Probabilistic Graphical Models

Case Study: Topic Models

Eric Xing
Lecture 15, March 4, 2015

Reading: See class website
Probabilistic Topic Models

- Humans cannot afford to deal with (e.g., search, browse, or measure similarity) a huge number of text documents
- We need computers to help out …
How to get started?

- Here are some important elements to consider before you start:
  - Task:
    - Embedding? Classification? Clustering? Topic extraction? …
  - Data representation:
    - Input and output (e.g., continuous, binary, counts, …)
  - Model:
    - BN? MRF? Regression? SVM?
  - Inference:
    - Exact inference? MCMC? Variational?
  - Learning:
    - MLE? MCLE? Max margin?
  - Evaluation:
    - Visualization? Human interpretability? Perplexity? Predictive accuracy?

- It is better to consider one element at a time!
Tasks: document embedding

- Say, we want to have a mapping ..., so that

- Compare similarity
- Classify contents
- Cluster/group/categorizing
- Distill semantics and perspectives
- ..
### Summarizing the data using topics

<table>
<thead>
<tr>
<th>Bayesian modeling</th>
<th>Visual cortex</th>
<th>Education</th>
<th>Market</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bayesian model</td>
<td>cortex</td>
<td>students</td>
<td>market</td>
</tr>
<tr>
<td>inference</td>
<td>cortical</td>
<td>education</td>
<td>economic</td>
</tr>
<tr>
<td>models</td>
<td>areas</td>
<td>learning</td>
<td>financial</td>
</tr>
<tr>
<td>probability</td>
<td>visual</td>
<td>educational</td>
<td>economics</td>
</tr>
<tr>
<td>probabilistic</td>
<td>area</td>
<td>teaching</td>
<td>markets</td>
</tr>
<tr>
<td>Markov prior</td>
<td>primary</td>
<td>school</td>
<td>returns</td>
</tr>
<tr>
<td>hidden</td>
<td>connections</td>
<td>student</td>
<td>price</td>
</tr>
<tr>
<td>approach</td>
<td>ventral</td>
<td>skills</td>
<td>stock</td>
</tr>
<tr>
<td></td>
<td>cerebral</td>
<td>teacher</td>
<td>value</td>
</tr>
<tr>
<td></td>
<td>sensory</td>
<td>academic</td>
<td>investment</td>
</tr>
</tbody>
</table>
See how data changes over time
User interest modeling using topics

User interest profile (adjustable with sliders---Changing these changes recommendations.)

Weight

User preferred topics

1: learning machine training vector learn machines kernel learned classifiers classifier
2: online classification digital library libraries browsing classify classifying labels catalog
3: two differences active hypothesis arise difference evolved morphological modify morphology
4: experiments ability demonstrated produced contexts situations instances fail recognize string
5: features class classes subset java characteristic earlier represented defines separate
6: process making presents objective steps reports distinguish exploit maintaining select
7: algorithm signal input signals output exact performs music sound iterative
8: database databases contains version list comprehensive release stored update curated
9: applications application provide built numerous proven providing discusses tremendous presents
10: text literature discovery mining biomedical full extract discovering texts themes

http://cogito-demos.ml.cmu.edu/cgi-bin/recommendation.cgi
Representation:

- **Data:** Bag of Words Representation

  As for the Arabian and Palestinian voices that are against the current negotiations and the so-called peace process, they are not against peace per se, but rather for their well-founded predictions that Israel would NOT give an inch of the West bank (and most probably the same for Golan Heights) back to the Arabs. An 18 months of "negotiations" in Madrid, and Washington proved these predictions. Now many will jump on me saying why are you blaming Israelis for no-result negotiations. I would say why would the Arabs stall the negotiations, what do they have to loose?

- Each document is a vector in the word space
- Ignore the order of words in a document. Only count matters!

- A high-dimensional and sparse representation (\(|V| \gg D\))
  - Not efficient text processing tasks, e.g., search, document classification, or similarity measure
  - Not effective for browsing
How to Model Semantic?

- Q: What is it about?
- A: Mainly MT, with syntax, some learning

### A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses hierarchical phrases—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the formal machinery of syntax based translation systems without any linguistic commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative Improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.

<table>
<thead>
<tr>
<th>Topics</th>
<th>Source</th>
<th>Target</th>
<th>SMT Alignment</th>
<th>Parse Tree</th>
<th>Noun Phrase</th>
<th>Grammar CFG</th>
<th>likelihood</th>
<th>EM Hidden</th>
<th>Parameters Estimation</th>
<th>argMax</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
<td>Unigram over vocabulary</td>
</tr>
</tbody>
</table>

© Eric Xing @ CMU, 2005-2015
Why this is Useful?

- Q: What is it about?
  - A: Mainly MT, with syntax, some learning

<table>
<thead>
<tr>
<th>MT</th>
<th>Syntax</th>
<th>Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.6</td>
<td>0.3</td>
<td>0.1</td>
</tr>
</tbody>
</table>

- Q: give me similar document?
  - Structured way of browsing the collection

- Other tasks
  - Dimensionality reduction
    - TF-IDF vs. topic mixing proportion
    - Classification, clustering, and more ...

A Hierarchical Phrase-Based Model for Statistical Machine Translation

We present a statistical phrase-based Translation model that uses hierarchical phrases—phrases that contain sub-phrases. The model is formally a synchronous context-free grammar but is learned from a bitext without any syntactic information. Thus it can be seen as a shift to the formal machinery of syntax based translation systems without any linguistic commitment. In our experiments using BLEU as a metric, the hierarchical Phrase based model achieves a relative Improvement of 7.5% over Pharaoh, a state-of-the-art phrase-based system.
Words in Contexts

• “It was a nice shot.”
the opposition Labor Party fared even worse, with a predicted 35 seats, seven less than last election.
Topic Models: The Big Picture

Unstructured Collection

Structured Topic Network

Topic Discovery

Dimensionality Reduction

Word Simplex

Topic Simplex

© Eric Xing @ CMU, 2005-2015
LSI versus Topic Model (probabilistic LSI)

\[ \tilde{x} = W' \tilde{d} \]

- **LSI**
  - \( X = W \Lambda D^T \)
  - \( \sum P(w) = \sum P(w|z) P(z) \)
- **Topic models**
  - Topic-Mixing is via repeated word labeling

\( \sum P(w) = \sum P(w|z) P(z) \)
Subspace Analysis

- Clustering: \((0,1)\) matrix
- LSI/NMF: “arbitrary” matrices
- **Topic Models**: stochastic matrix
- Sparse coding: “arbitrary” **sparse** matrices
- “Deep Learning”: do the above for multiple layers
Objects are **bags** of elements

Mixtures are **distributions** over elements

Objects have **mixing vector** $\theta$
  - Represents each mixtures’ contributions

Object is **generated** as follows:
  - Pick a mixture component from $\theta$
  - Pick an element from that component
Aka: Topic Models

Generating a document

- Draw \( \theta \) from the prior
  For each word \( n \)
    - Draw \( z_n \) from \( \text{multinomial}(\theta) \)
    - Draw \( w_n \mid z_n, \{\beta_{1:k}\} \) from \( \text{multinomial}(\beta_{z_n}) \)

Which prior to use?
Choices of Priors

- Dirichlet (LDA) (Blei et al. 2003)
  - Conjugate prior means efficient inference
  - Can only capture variations in each topic’s intensity independently

  - Capture the intuition that some topics are highly correlated and can rise up in intensity together
  - Not a conjugate prior implies hard inference
Generative Semantic of LoNTAM

Generating a document

- Draw $\theta$ from the prior

For each word $n$
- Draw $z_n$ from multinomial $l(\theta)$
- Draw $w_n \mid z_n, \{\beta_{1:k}\}$ from multinomial $l(\beta_{z_n})$

$$\begin{align*}
\theta & \sim LN_K(\mu, \Sigma) \\
\gamma & \sim N_{K-1}(\mu, \Sigma) \quad \gamma_K = 0 \\
\theta_i &= \exp \left\{ \gamma_i - \log \left( 1 + \sum_{i=1}^{K-1} e^{\gamma_i} \right) \right\} \\
C(\gamma) &= \log \left( 1 + \sum_{i=1}^{K-1} e^{\gamma_i} \right)
\end{align*}$$

Problem

- Log Partition Function
- Normalization Constant

© Eric Xing @ CMU, 2005-2015
Posterior inference

Learning and inference in the brain.
Friston K.
The Wellcome Department of Imaging Neuroscience, Institute of Neur London WC1N 3BG, UK k.friston@filion.ucl.ac.uk

Abstract
This article is about how the brain data mines its sensory principles of functional brain anatomy that have emerged f over the past century. These principles are considered in 1
Posterior inference results
Joint probability of all variables

\[
p(\beta, \theta, z, w) = \prod_{k=1}^{K} p(\beta_k | \eta) \prod_{d=1}^{D} p(\theta_d | \alpha) \prod_{n=1}^{N} p(z_{dn} | \theta_d) p(w_{dn} | z_{dn}, \beta)
\]

We are interested in computing the posterior, and the data likelihood!
Inference and Learning are both intractable

- A possible query:
  \[ p(\theta_n | D) = ? \]
  \[ p(z_{n,m} | D) = ? \]

  Close form solution?
  \[ p(\theta_n | D) = \frac{p(\theta_n, D)}{p(D)} \]
  \[ = \sum_{\{z_{n,m}\}} \int \left( \prod_n \left( \prod_m p(w_{n,m} | \beta_{z_n}) p(z_{n,m} | \theta_n) \right) p(\theta_n | \alpha) \right) p(\beta | \eta) d\theta_i d\beta \]

  \[ p(D) = \sum_{\{z_{n,m}\}} \int \cdots \int \left( \prod_n \left( \prod_m p(x_{n,m} | \beta_{z_n}) p(z_{n,m} | \theta_n) \right) p(\theta_n | \alpha) \right) p(\beta | \eta) d\theta_1 \cdots d\theta_N d\beta \]

- Sum in the denominator over \( T^n \) terms, and integrate over \( n \) \( k \)-dimensional topic vectors

- Learning: What to learn? What is the objective function?
Approximate Inference

- Variational Inference
  - Mean field approximation (Blei et al)
  - Expectation propagation (Minka et al)
  - Variational 2nd-order Taylor approximation (Ahmed and Xing)

- Markov Chain Monte Carlo
  - Gibbs sampling (Griffiths et al)
Mean-field assumption

- The fully factorized variational distribution

\[ q(\beta, z_{1:n}) = q(\beta | \lambda) \prod_{i=1}^{n} q(z_i | \phi_i) \]

- Closed-form updates for the mean-field approach with conditional conjugate assumptions.
Mean-field assumption

- True posterior

\[ p(\beta, \theta, z|w) = \frac{p(\beta, \theta, z, w)}{p(w)} \]

- Break the dependency using the fully factorized distribution

\[ q(\beta, \theta, z) = \prod_k q(\beta_k) \prod_d q(\theta_d) \prod_n q(z_{dn}) \]

- Mean-field family usually does NOT include the true posterior.
Minimizing the KL-divergence

- We intend to optimize…

\[ q = \arg \min_q KL(q \| p) \]

- Alternatively, let latent variables be \( h = \{\beta, \Theta, z\} \)

\[
\log p(\omega) = \log \int p(\omega, h) dh = \log \mathbb{E}_{q(h)} \left[ \frac{p(\omega, h)}{q(h)} \right] \\
\geq \mathbb{E}_{q}[\log p(\omega, h)] + \mathcal{H}(q(h)) \\
\triangleq \mathcal{L}(q(h)) \leq \text{Lower bound}
\]

- We can verify \( \log p(\omega) = \mathcal{L}(q) + KL(q \| p) \)
Maximize the lower bound

- The lower bound

\[ \mathcal{L}(q(h)) = \mathbb{E}_q[\log p(w, h)] + \mathcal{H}(q(h)) \]

- The factorized distribution

\[ h = \{\beta, \theta, z\} \]
\[ q(\beta, \theta, z) = \prod_k q(\beta_k) \prod_d q(\theta_d) \prod_n q(z_{dn}) \]

- To be a little more general,

\[ h = \{h_1, h_2, \ldots, h_M\} \]
\[ q(h) = \prod_i q(h_i) \]
A coordinate ascent algorithm

- Let us find the best $q(h_i)$ given $q(h_j), \ j \neq i$ fixed.

- The objective function is

$$\mathcal{L}(q(h_i)) = \int q(h_i) \mathbb{E}_{q_{-i}} [\log p(w, h)] + \mathcal{H}(q(h_i)) + C$$

- The optimal solution is (Bishop, 2006)

$$q(h_i) \propto \exp \{ \mathbb{E}_{q_{-i}} [\log p(w, h)] \}$$

- We iterate over all hidden variables until convergence.
Update each marginals

- Update

\[
q(\theta_d) \propto \exp \left\{ \mathbb{E}_{\prod_n q(z_{dn})} \left[ \log p(\theta_d|\alpha) + \sum_n \log p(z_{dn}|\theta_d) \right] \right\}
\]

- In LDA,

\[
p(\theta_d|\alpha) \propto \exp \left\{ \sum_{k=1}^{K} (\alpha_k - 1) \log \theta_{dk} \right\} \quad \text{---Dirichlet}
\]

\[
p(z_{dn}|\theta_d) = \exp \left\{ \sum_{k=1}^{K} 1[z_{dn} = k] \log \theta_{dk} \right\} \quad \text{---Multinomial}
\]

- We obtain

\[
q(\theta_d) \propto \exp \left\{ \sum_{k=1}^{K} \left( \sum_{n=1}^{N} q(z_{dn} = k) + \alpha_k - 1 \right) \log \theta_{dk} \right\}
\]

This is also a Dirichlet---the same as its prior!
Coordinate ascent algorithm for LDA

1: Initialize variational topics $q(\beta_k)$, $k = 1, ..., K$.
2: repeat
3: for each document $d \in \{1, 2, ..., D\}$ do
4: Initialize variational topic assignments $q(z_{dn})$, $n = 1, ..., N$
5: repeat
6: Update variational topic proportions $q(\theta_d)$
7: Update variational topic assignments $q(z_{dn})$, $n = 1, ..., N$
8: until Change of $q(\theta_d)$ is small enough
9: end for
0: Update variational topics $q(\beta_k)$, $k = 1, ..., K$.
1: until Lower bound $L(q)$ converges
Drawback of coordinate ascent

- Let’s use $q(\beta | \lambda) \triangleq q(\beta)$ to indicate the variational topics.
- The previous algorithm can be summarized in a high level,

1: Initialize global parameters $\lambda$
2: repeat
3: for each document $d \in \{1, 2, ..., D\}$ do
4: Update document-specific variational distributions
5: end for
6: Update global parameters $\lambda$.
7: until Convergence

- What if we have millions of documents? This could be very slow.
The lower bound in a different form

- Some algebra shows the lower bound is (verify yourself)

\[ \mathcal{L}(\lambda, \phi_{1:n}) = E_q[\log p(\beta) - \log q(\beta|\lambda)] + \sum_{i=1}^{n} \left\{ E_q[\log p(x_i, z_i|\beta) - \log q(z_i|\phi_i)] \right\} \]

- This can be simplified as

\[ \mathcal{L}(\lambda, \phi_{1:n}) = f(\lambda) + \sum_{i=1}^{n} g_i(\lambda, \phi_i). \]
The one-parameter lower bound

- Let us maximize the objective w.r.t. to parameter $\phi_{1:n}$ first

$$\mathcal{L}(\lambda) = f(\lambda) + \sum_{i=1}^{n} \max_{\phi_i} g_i(\lambda, \phi_i).$$

- Let

$$\phi_i^* = \max_{\phi_i} g_i(\lambda, \phi_i)$$

- The gradient of $\mathcal{L}(\lambda)$ has the following form,

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = \frac{\partial f(\lambda)}{\partial \lambda} + \sum_{i=1}^{n} \frac{\partial g_i(\lambda, \phi_i^*)}{\partial \lambda}.$$  

- This allows us to stochastic gradient algorithms to estimate $\lambda$.

- Once $\lambda$ is estimated, each $\phi_i$ can be estimated online if needed.
Natural gradient

- But remember our parameter describes a distribution.

- Gradient $\frac{\partial L(\lambda)}{\partial \lambda}$ is usually not the steepest direction.

(from Honkela et al., 2010)
For distributions, natural gradient is the steepest direction.

Since our model is conditional conjugate, variational distribution is also in exponential family,

\[ q(\beta | \lambda) = h(\beta) \exp \left\{ \lambda^\top t(\beta) - a(\lambda) \right\} \]

The Riemannian metric describes the local curvature,

\[ G(\lambda) = \mathbb{E}_q \left[ \frac{\partial \log q(\beta | \lambda)}{\partial \lambda} \frac{\partial \log q(\beta | \lambda)}{\partial \lambda^\top} \right] = \nabla^2 a(\lambda). \]

The natural gradient is as follows (please verify)

\[ g(\lambda) = G(\lambda) \frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} = -\lambda + \gamma + \sum_{i=1}^{n} \psi_i^x(x_i) \]

Setting \( g(\lambda) = 0 \) gives the traditional mean-field update.
Stochastic variational inference using natural inference

1: Initialize global parameters $\lambda_0$, $t = 0$.
2: Set step-size schedule $\rho_t$.
3: **for** $t = 1, ..., \infty$ **do**
4: Sample a data point $i \sim \text{Unif}(1, ..., n)$.
5: Compute the optimal local parameter $\phi_i^*(\lambda_t)$.
6: Perform natural gradient ascent on global parameters $\lambda$,

\[
\lambda_{t+1} = \lambda_t + \rho_t g(\lambda_t) = (1 - \rho_t)\lambda_t + \rho_t \left( \eta + nt\phi_i^*(x_i) \right)
\]

7: **end for**
Choice of $q()$ does matter

$P(\gamma, \{z\}|D)$

$\Sigma^*$ is full matrix

Multivariate Quadratic Approx.

Closed Form Solution for $\mu^*, \Sigma^*$

Log Partition Function

$\log \left( 1 + \sum_{i=1}^{K-1} e^{\gamma_i} \right)$

Tangent Approx.

$\Sigma^*$ is assumed to be diagonal

Numerical Optimization to fit $\mu^*$, $\text{Diag}(\Sigma^*)$

Ahmed&Xing

Blei&Lafferty
Tangent Approximation
How to evaluate?

- Empirical Visualization: e.g., topic discovery on New York Times

The 5 most frequent topics from the HDP on the *New York Times*.

<table>
<thead>
<tr>
<th>game</th>
<th>life</th>
<th>film</th>
<th>book</th>
<th>wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>season</td>
<td>know</td>
<td>movie</td>
<td>life</td>
<td>street</td>
</tr>
<tr>
<td>team</td>
<td>school</td>
<td>show</td>
<td>books</td>
<td>hotel</td>
</tr>
<tr>
<td>coach</td>
<td>street</td>
<td>life</td>
<td>novel</td>
<td>house</td>
</tr>
<tr>
<td>play</td>
<td>man</td>
<td>television</td>
<td>story</td>
<td>room</td>
</tr>
<tr>
<td>points</td>
<td>family</td>
<td>films</td>
<td>man</td>
<td>night</td>
</tr>
<tr>
<td>games</td>
<td>says</td>
<td>director</td>
<td>author</td>
<td>place</td>
</tr>
<tr>
<td>giants</td>
<td>house</td>
<td>man</td>
<td>house</td>
<td>restaurant</td>
</tr>
<tr>
<td>second</td>
<td>children</td>
<td>story</td>
<td>war</td>
<td>park</td>
</tr>
<tr>
<td>players</td>
<td>night</td>
<td>says</td>
<td>children</td>
<td>garden</td>
</tr>
</tbody>
</table>
How to evaluate?

- Test on Synthetic Text where ground truth is known:
Comparison: accuracy and speed

L2 error in topic vector est. and # of iterations

- Varying Num. of Topics
- Varying Voc. Size
- Varying Num. Words Per Document
Comparison: perplexity
Classification Result on PNAS collection

- PNAS abstracts from 1997-2002
  - 2500 documents
  - Average of 170 words per document
- Fitted 40-topics model using both approaches
- Use low dimensional representation to predict the abstract category
  - Use SVM classifier
  - 85% for training and 15% for testing

### Classification Accuracy

<table>
<thead>
<tr>
<th>Category</th>
<th>Doc</th>
<th>BL</th>
<th>AX</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetics</td>
<td>21</td>
<td>61.9</td>
<td>61.9</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>86</td>
<td>65.1</td>
<td>77.9</td>
</tr>
<tr>
<td>Immunology</td>
<td>24</td>
<td>70.8</td>
<td>66.6</td>
</tr>
<tr>
<td>Biophysics</td>
<td>15</td>
<td>53.3</td>
<td>66.6</td>
</tr>
<tr>
<td>Total</td>
<td>146</td>
<td>64.3</td>
<td>72.6</td>
</tr>
</tbody>
</table>

- Notable Difference
- Examine the low dimensional representations below

© Eric Xing @ CMU, 2005-2015
What makes topic models useful -- The Zoo of Topic Models!

- It is a building block of many models.

Williamson et al. 2010  Chang & Blei, 2009  Titov & McDonald, 2008

Conclusion

- GM-based topic models are cool
  - Flexible
  - Modular
  - Interactive
- There are many ways of implementing topic models
  - unsupervised
  - supervised
- Efficient Inference/learning algorithms
  - GMF, with Laplace approx. for non-conjugate dist.
  - MCMC
- Many applications
  - ...
  - Word-sense disambiguation
  - Image understanding
  - Network inference
Summary on VI

- Variational methods in general turn inference into an optimization problem via exponential families and convex duality.

- The exact variational principle is intractable to solve; there are two distinct components for approximations:
  - Either inner or outer bound to the marginal polytope
  - Various approximation to the entropy function

- **Mean field**: non-convex inner bound and exact form of entropy
- **BP**: polyhedral outer bound and non-convex Bethe approximation
- **Kikuchi and variants**: tighter polyhedral outer bounds and better entropy approximations (Yedidia et. al. 2002)