Befriending LDA

Nisheeth
LDA in plate notation
Generative model

• For all documents
  – Generate $\theta \sim \text{Dirichlet}(\alpha)$
  – Generate all $K \psi \sim \text{Dirichlet}(\beta)$

• For all words in each document
  – Generate $t \sim \text{Multinomial}(\theta)$
  – Generate $w \sim \text{Multinomial}(\psi_t)$
LDA math – the Dirichlet distribution

- A $k$-dimensional Dirichlet random variable $\theta$ can take values in the $(k-1)$-simplex, and has the following probability density on this simplex:

$$p(\theta | \alpha) = \frac{\Gamma \left( \sum_{i=1}^{k} \alpha_i \right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1-1} \cdots \theta_k^{\alpha_k-1}$$

- Easier to understand
  - Prior $\text{Dir}(\alpha_1, \alpha_2)$
  - Likelihood $\text{Multi}(\theta_1, \theta_2)$
  - Outcome $\{n_1, n_2\}$
  - Posterior $\text{Dir}(\alpha_1 + n_1, \alpha_2 + n_2)$

- Ignoring the normalization constant, what is the Dirichlet probability of a multinomial sample $[0.1, 0.5, 0.4]$ with parameter 10
  - $(0.1)^9 (0.5)^9 (0.4)^9 = 5e-16$

- What would it be for parameter 0.2?
  - 22
Dirichlet update – dice roll

- Data $d = (2, 5, 4, 2, 6)$
LDA math – the multinomial distribution

- For \( n \) independent trials that could yield exactly one of \( k \) possible results, the multinomial distribution gives the probability of seeing any particular combination of outcomes

\[
p(x, \gamma) = \frac{n!}{x_1! x_2! \ldots x_k!} \gamma_1^{x_1} \gamma_2^{x_2} \ldots \gamma_k^{x_k}
\]

- Parameterized by \( \gamma \) and \( n \)
- Easier to understand
  - Tracks word frequencies
  - Given a vocabulary of 3 words A,B,C with normalized empirical frequencies [0.3, 0.4, 0.3] in a corpus and a document AABB

  \[
p(d) = \frac{4!}{2!2!0!} (0.3)^2 (0.4)^2 = 0.0864
\]
  - Given normalized empirical frequencies [0.1,0.1,0.8], what would the probability of the same document be?
  - Given normalized empirical frequencies [0.3, 0.4, 0.3] and a document A, what would its probability be?
\( p(w | t) \) is high when many words in a document show up as high frequency terms in the corresponding topic word distribution

\( \psi \) is distribution of words in a topic

\( p(t | \theta) \) is high when many words in a topic show up as high frequency terms in the document topic distribution

\( \theta \) is distribution of topics in a document
Compare with Gaussian mixture model

• $p(K|\phi)$ is high when the $\phi$ value is high for the $K^{th}$ label
• $p(x|K)$ is high when $x$ is statistically likely to be drawn from the Gaussian with the Kth summary statistics
LDA inference

• Latent variable inference

\[ p(\theta, t|w, \alpha, \psi) = \frac{p(\theta, z, w|\alpha, \psi)}{p(w|\alpha, \psi)} \]

• From the graphical model

\[ p(\theta, z, w|\alpha, \psi) = p(w|t, \psi)p(t|\theta)p(\theta|\alpha) \]

• What are these terms?

1. \( p(w|t, \psi) = \prod_{n=1}^{d} \psi_{tn, wn} \)
2. \( p(t|\theta) = \prod_{n=1}^{d} \theta_{tn} \)
3. \( p(\theta|\alpha) = C(\alpha) \sum_{i=1}^{K} \theta_i^{\alpha_i - 1} \)
LDA intuition

• Given the optimal denominator, the correct partitioning of the data into topics is determined by the numerator
• What does the numerator say about what constitutes a good topic partitioning?
  1. \( p(w|t, \psi) \) will have high values iff \( \psi \) is sparse
  2. \( p(t|\theta) \) will have high values iff \( \theta \) is concentrated
  3. \( p(\theta|\alpha) \) will have high values if \( \alpha \) is small
• Implications
  1. Better to have non-overlapping topics
  2. Better to have fewer topics per document
  3. Better to be biased towards few topics in general
• Net upshot: make clusters with co-occurring terms
LDA inference

• From these building blocks we get the full numerator

• Denominator obtained by marginalizing over the latent variables
  – Involves an intractable integral
  – Have to use approximate inference methods
    • Variational EM
    • Gibbs sampling

• MLE inference is standard

$$\ell(\alpha, \beta) = \sum_{d=1}^{N} \log p(w_d | \alpha, \beta)$$
Model selection

\[ \log P(w \mid T) \]

Number of topics (T)

\[ 10^7 \times \]

\[ 1.55 \]

\[ 1.6 \]

\[ 1.65 \]

\[ 1.7 \]

\[ 1.75 \]

\[ 1.8 \]
Document modeling

- Unlabeled data – our goal is density estimation.
- Compute the *perplexity* of a held-out test to evaluate the models – lower perplexity score indicates better generalization.

\[
\text{perplexity}(D_{test}) = \exp \left\{ - \frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d} \right\}
\]
Document Modeling – cont.

data used

- C. Elegans Community abstracts
  - 5,225 abstracts
  - 28,414 unique terms
- TREC AP corpus (subset)
  - 16,333 newswire articles
  - 23,075 unique terms
- Held-out data – 10%
- Removed terms – 50 stop words, words appearing once (AP)
What can you get from it?

<table>
<thead>
<tr>
<th>“Arts”</th>
<th>“Budgets”</th>
<th>“Children”</th>
<th>“Education”</th>
</tr>
</thead>
<tbody>
<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>

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.
Topic membership

Original article

Most likely words from top topics

sequence
genome
genes
sequences
human
gene
dna
sequencing
cromosome
regions
analysis
data
genomic
number
devices
device
materials
current
high
gate
light
silicon
material
technology
electrical
fiber
power
based
data
information
network
web
computer
language
networks
time
software
system
words
algorithm
number
internet
Document similarity

\[ d_{ij} = \mathbb{E} \left[ \sum_{k=1}^{K} \left( \sqrt{\theta_{i,k}} - \sqrt{\theta_{j,k}} \right)^2 \mid \mathbf{w}_i, \mathbf{w}_j \right] \]
Topic similarity

Thomas L. Griffiths, and Mark Steyvers PNAS 2004;101:5228-5235

©2004 by National Academy of Sciences
Document tagging, relevance scoring

**Tax Innovation in the States: Capitalizing on Political Opportunity**
Frances Stokes Berry, William D. Berry, American Journal of Political Science (1992), pp. 715-742

*Journal Disciplines:*
- Political Science

**Chaos and Nonlinear Forecastability in Economics and Finance**
Blake LeBaron, Philosophical Transactions: Physical Sciences and Engineering (1994), pp. 397-404

*Journal Disciplines:*
- Mathematics
- Biological Sciences
- General Science

**Reply: Theory Is Not a Social Dilemma**
Gerald Marwell, Pamela Oliver, Social Psychology Quarterly (1994), pp. 373

*Journal Disciplines:*
- Psychology
- Sociology
Extension: correlated topic models

Logit plot

\[ f(x) = \log \left( \frac{x}{1-x} \right) \]

\[ x \sim N(\mu, \Sigma) \]

\[ \theta \propto \exp(x_i) \]
Topic hierarchies
Extension: dynamic topic modeling
Time-drifting topic distributions

- Use a logistic normal distribution to model topics evolving over time (Aitchison, 1980)

- A state-space model on the natural parameter of the topic multinomial (West and Harrison, 1997)

\[ \beta_{t,k} | \beta_{t-1,k} \sim \mathcal{N}(\beta_{t-1,k}, I\sigma^2) \]

\[ p(w | \beta_{t,k}) \propto \exp \{ \beta_{t,k} \} \]
Temporal changes

1880 electric machine
power engine
steam two machines
iron battery wire

1890 electric power company
steam electrical machine
two system motor engine

1900 apparatus steam
engineering water
construction engineer
room feet

1910 air water engineering
apparatus room
laboratory engineer
made gas tube

1920 apparatus tube
air pressure water
glass gas made
laboratory mercury

1930 tube apparatus glass
air mercury laboratory
pressure made gas small

1940 air tube apparatus
glass laboratory rubber
pressure small mercury gas

1950 tube apparatus glass
air chamber instrument
small laboratory pressure rubber

1960 tube system temperature
air heat chamber power
high instrument control

1970 air heat power system
temperature chamber high
flow tube design

1980 high power design
heat system devices
instruments control large

1990 materials high power
current applications technology
devices design device heat

2000 devices device materials
current gate high light silicon material technology
Trends

"Theoretical Physics"

FORCE
RELATIVITY
LASER

"Neuroscience"

OXGEN
NERVE
NEURON
Other uses

Corr-LDA:
TREE, LIGHT, SUNSET, WATER, SKY

GM-Mixture:
CLOSE-UP, TREE, PEOPLE, MUSHROOMS, LICHEN

GM-LDA:
WATER, SKY, TREE, PEOPLE, GRASS

Corr-LDA:
TREE, WATER, GRASS, FLOWERS, BIRDS

GM-Mixture:
TREE, WATER, GRASS, SKY, FIELD

GM-LDA:
WATER, SKY, TREE, PEOPLE, GRASS

Corr-LDA:
1. PEOPLE, TREE
2. SKY, JET
3. SKY, CLOUDS
4. SKY, MOUNTAIN
5. PLANE, JET
6. PLANE, JET

GM-LDA:
1. HOTEL, WATER
2. PLANE, JET
3. TUNDRA, PENGUIN
4. PLANE, JET
5. WATER, SKY
6. BOATS, WATER