LDA AND OTHER DIRECTED MODELS FOR MODELING TEXT
DGMs to describe parameter learning
Recap: Naïve Bayes as a DGM

• For each document \(d\) in the corpus (of size \(D\)):
  - Pick a label \(y_d\) from \(Pr(Y)\)
  - For each word in \(d\) (of length \(N_d\)):
    • Pick a word \(x_{id}\) from \(Pr(X|Y=y_d)\)

for every \(X\)

| \(Y\)     | \(X\)     | \(Pr(X|y=y)\) |
|-----------|-----------|---------------|
| onion     | aardvark  | 0.034         |
| onion     | ai        | 0.0067        |
| \ldots   | \ldots    | \ldots        |
| economist | aardvark  | 0.0000003     |
| \ldots   | \ldots    | \ldots        |
| economist | zymurgy   | 0.01000       |

Plate diagram

\[Pr(Y=y)\]

onion 0.3

economist 0.7
Recap: Naïve Bayes as a DGM

• For each document \(d\) in the corpus (of size \(D\)):
  - Pick a label \(y_d\) from \(Pr(Y)\)
  - For each word in \(d\) (of length \(N_d\)):
    • Pick a word \(x_{id}\) from \(Pr(X|Y=y_d)\)

Not described: how do we smooth for classes? For multinomials? How many classes are there? ....
Recap: smoothing for a binomial

**MLE:** maximize $\text{Pr}(D|\theta)$

**MAP:** maximize $\text{Pr}(D|\theta)\text{Pr}(\theta)$

**Tom:** estimate $\Theta = \text{P(heads)}$ for a binomial with MLE as:

$$\hat{\theta} = \frac{\alpha_1}{\alpha_1 + \alpha_0}$$

and with MAP as:

$$\hat{\theta} = \frac{(\alpha_1 + \gamma_1)}{(\alpha_1 + \gamma_1) + (\alpha_0 + \gamma_0)}$$

**Smoothing = prior over the parameter $\theta$**
Smoothing for a binomial as a DGM

MAP for dataset D with $\alpha_1$ heads and $\alpha_2$ tails:

$$\hat{\theta} = \frac{(\alpha_1 + \gamma_1)}{(\alpha_1 + \gamma_1) + (\alpha_0 + \gamma_0)}$$

MAP is a Viterbi-style inference: want to find max probability parameter $\theta$, to the posterior distribution

Also: inference in a simple graph can be intractable if the conditional distributions are complicated
Smoothing for a binomial as a DGM

MAP for dataset D with $\alpha_1$ heads and $\alpha_2$ tails:

$$\hat{\theta} = \frac{(\alpha_1 + \gamma_1)}{(\alpha_1 + \gamma_1) + (\alpha_0 + \gamma_0)}$$

Final recap: conjugate for a multinomial is called a **Dirichlet**
Recap: Naïve Bayes as a DGM

Now: let’s turn Bayes up to 11 for naïve Bayes....

Plate diagram
A more Bayesian Naïve Bayes

• From a Dirichlet $\alpha$:
  - Draw a multinomial $\pi$ over the $K$ classes

• From a Dirichlet $\eta$
  - For each class $y=1...K$
    • Draw a multinomial $\beta[y]$ over the vocabulary

• For each document $d=1..D$:
  - Pick a label $y_d$ from $\pi$
  - For each word in $d$ (of length $N_d$):
    • Pick a word $x_{id}$ from $\beta[y_d]
Pros and cons (as William sees them)

• From Dirichlet $\alpha$:
  
  - Draw a multinomial $\pi$ over the $K$ classes

• From a Dirichlet $\eta$
  
  - For each class $y=1...K$
    
    • Draw a multinomial $\beta[y]$ over the vocabulary

• For each document $d=1..D$:
  
  - Pick a label $y_d$ from $\pi$
  
  - For each word in $d$ (of length $N_d$):
    
    • Pick a word $x_{id}$ from $\beta[y_d]$
Pros and cons (as William sees them)

- From Dirichlet $\alpha$:
  - Draw a multinomial $\pi$ over the $K$ classes

- From a Dirichlet $\eta$
  - For each class $y=1...K$
    - Draw a multinomial $\beta[y]$ over the vocabulary

- For each document $d=1..D$:
  - Pick a label $y_d$ from $\pi$
  - For each word in $d$ (of length $N_d$):
    - Pick a word $x_{id}$ from $\beta[y_d]$
## Naïve Bayes-like DGMs covered so far

| P(X|Y)    | Y observed       | Y mixed          | Y hidden                     |
|----------|------------------|------------------|------------------------------|
| Multinomial | naïve Bayes ✔️ | SS naïve Bayes ✔️ | Mixture of multinomials      |
| Gaussian | Gaussian naïve Bayes | SS Gaussian naïve Bayes | Mixture of Gaussians ✔️ |
Supervised versus unsupervised NB
Unsupervised naïve Bayes (mixture of multinomials)

• From Dirichlet $\alpha$:
  - Draw a multinomial $\pi$ over the $K$ classes

• From a Dirichlet $\eta$
  - For each class $y=1...K$
    - Draw a multinomial $\beta[y]$ over the vocabulary

• For each document $d=1..D$:
  - Pick a label $y_d$ from $\pi$
  - For each word in $d$ (of length $N_d$):
    - Pick a word $x_{id}$ from $\beta[y_d]$
Unsupervised naïve Bayes (mixture of multinomials)

Learning: via EM

- Pick $\pi^0, \beta[1]^0, ..., \beta[K]^0$
- For $t=1,...,T$:
  - For $d=1,...,D$:
    - Find $P(Y_d|\pi^{t-1}, \beta[1]^{t-1}, ..., \beta[K]^{t-1})$
    - Maximize $\pi^t, \beta[1]^t, ..., \beta[K]^t$ on the weighted data
LATENT DIRICHLET ANALYSIS
(LDA)
The LDA Topic Model

Latent Dirichlet Allocation

David M. Blei
Computer Science Division
University of California
Berkeley, CA 94720, USA

Andrew Y. Ng
Computer Science Department
Stanford University
Stanford, CA 94305, USA

Michael I. Jordan
Computer Science Division and Department of Statistics
University of California
Berkeley, CA 94720, USA
Latent dirichlet allocation
DM Blei, AY Ng, ML Jordan - the Journal of machine Learning research, 2003 - dl.acm.org
Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for
collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian
model, in which each item of a collection is modeled as a finite mixture over an underlying ...
Cited by 14019  Related articles  All 141 versions  Web of Science: 3390  Cite  Saved

[PDF] Latent dirichlet allocation
DM Blei, AY Ng, ML Jordan - Advances in neural …, 2001 - machinelearning.wustl.edu
Abstract We develop an online variational Bayes (VB) algorithm for Latent Dirichlet
Allocation (LDA). Online LDA is based on online stochastic optimization with a natural
gradient step, which we show converges to a local optimum of the VB objective function. It ...
Cited by 262  Related articles  All 8 versions  Cite  Save  More

[HTML] Online learning for latent dirichlet allocation
M Hoffman, FR Bach, DM Blei - advances in neural information …, 2010 - papers.nips.cc
Abstract We develop an online variational Bayes (VB) algorithm for Latent Dirichlet
Allocation (LDA). Online LDA is based on online stochastic optimization with a natural
gradient step, which we show converges to a local optimum of the VB objective function. It ...
Cited by 559  Related articles  All 18 versions  Cite  Save
Unsupervised NB vs LDA

One class prior

One Y per doc

One Y per class

Different class distrib for each doc

Different Y per word

\[ \alpha \rightarrow \pi \rightarrow Y \rightarrow W \rightarrow \eta \rightarrow \beta \]
Unsupervised NB vs LDA

- **one class prior**
  - \( \alpha \rightarrow \pi \)
  - \( \pi \rightarrow Y \)
  - \( Y \rightarrow W \)
  - \( W \rightarrow N_d \)
  - \( N_d \rightarrow D \)
  - \( D \rightarrow \eta \)
  - \( \eta \rightarrow \beta \)

- **one Y per doc**

- **different class distrib \( \theta_d \) for each doc**
  - \( \alpha \rightarrow \theta_d \)
  - \( \theta_d \rightarrow Z_{di} \)
  - \( Z_{di} \rightarrow W_{di} \)
  - \( W_{di} \rightarrow N_d \)
  - \( N_d \rightarrow D \)
  - \( D \rightarrow \eta \)
  - \( \eta \rightarrow \beta_k \)
  - \( \beta_k \rightarrow K \)

- **one Z per word**
Blei’s motivation: start with BOW assumption

Assumptions: 1) documents are i.i.d 2) within a document, words are i.i.d. (bag of words)

• For each document $d = 1, \cdots, M$
  • Generate $\theta_d \sim D_1(\cdot)$
  • For each word $n = 1, \cdots, N_d$
    • generate $w_n \sim D_2(\cdot | \theta_{dn})$

Now pick your favorite distributions for $D_1, D_2$
Unsupervised NB vs LDA

- Unsupervised NB clusters documents into latent classes
- LDA clusters word occurrences into latent classes (topics)
- The smoothness of $\beta_k \Rightarrow$ same word suggests same topic
- The smoothness of $\theta_d \Rightarrow$ same document suggests same topic
LDA’s view of a document

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.

“Arts” “Budgets” “Children” “Education”
• LDA topics: top words \( w \) by \( \Pr(w|Z=k) \)

<table>
<thead>
<tr>
<th>( Z=13 )</th>
<th>( Z=22 )</th>
<th>( Z=27 )</th>
<th>( Z=19 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>“Arts”</strong></td>
<td><strong>“Budgets”</strong></td>
<td><strong>“Children”</strong></td>
<td><strong>“Education”</strong></td>
</tr>
<tr>
<td>NEW</td>
<td>MILLION</td>
<td>CHILDREN</td>
<td>SCHOOL</td>
</tr>
<tr>
<td>FILM</td>
<td>TAX</td>
<td>WOMEN</td>
<td>STUDENTS</td>
</tr>
<tr>
<td>SHOW</td>
<td>PROGRAM</td>
<td>PEOPLE</td>
<td>SCHOOLS</td>
</tr>
<tr>
<td>MUSIC</td>
<td>BUDGET</td>
<td>CHILD</td>
<td>EDUCATION</td>
</tr>
<tr>
<td>MOVIE</td>
<td>BILLION</td>
<td>YEARS</td>
<td>TEACHERS</td>
</tr>
<tr>
<td>PLAY</td>
<td>FEDERAL</td>
<td>FAMILIES</td>
<td>HIGH</td>
</tr>
<tr>
<td>MUSICAL</td>
<td>YEAR</td>
<td>WORK</td>
<td>PUBLIC</td>
</tr>
<tr>
<td>BEST</td>
<td>SPENDING</td>
<td>PARENTS</td>
<td>TEACHER</td>
</tr>
<tr>
<td>ACTOR</td>
<td>NEW</td>
<td>SAYS</td>
<td>BENNETT</td>
</tr>
<tr>
<td>FIRST</td>
<td>STATE</td>
<td>FAMILY</td>
<td>MANIGAT</td>
</tr>
<tr>
<td>YORK</td>
<td>PLAN</td>
<td>WELFARE</td>
<td>NAMPHY</td>
</tr>
<tr>
<td>OPERA</td>
<td>MONEY</td>
<td>MEN</td>
<td>STATE</td>
</tr>
<tr>
<td>THEATER</td>
<td>PROGRAMS</td>
<td>PERCENT</td>
<td>PRESIDENT</td>
</tr>
<tr>
<td>ACTRESS</td>
<td>GOVERNMENT</td>
<td>CARE</td>
<td>ELEMENTARY</td>
</tr>
<tr>
<td>LOVE</td>
<td>CONGRESS</td>
<td>LIFE</td>
<td>HAITI</td>
</tr>
</tbody>
</table>
SVM using 50 features: Pr(Z=k|\theta_d)

50 topics vs all words, SVM

Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.
Gibbs Sampling for LDA
LDA

- Latent Dirichlet Allocation
  - Parameter learning:
    - Variational EM
      - Not covered in 601-B
    - Collapsed Gibbs Sampling
      - Wait, why is sampling called “learning” here?
      - Here’s the idea....
LDA

• Gibbs sampling – works for *any* directed model!
  – Applicable when joint distribution is hard to evaluate but conditional distribution is known
  – Sequence of samples comprises a Markov Chain
  – Stationary distribution of the chain is the joint distribution

1. Initialise $x_{0:1:n}$.
2. For $i = 0$ to $N - 1$
   - Sample $x_1^{(i+1)} \sim p(x_1|x_2^{(i)}, x_3^{(i)}, \ldots, x_n^{(i)})$.
   - Sample $x_2^{(i+1)} \sim p(x_2|x_1^{(i+1)}, x_3^{(i)},\ldots, x_n^{(i)})$.
   
   $\vdots$

   - Sample $x_j^{(i+1)} \sim p(x_j|x_1^{(i+1)}, \ldots, x_{j-1}^{(i+1)}, x_{j+1}^{(i)}, \ldots, x_n^{(i)})$.
   
   $\vdots$

   - Sample $x_n^{(i+1)} \sim p(x_n|x_1^{(i+1)}, x_2^{(i+1)}, \ldots, x_{n-1}^{(i+1)})$.

Key capability: estimate distribution of one latent variables given the other latent variables and observed variables.
I’ll assume we know parameters for Pr(X|Z) and Pr(Y|X)
Initialize all the hidden variables randomly, then….

Pick $Z_1 \sim \Pr(Z|x_1, y_1, \theta)$

Pick $\theta \sim \Pr(z_1, z_2, \alpha)$

*pick from posterior*

Pick $Z_2 \sim \Pr(Z|x_2, y_2, \theta)$

Pick $Z_1 \sim \Pr(Z|x_1, y_1, \theta)$

Pick $\theta \sim \Pr(z_1, z_2, \alpha)$

*pick from posterior*

Pick $Z_2 \sim \Pr(Z|x_2, y_2, \theta)$

So we will have (a sample of) the true $\theta$

in a broad range of cases eventually these will converge to samples from the **true joint distribution**
Why does Gibbs sampling work?

• Basic claim: when you sample $x \sim P(X|y_1,\ldots,y_k)$ then if $y_1,\ldots,y_k$ were sampled from the true joint then $x$ will be sampled from the true joint.

• So the true joint is a “fixed point”
  – you tend to stay there if you ever get there.

• How long does it take to get there?
  – depends on the structure of the space of samples: how well-connected are they by the sampling steps?
LDA

• Latent Dirichlet Allocation
  – Parameter learning:
    • Variational EM
      – Not covered in 601-B
    • Collapsed Gibbs Sampling
      – What is collapsed Gibbs sampling?
Initialize all the Z’s randomly, then....

Pick $Z_1 \sim \Pr(Z_1 | z_2, z_3, \alpha )$

Pick $Z_2 \sim \Pr(Z_2 | z_1, z_3, \alpha )$

Pick $Z_3 \sim \Pr(Z_3 | z_1, z_2, \alpha )$

Pick $Z_1 \sim \Pr(Z_1 | z_2, z_3, \alpha )$

Pick $Z_2 \sim \Pr(Z_2 | z_1, z_3, \alpha )$

Pick $Z_3 \sim \Pr(Z_3 | z_1, z_2, \alpha )$

. . .

Converges to samples from the true joint … and then we can estimate $\Pr(\theta | \alpha, \text{sample of Z’s})$
Initialize all the Z’s randomly, then….

Pick $Z_1 \sim \Pr(Z_1 | z_2, z_3, \alpha)$

What’s this distribution?
Initialize all the $Z$'s randomly, then....

Pick $Z_1 \sim \text{Pr}(Z_1|z_2,z_3, \alpha )$

Simpler case: What's this distribution?
called a Dirichlet-multinomial and it looks like this:

$$
\Pr(Z_1 = k_1, Z_2 = k_2, Z_3 = k_3 | \alpha ) = \int_\theta \Pr(Z_1 = k_1, Z_2 = k_2, Z_3 = k_3 | \theta ) \Pr(\theta | \alpha ) d\theta
$$

If there are $k$ values for the $Z$'s and $n_k = \# Z$'s with value $k$, then it turns out:

$$
\Pr(Z | \alpha ) = \int_\theta \Pr(Z | \theta ) \Pr(\theta | \alpha ) d\theta = \frac{\Gamma \left( \sum_k \alpha_k \right)}{\Gamma \left( \sum_k \alpha_k + \sum_k n_k \right)} \prod_k \frac{\Gamma (n_k + \alpha_k)}{\Gamma (\alpha_k)}
$$
Initialize all the $Z$'s randomly, then....

Pick $Z_1 \sim Pr(Z_1|z_2,z_3,\alpha)$

It turns out that sampling from a Dirichlet-multinomial is very easy!

Notation:
- $k$ values for the $Z$'s
- $n_k = \# Z$'s with value $k$
- $Z = (Z_1, ..., Z_m)$
- $Z^{(-i)} = (Z_1, ..., Z_{i-1}, Z_{i+1}, ..., Z_m)$
- $n_{k}^{(-i)} = \# Z$'s with value $k$ excluding $Z_i$

\[
Pr(Z_i = k \mid Z^{(-i)}, \alpha) \propto n_{k}^{(-i)} + \alpha_k
\]

\[
Pr(Z \mid \alpha) = \int_\theta Pr(Z \mid \theta) Pr(\theta \mid \alpha) d\theta = \frac{\Gamma\left(\sum_k \alpha_k\right)}{\Gamma\left(\sum_k \alpha_k + \sum_k n_k\right)} \prod_k \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)}
\]
What about with downstream evidence?

\[
\Pr(Z_i = k \mid Z^{(-i)}, \alpha) \propto n_k^{(-i)} + \alpha_k
\]
captures the constraints on \(Z_i\) via \(\theta\) (from “above”, “causal” direction)

what about via \(\beta\)?

\[
\Pr(Z) = \frac{1}{c} \cdot \Pr(Z \mid E^+) \cdot \Pr(E^+ \mid Z)
\]

\[
\Pr(Z_i = k \mid Z^{(-i)}, X_i = x, \eta) \propto \frac{\#(X = x \text{ with } Z = k \text{ in } Z^{(-i)}) \text{ + smoothing}}{\#(X = x) \text{ + smoothing}}
\]
Sampling for LDA

Notation:
- $k$ values for the $Z_{d,i}$’s
- $Z^{(-d,i)} = \text{all the } Z\text{'s but } Z_{d,i}$
- $n_{w,k} = \# \text{ } Z_{d,i} \text{'s with value } k \text{ paired with } W_{d,i} = w$
- $n_{*,k} = \# \text{ } Z_{d,i} \text{'s with value } k$
- $n_{w,k}^{(-d,i)} = n_{w,k} \text{ excluding } Z_{i,d}$
- $n_{*,k}^{(-d,i)} = n_{*,k} \text{ excluding } Z_{i,d}$
- $n_{*,k}^{d,(-i)} = n_{*,k} \text{ from doc } d \text{ excluding } Z_{i,d}$

\[
\Pr(Z_{d,i} = k \mid Z^{(-d,i)}, W_{d,i} = w, \alpha) \propto \frac{\Pr(Z \mid E+)}{\left(n_{*,k}^{d,(-i)} + \alpha_k\right) \left(\frac{n_{w,k}^{(-d,i)} + \eta_w}{\sum_{w',k} n_{w',k}^{(-d,i)} + \eta_{w'}}\right)}
\]

- Fraction of time $Z=k$ in doc $d$
- Fraction of time $W=w$ in topic $k$
Unsupervised NB vs LDA

- Unsupervised NB clusters documents into latent classes
- LDA clusters word occurrences into latent classes (topics)
- The smoothness of $\beta_k$ suggests same word suggests same topic
- The smoothness of $\theta_d$ suggests same document suggests same topic
EVEN MORE DETAIL ON LDA...
Way way more detail

# topic k, docId d, and wordId w are integer indices
#
# x[d][j] = w, index of j-th word in doc d
# z[d][j] = k, index of latent topic of j-th word in doc d
# vocab[w] = string for the word with index w
#
# totalTopicCount[k] = number of words in topic k
# docTopicCount[d][k] = number of words in topic k for document d
# wordTopicCount[w][k] = number of occurrences of word w in topic k
# totalWords = number of words in the corpus

\[
\Pr(Z_{d,i} = k \mid Z^{(-d,i)}, W_{d,i} = w, \alpha) \propto \left( n_{*,k}^{d,(-i)} + \alpha_k \right) \frac{n_{w,k}^{(-d,i)} + \eta_w}{\sum_{w',k} n_{w',k}^{(-d,i)} + \eta_{w'}}
\]
# topic k, docId d, and wordId w are integer indices
#
# x[d][j] = w, index of j-th word in doc d
# z[d][j] = k, index of latent topic of j-th word in doc d
# vocab[w] = string for the word with index w
#
# totalTopicCount[k] = number of words in topic k
# docTopicCount[d][k] = number of words in topic k for document d
# wordTopicCount[w][k] = number of occurrences of word w in topic k
# totalWords = number of words in the corpus

def initGibbs(self):
    # initializing latent vars'
    self.totalTopicCount = self.topicCounter()
    self.docTopicCount = [self.topicCounter() for d in xrange(len(self.x))]
    self.wordTopicCount = [self.topicCounter() for w in xrange(len(self.vocab))]
    self.z = [[-1 for j in xrange(len(self.x[d]))] for d in xrange(len(self.x))]
    for d in xrange(len(self.x)):
        if (d+1)%self.dstep==0: print '..doc', d+1, 'of', len(self.x)
        for j in xrange(len(self.x[d])):
            w = self.x[d][j]
            k = random.randint(0, self.numTopics-1)
            self.z[d][j] = k
            self.docTopicCount[d].add(k, 1)
            self.wordTopicCount[w].add(k, 1)
            self.totalTopicCount.add(k, 1)
    #reasonable parameters
    self.alpha = 1.0/self.numTopics
    self.beta = 1.0/len(self.vocab)
    print "alpha: ", self.alpha, "beta: ", self.beta
def runGibbs(self, maxT):
    for t in xrange(maxT):
        print '.iteration', t+1, 'of', maxT
        for d in xrange(len(self.x)):
            if (d+1)%self.dstep==0: print '...doc', d+1, 'of', len(self.x)
            for j in xrange(len(self.x[d])):
                k = self.resample(d, j)
                self.flip(d, j, self.z[d][j], k)

def flip(self, d, j, k_old, k_new):
    """update counts to reflect a changed value of z[d][j]""
    if k_old != k_new:
        w = self.x[d][j]
        self.docTopicCount[d].add(k_old, -1)
        self.wordTopicCount[w].add(k_old, -1)
        self.wordTopicCount[w].add(k_new, +1)
        self.totalTopicCount.add(k_old, -1)
        self.totalTopicCount.add(k_new, +1)
        self.z[d][j] = k_new
```python
def runGibbs(self, maxT):
    for t in xrange(maxT):
        print '.iteration', t+1, 'of', maxT
        for d in xrange(len(self.x)):
            if (d+1) % self.dstep==0: print '...doc', d+1, 'of', len(self.x)
            for j in xrange(len(self.x[d])):
                k = self.resample(d, j)
                self.flip(d, j, self.z[d][j], k)

def resample(self, d, j):
    """sample a new value of z[d][j]""
    p = []
    norm = 0.0
    # compute pk = Pr(z_dj=k | everything else)
    for k in xrange(self.numTopics):
        w = self.x[d][j]
        z_dj_equal_k = 1 if self.z[d][j] == k else 0
        ak = (self.docTopicCount[d][k] - z_dj_equal_k + self.alpha)
        bk = ((self.wordTopicCount[w][k] - z_dj_equal_k * self.beta) / (self.totalTopicCount[k] - z_dj_equal_k + self.totalWords * self.beta))
        pk = ak*bk
        p.append(pk)
        norm += pk
    # pick randomly from the normalized pk
    sum_p_up_to_k = 0.0
    r = random.random()
    for k in xrange(self.numTopics):
        sum_p_up_to_k += p[k]/norm
        if r < sum_p_up_to_k:
            return k
```
What gets learned…..

def phi(self, w, k):
    """weight of word w under topic k""
    num = (self.wordTopicCount[w][k] + self.beta)
    denom = (self.totalTopicCount[k] + self.totalWords * self.beta)
    return num/denom

def theta(self, d, k):
    """weight of doc under topic k""
    num = (self.docTopicCount[d][k] + self.alpha)
    denom = (sum(self.docTopicCount[d]) + self.numTopics*self.alpha)
    return num/denom

Figure 1: Graphical model for LDA.
Some comments on LDA

• Very widely used model
• Also a component of many other models
LDA is not just for text

An example: modeling urban neighborhoods

- Grid cell in lat/long space ~ “document”
- 4Square check-in categories ~ “words”
- Prototypical neighborhood category ~ set of related “words” ~ topic
<table>
<thead>
<tr>
<th>Topic</th>
<th>Top categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Home, Pizza, Salon/Barbershop, Beauty/Cosmetic, Bank/Financial, Other - Buildings, Food:Chinese</td>
</tr>
<tr>
<td>1</td>
<td>Doctor’s Office, School, Medical, Dentist’s Office, Hospital, Religious, Gym/Fitness</td>
</tr>
<tr>
<td>2</td>
<td>Boat/Ferry, Golf Course, Harbor/Marina, Farm, Campground, Lake/Pond, Arcade</td>
</tr>
<tr>
<td>3</td>
<td>Pub, Resort, Community College, Racetrack, Motel, Beer Garden, Hotel Bar</td>
</tr>
<tr>
<td>4</td>
<td>Beach, Seafood, Scenic Lookout, Surf Spot, Harbor/Marina, Landmark, Skate/Surf/Snowboard</td>
</tr>
<tr>
<td>5</td>
<td>Parks&amp;Outdoor, Entertainment, Baseball Field, Theme Park, Zoo/Aquarium, Baseball, Stadium</td>
</tr>
<tr>
<td>6</td>
<td>Hiking Trail, History Museum, Hostel, Post Office, Casino, Parks&amp;Outdoor, Scenic Lookout</td>
</tr>
<tr>
<td>7</td>
<td>🌟 Bar, Event Space, Art Gallery, Entertainment, Discotheque, Taxi, Speakeasy/Secret Spot</td>
</tr>
<tr>
<td>8</td>
<td>Corporate/Office, Home, Coworking Space, Buildings, Tech Startup, Gym, Hotel</td>
</tr>
<tr>
<td>9</td>
<td>Grocery/Supermarket, Flower Shop, Drug Store, Post Office, Italian, Flea Market, Breakfast/Brunch</td>
</tr>
<tr>
<td>10</td>
<td>Boat/Ferry, Golf Course, Harbor/Marina, Farm, Campground, Lake/Pond, Arcade</td>
</tr>
<tr>
<td>11</td>
<td>Automotive Shop, Mexican, Asian, Chinese, Vietnamese, Korean, Hardware</td>
</tr>
<tr>
<td>12</td>
<td>🌟 Airport Gate, Plane/In-flight, Plane, Airport, Terminal, Other - Travel, Hotel</td>
</tr>
<tr>
<td>13</td>
<td>Fast Food, Mexican, Gas Station, Automotive Shop, Pizza, Other - Buildings, Burgers</td>
</tr>
<tr>
<td>14</td>
<td>Other - Buildings, Government, Train Station, Bus, Light Rail, Courthouse</td>
</tr>
<tr>
<td>15</td>
<td>Highway/Traffic, Golf Course, Bridge, Farm, Cemetery, BBQ, Field, Subway</td>
</tr>
<tr>
<td>16</td>
<td>Other - Travel, Coffee Shop, Truck/Street Food, Sandwiches, Garden, Steakhouse, Karaoke</td>
</tr>
<tr>
<td>17</td>
<td>Home, Playground, Pool, Speakeasy/Secret Spot, College&amp;Educational, Gym</td>
</tr>
</tbody>
</table>
Topic 7 - “Night Life”
On Beyond LDA....
a sample of some of the many extensions
Correspondence LDA (for images and captions)

Generate image “region” from a Gaussian (of colors, sizes)

Generate the image Z’s first
Pick y uniformly from 1...N
Pick the Z_y from image

Generate word from a caption
Wang and McCallum, KDD 2006
Topics over Time

(a) LDA

Beta distribution for each topic to generate timestamps

(c) TOT model, for Gibbs sampling
### Topics over Time

<table>
<thead>
<tr>
<th>Mexican War</th>
<th>Panama Canal</th>
<th>Cold War</th>
<th>Modern Tech</th>
</tr>
</thead>
<tbody>
<tr>
<td>states 0.02032</td>
<td>government 0.02928</td>
<td>world 0.01875</td>
<td>energy 0.03902</td>
</tr>
<tr>
<td>mexico 0.01832</td>
<td>united 0.02132</td>
<td>states 0.01717</td>
<td>national 0.01534</td>
</tr>
<tr>
<td>government 0.01670</td>
<td>states 0.02067</td>
<td>security 0.01710</td>
<td>development 0.01448</td>
</tr>
<tr>
<td>united 0.01521</td>
<td>islands 0.01167</td>
<td>soviet 0.01664</td>
<td>space 0.01436</td>
</tr>
<tr>
<td>war 0.01059</td>
<td>canal 0.01014</td>
<td>united 0.01491</td>
<td>science 0.01227</td>
</tr>
<tr>
<td>congress 0.00951</td>
<td>american 0.00872</td>
<td>nuclear 0.01454</td>
<td>technology 0.01227</td>
</tr>
<tr>
<td>country 0.00906</td>
<td>cuba 0.00834</td>
<td>peace 0.01408</td>
<td>oil 0.01178</td>
</tr>
<tr>
<td>texas 0.00852</td>
<td>made 0.00747</td>
<td>nations 0.01069</td>
<td>make 0.00994</td>
</tr>
<tr>
<td>made 0.00727</td>
<td>general 0.00731</td>
<td>international 0.01024</td>
<td>effort 0.00969</td>
</tr>
<tr>
<td>great 0.00611</td>
<td>war 0.00660</td>
<td>america 0.00987</td>
<td>administration 0.00957</td>
</tr>
</tbody>
</table>
Figure 1. Graphical representation of a dynamic topic model (for three time slices). Each topic’s natural parameters $\beta_{t,k}$ evolve over time, together with the mean parameters $\alpha_t$ of the logistic normal distribution for the topic proportions.
1881 On Matter as a form of Energy
1892 Non-Euclidean Geometry
1900 On Kathode Rays and Some Related Phenomena
1917 "Keep Your Eye on the Ball"
1920 The Arrangement of Atoms in Some Common Metals
1933 Studies in Nuclear Physics
1943 Aristotle, Newton, Einstein. II
1950 Instrumentation for Radioactivity
1965 Lasers
1975 Particle Physics: Evidence for Magnetic Monopole Obtained
1985 Fermilab Tests its Antiproton Factory
1999 Quantum Computing with Electrons Floating on Liquid Helium

Atomic Physics
Choice of topics constrained by a set of possible labels for the document
<table>
<thead>
<tr>
<th>Tag (Labeled LDA)</th>
<th>(LDA) Topic ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>web</td>
<td>8</td>
</tr>
<tr>
<td>web search site blog css content</td>
<td></td>
</tr>
<tr>
<td>google list page posted great work</td>
<td></td>
</tr>
<tr>
<td>comments read nice post great april</td>
<td></td>
</tr>
<tr>
<td>blog march june wordpress</td>
<td></td>
</tr>
<tr>
<td>books</td>
<td>13</td>
</tr>
<tr>
<td>book image pdf review library posted</td>
<td></td>
</tr>
<tr>
<td>read copyright books title</td>
<td></td>
</tr>
<tr>
<td>news information service web online</td>
<td></td>
</tr>
<tr>
<td>project site free search home</td>
<td></td>
</tr>
<tr>
<td>science</td>
<td>19</td>
</tr>
<tr>
<td>works water map human life work</td>
<td></td>
</tr>
<tr>
<td>science time world years sleep</td>
<td></td>
</tr>
<tr>
<td>web images design content java css</td>
<td></td>
</tr>
<tr>
<td>website articles page learning</td>
<td></td>
</tr>
<tr>
<td>computer</td>
<td>4</td>
</tr>
<tr>
<td>windows file version linux computer</td>
<td></td>
</tr>
<tr>
<td>free system software mac</td>
<td></td>
</tr>
<tr>
<td>jun quote pro views added check</td>
<td></td>
</tr>
<tr>
<td>anonymous card core power ghz</td>
<td></td>
</tr>
<tr>
<td>religion</td>
<td>3</td>
</tr>
<tr>
<td>comment god jesus people gospel</td>
<td></td>
</tr>
<tr>
<td>bible reply lord religion written</td>
<td></td>
</tr>
<tr>
<td>life written jesus words made man</td>
<td></td>
</tr>
<tr>
<td>called mark john person fact name</td>
<td></td>
</tr>
<tr>
<td>java</td>
<td>2</td>
</tr>
<tr>
<td>applications spring open web java</td>
<td></td>
</tr>
<tr>
<td>pattern eclipse development ajax</td>
<td></td>
</tr>
<tr>
<td>house light radio media photography</td>
<td></td>
</tr>
<tr>
<td>news music travel cover</td>
<td></td>
</tr>
<tr>
<td>culture</td>
<td>12</td>
</tr>
<tr>
<td>people day link posted time comments</td>
<td></td>
</tr>
<tr>
<td>music jane permalink</td>
<td></td>
</tr>
<tr>
<td>game review street public art health</td>
<td></td>
</tr>
<tr>
<td>food city history science</td>
<td></td>
</tr>
<tr>
<td>health food city history science</td>
<td></td>
</tr>
</tbody>
</table>
Linear regression using latent topics as features
### Abstract

BACKGROUND: Molecular chaperone Hsp40 in protein refolding. How Hsp40 and other unfolded states of proteins to bind nonnal.

<table>
<thead>
<tr>
<th>Method</th>
<th>$F_1$</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block-LDA</td>
<td>0.249</td>
<td>0.247</td>
<td>0.250</td>
</tr>
<tr>
<td>Sparse Block model</td>
<td>0.161</td>
<td>0.224</td>
<td>0.126</td>
</tr>
<tr>
<td>Link LDA</td>
<td>0.152</td>
<td>0.150</td>
<td>0.155</td>
</tr>
<tr>
<td>MMSB</td>
<td>0.165</td>
<td>0.166</td>
<td>0.164</td>
</tr>
<tr>
<td>Random</td>
<td>0.145</td>
<td>0.155</td>
<td>0.137</td>
</tr>
</tbody>
</table>

Table 4: Functional category prediction
Summary: key points

• DGMs with random variables that are parameters
  – Inferring a MAP value is parameter learning
  – Can often be used to model EM algorithms in a Bayesian setting
• Examples
  – Unsupervised naïve Bayes, mixtures of Gaussians, ...
  – Latent Dirichlet allocation
    • LDA is a “mixed membership” model
• Gibbs sampling and collapsed Gibbs sampling
• Collapsed Gibbs for LDA
• Potential applications of LDA
• Plausible extensions to LDA