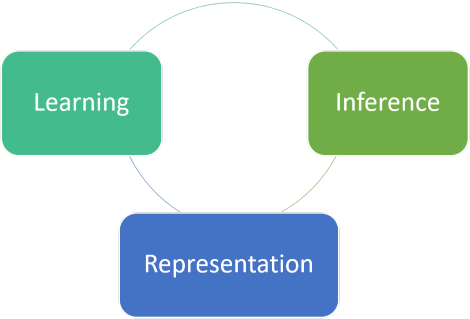
Spring 2018

∗Contents are based on materials created by Noah Smith, Xiaojin (Jerry) Zhu, Chengxiang Zhai

Reference Content

- Noah Smith. CSE 517: Natural Language Processing https://courses.cs.washington.edu/courses/cse517/16wi/
- Xiaojin (Jerry) Zhu. CS 769: Advanced Natural Language Processing. http://pages.cs.wisc.edu/~jerryzhu/cs769.html
- Chengxiang Zhai. CS598CXZ Advanced Topics in Information Retrieval. http://times.cs.uiuc.edu/course/598f16/

Course Topics



- Representation: language models, word embeddings, topic models
- Learning: supervised learning, semi-supervised learning, sequence models, deep learning, optimization techniques
- Inference: constraint modeling, joint inference, search algorithms

NLP applications: tasks introduced in Lecture 1

Overview

1. EM Algorithm
2. Language Models: Recap
3. Topic Models
4. Probabilistic Latent Semantic Analysis (PLSA)

Naive Bayes and Mixture Model

In naive Bayes, both y_m and $\mathbf{x}_m = (x_m^1, \ldots, x_m^V)^T$ are observed variables; $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_k$ are parameters

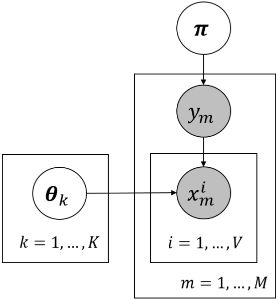


Figure: Native Bayes

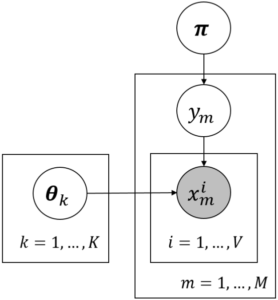


Figure: Mixture Model

However, in clustering problems, y_m is not observed (labeled before feeding into machine learning algorithm)

Expectation Maximization (EM) Algorithm

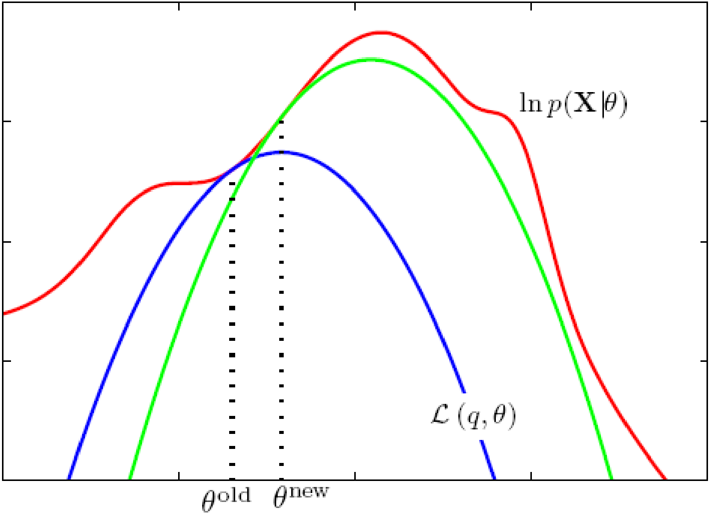
- EM might look like a heuristic method. However, it is not.
- EM is guaranteed to find a local optimum of data log likelihood
- Recall if we have complete data set $\{\mathbf{x}_m, y_m\}_{m=1}^{M}$ and denote parameter set as $\Theta = \{\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}\}$, the likelihood estimation of native Bayes is

$$\mathcal{J}_{NB}(\Theta) = \log \prod_{m=1}^{M} P_{\boldsymbol{\pi},\{\boldsymbol{\theta}_k\}}(\mathbf{x}_m, y_m) = \log P(\{\mathbf{x}_m, y_m\}_{m=1}^{M} | \Theta)$$

- However, now $\{y_m\}_{m=1}^{M}$ are not observed (labeled), so we treat them as hidden variables
- We instead maximize the marginal log likelihood:

$$\mathcal{J}(\Theta) = \log P(\{\mathbf{x}_m\}_{m=1}^{M} | \Theta)$$

EM Algorithm: General Idea



Lower Bound $Q(\Theta, \Theta^t)$ (Cont'd)

- The lower bound is obtained via Jensens inequality (concavity of log function)

$$\log \sum_i P_i f_i(x) \geq \sum_i P_i \log f_i(x)$$

which holds if the p_i's form a probability distribution

- Then the lower bound can be derived:

$$\begin{aligned}
\mathcal{J}(\Theta^t) &= \sum_{m=1}^{M} \log \sum_{y=1}^{K} P(\mathbf{x}_m, y|\Theta^t) \\
&= \sum_{m=1}^{M} \log \sum_{y=1}^{K} q_{\mathbf{x}_m,y}(\Theta) \frac{P(\mathbf{x}_m, y|\Theta^t)}{q_{\mathbf{x}_m,y}(\Theta)} \\
&\geq \sum_{m=1}^{M} \sum_{y=1}^{K} q_{\mathbf{x}_m,y}(\Theta) \log \frac{P(\mathbf{x}_m, y|\Theta^t)}{q_{\mathbf{x}_m,y}(\Theta)} \\
&\doteq Q(\Theta, \Theta^t)
\end{aligned}$$

where $\sum_{y=1}^{K} q_{\mathbf{x}_m,y}(\Theta) = 1$ is some distribution

E-step

$$\sum_{m=1}^{M} \log \sum_{y=1}^{K} q_{\mathbf{x}_m,y}(\Theta) \frac{P(\mathbf{x}_m, y|\Theta^t)}{q_{\mathbf{x}_m,y}(\Theta)} \geq \sum_{m=1}^{M} \sum_{y=1}^{K} q_{\mathbf{x}_m,y}(\Theta) \log \frac{P(\mathbf{x}_m, y|\Theta^t)}{q_{\mathbf{x}_m,y}(\Theta)}$$

- To make the bound tight for a particular value of Θ, we need for the step involving Jensens inequality in our derivation above to hold with equality
- For this to be true, we know it is sufficient that the expectation be taken over a constant-valued random variable $\frac{P(\mathbf{x}_m,y|\Theta^t)}{q_{\mathbf{x}_m,y}(\Theta)} = c$
- This is easily done by choosing $q_{\mathbf{x}_m,y}(\Theta) \propto P(\mathbf{x}_m, y|\Theta^t)$
- Since $\sum_{y=1}^{K} q_{\mathbf{x}_m,y}(\Theta) = 1$, we have (considered as E-step)

$$q_{\mathbf{x}_m,y}(\Theta) = \frac{P(\mathbf{x}_m, y|\Theta^t)}{\sum_{y=1}^{K} P(\mathbf{x}_m, y|\Theta^t)} = P(y|\mathbf{x}_m, \Theta^t)$$

- The equation holds in the inequality iff $q_{\mathbf{x}_m,y} = P(y|\mathbf{x}_m, \Theta^t)$

M-step

- In M-step, we maximize the lower bound

$$
\begin{aligned}
Q(\Theta^t, \Theta) &= \sum_{m=1}^{M} \sum_{y=1}^{K} q_{\mathbf{x}_m,y} \log \frac{P(\mathbf{x}_m, y|\Theta)}{q_{\mathbf{x}_m,y}} \\
&= \sum_{m=1}^{M} \sum_{y=1}^{K} q_{\mathbf{x}_m,y} \log \frac{P(y_m|\boldsymbol{\pi}) P(\mathbf{x}_m|y_m, \boldsymbol{\theta}_{*|y_m})}{q_{\mathbf{x}_m,y}}
\end{aligned}
$$

- Now we can set the gradient of Q w.r.t. $\boldsymbol{\pi}$ and $\boldsymbol{\theta}_k$'s to zero and obtain a closed form solution

$$
\begin{aligned}
\pi_k &= \frac{\sum_m q_{\mathbf{x}_m,y}}{M} \\
\theta_k^j &= \frac{\sum_m q_{\mathbf{x}_m,y} x_m^j}{\sum_m \sum_{j=1}^{d} q_{\mathbf{x}_m,y} x_m^j}
\end{aligned}
$$

- Compared to naive Bayes:

$$
\begin{aligned}
\pi_k &= \frac{|\{y_m = k\}|}{M} \\
\theta_k^j &= \frac{\sum_{m, y_m = k} x_m^j}{\sum_{m, y_m = k} \sum_{j=1}^{d} x_m^j}
\end{aligned}
$$

Convergence of EM Algorithm

- E-step: With $q_{\mathbf{x}_m,y}(\Theta) = P(y|\mathbf{x}_m, \Theta^t)$, the equation holds, which leads
$$Q(\Theta^t, \Theta^t) = \mathcal{J}(\Theta^t)$$
- M-step: Since Θ^{t+1} maximizes $Q(\Theta^t, \Theta)$, we have
$$Q(\Theta^t, \Theta^{t+1}) \geq Q(\Theta^t, \Theta^t) = \mathcal{J}(\Theta^t)$$
- On the other hand, Q is lower bound of $\mathcal{J}$, we have:
$$\mathcal{J}(\Theta^{t+1}) \geq Q(\Theta^t, \Theta^{t+1}) \geq Q(\Theta^t, \Theta^t) = \mathcal{J}(\Theta^t)$$
- This shows EM algorithm always increase the objective function (log likelihood)
- By iterating, we arrive at a local maximum of it

A More General View of EM

- EM is general and applies to joint probability models whenever some random variables are missing
- EM is advantageous when the marginal is difficult to optimize, but the joint is easy
- To be general, consider a joint distribution $P(X, Z|\Theta)$, where X is the collection of observed variables, and Z unobserved variables
- The quantity we want to maximize is the marginal log likelihood

$$\mathcal{J}(\Theta) = \log P(X|\Theta) = \log \sum_{Z} P(X, Z|\Theta)$$

which we assume difficult to optimize

A More General View of EM (Cont'd)

- One can introduce an arbitrary distribution over hidden variables $Q(Z)$

$$
\begin{aligned}
\mathcal{J}(\Theta) &= \log P(X|\Theta) = \log \sum_Z P(X,Z|\Theta) \\
&= \sum_Z Q(Z) \log P(X|\Theta) \\
&= \sum_Z Q(Z) \log \frac{P(X|\Theta)\ \ Q(Z)\ \ P(X,Z|\Theta)}{P(X,Z|\Theta)\ \ Q(Z)} \\
&= \sum_Z Q(Z) \log \frac{P(X,Z|\Theta)}{Q(Z)} + \sum_Z Q(Z) \log \frac{P(X|\Theta) Q(Z)}{P(X,Z|\Theta)} \\
&= \sum_Z Q(Z) \log \frac{P(X,Z|\Theta)}{Q(Z)} + \sum_Z Q(Z) \log \frac{Q(Z)}{P(Z|X,\Theta)} \\
&= F(Q,\Theta) + KL[Q(Z)||P(Z|X,\Theta)]
\end{aligned}
$$

- Note $F(Q,\Theta)$ is the right hand side of Jensen's inequality
  - If $KL > 0$, $F(Q,\Theta)$ is a lower bound of $\mathcal{J}(\Theta)$
- First consider the maximization of F on Q with Θ^t fixed
  - $F(Q,\Theta)$ is maximized by $Q(Z) = P(Z|X,\Theta^t)$ since $\mathcal{J}(\Theta)$ is fixed and KL attends its minimum zero (E-Step)
- Next consider the maximization of F on Θ with Q fixed as above
  - Note in this case $F(Q,\Theta) = Q(\Theta^t,\Theta)$ (M-Step)

Illustration of EM

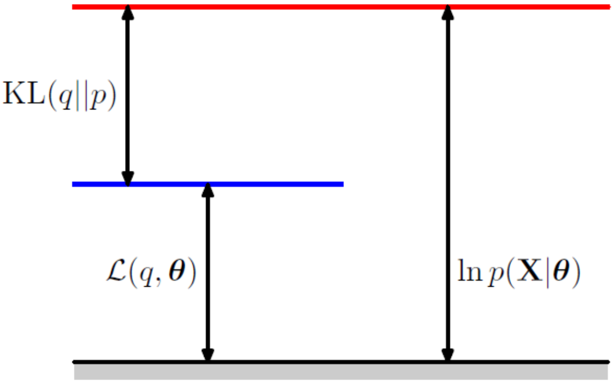


Figure: EM Algorithm

Illustration of EM

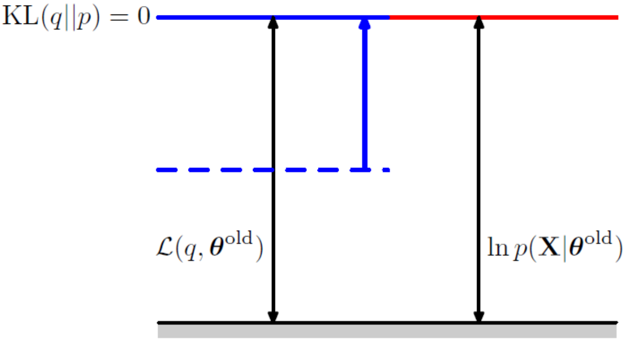


Figure: EM Algorithm

Illustration of EM

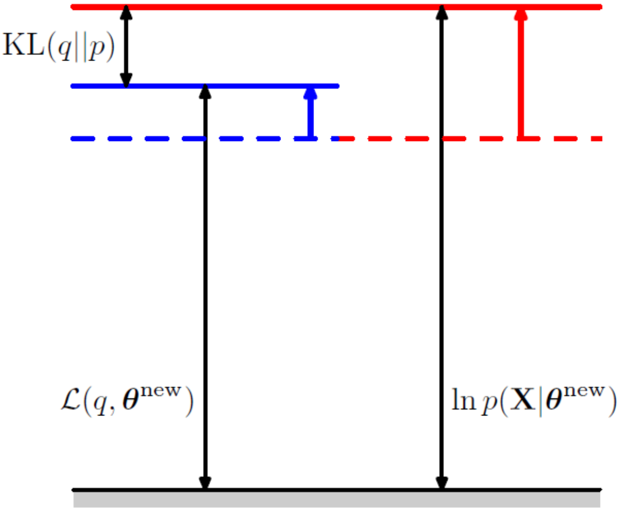


Figure: EM Algorithm

Variations of EM

- Generalized EM (GEM) finds Θ that improves, but not necessarily maximizes, $F(Q, \Theta) = Q(\Theta, \Theta^t)$ in the M-step. This is useful when the exact M-step is difficult to carry out. Since this is still coordinate ascent, GEM can find a local optimum.
- Stochastic EM: The E-step is computed with Monte Carlo sampling. This introduces randomness into the optimization, but asymptotically it will converge to a local optimum.
- Variational EM: $Q(Z)$ is restricted to some easy-to-compute subset of distributions, for example the fully factorized distributions $Q(Z) = \prod_i Q(z_i)$. In general $P(Z|X, \Theta)$, which might be intractable to compute, will not be in this subset. There is no longer guarantee that variational EM will find a local optimum.

Overview

Paragraphs of Text

- A language model is a probability distribution over $\mathcal{V}^\dagger$
- Typically P decomposes into probabilities $P(x_i|\mathbf{h}_i)$
  - We considered n-gram, log-linear, and neural language models, etc.
- Today: probabilistic models that relate a word and its cotext (the linguistic environment of the word)
- This might help us learn to represent words, contexts, or both

Three Kinds of Cotext

If we consider a word token at a particular position i in text to be the observed value of a random variable X_i, what other random variables are predictive of/related to X_i?

- The words that occur within a small "window" around i (e.g., $x_{i-2}, x_{i-1}, x_{i+1}, x_{i+2}$, or maybe the sentence containing i) $\rightarrow$ distributional semantics
- The document containing i (a moderate-to-large collection of other words) $\rightarrow$ topic models
- A sentence known to be a translation of the one containing i $\rightarrow$ translation models

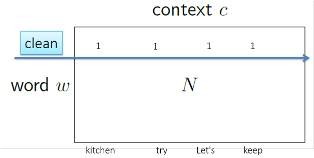
Overview

Topic Models

- Words are not independent and identically distributed (i.i.d.)!
  - Predictable given history: n-gram/Markov models
  - Predictable given other words in the document: topic models
- Let $Z = \{1, \ldots, k\}$ be a set of "topics" or "themes" that will help us capture the interdependence of words in a document
  - Usually these are not named or characterized in advance; they are just k different values with no a priori meaning

The Term-Document Matrix

- Let $\mathbf{A} \in \mathbb{R}^{V \times M}$ contain statistics of association between words in $\mathcal{V}$ and M documents. N is the total number of word tokens.
- Comparison of contexts
  - Local context (Let's try to keep the kitchen clean.)



  - Document-level context ($[\mathbf{A}]_{v,d} = c_{\mathbf{x}_d}(v)$)
    - d1: "yes, we have no bananas"
    - d2: "say yes for bananas"
    - d3: "no bananas , we say"

	1	2	3
,	1	0	1
bananas	1	1	1
for	0	1	0
have	1	0	0
no	1	0	1
say	0	1	1
we	1	0	1
yes	1	1	0

Association Score

- What we really want here is some way to get at "surprise"
- One way to think about this is, is the occurrence of word v in document d surprisingly high (or low), given what we'd expect due to chance?
- Chance would be $\frac{c_{\mathbf{x}_{1:M}}(v)}{N}$ words out of the $len(d)$ (length of document) words in document d
- Intuition: consider the ratio of observed frequency $c_{\mathbf{x}_d}(v)$ to "chance" under independence $\frac{c_{\mathbf{x}_{1:M}}(v)}{N} \cdot len(d)$

Pointwise Mutual Information

- A common starting point is positive pointwise mutual information:

$$[\mathbf{A}]_{v,d} = \left[\log \frac{c_{\mathbf{x}_d}(v)}{\frac{c_{\mathbf{x}_{1:M}}(v)}{N} \cdot len(d)}\right]_+ = \left[\log \cdot \frac{N \cdot c_{\mathbf{x}_d}(v)}{c_{\mathbf{x}_{1:M}}(v) \cdot len(d)}\right]_+$$

- For our problem
  - $[\mathbf{A}]_{banana,d1} = \log \frac{15 \cdot 1}{3 \cdot 6} \approx -0.18 \rightarrow 0$
  - $[\mathbf{A}]_{for,d2} = \log \frac{15 \cdot 1}{1 \cdot 4} \approx 0.32$

	1	2	3
,	1	0	1
bananas	1	1	1
for	0	1	0
have	1	0	0
no	1	0	1
say	0	1	1
we	1	0	1
yes	1	1	0

A Nod to Information Theory

- Pointwise mutual information for two random variables A and B:

$$\begin{aligned}\mathrm{PMI}(a,b) &= \log \frac{P(A=a,B=b)}{P(A=a)\cdot P(B=b)} \\ &= \log \frac{P(A=a|B=b)}{P(A=a)} \\ &= \log \frac{P(B=b|A=a)}{P(B=b)}\end{aligned}$$

- The average mutual information is given by

$$\mathrm{MI}(A,B)) = \sum_{a,b} P(A=a,B=b) \log \frac{P(A=a,B=b)}{P(A=a)\cdot P(B=b)}$$

This comes from information theory; it is the amount of information each r.v. offers about the other.

Pointwise Mutual Information

$$[\mathbf{A}]_{v,d} = \left[\log \frac{c_{\mathbf{x}_d}(v)}{\frac{c_{\mathbf{x}_{1:M}}(v)}{N} \cdot len(d)} \right]_+ = \left[\log \cdot \frac{N \cdot c_{\mathbf{x}_d}(v)}{c_{\mathbf{x}_{1:M}}(v) \cdot len(d)} \right]_+$$

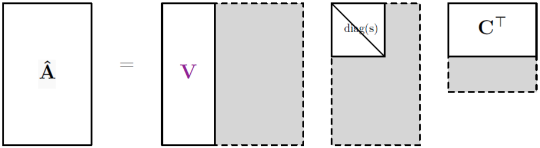
- If a word v appears with nearly the same frequency in every document, its row $[\mathbf{A}]_{v,\cdot}$ will be all nearly zero ($\approx \log$).
- If a word v occurs only in document d, PMI will be large and positive.
- PMI is very sensitive to rare occurrences; usually we smooth the frequencies and filter rare words.
- One way to think about PMI: it's telling us where a unigram model is most wrong.
- We could use $\mathbf{A}$ as feature representation of documents,

Topic Models: Latent Semantic Indexing/Analysis (Deerwester et al. (1990))

- LSI/A seeks to solve:

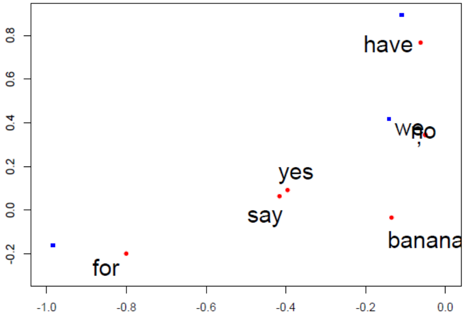
$$\underset{V\times M}{\mathbf{A}} \approx \underset{V\times d}{\mathbf{V}} \times \underset{d\times d}{\mathrm{diag}(\mathbf{s})} \times \underset{d\times M}{\mathbf{C}}^{\top}$$

  where **V** contains embeddings of words and **C** contains embeddings of documents
- This can be solved by applying singular value decomposition to **A**



LSI/A Example

- $d = 2$: Words and documents in two dimensions.



	1	2	3
,	1	0	1
bananas	1	1	1
for	0	1	0
have	1	0	0
no	1	0	1
say	0	1	1
we	1	0	1
yes	1	1	0

Note how "no", "we", and "," are all in the exact same spot. Why?

Understanding LSI/A

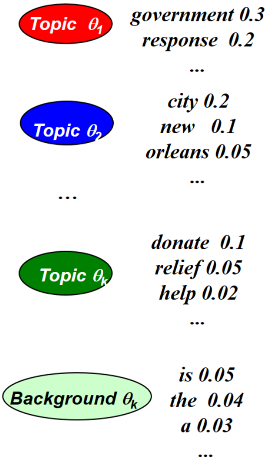
- Mapping words and documents into the same *d*-dimensional space.
- Bag of words assumption (Salton et al. (1975)): a document is nothing more than the distribution of words it contains.
- Distributional hypothesis (Harris (1954); Firth (1957)): words are nothing more than the distribution of contexts (here, documents) they occur in. Words that occur in similar contexts have similar meanings.
- $\mathbf{A}$ is sparse and noisy; LSI/A "fills in" the zeroes and tries to eliminate the noise.

Probabilistic Topic Models

- LSI/A: assumes the elements of $\mathbf{A}$ are the result of Gaussian noise.
- Probabilistic Latent Semantic Analysis (PLSA) (Hofmann (1999)) model the probability distribution $p(\mathbf{x}_d|d)$
  - This is a particular kind of conditional language model
- Latent Dirichlet Allocation (Blei et al. (2003))
  - Introduce Bayesian inference to PLSA

Overview

Document as a Sample of Mixed Topics



[Criticism of government response to the hurricane primarily consisted of criticism of its response to the approach of the storm and its aftermath, specifically in the delayed response] to the [flooding of New Orleans. … 80% of the 1.3 million residents of the greater New Orleans metropolitan area evacuated] …[Over seventy countries pledged monetary donations or other assistance]. …

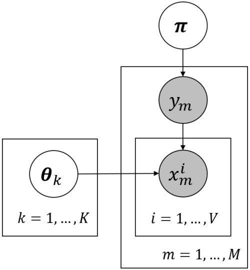
Probabilistic Topic Models

- As a language model, LSI/A is kind of broken.
  - It assumes the elements of $\mathbf{A}$ are the result of Gaussian noise.
- Hofmann (1999) proposed to model the probability distribution $P(d, w)$ based on each topic of a word w in a document d
  - This is a particular kind of conditional language model.

Mixture Models

- Recall naive Bayes based mixture models for a document collection by K topics (classes)
- Each topic is a multinomial over words, and each document is generated from a single topic

Naive Bayes from Class Conditional Unigram Model

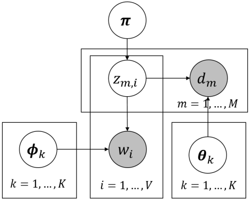


- For $m = 1, \dots, M$
  - Choose $y_m \sim Multinomial(y_m|1, \boldsymbol{\pi})$
  - Choose $N_m = \sum_j^d x_m^j \sim Poisson(\xi)$
  - For $n = 1, \dots, N_m$
    - Choose $v \sim Multinomial(v|1, \boldsymbol{\theta}_{*|y_m}) = \prod_{j=1}^d (\theta_{*|y_m}^j)^{v=j}$

It assumes "one document, one topic."

Probabilistic Latent Semantic Analysis (PLSA)

- PLSA assumes that each document d (with word vector w) is generated from all topics, with documentspecific topic weights.



- Choose a $z_{m,i} = k$ from topic distribution $\boldsymbol{\pi}$
- Choose a document from $d_m \sim Multinomial(d_m|1, \boldsymbol{\theta}_k)$
- Choose a word w_i from $w_i \sim Multinomial(w_i|1, \phi_k)$
- Add one count of word w_i to document d_m
- Repeat until we generate the document-word matrix

Under this process, the probability of picking the corpus is:

$$\begin{aligned} P(\mathcal{D}, \mathcal{W}) &= \prod_{m=1}^{M} \prod_{i=1}^{N_m} \sum_{k=1}^{K} P(z_{m,i} = k) P(d_m|\boldsymbol{\theta}_k) P(w_i|\phi_k) \\ &= \prod_{m=1}^{M} \prod_{i=1}^{V} \left(\sum_{k=1}^{K} P(z_{m,i} = k) P(d_m|\boldsymbol{\theta}_k) P(w_i|\phi_k) \right)^{c_{d_m}(w_i)} \end{aligned}$$

Maximize Log Likelihood

- Log likelihood:

$$P(\mathcal{D}, \mathcal{W}) = \prod_{m=1}^{M} \prod_{i=1}^{V} \left(\sum_{k=1}^{K} P(z_{m,i} = k) P(d_m|\boldsymbol{\theta}_k) P(w_i|\phi_k) \right)^{c_{d_m}(w_i)}$$

- To reduce the notation complexity, we denote:

$$\log P(\mathcal{D}, \mathcal{W}) = \sum_{d=1}^{M} \sum_{w=1}^{V} c_d(w) \log \left(\sum_{k=1}^{K} P(z) P(d|z) P(w|z) \right)$$

- We denote the parameters as
  $\Theta = \{\boldsymbol{\pi}, \phi_k, \boldsymbol{\theta}_k, k = 1, \ldots, K\} = \{P(z), P(d|z), P(w|z)\}$
- Note here z is a hidden variable, and note that the sum is inside the log
- We can apply EM algorithm to maximize the likelihood

Lower Bound and E-Step

- Remember Jensens inequality

$$\log \sum_i P_i f_i(x) \geq \sum_i P_i \log f_i(x)$$

- We first compute the lower bound of the log likelihood:

$$\begin{aligned}
& \log P(\mathcal{D}, \mathcal{W}) \\
= \; & \sum_{d=1}^{M} \sum_{w=1}^{V} c_d(w) \log \left(\sum_{k=1}^{K} P(z)P(d|z)P(w|z) \right) \\
= \; & \sum_{d=1}^{M} \sum_{w=1}^{V} c_d(w) \log \left(\sum_{k=1}^{K} q_{z,d,w}(\Theta) \frac{P(z)P(d|z)P(w|z)}{q_{z,d,w}(\Theta)} \right) \\
\geq \; & \sum_{d=1}^{M} \sum_{w=1}^{V} c_d(w) \sum_{k=1}^{K} q_{z,d,w}(\Theta) \left(\log \frac{P(z)P(d|z)P(w|z)}{q_{z,d,w}(\Theta)} \right)
\end{aligned}$$

- Note Jensen's inequality involves computing $q_{z,d,w}(\Theta) = P(z|d, w, \Theta^t)$, which computes the probability of topics separately for each cell, under the current parameters Θ^t
- This is exactly the E-step:

$$P(z|d, w, \Theta^t) \propto P(z|\Theta^t)P(d|z, \Theta^t)P(w|z, \Theta^t)$$

M-Step

$$
\begin{aligned}
& \log P(\mathcal{D}, \mathcal{W}) \\
= \;& \sum_{d=1}^{M} \sum_{w=1}^{V} c_d(w) \log \left(\sum_{k=1}^{K} P(z) P(d|z) P(w|z) \right) \\
= \;& \sum_{d=1}^{M} \sum_{w=1}^{V} c_d(w) \log \left(\sum_{k=1}^{K} P(z|d, w, \Theta^t) \frac{P(z)P(d|z)P(w|z)}{P(z|d,w,\Theta^t)} \right) \\
= \;& \sum_{d=1}^{M} \sum_{w=1}^{V} c_d(w) \sum_{k=1}^{K} P(z|d, w, \Theta^t) \left(\log \frac{P(z)P(d|z)P(w|z)}{P(z|d,w,\Theta^t)} \right)
\end{aligned}
$$

- Maximizing the right of the above inequality by setting the gradient to zero amounts to the M-step, which gives
  - $P(z) \propto \sum_d \sum_w c_d(w) P(z|d, w, \Theta^t)$
  - $P(d|z) \propto \sum_w c_d(w) P(z|d, w, \Theta^t)$
  - $P(w|z) \propto \sum_d c_d(w) P(z|d, w, \Theta^t)$

Illustration of EM

c(w, d)

d_1

aid	7
price	5
oil	6

d_2

aid	8
price	7
oil	5

$P(w\mid \theta)$	*Topic 1*	*Topic 2*
aid		
price	Initial value	
oil		

$\pi_{d1,1}$ $(P(\theta_1|d_1))$ $\pi_{d1,2}$ $(P(\theta_2|d_1))$ **Topic coverage**

Initial value

$1-\lambda_B$ λ_B

$\pi_{d2,1}$ $(P(\theta_1|d_2))$ $\pi_{d2,2}$ $(P(\theta_2|d_2))$

Initial value

$1-\lambda_B$ λ_B

Initializing $\pi_{d,j}$ and $P(w|\theta_j)$ with random values

Illustration of EM

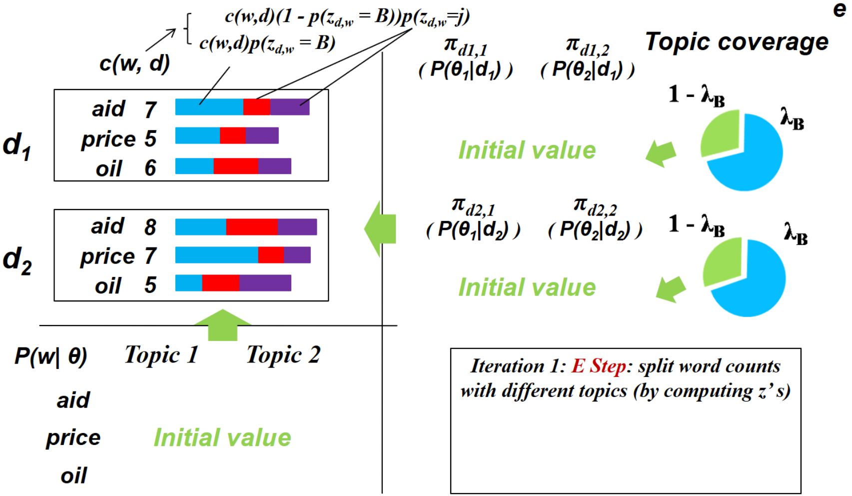


Illustration of EM

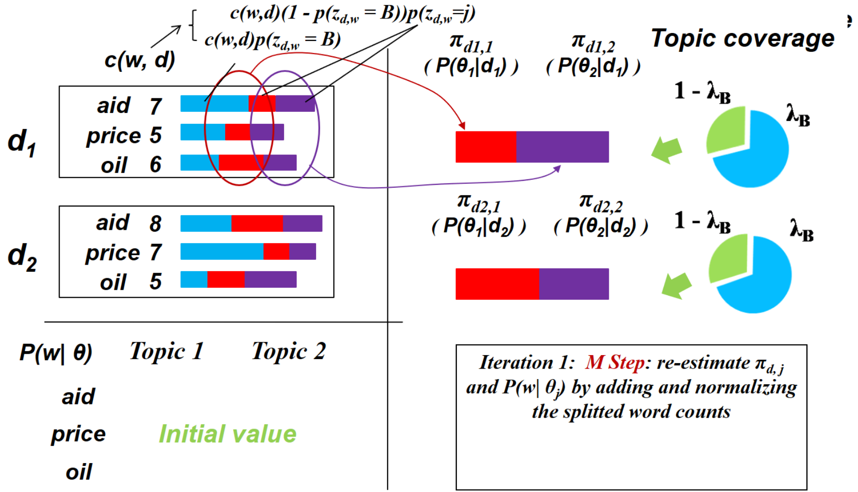
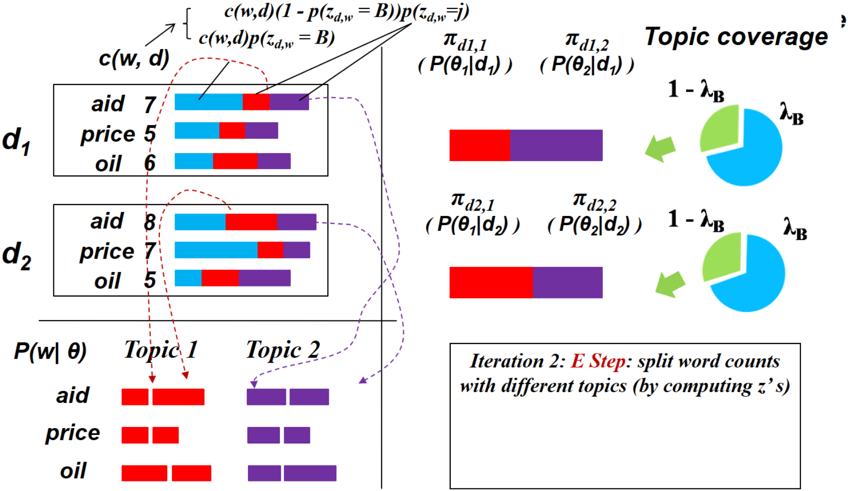
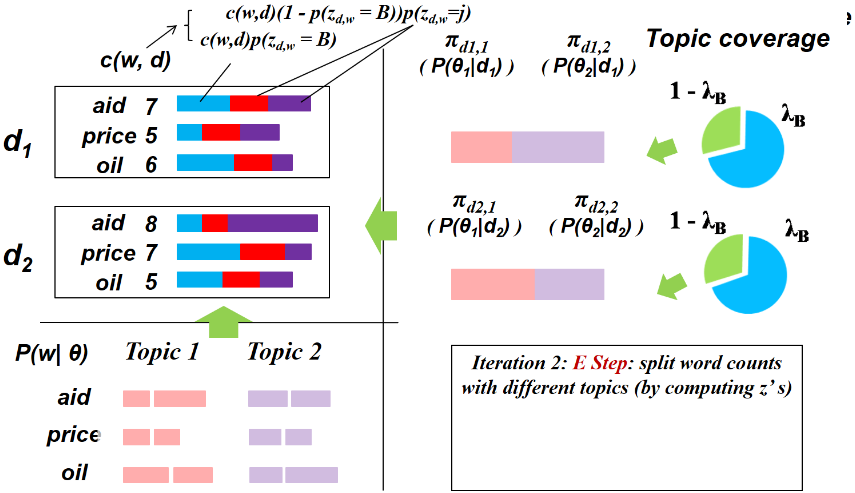


Illustration of EM



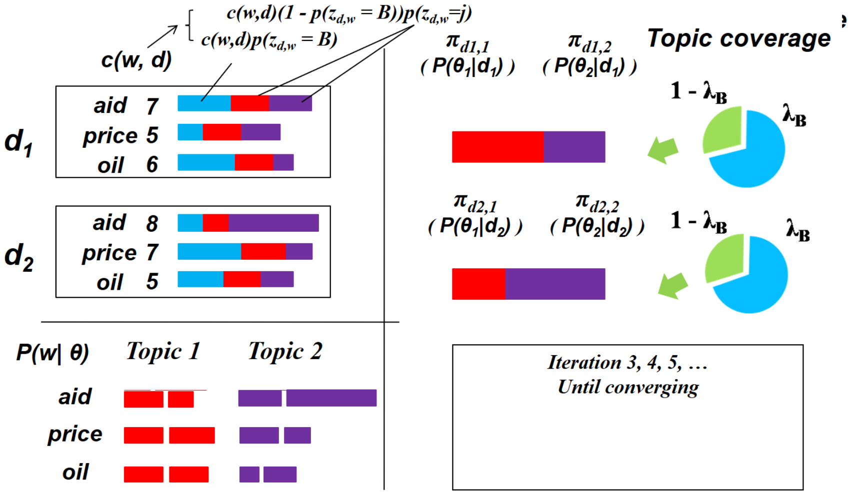
Iteration 2: E Step: split word counts with different topics (by computing z's)

Illustration of EM



Iteration 2: E Step: split word counts with different topics (by computing z's)

Illustration of EM



Use of PLSA

- Once the model is trained, we can look at it in the following way
  - $P(w|z)$ are the topics. Each topic is defined by a word multinomial. Often people find that the topics seem to have distinct semantic meanings.
  - From $P(d|z)$ and $P(z)$, we can compute $P(z|d) \propto p(d|z)p(z)$. $P(z|d)$ is the topic wights for document d.
- One drawback of PLSA is that it is transductive in nature. That is, there is no easy way to handle a new document that is not already in the collection

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Example of topics found from a Science Magazine papers collection

universe	0.0439	drug	0.0672	cells	0.0675	sequence	0.0818	years	0.156
galaxies	0.0375	patients	0.0493	stem	0.0478	sequences	0.0493	million	0.0556
clusters	0.0279	drugs	0.0444	human	0.0421	genome	0.033	ago	0.045
matter	0.0233	clinical	0.0346	cell	0.0309	dna	0.0257	time	0.0317
galaxy	0.0232	treatment	0.028	gene	0.025	sequencing	0.0172	age	0.0243
cluster	0.0214	trials	0.0277	tissue	0.0185	map	0.0123	year	0.024
cosmic	0.0137	therapy	0.0213	cloning	0.0169	genes	0.0122	record	0.0238
dark	0.0131	trial	0.0164	transfer	0.0155	chromosome	0.0119	early	0.0233
light	0.0109	disease	0.0157	blood	0.0113	regions	0.0119	billion	0.0177
density	0.01	medical	0.00997	embryos	0.0111	human	0.0111	history	0.0148

bacteria	0.0983	male	0.0558	theory	0.0811	immune	0.0909	stars	0.0524
bacterial	0.0561	females	0.0541	physics	0.0782	response	0.0375	star	0.0458
resistance	0.0431	female	0.0529	physicists	0.0146	system	0.0358	astrophys	0.0237
coli	0.0381	males	0.0477	einstein	0.0142	responses	0.0322	mass	0.021
strains	0.025	sex	0.0339	university	0.013	antigen	0.0263	disk	0.0173
microbiol	0.0214	reproductive	0.0172	gravity	0.013	antigens	0.0184	black	0.0161
microbial	0.0196	offspring	0.0168	black	0.0127	immunity	0.0176	gas	0.0149
strain	0.0165	sexual	0.0166	theories	0.01	immunology	0.0145	stellar	0.0127
salmonella	0.0163	reproduction	0.0143	aps	0.00987	antibody	0.014	astron	0.0125
resistant	0.0145	eggs	0.0138	matter	0.00954	autoimmune	0.0128	hole	0.00824

Remarks

- Like LSI/A, PLSA "squeezes" the relationship between words and contexts (documents) through topics.
- A document is now characterized as a mixture of corpus-universal topics (each of which is a unigram model).
- Topic mixtures can be incorporated into language models; see Iyer and Ostendorf (1999), for example.
- Compared to LSI/A: PLSA is more interpretable (e.g., LSI/A can give negative values!).
- PLSA cannot assign probability to a text not in $\mathcal{W}$; it only defines conditional distributions over words given texts in $\mathcal{W}$.
- The next model overcomes this problem by adding another level of randomness: $P(z|d)$ becomes a random variable, not a parameter.

References I

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., and Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science (JASIS)*, 41(6):391–407.

Firth, J. R. (1957). *A synopsis of linguistic theory 1930-55.*, volume 1952-59, pages 1–32. The Philological Society, Oxford.

Harris, Z. (1954). Distributional structure. *Word*, 10(23):146–162.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *UAI*, pages 289–296.

Salton, G., Wong, A., and Yang, C. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.