Latent Dirichlet Allocation (LDA)

Following slides borrowed (with heavily modification) from:
Jonathan Huang (jch1@cs.cmu.edu)

Naïve Bayes: each doc is on a topic

Model: For each document:
- Choose a topic $z_d$ with $p(topic_i) = \theta$
- Choose $N$ words $w_n$ by drawing each one independently from a multinomial conditioned on $z_d$ with $p(w_n = word_j | topic_i = z) = \beta_z$
  - *Multinomial*: take a (non-uniform prior) dice with a word on each side; roll the dice $N$ times and count how often each word comes up

In NB, we have exactly one topic per document
LDA: Each doc is a mixture of topics

- **LDA**: each document is a (different) mixture of topics
  - Naïve Bayes assumes each *document* is on a single topic
  - LDA lets each *word* be on a different topic
  - For each document, \(d\):
    - Choose a *multinomial distribution* \(\theta_d\) over topics for that document
    - For each of the \(N\) words \(w_n\) in the document
      - Choose a topic \(z_n\) with \(p(topic) = \theta_d\)
      - Choose a word \(w_n\) from a multinomial conditioned on \(z_n\) with \(p(w=w_n|\text{topic}=z_n)\)
      - Note: each topic has a different probability of generating each word
Dirichlet Distributions

- In the LDA model, we want the *topic mixture proportions* for each document to be drawn from some *distribution*.
  - *distribution* = “probability distribution”, so it sums to one
- So, we want to put a prior distribution on multinomials. That is, k-tuples of non-negative numbers that sum to one.
  - We want probabilities of probabilities
  - These multinomials lie in a *(k-1)*-simplex
    - *Simplex* = generalization of a triangle to *(k-1)* dimensions.
- Our prior:
  - Defined for a *(k-1)*-simplex.
  - Conjugate to the multinomial
3 Dirichlet Examples (over 3 topics)

Corners: only one topic
Center: uniform mixture of topics
Colors indicate probability of seeing the topic distribution
Dirichlet Distribution

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

**Dirichlet distribution**

- is defined over a \((k-1)\)-simplex. I.e., it takes \(k\) non-negative arguments which sum to one.
- is the conjugate prior to the multinomial distribution.
  - I.e. if our likelihood is multinomial with a Dirichlet prior, then the posterior is also Dirichlet
- The Dirichlet parameter \(\alpha_i\) can be thought of as the prior count of the \(i^{th}\) class.
The LDA Model

For each document,
- Choose the topic distribution $\theta \sim \text{Dirichlet}(\alpha)$
- For each of the N words $w_n$:
  - Choose a topic $z \sim \text{Multinomial}(\theta)$
  - Choose a word $w_n \sim \text{Multinomial}(\beta_z)$
    - Where each topic has a different parameter vector $\beta$ for the words
The LDA Model: “Plate representation”

For each of M documents,
- Choose the topic distribution \( \theta \sim \text{Dirichlet}(\alpha) \)
- For each of the N words \( w_n \):
  - Choose a topic \( z \sim \text{Multinomial}(\theta) \)
  - Choose a word \( w_n \sim \text{Multinomial}(\beta_z) \)
Parameter Estimation

- Given a corpus of documents, find the parameters $\alpha$ and $\beta$ which maximize the likelihood of the observed data (words in documents), marginalizing over the hidden variables $\theta$, $z$.

- **E-step:**
  - Compute $p(\theta, z|w, \alpha, \beta)$, the posterior of the hidden variables $(\theta, z)$ given each document $w$, and parameters $\alpha$ and $\beta$.

- **M-step**
  - Estimate parameters $\alpha$ and $\beta$ given the current hidden variable distribution estimates.

- **Unfortunately, the E-step cannot be solved in a closed form**
  - So people use a “variational” approximation.
Variational Inference

You don’t need to know the details; only what is hidden and what observed; and that EM works here.

In variational inference, we consider a simplified graphical model with variational parameters $\gamma$, $\phi$ and minimize the KL Divergence between the variational and posterior distributions.

- $q$ approximates $p$

\[
(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} KL(q(\theta, z | \gamma, \phi), p(\theta, z | w, \alpha, \beta))
\]
Parameter Estimation: Variational EM

- Given a corpus of documents, find the parameters $\alpha$ and $\beta$ which maximize the likelihood of the observed data.

- **E-step:**
  - Estimate the variational parameters $\gamma$ and $\phi$ in $q(\gamma, \phi; \alpha, \beta)$ by minimizing the KL-divergence to $p$ (with $\alpha$ and $\beta$ fixed)

- **M-step**
  - Maximize (over $\alpha$ and $\beta$) the lower bound on the log likelihood obtained using $q$ in place of $p$ (with $\gamma$ and $\phi$ fixed)

*You don’t need to know the details; only what is hidden and what observed; and that EM works here.*
<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW MILLION</td>
<td>CHILDREN SCHOOL</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FILM TAX</td>
<td>WOMEN STUDENTS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SHOW PROGRAM</td>
<td>PEOPLE SCHOOLS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUSIC BUDGET</td>
<td>CHILD EDUCATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOVIE BILLION</td>
<td>YEARS TEACHERS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLAY FEDERAL</td>
<td>FAMILIES HIGH</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MUSICAL YEAR</td>
<td>WORK PUBLIC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BEST SPENDING</td>
<td>PARENTS TEACHER</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTOR NEW</td>
<td>SAYS BENNETT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIRST STATE</td>
<td>FAMILY MANIGAT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YORK PLAN</td>
<td>WELFARE NAMPHY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OPERA MONEY</td>
<td>MEN STATE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>THEATER PROGRAMS</td>
<td>PERCENT PRESIDENT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTRESS GOVERNMENT</td>
<td>CARE ELEMENTARY</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LOVE CONGRESS</td>
<td>LIFE HAITI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.
LDA topics can be used for semi-supervised learning
LDA requires fewer topics than NB

perplexity = $2^H(p)$ per word

I.e., $\log_2(\text{perplexity}) = H$
There are many LDA extensions

The author-topic model

You don’t need to know the details; only that there are different hidden and observed features; and that EM works here.
Ailment Topic Aspect Model

- Set the background switching binomial $\lambda$
- Draw an ailment distribution $\eta \sim \text{Dir}(\sigma)$
- Draw word multinomials $\phi \sim \text{Dir}(\beta)$ for the topic, ailment, and background distributions
- For each message $1 \leq m \leq D$:
  - Draw a switching distribution $\pi \sim \text{Beta}(\gamma_0, \gamma_1)$
  - Draw an ailment $a \sim \text{Mult}(\eta)$
  - Draw a topic distribution $\theta \sim \text{Dir}(\alpha_a)$
  - For each word $w_i \in N_m$:
    - Draw aspect $y_i \in \{0, 1, 2\}$ (observed)
    - Draw background switcher $\ell \in \{0, 1\} \sim \text{Bi}(\lambda)$
    - If $\ell == 0$:
      - Draw $w_i \sim \text{Mult}(\phi_{B,y})$ (a background)
    - Else:
      - Draw $x_i \in \{0, 1\} \sim \text{Bi}(\pi)$
      - If $x_i == 0$: (draw word from topic $z$)
        - Draw topic $z_i \sim \text{Mult}(\theta)$
        - Draw $w_i \sim \text{Mult}(\phi_z)$
      - Else: (draw word from ailment $a$ aspect $y$)
        - Draw $w_i \sim \text{Mult}(\phi_{a,y})$

Paul & Dredze
What you should know about LDA

- Each document is a mixture over topics
- Each topic looks like a Naïve Bayes model
  - It produces words with some probability
- Estimation of LDA is messy
  - Requires variational EM or Gibbs sampling
- In a plate model, each “plate” represents repeated nodes in a network
  - The plate model shows conditional independence, but not the form of the statistical distribution (e.g. Gaussian, Poisson, Dirichlet, ….)
LDA generation - example

- Topics = \{sports, politics\}
- Words = \{football, baseball, TV, win, president\}
- \(\alpha = (0.8,0.2)\)
- \(\beta = \begin{pmatrix}
sports & politics \\ football & 0.30 & 0.01 \\ baseball & 0.25 & 0.01 \\ TV & 0.10 & 0.15 \\ win & 0.30 & 0.25 \\ president & 0.01 & 0.20 \\ OOV & 0.04 & 0.38 \\ \end{pmatrix}\)
LDA generation - example

For each document, d
- Pick a topic distribution, $\theta_d$ using $\alpha$
- For each word in the document
  - pick a topic, z
  - given that topic, pick a word using $\beta$

$\alpha = (0.8, 0.2)$
$\beta =$

<table>
<thead>
<tr>
<th></th>
<th>sports</th>
<th>politics</th>
</tr>
</thead>
<tbody>
<tr>
<td>football</td>
<td>0.30</td>
<td>0.01</td>
</tr>
<tr>
<td>baseball</td>
<td>0.25</td>
<td>0.01</td>
</tr>
<tr>
<td>TV</td>
<td>0.10</td>
<td>0.15</td>
</tr>
<tr>
<td>win</td>
<td>0.30</td>
<td>0.25</td>
</tr>
<tr>
<td>president</td>
<td>0.01</td>
<td>0.20</td>
</tr>
<tr>
<td>OOV</td>
<td>0.04</td>
<td>0.38</td>
</tr>
</tbody>
</table>