Topic Models, Latent Space Models, Sparse Coding, and All That

A systematic understanding of probabilistic semantic extraction in large corpora

Eric Xing
Carnegie Mellon University

We are inundated with data …

- Humans cannot afford to deal with (e.g., search, browse, or measure similarity) a huge number of text and media documents
- We need computers to help out …
To get started on intelligent systems for automated processing and management of large text or media corpora …

- Here are some important elements to consider before you start:
  - **Task:**
    - Embedding (visualization)? Classification? Clustering? Topic extraction? …
  - **Data representation:**
    - Input and output (e.g., continuous, binary, counts, …)
  - **Model:**
  - **Inference:**
    - MCMC? Variational? Spectrum Analysis?
  - **Learning:**
    - MLE? MCLE? Max margin?
  - **Computation:**
    - Desktop? Hadoop? MPI?
  - **Evaluation:**
    - Visualization? Human interpretability? Perplexity? Predictive accuracy?

- It is better to consider one element at a time!

Tasks:

- Say, we want to have a mapping …, so that
  - Compare similarity
  - Classify contents
  - Cluster/group/categorizing
  - Distill semantics and perspectives
  - …
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

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>SMT Alignment Score</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parse Tree Noun Phrase Grammar CFG</td>
<td>likelihood EM Hidden Parameters Estimation argMax</td>
<td>0.6</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Unigram over vocabulary

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 bilab 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.
**Topic Models: The Big Picture**

- **Unstructured Collection**
  - Topic Discovery

- **Structured Topic Network**
  - Dimensionality Reduction

**Word Simplex**

**Topic Simplex**

---

**Topic Model as a graphical model**

**Generating a document**

- Draw $\theta$ from the prior
  - For each word $n$
    - Draw $z_n$ from $\text{multinomial}(\theta)$
    - Draw $w_n | z_n, \{\beta_k\}$ from $\text{multinomial}(\beta_{z_n})$

We can go beyond this by adding more variables and structures to the graph!
Learning and Inference

Advanced issues:
- Objective function: likelihood? Margin? RSS? ...
- Data: iid docs, streaming text, multimodal media ...
- Algorithm: direct optimization, Monte Carlo, variational methods ...
- System: single machine, multi-care machine, distributed system ...

Deliverables:

We want:
- Topics and categorization of documents
- Semantic-based ranking of docs
- Multimedia inference
- Automatic translation
- Predict how topics evolve
- ...
Questions:

- What is the mathematical and computational basis of all these?
- How to do it right, modular, fast, and real time?
- How to build other related applications on topic models?
- How to scale up?

Plan of this tutorial

- 1. Overview of basic topic models
- 2. Computational Challenges and two classical algorithmic paths
- 3. Scenario I: Multimodal data
- 4. Scenario II: when supervision is available
- 5. Scenario III: what if I don't know the total number of topics
- 7. Advanced subjects: Sparsity in topic modeling (Optional)
- 8. Scalability, Complexity, and Fast algorithms (Optional)
- 9. Other applications (Optional)
1. Overview of topic models

Understanding document corpora

- A document collection is a dataset where each data point is itself a collection of simpler data.
  - Text documents are collections of words.
  - Segmented images are collections of regions.
  - User histories are collections of purchased items.
- Many modern problems ask questions on such data.
  - What topics do these documents “span”?
  - Is this text document relevant to my query?
  - Which category is this text/image in?
  - How have topics changed over time?
  - Who wrote this specific document?
  - What will author X write about?
  - and so on.....
The Vector Space Model

- Represent each document by a high-dimensional vector in the space of words

![Diagram of document vectors]

Latent Semantic Indexing

- LSI does not define a properly normalized probability distribution of observed and latent entities
- Does not support probabilistic reasoning under uncertainty and data fusion

\[ \vec{w} = \sum_{k=1}^{K} d_k \lambda_k \vec{t}_k \]
How our brain might work ... 

Apoptosis + Medicine

probabilistic generative model
How our brain might work …

Apoptosis + Medicine

statistical inference

What is Learning

Learning is about seeking a predictive and/or executable understanding of natural/artificial subjects, phenomena, or activities from …

Apoptosis + Medicine

Grammatical rules
Manufacturing procedures
Natural laws
...

Inference
Connecting Probability Models to Data

(Generative Model)
P(Data | Parameters)

Probabilistic Model

Real World Data

P(Parameters | Data)
(Inference)

What is a Graphical Model?
--- example from a signal transduction pathway

- A possible world for cellular signal transduction:

Receptor A
Kinase C
Gene G

TF F

Receptor B
Kinase D
Gene H

Kinase E

A total of $2^8$ joint state configurations
No "structured insight" of the domain
Recap of Basic Prob. Concepts

- **Representation:** what is the joint probability distribution on multiple variables?
  \[ P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \]
  - How many state configurations in total? --- \(2^8\)
  - Are they all needed to be represented?
  - Do we get any scientific/medical insight?

- **Learning:** where do we get all this probabilities?
  - Maximal-likelihood estimation? but how much data do we need?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- **Inference:** If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?

GM: Structure Simplifies Representation

- **Dependencies among variables**

  ![Diagram](https://example.com/diagram.png)

© Eric Xing @ CMU, ACL Tutorial 2012
Probabilistic Graphical Models

- Represent dependency structure with a graph
  - Node <-> random variable
  - Edges encode dependencies
    - Absence of edge -> conditional independence
  - Directed and undirected versions

- Why is this useful?
  - A language for communication
  - A language for computation
  - A language for development

- Origins:
  - Wright 1920’s
  - Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980’s

If \( X_i \)'s are conditionally independent (as described by a PGM), the joint can be factored to a product of simpler terms, e.g.,

\[
P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) = P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6)
\]

- Why we may favor a PGM?
  - Representation cost: how many probability statements are needed?
  - Algorithms for systematic and efficient inference/learning computation
    - Exploring the graph structure and probabilistic (e.g., Bayesian, Markovian) semantics
  - Incorporation of domain knowledge and causal (logical) structures
Computing statistical queries regarding the network, e.g.:
- Is node X independent on node Y given nodes Z,W?
- What is the probability of X=true if (Y=false and Z=true)?
- What is the joint distribution of (X,Y) if Z=false?
- What is the likelihood of some full assignment?
- What is the most likely assignment of values to all or a subset the nodes of the network?

General purpose algorithms exist to fully automate such computation
- Computational cost depends on the topology of the network
- Exact inference:
  - The junction tree algorithm
- Approximate inference:
  - Loopy belief propagation, variational inference, Monte Carlo sampling

Probabilistic Inference

An (incomplete) genealogy of graphical models

(Picture by Zoubin Ghahramani and Sam Roweis)
Latent Semantic Structure in GM

- Latent Structure $\ell$
- Words $W$
- Distribution over words
  \[ P(w) = \sum_{\ell} P(w, \ell) \]
- Inferring latent structure
  \[ P(\ell | w) = \frac{P(w | \ell) P(\ell)}{P(w)} \]
- Prediction
  \[ P(w_{n+1} | w) = \ldots \]

How to Model Semantics?

- 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>

AdMixing Proportion

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.
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.

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>

- A: Structured way of browsing the collection
- Other tasks
  - Dimensionality reduction
    - TF-IDF vs. topic mixing proportion
    - Classification, clustering, and more …

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.
A possible generative process of a document

Method One:

- Hierarchical Bayesian Admixture (a.k.a. probabilistic Topic Models)
Probabilistic LSI

- A "generative" model
- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- A topic is characterized by a distribution over words.
- Each document is represented as a list of admixing proportions for the components (i.e. topic vector $\theta$).
Latent Dirichlet Allocation

Blei, Ng and Jordan (2003)

Essentially a Bayesian pLSI:

\[ \theta \sim \text{Dir}(\alpha) \]
\[ z_n \sim \text{Mult}(\theta) \]
\[ w_n \sim p(w_n|z_n, \beta) \]

\[
p(w) = \sum_z \int p(\theta) p(\beta) \left( \prod_{n=1}^{N} p(z_n|\theta) p(w_n|\beta_{z_n}) \right) d\theta d\beta
\]

LDA

- Generative model
- Models each word in a document as a sample from a mixture model.
- Each word is generated from a single topic, different words in the document may be generated from different topics.
- A topic is characterized by a distribution over words.
- Each document is represented as a list of admixing proportions for the components (i.e. topic vector).
- The topic vectors and the word rates each follows a Dirichlet prior --- essentially a Bayesian pLSI
"Correlated" Topic Model


\[ \theta \sim \text{Dir}(\alpha) \]
\[ z_n \sim \text{Mult}(\theta) \]
\[ w_n \sim p(w_n|z_n, \beta) \]

\[
p(w) = \sum_z \int p(\theta)p(\beta) \left( \prod_{n=1}^{N} p(z_n|\theta)p(w_n|\beta_{z_n}) \right) \theta d\beta
\]

Topic Models = Mixed Membership Models

Generating a document

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

- Nested CRP (Blei et al. 2005)
  - Defines hierarchy on topics
  - ...

Generative Semantic of LoNTAM

Generating a document

- Draw $\phi$ from the prior
- For each word $n$
  - Draw $z_n$ from $\text{multinomial}(\phi)$
  - Draw $w_n | z_n, \phi_{z_n}$ from $\text{multinomial}(\phi_{z_n})$

$\phi \sim LN_K(\mu, \Sigma)$
$\gamma \sim N_{k\times 1}(\mu, \Sigma), \gamma_k = 0$
$y = \exp\{y - \log(1 + \sum_{i=1}^{k-1} e^{y_i})\}$
$C(y) = \log(1 + \sum_{i=1}^{k} e^{y_i})$

- Log Partition Function
- Normalization Constant

Problem
Outcomes from a topic model

- The “topics” $\beta$ in a corpus:

- There is no name for each “topic”, you need to name it!
- There is no objective measure of good/bad
- The shown topics are the “good” ones, there are many many trivial ones, meaningless ones, redundant ones, ... you need to manually prune the results
- How many topics? …

Outcomes from a topic model

- The “topic vector” $\theta$ of each doc

- Create an embedding of docs in a “topic space”
- Their no ground truth of $\theta$ to measure quality of inference
- But on $\theta$ it is possible to define an “objective” measure of goodness, such as classification error, retrieval of similar docs, clustering, etc., of documents
- But there is no consensus on whether these tasks bear the true value of topic models …
Outcomes from a topic model

- The per-word topic indicator $z$:

> The William Randolph Hearst Foundation will give $1.56 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 as they are 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 $250,000 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.

- Not very useful under the bag of word representation, because of loss of ordering
- But it is possible to define simple probabilistic linguistic constraints (e.g., bi-grams) over $z$ and get potentially interesting results [Griffiths, Steyvers, Blei, & Tenenbaum, 2004]

Outcomes from a topic model

- The topic graph $\Sigma$ (when using CTM):

Kind of interesting for understand/visualizing large corpora
Method Two:

- Layered Boltzmann machines (an undirected Topic Model)

The Harmonium

Boltzmann machines:

\[ p(x, h | \theta) = \exp \left\{ \sum_i \theta_i \phi_i(x_i) + \sum_j \theta_j \phi_j(h_j) + \sum_{i,j} \theta_{ij} \phi_{ij}(x_i, h_j) - A(\theta) \right\} \]
A Binomial Word-count Model

$$h_j = 3: \text{topic } j \text{ has strength } 3$$
$$h_j \in \mathbb{R}, \quad \langle h_j \rangle = \sum_i W_{i,j} x_i$$

$$x_i = n: \text{word } i \text{ has count } n$$
$$x_i \in \mathbb{I}$$

$$p(h \mid x) = \prod_j \text{Normal}_{h_j} \left[ \sum_i \hat{W}_{ij} x_i, 1 \right]$$

$$p(x \mid h) = \prod_i \text{Bi}_{\alpha_i} \left[ N, \frac{\exp(\alpha_j + \sum_j W_{ij} h_j)}{1 + \exp(\alpha_j + \sum_j W_{ij} h_j)} \right]$$

$$\Rightarrow p(x) \exp \left\{ \left( \sum_i \alpha_i x_i - \log \Gamma(x_i) - \log \Gamma(N - x_i) \right) + \frac{1}{2} \sum_j \left( \sum_i W_{ij} x_i \right)^2 \right\}$$

The Computational Trade-off

**Undirected model:** Learning is hard, inference is easy.

**Directed Model:** Learning is "easier", inference is hard.

Example: Document Retrieval.

Retrieval is based on comparing (posterior) topic distributions of documents.
- **Directed models:** inference is slow. Learning is relatively "easy".
- **Undirected model:** inference is fast. Learning is slow but can be done offline.
Method Three:

**Sparse topic coding (a non-probabilistic Topic Model)**
- And in this category recently there is also nonnegative matrix factorization (NMF)

**Sparse Coding**
- Let $X$ be a signal, e.g., speech, image, etc.
- Let $\beta$ be a set of normalized “basis vectors”
  - We call it dictionary
- $\beta$ is “adapted” to $x$ if it can represent it with a few basis vectors
  - There exists a sparse vector $\theta$ such that $x \approx \beta \theta$
  - We call $\theta$ the sparse code
**Primer on Sparse Coding**

- **Sparse Coding** with appropriate constraints:
  \[
  \min_{\theta, \beta} \sum_d \ell(d, \beta|x_d) + \lambda \Psi(\theta)
  \]
  s.t. : \( \beta \in \Omega_1; \theta \in \Omega_2 \).

- Reconstruction loss can be:
  - the general log-likelihood loss of an exponential family distribution (Lee et al., 2010)
  - Sparsity-inducing regularizer can be:
    - the \( \ell_0 \) "pseudo-norm": \( \|\theta\|_0 := \sum_i \delta(\theta_i, 0) \)
    - the \( \ell_1 \) norm: \( \|\theta\|_1 := \sum_i |\theta_i| \)
    - Structured regularizers, e.g., group Lasso (Bengio et al., 2009) \( \|\theta\|_{1/2} := \|\theta_{x_k}\|_2 \).
    - Suggests an alternating optimization procedure.

**Sparse Topical Coding**

- **Goal:** design a non-probabilistic topic model that is amenable to
  - direct control on the posterior sparsity of inferred representations
  - avoid dealing with normalization constant when considering supervision or rich features
  - seamless integration with a convex loss function (e.g., svm hinge loss)

- We extend sparse coding to hierarchical sparse topical coding
  - word code \( \theta \)
  - document code \( s \)
    \[
    \min_{\{s, \theta_k\}, \theta} \sum_{d} \ell(d, \theta|x_d) + \lambda \sum_d \|\theta_d\|_1 + \sum_{k} \|\theta_k\|_2
    \]
    s.t.: \( \theta_d \geq 0, s_d \geq 0, \forall d \in D; \beta_k \in \mathbb{P}^k, \forall k \).

J. Zhu & E.P. Xing, UAI, 2011
Summary: Latent Sub-space Models

The Model:

\[ P(w, \theta; \beta) \]

Inferring latent representation:

\[ P(\theta | w) = \frac{P(w, \theta)}{P(w)} \]

Learning the subspace:

\[ \beta = \text{arg min } f_\beta (w, \theta) \]

The Big Picture

Unstructured Collection → Structured Topic Network

Word Simplex → Topic Space (e.g., a Simplex)

Topic Discovery → Dimensionality Reduction
## Comparison of model semantics

<table>
<thead>
<tr>
<th>Documents</th>
<th>Words</th>
<th>Topic</th>
<th>Words</th>
<th>Topical Distribution</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>$W$</td>
<td>B</td>
<td>$\Sigma$</td>
<td>$\Theta$</td>
<td>$\mathbb{W} = B' \bar{\theta}$</td>
<td></td>
</tr>
</tbody>
</table>

STC/NMF/LSI

$P(W) = \sum_{\zeta} P(w | \zeta) \cdot P(\zeta)$

LDA

$p(\mathbb{W}) \propto z \theta$

Harmonium

$p(\mathbb{W}) \propto B' \bar{\theta}$

Topic-Mixing is via marginalizing over word labeling.

Mixing is via determining individual word rate.

## Comparison of topic space

<table>
<thead>
<tr>
<th>Word Simplex</th>
<th>Topic Simplex</th>
<th>Word Count Quadrant</th>
</tr>
</thead>
</table>

© Eric Xing @ CMU, ACL Tutorial 2012
Comparisons

LDA vs. Harmonium
[Xing, Yan, and Hauptman, UAI 2005]

- LDA is actually doing very poor on several "objectively" evaluable predictive tasks

LDA vs. STC
[Zhu and Xing, UAI 2011]
Sparse word codes

- Sparsity ratio: percentage of non-zeros

![Graph showing sparsity ratio vs. topics for different methods](image)

- NMF: non-negative matrix factorization
- MedLDA (Zhu et al., 2009)
- regLDA: LDA with entropic regularizer
- gaussSTC: use L2 rather than L1-norm

2. Computational Challenges and three algorithmic paths
Computation on LDA

- **Inference**
  - Given a Document $D$
    - Posterior: $P(\Theta | \mu, \Sigma, \beta, D)$
    - Evaluation: $P(D | \mu, \Sigma, \beta)$

- **Learning**
  - Given a collection of documents $\{D_i\}$
    - Parameter estimation

\[
\arg \max_{(\mu, \Sigma, \beta)} \sum \log(P(D | \mu, \Sigma, \beta))
\]

Exact Bayesian inference on LDA is intractable

- A possible query:
  \[
p(\pi_x | D) = ?
\]
  \[
p(\pi_{x,m} | D) = ?
\]
  - Close form solution?

\[
p(D) = \sum_{\{\pi_x\}} \int \int \left( \prod_m p(X_{x,m} | \phi_x) p(\pi_x | \alpha) \right) p(\phi | G) d\pi_x d\phi
\]

- Sum in the denominator over $T^h$ terms, and integrate over $n \lambda$-dimensional topic vectors
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)

Collapsed Gibbs sampling
(Tom Griffiths & Mark Steyvers)

- Collapsed Gibbs sampling
  - Integrate out $\theta$

For variables $z = z_1, z_2, \ldots, z_n$

Draw $z_i^{(t+1)}$ from $P(z_i | z_{-i}, w)$

$z_{-i} = z_1^{(t+1)}, z_2^{(t+1)}, \ldots, z_{i-1}^{(t+1)}, z_{i+1}^{(t)}, \ldots, z_n^{(t)}$

$\{z^{(1)}, z^{(2)}, \ldots, z^{(T)}\}$

$\theta = \frac{1}{T} \sum_t z^{(t)}$
Gibbs sampling

- Need full conditional distributions for variables
- Since we only sample $z$ we need

$$P(z_i = j | z_{-i}, w) \propto P(w_i | z_i = j, z_{-i}, w) P(z_i = j | z_{-i})$$

$$= \frac{n_{-i,j}^{(w_i)} + \beta}{n_{-i,j}^{(z_i)} + W \beta} \frac{n_{-i,j}^{(d_i)} + \alpha}{n_{-i,j}^{(d_i)} + T \alpha}$$

$n_j^{(w)}$ number of times word $w$ assigned to topic $j$

$n_j^{(d)}$ number of times topic $j$ used in document $d$

---

Gibbs sampling

<table>
<thead>
<tr>
<th>iteration</th>
<th>$i$</th>
<th>$w_i$</th>
<th>$d_i$</th>
<th>$z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>WORK</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>WORK</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>JOY</td>
<td>5</td>
<td>2</td>
</tr>
</tbody>
</table>
### Gibbs sampling

<table>
<thead>
<tr>
<th>iteration</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$i$</td>
<td>$w_i$</td>
<td>$d_i$</td>
</tr>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
</tr>
</tbody>
</table>

$P(z_i = j | z_{-i}, w) \propto \frac{n_{z_i, j}^{(w_i)} + \beta}{n_{z_i, j}^{(d_i)} + \alpha} \frac{n_{-z_i, j}^{(d_i)} + \alpha}{n_{-z_i, j}^{(d_i)} + \beta W}$
Gibbs sampling

\[ P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(w_i)}_{i,j} + \beta}{n^{-1}_{-i,j}} + \frac{n^{(d_i)}_{i,j} + \alpha}{W \beta n^{-1}_{-i,i} + T \alpha} \]

© Eric Xing @ CMU, ACL Tutorial 2012
**Gibbs sampling**

<table>
<thead>
<tr>
<th>iteration</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
<td>( w_i )</td>
<td>( d_i )</td>
</tr>
<tr>
<td>1</td>
<td>MATHEMATICS</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>KNOWLEDGE</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>RESEARCH</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>WORK</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>MATHEMATICS</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>RESEARCH</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>WORK</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>SCIENTIFIC</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>MATHEMATICS</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>WORK</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>SCIENTIFIC</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>KNOWLEDGE</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>50</td>
<td>JOY</td>
<td>5</td>
</tr>
</tbody>
</table>

\[
P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n^{(w_i)}_{-1,j} + \beta}{n^{(j)}_{-1,j} + W \beta n^{(d_i)}_{-1,j}} + T \alpha
\]

© Eric Xing @ CMU, ACL Tutorial 2012
Gibbs sampling

\[ P(z_i = j | \mathbf{z}_{-i}, \mathbf{w}) \propto \frac{n_{-1,j}^{(w_i)} + \beta}{n_{-1,j}} + W \beta \frac{n_{-1,j}^{(d_i)} + \alpha}{n_{-1,i}} + T \alpha \]

© Eric Xing @ CMU, ACL Tutorial 2012
Variational Inference

(e.g., MF, Jordan et al 1999, GMF, Xing et al 2004)

- Variational approximation
  \[ q(\theta, z) = q_\theta(\theta)q_z(z) \]
  \[ = \text{Dir}(\theta | \gamma = f(\alpha, \{z\})) \times \]
  \[ \text{Mult}(z | \phi = f(\beta, \{\ln \phi\})) \]

- Data set:
  - 15,000 documents
  - 90,000 terms
  - 2.1 million words

- Model:
  - 100 factors
  - 9 million parameters

- On a single machine MCMC could converge too slowly for this problem

Learning a TM

- Maximum likelihood estimation:
  \[
  \{\beta_1, \beta_2, \ldots, \beta_K\}, \alpha = \arg \max_{(\alpha, \beta)} \sum \log P(D | \alpha, \beta)
  \]

- Need statistics on topic-specific word assignment (due to \(z\)), topic vector distribution (due to \(\theta\)), etc.
  - E.g., this is the formula for topic \(k\):
    \[
    \beta_k = \frac{1}{\sum_d N_d} \sum_{d=1}^{D} \sum_{d_n=1}^{N_d} \delta(z_{d,d_n}, k) w_{d,d_n}
    \]

- These are hidden variables, therefore need an EM algorithm (also known as data augmentation, or DA, in Monte Carlo paradigm)

- This is a “reduce” step in parallel implementation
The Correlated Topic Model

Two approaches to approximate it:
- Blei and Lafferty use tangent
- (Xing 2005) uses second order truncated Taylor approximation

Variational Inference of CTM

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

Ahmed&Xing 05
Variational Inference With no Tears

Iterate until Convergence

- Pretend you know $E[Z_{1:n}]$
  - $P(\gamma|E[z_{1:n}], \mu, \Sigma)$
- Now you know $E[\gamma]$
  - $P(z_{1:n}|E[\gamma], w_{1:n}, \beta_{1:k})$

More Formally:

$$q^*(X_C) = P(X_C|\langle S_C \rangle_{q_\gamma} : \forall y \in X_{MB})$$

Message Passing Scheme (GMF)
Equivalent to previous method (Xing et al. 2003)

LoNTAM Variations Inference

- Fully Factored Distribution
  $$q(\gamma, z_{1:n}) = q(\gamma) \prod q(z_n)$$
- Two clusters: $\lambda$ and $Z_{1:n}$
  $$q^*(X_C) = P(X_C|\langle S_C \rangle_{q_\gamma}, : \forall y \in X_{MB})$$
- Fixed Point Equations
  $$q_\gamma^*(\gamma) = P(\gamma|\langle S_C \rangle_{q_\gamma}, \mu, \Sigma)$$
  $$q_z^*(z) = P(z|\langle S_C \rangle_{q_\gamma}, \beta_{1:k})$$

© Eric Xing @ CMU, ACL Tutorial 2012
Variational $\gamma$

$$q_\lambda^* (\gamma) = \mathcal{P}(\gamma | S_{z_{q_\lambda}}, \mu, \Sigma_{q_\lambda})$$

$$\propto \mathcal{P}(\gamma, \mu, \Sigma) \mathcal{P}(S_{z_{q_\lambda}} | \gamma)$$

$$S_z = m = \left[ \sum_q I(z_q = 1), ..., \sum_q I(z_q = k) \right]$$

$$\propto N(\gamma, \mu, \Sigma) \exp \left\{ m \gamma - N \times C(\gamma) \right\}$$

$$\propto \exp \left\{ -\frac{1}{2} \gamma \Sigma^{-1} \gamma + \gamma \Sigma^{-1} \mu + \langle m \rangle \gamma - N \times C(\gamma) \right\}$$

$$C(\gamma) = C(\gamma,) + g' (\gamma - \gamma,) + .5 (\lambda - \lambda,) H(\gamma - \gamma,)$$

$$q_\lambda^* (\gamma) = \mathcal{N}(\mu, \Sigma)$$

$$\mu = \Sigma \Sigma^{-1} \mu + N H (\gamma, - \lambda, H (\gamma, - \lambda,)$$

Variational $Z$

$$q_z^* (z) = \mathcal{P}(z | S_{z_{q_z}}, \beta, w)$$

$$\propto \mathcal{P}(z^* | S_{z_{q_z}}) \mathcal{P}(\beta, \gamma)$$

$$\propto \mathcal{P}(z^* | \gamma) \mathcal{P}(\beta)$$

$$\propto \exp \left\{ -\frac{1}{2} \mu \Sigma^{-1} \mu + N H (\gamma, - \lambda, H (\gamma, - \lambda,)$$

Now what is $\langle S_z \rangle_{q_z}^*$?
Tangent Approximation

Different Learning/Inference deliver different performance

Test on Synthetic Text (of “known” ground truth):
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

Result on NIPS collection

- NIPS proceeding from 1988-2003
- 14036 words
- 2484 docs
- 80% for training and 20% for testing
- Fit both models with 10,20,30,40 topics
- Compare perplexity on held out data
  - The perplexity of a language model with respect to text x is the reciprocal of the geometric average of the probabilities of the predictions in text x. So, if text x has k words, then the perplexity of the language model with respect to that text is

  \[ Pr(x)^{-1/k} \]
Comparison: perplexity

![Comparison: perplexity](image)

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

<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
**Computation on undirected TM**

[Welling et al NIPS 04, Xing et al, UAI 05]

**Undirected model**: Learning is hard, inference is easy.

**Directed Model**: Learning is "easier", inference is hard.

Example: Document Retrieval.

- **directed models**: inference is slow. Learning is relatively "easy".
- **undirected model**: inference is fast. Learning is slow but can be done offline.

---

**Properties of Directed Networks**

- Factors are marginally *independent*.
- Factors are conditionally *dependent* given observations on the visible nodes.
  \[
  P(\ell | w) = \frac{P(w | \ell) P(\ell)}{P(w)}
  \]
- Easy ancestral sampling.
- Learning with (variational) EM

\[
\max_{\theta_t} Q(\theta | \theta_{old})
\]
Properties of Harmoniums

- Factors are marginally dependent.
- Factors are conditionally independent given observations on the visible nodes.
  \[ P(\ell | w) = \prod_i P(\ell_i | w) \]
- Iterative Gibbs sampling.
- Learning with contrastive divergence

Learning and Inference

- Maximal likelihood learning based on gradient ascent.
  \[ \delta \theta_i \propto \left\{ f_i(x_i) \right\}_{\text{data}} - \left\{ f_i(x_i) \right\}_p \]
  - gradient computation requires model distribution \( p(.) \)
  - \( p(.) \) is intractable

- Contrastive Divergence
  - approximate \( p(.) \) with Gibbs sampling

- Variational approximation
  - GMF approximation
  \[ q(x, z, h) = \prod_i q(x_i | v_i) \prod_i q(z_i | \mu_z, \alpha_z) \prod_j q(h_j | \gamma_j) \]
Performance

Opt. Algorithm for Sparse Coding

- Much research has been done for optimizing a convex, but non-smooth objective (may subject to some constraints, e.g., non-negativity)
- Greedy algorithm for the non-convex $L_0$ “pseudo-norm”:
  - select the element with maximum correlation with the residual
  - known as “matching pursuit” (Mallat & Zhang, 1993)
- For the convex $L_1$ norm, many algorithms:
  - Soft-thresholding with coordinate descent (Friedman et al., 2007; Fu, 1998; Zhu & Xing, 2011)
  - Proximal methods (Nesterov, 2007; Jenatton et al., 2010)
  - Active-set methods (Roth & Fischer, 2008)
  - Iterative Re-weighted Least Squares (Daubechies et al., 2008)
  - LARS (Efron et al., 2004) solves for regularization path
  - Online/stochastic variants
  - …
Opt. Algorithm for Dictionary Learning

- Optimize a convex and usually smooth objective w/o (convex) constraints

- General optimization procedure can be applied, less research has been done for this step
  - Projected gradient descent
  - Block-wise coordinate descent
  - …
- A recent progress is made on online/stochastic optimization method (Mairal et al., 2010)

Computation on STC [Zhu and Xing, UAI 11]

- Hierarchical sparse coding
  - for each document
    $$\min_{\delta, s} \sum_{n \in I} f(w_n, s_n^T \theta_n) + \lambda \|\theta\|_1 + \sum_{n \in I} (\gamma \|s_n - \theta\|_2^2 + \rho \|s_n\|_1)$$
  - s.t. $$\vartheta \geq 0; \quad s_n \geq 0, \quad \forall n \in I$$
  - Word code
    $$s_{nk} = \max(0, \nu_k)$$
    where $$2\gamma \beta_{kn} \nu_k^2 + (2\gamma \mu + \beta_{kn} \eta) \nu_k + \mu \eta - w_n \beta_{kn} = 0$$
  - Document code (truncated averaging)
    $$\theta_k = \max(0, \bar{s}_k - \frac{\lambda}{2\gamma I})$$ where $$\bar{s}_k = \frac{1}{|I|} \sum_{n \in I} s_{nk}$$

- Dictionary learning
  - projected gradient descent
  - any faster alternative method can be used
Performance

3. Scenario I: Multimodal data
Annotated data

- Many examples of multi-type data where one type serves as a description of another type.
  - Images and their captions
  - Scholarly articles and their references
  - Genes and their functions
- What can we do with annotated data:
  - Which class does this image/caption belong in?
  - Describe this image with words.
  - Is this image relevant to this query of words?
- Joint probabilistic models to answer these questions:
  - Provide a good joint distribution (as before)
  - Provide good conditional distributions of the description type conditioned on the primary type.

Latent Space Models for Images

“beach”

Latent Dirichlet Allocation (LDA)

© Eric Xing @ CMU, ACL Tutorial 2012
To Generate an Image ...
Annotated images

- Forsyth et. al. (2001): images as documents where region-specific feature vectors are like visual words.
- A captioned image can be thought of as annotated data: two documents, one of which describes the other.

Gaussian-multinomial LDA

- A natural next step is to glue two LDA models together.
- Bottom: a traditional LDA model on captions
- Top: a Gaussian-LDA model on images
  - each region is a multivariate Gaussian
- Does not work well
Exchangeability

- Like LDA, GM-LDA implicitly makes an *exchangeability* assumption about words and regions, and their corresponding topics.
- The order in which words and regions are generated does not matter.
- But this is goes against the way we’re thinking about the data!
- The words are chosen to describe the image.
- The implicit exchangeability assumptions in the model should reflect this. In other words, we want to model *partial exchangeability*.

Corr-LDA

- Since, w is conditioned on z, the image must be generated first.
- Unlike GM-LDA, the caption is guaranteed to be generated by a subset of the same hidden factors which generated the image.
- The model enforces a correspondence between the latent space associated with images and the latent space associated with captions.
Automatic annotation

True caption
birds tree
Corr–LDA
birds nest leaves branch tree
GM–LDA
water birds nest tree sky
GM–Mixture
tree ocean fungus mushrooms coral

True caption
fish reefs water
Corr–LDA
fish water ocean tree coral
GM–LDA
water sky vegetables tree people
GM–Mixture
fungus mushrooms tree flowers leaves

Text-based image retrieval

Candy
Sunset
People & Fish
Multi-view social media data

Friendship Network

User Text

Interest

Friends

Interest Labels

User image

Latent space models for network

- Micro-inference vs. Meso- or Macro-inference
- Multi-role of every node
- Context dependent role-instantiation
- Role dynamics
Example:

Mixed Membership Stochastic Blockmodel

[airoldi, blei, fienberg and xing, 2008]

1. \( \{ \theta_i \}_{i=1}^N \sim p(\theta | \alpha) \equiv \text{Dirichlet}(\theta; \alpha) \) sample mixed membership vectors.

2. For each actor \( v_j \) that actor \( v_i \) possibly interacts with:
   - \( z_{i \rightarrow j} \sim \text{Multinomial}(z | \theta_i) \) sample an indicator for \( v_i \);
   - \( z_{i \leftarrow j} \sim \text{Multinomial}(z | \theta_j) \) sample an indicator for \( v_j \);
   - \( e_{ij} \sim \text{Bernoulli}(e | z_{i \leftarrow j}, B z_{i \rightarrow j}) \) sample a link.
In the mixed-membership simplex [Airoldi, Blei, Fienberg and Xing, 2008]

The “Facebook” model

© Eric Xing @ CMU, ACL Tutorial 2012
A peep of the Facebook community
The Harmonium Counterpart

\[ p(z | h) = \prod_{i} \text{Normal} \left( \sigma^{2}(\alpha_j + \sum_j W_{ij}^2), \sigma^{2} \right), \quad p(h | x, z) = \prod_{j} \text{Normal} \left( \sum_j W_{ij}^2 x_i + \sum_k U_{kj}^2 z_k, \sigma^2 \right) \]

Inter-Source Associations

Z and X are marginally dependent (same as GH-LDA)
Multi-wing Harmoniums

Examples of Latent Topics

- $T_1$: storms gulf hawaii low forecast southeast showers
- $T_2$: rebounds 14 shouting tests guard cut hawks
- $T_3$: engine flying craft asteroid say hour aerodynamic
- $T_4$: safe cross red sure dry providing services
- $T_5$: losing jersey sixth antonio david york orlando
Are we done?

- What was our task?
  - Embedding (lower dimensional representation): yes, Dec $\rightarrow \theta$
  - Distillation of semantics: kind of, we've learned "topics" $\beta$
  - Classification: is it good?
  - Clustering: is it reasonable?
  - Other predictive tasks?

4. Scenario II: when supervision is available
Problem 1: Discriminative topic models for text classification/scoring

- Democratic or republican?
- Movie review/scoring

We want to answer ...

- Are we satisfying with the conventional topic models and the MLE method for PREDICTION?
- Can we learn a PREDICTIVE model better?
The shocking results on LDA

- LDA is actually doing very poor on several "objectively" evaluable predictive tasks

Why?

- LDA is not designed, nor trained for such tasks, such as classification, there is not warrantee that the estimated topic vector $\theta$ is good at discriminating documents
Unsupervised Latent Subspace Discovery

- Finding latent subspace representations (an old topic)
  - Mapping a high-dimensional representation into a latent low-dimensional representation, where each dimension can have some interpretable meaning, e.g., a semantic topic

- Examples:
  - Topic models (aka LDA) [Blei et al 2003]
  - Total scene latent space models [Li et al 2009]
  - Multi-view latent Markov models [Xing et al 2005]
  - PCA, CCA, ...

Unsupervised latent subspace representations are generic but can be sub-optimal for predictions.

Many datasets are available with supervised side information:
- Tripadvisor Hotel Review (http://www.tripadvisor.com)
- LabelMe (http://labelme.csail.mit.edu/)
- Flickr (http://www.flickr.com/)
- Many others

Can be noisy, but not random noise (Ames & Naaman, 2007)
- Labels & rating scores are usually assigned based on some intrinsic property of the data
- Helpful to suppress noise and capture the most useful aspects of the data

Goals:
- Discover latent subspace representations that are both predictive and interpretable by exploring weak supervision information

Predictive Subspace Learning with Supervision

Unsupervised latent subspace representations are generic but can be sub-optimal for predictions.

Many datasets are available with supervised side information:
- Tripadvisor Hotel Review (http://www.tripadvisor.com)
- LabelMe (http://labelme.csail.mit.edu/)
- Flickr (http://www.flickr.com/)
- Many others
I. Supervised Topic Model

- How to integrate the max-margin principle into a probabilistic latent variable model?

<table>
<thead>
<tr>
<th>Max-Likelihood Estimation</th>
<th>Max-Margin and Max-Likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>sLDA</td>
<td>MedLDA</td>
</tr>
</tbody>
</table>

© Eric Xing @ CMU, ACL Tutorial 2012

Supervised Topic Model

- LDA ignores documents’ side information (e.g., categories or rating score), thus lead to suboptimal topic representation for supervised tasks

- Supervised Topic Models handle such problems, e.g., sLDA (Blei & McAuliffe, 2007) and DiscLDA (Simon et al., 2008)

- Generative Procedure (sLDA):
  - For each document $d$:
    - Sample a topic proportion $\theta_d \sim \text{Dir}(\alpha)$
    - For each word:
      - Sample a topic $Z_{d,n} \sim \text{Mult}(\theta_d)$
      - Sample a word $W_{d,n} \sim \text{Mult}(\beta_d,n)$
  - Sample $y_d$:
    - $y_d \sim N(\eta^T Z_d, \delta^2)$
    - $y_d \sim \text{GLM}(Z_d, \eta^T, \delta^2)$

© Eric Xing @ CMU, ACL Tutorial 2012
How to train sLDA?

- Maximize $P(Y, W)$?

- Maximize $P(Y|W)$?

- Support vector machines

Support vector machines

$$
\min_{w, b} \frac{1}{2} w^T w + C \sum_{i} \xi_i \\
\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i \\
\xi_i \geq 0, \quad \forall i
$$
SVM using VC-dimension

VC Theory
(Vapnik, 1982)
Given \( x_1, \ldots, x_n \in \mathbb{R}^d \) iid and \( \|x_i\|_2 \leq D \), if \( \mathcal{H}_\gamma \) is the hypothesis space of linear classifiers in \( \mathbb{R}^d \) with margin \( \gamma \),

\[
\text{VC}(\mathcal{H}_\gamma) \leq \min \left\{ d, \left\lceil \frac{4D^2}{\gamma^2} \right\rceil \right\}.
\]

\[
\text{error}_{\text{true}}(h) < \text{error}_{\text{train}}(h) + \sqrt{\frac{\text{VC}(H)(\ln \frac{2m}{\text{VC}(H)} + 1) + \ln \frac{4}{\delta}}{m}}
\]
MLE versus max-margin learning

- Likelihood-based estimation
  - Probabilistic (joint/conditional likelihood model)
  - Easy to perform Bayesian learning, and incorporate prior knowledge, latent structures, missing data
  - Bayesian regularization!!

- Max-margin learning
  - Non-probabilistic (concentrate on input-output mapping)
  - Not obvious how to perform Bayesian learning or consider prior, and missing data
  - Sound theoretical guarantee with limited samples

- Maximum Entropy Discrimination (MED) (Jaakkola, et al., 1999)
  - Model averaging
    \[ y = \text{sign} \int p(w) F(x; w) \, dw \quad (y \in \{+1, -1\}) \]
  - The optimization problem (binary classification)
    \[ \min_L KL[p(\Theta)||p_0(\Theta)] \]
    \[ \text{s.t. } \int p(\Theta)|y, F(x; w) - \xi|d\Theta \geq 0, \forall i \]

where \( \Theta \) is the parameter \( w \) when \( \xi \) are kept fixed or the pair \( (w, \xi) \) when we want to optimize over \( \xi \)

© Eric Xing @ CMU, ACL Tutorial 2012

A road map for max-margin learning

- SVM
  \[ y = \text{sign}(w^T x + b) \]
  \[ \min_w \frac{1}{2}||w||^2 + C \sum_i \xi_i \]
  \[ y'(w^T x'_i + b) \geq 1 - \xi_i, \forall i \]

- MED
  \[ y = \text{sign}(f(x, w), \xi) \]
  \[ \min_w KL(Q||Q_w) \]
  \[ y'(f(x', w)) \geq \xi_i, \forall i \]

- MED-MN
  \[ y = \text{arg} \max_{y \in \{+1\}} F(x, y; w) \]
  \[ \min_w \frac{1}{2}||w||^2 + C \sum_i \xi_i \rightarrow w^T [f(x) - f(x', y)] \geq f(y', y) - \xi_i, \forall i, y \neq y' \]

- M^3N
  \[ y = \text{arg} \max_{y \in \{+1\}} F(x, y; w) \]
  \[ \min_w \frac{1}{2}||w||^2 + C \sum_i \xi_i \rightarrow w^T [f(x) - f(x', y)] \geq f(y', y) - \xi_i, \forall i, y \neq y' \]

MED-MN ? = SMED + "Bayesian" M^3N

Primal and Dual Sparse!
MaxEnt Discrimination Markov Network

- Structured MaxEnt Discrimination (SMED):

\[ P_1 : \min_{p(w) : \xi} KL(p(w)||p_0(w)) + U(\xi) \]

s.t. \( p(w) \in \mathcal{F}_1 \), \( \xi \geq 0, \forall i \).

*generalized maximum entropy or regularized KL-divergence

- Feasible subspace of weight distribution:

\[ \mathcal{F}_1 = \{ p(w) : \int p(w) [\Delta F(y; w) - \Delta \ell_i(y)] dw \geq -\xi, \forall i, \forall y \neq y' \}. \]

*expected margin constraints.

- Average from distribution of \( \mathbb{M}^3 \)Ns

\[ h_1(x; p(w)) = \arg \max_{y \in \mathcal{Y}(x)} \int p(w) F(x, y; w) dw \]

MedLDA: a max-margin approach

- Big picture of supervised topic models
  - sLDA: optimizes the joint likelihood for regression and classification
  - DiscLDA: optimizes the conditional likelihood for classification ONLY
  - MedLDA: based on max-margin learning for both regression and classification
MedLDA Regression Model

- **Bayesian sLDA:**

- **MED Estimation:**

  \[
  \text{P1} (\text{MedLDA}) : \min_{q, \alpha, \beta, \delta^2, \xi, \xi_d^*} \mathcal{L}(q) + C \sum_{d=1}^{D} (\xi_d + \xi_d^*)
  \]

  \[
  \text{s.t.} \forall d : \begin{cases}
    y_d - E[\eta^T Z_d] \leq \epsilon + \xi_d, \mu_d \\
    -y_d + E[\eta^T Z_d] \leq \epsilon + \xi_d^*, \mu_d \\
    \xi_d \geq 0, v_d \\
    \xi_d^* \geq 0, \gamma_d
  \end{cases}
  \]

- **Variational bound**

  \[
  q(\theta, z, \eta; \gamma, \phi) \sim p(\theta, z, \eta|\alpha, \beta, \delta^2, y, W)
  \]

  \[
  \mathcal{L}(q) \doteq -E[\log p(\theta, z, \eta; y, y', W|\alpha, \beta, \delta^2)] - H(q(\theta, \eta)) \geq -\log p(\eta, W|\alpha, \beta, \delta^2)
  \]

- **Predictive Rule:**

  \[
  \bar{y} = E[Y|w_{1:N}, \alpha, \beta, \delta^2] = \sum_{z} \sum_{\eta} \eta^T \bar{Z} | w_{1:N}, \alpha, \beta, \delta^2
  \]

MedLDA Classification Model

- **Bayesian sLDA:**

- **Multiclass MedLDA Classification Model:**

  \[
  \text{P2} (\text{MedLDA}) : \min_{q(\theta), \alpha, \beta, \delta^2} \mathcal{L}(q) + C \sum_{d=1}^{D} \xi_d
  \]

  \[
  \text{s.t.} \forall d, y \neq y_d : E[\eta^T \Delta_{\theta}(y)] \geq 1 - \xi_d, \xi_d \geq 0.
  \]

- **Variational bound**

  \[
  q(\theta, z, \eta; \gamma, \phi) \sim p(\theta, z, \eta|\alpha, \beta, \delta^2, y, W)
  \]

  \[
  \mathcal{L}(q) \doteq -E[\log p(\theta, z, \eta; y, y', W|\alpha, \beta, \delta^2)] - H(q(\theta, \eta)) \geq -\log p(\eta, W|\alpha, \beta, \delta^2)
  \]

- **Predictive Rule:**

  \[
  y^* = \text{arg} \max \mathbb{E}[\eta^T \mathbf{f}(y, \bar{Z})|\alpha, \beta]
  \]
**Variational EM Alg.**

- **E-step**: infer the posterior distribution of hidden r.v. $(\theta, z, \eta)$
- **M-step**: estimate unknown parameters $(\alpha, \beta, \delta^2)$

- Independence assumption: $q(\theta, z, \eta | \gamma, \phi) = q(\eta) \prod_{d=1}^{D} q(\theta_d | \gamma_d) \prod_{n=1}^{N} q(z_{dn} | \phi_{dn})$

$$L(\gamma, \phi, q(\eta), \alpha, \beta, \delta^2, \xi, \xi^*, \mu^*, \nu^*, e, v^*) = L(q) + C \sum_{d=1}^{D} (\xi_d + \xi_d^*) - \sum_{d=1}^{D} \sum_{n=1}^{N} \phi_{dn} (\phi_{dn} - 1)$$

- The first two terms are the same as in LDA
- The third and fourth terms are similar to those of sLDA, but in expected version. The variance matters!
- The last term is a regularizer. Only support vectors affect the topic proportions

- Optimize $L$ over $\phi$:

$$\phi_{dn} \propto \exp \left( E[\log \theta_d | q(\gamma)] + E[\log p(w_{dn} | \theta_d)] + \frac{y_{dn}}{N_{d}} E[\eta_{d}] - 2 \frac{E[\eta^T \phi_{dn} - \eta]}{N} + E[v \circ \eta] + E[q] (\mu_d - \nu_d) \right)$$

- The first two terms are the same as in LDA
- The third and fourth terms are similar to those of sLDA, but in expected version. The variance matters!
- The last term is a regularizer. Only support vectors affect the topic proportions

- Optimize $L$ over other variables. See the paper for details!

**MedTM: a general framework**

- MedLDA can be generalized to arbitrary topic models:
  - Unsupervised or supervised
  - Generative or undirected random fields (e.g., Harmoniums)

- MED Topic Model (MedTM):

$$P(\text{MedTM}) : \min_{q(H), q(\gamma), q(\eta), q(\zeta)} \mathcal{L}(q(H)) + KL(q(\gamma) || p_0(\gamma)) + U(\zeta)$$

- $H$: hidden r.v.s in the underlying topic model, e.g., $(\hat{\theta}, z)$ in LDA
- $\gamma$: parameters in predictive model, e.g., $\eta$ in sLDA
- $\Psi$: parameters of the topic model, e.g., $\alpha$ in LDA
- $\mathcal{L}$: a variational upper bound of the log-likelihood
- $U$: a convex function over slack variables
Experiments

- **Goal:**
  - To qualitatively and quantitatively evaluate how the max-margin estimates of MedLDA affect its topic discovering procedure

- **Data Sets:**
  - 20 Newsgroups (classification)
    - Documents from 20 categories
    - ~20,000 documents in each group
    - Remove stop word as listed in UMass Mallet
  
  - Movie Review (regression)
    - 5006 documents, and 1.6M words
    - Dictionary: 5000 terms selected by tf-idf
    - Preprocessing to make the response approximately normal (Blei & McAuliffe, 2007)

---

Document Modeling

- **Data Set:** 20 Newsgroups
- **110 topics + 2D embedding with t-SNE** (van der Maaten & Hinton, 2008)

![MedLDA vs LDA plots](image-url)
**Document Modeling (cont’)**

<table>
<thead>
<tr>
<th>Topics</th>
<th>MedLDA</th>
<th>LDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>image</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
<tr>
<td>jpeg</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>gif</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>gif</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
<tr>
<td>file</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>file</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
<tr>
<td>fmt</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>fmt</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
<tr>
<td>pro</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>pro</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
<tr>
<td>text</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>text</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
<tr>
<td>images</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>images</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
<tr>
<td>meta</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>meta</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
<tr>
<td>tools</td>
<td>image</td>
<td>jpeg</td>
</tr>
<tr>
<td>tools</td>
<td>jpeg</td>
<td>jpeg</td>
</tr>
</tbody>
</table>

**Classification**

- **Data Set**: 20Newsgroups
  - Binary classification: “alt.atheism” and “talk.religion.misc” (Simon et al., 2008)
  - Multiclass Classification: all the 20 categories
- **Models**: DiscLDA, sLDA (Binary ONLY! Classification sLDA (Wang et al., 2009)), LDA+SVM (baseline), MedLDA, MedLDA+SVM
- **Measure**: Relative Improvement Ratio

\[ RR(M) = \frac{\text{precision}(M)}{\text{precision}(LDA + SVM)} - 1 \]
Regression

- **Data Set**: Movie Review (Blei & McAuliffe, 2007)
- **Models**: MedLDA(partial), MedLDA(full), sLDA, LDA+SVR
- **Measure**: predictive $R^2$ and per-word log-likelihood

$$pR^2 = 1 - \frac{\sum_d (y_d - \hat{y}_d)^2}{\sum_d (y_d - \bar{y}_d)^2}$$

Time Efficiency

- **Binary Classification**: MedLDA is comparable with LDA+SVM
- **Multiclass**: MedLDA is comparable with LDA+SVM
- **Regression**: MedLDA is comparable with sLDA
II. Supervised Multi-view MNs

- A probabilistic method with an additional view of response variables $Y$

\[
p(y|h) = \frac{\exp\{V^T f(h, y)\}}{Z(V, h)}
\]

- Parameters can be learned with maximum likelihood estimation, e.g., special supervised Harmonium (Yang et al., 2007)
  - contrastive divergence is the commonly used approximation method in learning undirected latent variable models (Welling et al., 2004; Salakhutdinov & Murray, 2008).

Max-margin learning of MNs

- Expected discriminant function:

\[
F(y; V) = \mathbb{E}_H[F(y, H; V)], \text{ where } F(y, H; V) = V_y^T H
\]

- Prediction rule:

\[
y^* = \arg\max_y \mathbb{E}_H[F(y, H; V)]
\]

- Hinge loss:

\[
\mathcal{R}_{\text{hinge}}(V) = \frac{1}{B} \sum_{y} \max_{\hat{y}}[\Delta L(y) - V^T \mathbb{E}_H[\Delta f_\hat{y}(y)]]
\]

- Joint max-margin and max-likelihood estimation:

\[
\min_{\Theta, V} - L(\Theta) + \frac{1}{2} C_1 \| V \|_2^2 + C_2 \mathcal{R}_{\text{hinge}}(V)
\]

  - where \( L(\Theta) := \sum_d \log p(z_d|q_d) \) is data likelihood

- The rationale is we want to find a latent representation and a prediction model, which on one hand tend to predict as accurate as possible on training data, while on the other hand tend to explain the data well.
Predictive Latent Representation

- t-SNE (van der Maaten & Hinton, 2008) 2D embedding of the discovered latent space representation on the TRECVID 2003 data
- Avg-KL: average pair-wise divergence
Classification Results

- Data Sets:
  - (Left) TRECVID 2003: (text + image features)
  - (Right) Flickr 13 Animal: (sift + image features)

- Models:
  - baseline(SVM), DWH+SVM, GM-Mixture+SVM, GM-LDA+SVM, TWH, MedLDA

Retrieval Results

- Data Set: TRECVID 2003
  - Each test sample is treated as a query, training samples are ranked based on the cosine similarity between a training sample and the given query
  - Similarity is computed based on the discovered latent topic representations

- Models: DWH, GM-Mixture, GM-LDA, TWH, MMH
III. Supervised STC

- Joint loss minimization

\[
\min_{\{\theta_d\}, \{s_d\}, \beta, \eta} \quad f(\{\theta_d\}, \{s_d\}, \beta) + C R_h(\{\theta_d\}, \eta) + \frac{1}{2} \|\eta\|^2_2
\]

s.t.:
\[\theta_d \geq 0, \forall d; \quad s_{dn} \geq 0, \forall d, n \in I_d; \quad \beta_k \in \mathcal{P}, \forall k;\]

- coordinate descent alg. applies with closed-form update rules
- No sum-exp function; seamless integration with non-probabilistic large-margin principle

Classification accuracy

- 20 newsgroup data:
Time efficiency

- training & testing time
  - No calls of digamma function
  - Converge faster with one additional dimension of freedom

Summary

- Max-margin, instead of max-likelihood learning of supervised topic models (MedLDA, MMH, MedSTC)
  - Explicit interpretation of effects by support vectors
  - MedLDA can discover discriminative topic representations that are more suitable for supervised tasks
  - The classification model is efficient and can avoid dealing with the normalization factor of a GLM

- The same principle can be applied to a wide variety of probabilistic (MedTM) and non-probabilistic latent variable models
Scenario III: what if I don't know the total number of topics?

Clustering
A Classical Approach

- Clustering as Mixture Modeling

Then "model selection"

- Model selection
  - "intelligent" guess: ???
  - cross validation: data-hungry 😐
  - information theoretic:
    - AIC
    - TIC
    - MDL :
      \[
      \text{arg min } KL(f(\cdot) | g(\cdot | \hat{\theta}_{ML}, K))
      \]
      Parsimony, Ockam’s Razor
    - Bayes factor: need to compute data likelihood

- Posterior inference:
  we want to handle uncertainty of model complexity explicitly
  \[
  p(M | D) \propto p(D | M)p(M)
  \]
  \[
  M = \{q, K\}
  \]
  - we favor a distribution that does not constrain \(M\) in a "closed" space!
Data point \((x, \theta)\).

\(a\) distribution

another distribution

\(\phi\)

\(\phi\)

\(\{\phi_6, \pi_6\}\)

\(\{\phi_5, \pi_5\}\)

Random Partition of Probability Space

Random partition of probability space.

\(\phi\)

\(\pi\)

\(\phi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)

\(\phi\)

\(\pi\)
Stick-breaking Process

\[ G \sim DP(\alpha, G_0) \]

\[ G = \sum_{k=1}^{\infty} \pi_k \delta(\theta_k) \]

\[ \theta_k \sim G_0 \]

\[ \sum_{k=1}^{\infty} \pi_k = 1 \]

\[ \pi_k = \frac{\beta_k}{\sum_{j=1}^{k-1} \beta_j} (1 - \beta_k) \]

\[ \beta_k \sim \text{Beta}(1, \alpha) \]

<table>
<thead>
<tr>
<th>[ \Pi_{j=1}^{k-1} (1 - \beta_j) ]</th>
<th>[ \beta_k ]</th>
<th>[ \pi_k ]</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>0.6</td>
<td>0.5</td>
<td>0.3</td>
</tr>
<tr>
<td>0.3</td>
<td>0.8</td>
<td>0.24</td>
</tr>
</tbody>
</table>

DP – a Pólya urn Process

\[ p = \frac{2}{5+\alpha} \]

\[ p = \frac{3}{5+\alpha} \]

\[ p = \frac{\alpha}{5+\alpha} \]

\[ G_0 := p(\bullet, \bullet, \ldots) \]

Joint: \[ G(\cdot) \sim DP(\alpha G_0) \]

Marginal: \[ \phi, r, \alpha, G_0 \sim \sum_{i=1}^{n} \frac{n_i}{i-1+\alpha} \delta_i + \frac{\alpha}{i-1+\alpha} G_0 \cdot \]

- Self-reinforcing property
- Exchangeable partition of samples
Clustering and DP Mixture

- We can associate mixture components with colors in the Pólya urn model and thereby define a clustering of the data.

Chinese Restaurant Process

- $P(c_i = k | c_{-i}) = \begin{cases} 
1 & \frac{1}{1+\alpha} \\
0 & \frac{\alpha}{1+\alpha} \\
0 & \frac{1}{2+\alpha} \\
0 & \frac{1}{3+\alpha} \\
0 & \frac{1}{4+\alpha} \\
0 & \frac{m_i}{i+\alpha-1} \\
0 & \frac{m_i}{i+\alpha-1} \\
0 & \frac{\alpha}{i+\alpha-1} \\
\end{cases}$
Gibbs sampling for exploring the posterior distribution under the proposed model

- Under the CRP metaphor, due to exchangeability, every sample can be treated as the LAST sample!

\[
p(c_i = k | c_{-i}, x, \theta) \propto p(c_i = k | c_{-i}) \cdot P(x_i | \theta_k, h_{-i}, c_{-i})
\]

- One can also integrate out the parameters such as \( \theta \) and perform collapse Gibbs sampling
- Gibbs sampling algorithm: draw samples of each random variable to be sampled given values of all the remaining variables

Convergence of Ancestral Inference

- On a single machine Gibbs sampling solution is not efficient enough to scale up to the large scale problems.
- Truncated stick-breaking approximation can be formulated in the space of explicit, non-exchangeable cluster labels.
- Variational inference can now be applied to such a finite-dimensional distribution

Variational Inference:
- For a complicated \( P(X_1, X_2, \ldots, X_n) \), approximate it with \( Q(X) \):
  \[
  Q(X) = \prod_i Q(X_{C_i})
  \]
  \[
  \{Q^*(X_{C_i})\} = \arg \min KL(Q(X)||P(X))
  \]

Approximations to DP

- Truncated stick-breaking representation
  
  \[
  v_i \sim B(v; 1, \alpha) \quad \text{for } i = 1, \ldots, T - 1
  \]
  \[
  \pi_T = 1
  \]
  \[
  \pi_i = v_i \prod_{j < i} (1 - v_j) \quad \text{for } i = 1, \ldots, T
  \]
  \[
  \pi_i = 0 \quad \text{for } i > T
  \]

- Finite symmetric Dirichlet approximation
  
  \[
  \pi \sim D(\pi; \frac{\alpha}{K}, \ldots, \frac{\alpha}{K})
  \]

The joint distribution can be expressed as:

\[
P(X, z, v, \eta) = \prod_{i=1}^T p(x_i|z_i) p(v_i|\pi(v)) \prod_{i=1}^T p(\eta_i|v_i; 1, \alpha)
\]

The joint distribution can be expressed as:

\[
P(X, z, \pi, \eta) = \prod_{i=1}^K p(x_i|z_i) p(\pi|\pi) \prod_{i=1}^K p(\eta_i|\pi) D(\pi; \frac{\alpha}{K}, \ldots, \frac{\alpha}{K})
\]
VB inference

- We can then apply the VB inference on the four approximations

\[ \{Q^*(X_{C_i})\} = \arg \min K L(Q(X)|P(X)) \]

The approximated posterior distribution for TSB and FSD are

\[ Q_{TSB}(z, \eta, \nu) = \left[ \prod_n q(z_n) \right] \left[ \prod_i q(\eta_v) q(\nu_i) \right] \]

\[ Q_{FSD}(z, \eta, \pi) = \left[ \prod_n q(z_n) \right] \left[ \prod_k q(\eta_k) \right] q(\pi) \]

Depending on marginalization or not, \( \nu \) and \( \pi \) may be integrated out.

LDA: The Generative Process

**Generative Process**

- For each document \( d \)
  - Sample \( \theta_d \sim \text{Dirichlet}(\alpha) \)
  - For each word \( w \) in \( d \)
    - Sample \( z \sim \text{Multi}(\theta_d) \)
    - Sample \( w \sim \text{Multi}(\phi_z) \)

Topics’ trends evolve over time? \( \times \)
Topics’ distributions evolve over time? \( \times \)
Number of topics grow with the data? \( \times \)
The Chinese Restaurant Franchise Process

- HDPM automatically determines number of topics in LDA
- We will focus on the Chinese Restaurant Franchise process construction
  - A set of restaurants that share a global menu
- Metaphor
  - Restaurant = documents
  - Customer = word
  - Dish = topic
  - Global Menu = Set of topics

HDPM  © Eric Xing @ CMU, ACL Tutorial 2012
The Chinese Restaurant Franchise Process

Global Menu

Restaurant 1
Restaurant 2
Restaurant 3

Generative Process
- For customer \( w \) in restaurant 3
  - Choose table \( j \sim \mathcal{N}_1 \)
  - Choose a new table \( b \sim \alpha \)
  - Sample a new dish for this table

\( w \sim \text{Multi}(\phi_3) \)
The Chinese Restaurant Franchise Process

- For customer $w$ in restaurant 3
  - Choose table $j \sim N_i$
  - Choose a new table $b \sim \alpha$
  - Sample a new dish for this table

- Exisiting dish $k \sim m_i$
- A new dish $\sim \gamma$

Global Menu

Restaurant 1
Restaurant 2
Restaurant 3
The Chinese Restaurant Franchise Process

For customer $w$ in restaurant 3
- Choose table $j \sim N_1$
- Choose a new table $b \sim \alpha$
  - Sample a new dish for this table
  - Existing dish $k \sim m_k$
  - A new dish $\gamma$

Generative Process

Global Menu

Restaurant 1
Restaurant 2
Restaurant 3

$w \sim \text{Multi}(L(\phi_3))$
The Chinese Restaurant Franchise Process

Hierarchical Dirichlet Process

- Two level Poisson urn scheme
  - At the i-th step in j-th "group",

- Choose $\theta_k$ with prob. $\frac{m_{jk}}{\sum_k m_{jk} + \alpha_0}$
- Go to the upper level DP with prob. $\frac{\alpha_0}{\sum_k m_{jk} + \alpha_0}$
- Choose $\theta_k$ with prob. $\frac{n_k}{\sum n_k + \gamma}$
- Draw a new sample with prob. $\frac{\gamma}{\sum n_k + \gamma}$
Hierarchical Dirichlet Process

- Two level Pólya urn scheme
  - At the $i$-th step in $j$-th "group",

  \[
  \theta_i | \theta, \gamma \sim \sum_{k=1}^{K} \frac{n_k}{i + \gamma} \delta(\theta_i) + \frac{\gamma}{i + \gamma} H(\theta_i)
  \]

- Conditioning on $DP(\gamma, H)$, the $m$th draw from the $m$th bottom-level urn also form a Dirichlet measure

\[
\theta_m | \theta, \gamma \sim \sum_{k=1}^{m} \frac{m_k}{m_m + \alpha} \delta(\theta_m) + \frac{\alpha}{m_m + \alpha} \frac{\gamma}{\gamma} H(\theta_m)
\]

\[= \sum_{j=1}^{\infty} \frac{p_j \delta(\theta_j)}{\gamma} + \frac{\gamma}{\gamma} H(\theta_m)\]

Infinite Topic Model for Image

- A single image with $k$ topic
  - An LDA

- A single image with inf-topic
  - A DP

- $J$ images with inf-topic
  - An HDP

© Eric Xing @ CMU, ACL Tutorial 2012
6. Scenario IV: Topic evolution in Streaming Corpus

How to model topic evolution?

Research topics

Nature papers from 1900-2000

1900

2000

?
Problem Statement

- Potentially infinite number of topics
  - With time-varying trends
  - And time-varying distributions
  - And variable durations
  - Topics can die
  - New topics can be born

The Big Picture

- LDA
- Dynamic LDA
- HDPM
- Infinite Dynamic Topic Models
The Big Picture

LDA

\[ \theta \]

\[ z \]

\[ w \]

\[ \beta \]

\[ N \]

\[ D \]

\[ \alpha \]

\[ K \]

Dynamic LDA

\[ \alpha' \]

\[ z' \]

\[ w' \]

\[ \beta' \]

\[ \beta'' \]

\[ \gamma' \]

\[ \gamma'' \]

Infinite Dynamic Topic Models

Model Dimension

Time

Text Stream

\[ \mu \]

\[ \Sigma \]

1990

1991

2004

2005

\[ H \]

\[ \gamma \]

\[ w \]
Text Stream

How to Model Topic Evolution

- Topic Trends
- Topic Keywords
- Topic correlations
- Number of topics

The Dynamic Correlated Topic model
Building Blocks

CTMs

Kalman Filters

X_t = AX_{t-1} + \xi
Y_t = CX_t + \delta_t
X_t|X_{t-1} \sim N(AX_{t-1}, \Phi)
Y_t|X_t \sim N(CX_t, \Gamma)

The Dynamic CTM

© Eric Xing @ CMU, ACL Tutorial 2012
Generalized Mean Field Inference

\[ q(X) = P\left( X \mid \left< S_y \right>_{q_y} : \forall y \in X_{MB} \right) \]

Experimental Results

- NIPS data set
  - 12 years
  - 14036 words
  - 2484 docs
  - 90% for training and 10% for testing
Topic Trends

Graphical Models

Neural Network

sensorimotor

Reinforcement Learning

Neuroscience

Genetic Learning

Topic Words over Time
Topic Correlations Over Time

Topics’ trends evolve over time? ✓
Topics’ distributions evolve over time? ✓
Number of topics grow with the data? ✗
The Big Picture

The Chinese Restaurant Franchise Process

- **HDPM automatically** determines number of topics in LDA
- We will focus on the **Chinese Restaurant Franchise** process construction
  - A set of restaurants that share a global menu
- **Metaphor**
  - Restaurant = documents
  - Customer = word
  - Dish = topic
  - Global Menu = Set of topics

We have covered it already!
The Big Picture

LDA

Dynamic LDA

HDP

Infinite Dynamic Topic Models

The Chinese Restaurant Franchise Process

Global Menu T=1

Global Menu T=2

Topics at end of epoch 1

- Height ($m_{k,1}$) represent topic k’s popularity
- $\phi_{k,1}$ represents topic k’s word distribution

Observations

- Popular topics at epoch 1 are likely to be popular at epoch 2
- $\phi_{k,2}$ is likely to smoothly evolve from $\phi_{k,1}$

Epoch 1

Documents in epoch 1 are generated as before

Pseudo counts

Decay factor

$\phi_{1,1}, \phi_{2,1}, \phi_{3,1}, \phi_{4,1}$

$\phi_{1,2}, \phi_{2,2}, \phi_{3,2}, \phi_{4,2}$

$\phi_{1,1}$

Exp $\frac{-1}{\lambda}$

$m_{1,1}' = \frac{m_{1,1}}{\exp \frac{-1}{\lambda}}$
The Chinese Restaurant Franchise Process

Global Menu $T=1$

- $\phi_{1,1} \phi_{2,1} \phi_{3,1} \phi_{4,1} \phi_{5,1}$

Epoch 1

Global Menu $T=2$

- $\phi_{2,2} \phi_{2,2}$

New real dish served

$\phi_{3,2} \sim \text{Normal}(\cdot | \phi_{3,1}, \rho)$

Inherited but not yet used

---

Generative Process

- For customer $w$ in restaurant 1
  - Choose table $j \sim N_j$
  - Choose a new table $b \sim \alpha$
  - Sample a new dish for this table
  - Existing and inherited dish $k \sim m_{\cdot,2} + m_{k,2}$
  - Existing but NOT inherited dish $k \sim m_{\cdot,2}$ Then $\phi_{k,2} \sim \text{Normal}(\cdot | \phi_{k,1}, \rho)$
  - A new dish $\gamma$ Then $\phi_{\gamma,2} \sim H$

© Eric Xing @ CMU, ACL Tutorial 2012
The Chinese Restaurant Franchise Process

Global Menu T=1

Global Menu T=2

Epoch 1

Generative Process

- For customer \( w \) in restaurant 1
  - [as in static case] Choose table \( j \sim N_i \)
  - Choose a new table \( b \sim \alpha \)
  - Sample a new dish for this table
    - Existing and inherited dish \( k \sim m_k \sim m_k^+ \sim m_k^- \)
    - Existing but NOT inherited dish \( k \sim m_k \) \( \text{Then } \phi_k \sim \text{Normal}(\phi_k^1, \rho) \)
    - A new dish \( \gamma \) \( \text{Then } \phi_{\text{new}} \sim H \)

\( \phi_{1,1} \phi_{2,1} \phi_{3,1} \phi_{4,1} \phi_{5,1} \)

\( \phi_{2,2} \phi_{3,2} \phi_{4,2} \phi_{5,2} \)

\( \phi_{1,2} \sim \text{Normal}(\phi_{1,1}, \rho) \)
The Chinese Restaurant Franchise Process

Epoch 1

- For customer \( w \) in restaurant 1
  - [as in static case] Choose table \( j \sim N_j \)
  - Choose a new table \( b \sim \alpha \)
  - Sample a new dish for this table
    - Existing and inherited dish \( k \sim m_{j,k} \pm m_{b,k} \)
    - Existing but NOT inherited dish \( k \sim m_{j,k} \)
    - A new dish \( \sim \gamma \)
      Then \( \phi_{new} \sim H \)

Generative Process

Global Menu T=1

Global Menu T=2

Global Menu T=3

© Eric Xing @ CMU, ACL Tutorial 2012
The Chinese Restaurant Franchise Process

Epoch 1
- Topics’ trends evolve over time?
- Topics’ distributions evolve over time?
- Number of topics grow with the data?

Epoch 2

The Chinese Restaurant Franchise Process

- We just described a first order RCRF process
- for a general $\Delta$-order process

\[ m'_{kt} = \sum_{\delta=1}^{\Delta} \exp \frac{-\delta}{\lambda} m_{k,t-\delta} \]
Inference

- **Gibbs Sampling**
  - Sample a table for each word
  - Sample a topic for each table
  - Sample the topic parameter over time
  - Sample hyper-parameters

- **How to deal with non-conjugacy**
  - Algorithm 8 in Neal’s 1998 + Metropolis-Hasting

- **Efficiency**
  - The Markov blanket contains the previous and following \( \Delta \) epochs

---

**Sampling a Topic for a Table**

\[
P(k_{i,db} = k | k_{t-\Delta t+\delta}^{t-\Delta t+\delta}, b_{i,db}, \phi, w_i) \propto \]

\[
P(k_{i,db} = k | k_{t-\Delta t}^{t-\Delta t}) P(w_{i,db} | \phi, k_{i,db} = k) \prod_{\delta=1}^{\Delta} P(k_{i+\delta-\Delta t-\Delta t+\delta-1}^{i+\delta-\Delta t})
\]

Past  | Emission  | Future
Non-Conjugacy  | Efficiency

© Eric Xing @ CMU, ACL Tutorial 2012
Sampling a Topic for a Table

\[ P(k_{i,db} = k | k_{i-1,db}^{\text{tb}}, b_{i,d}, \phi_i, w_i) \propto P(k_{i,db} = k | k_{i-1,db}^{\text{tb}}) \prod_{\delta=1}^{\Delta} P(k_{i,\delta}^{\text{tb}} | k_{i-1,\delta}^{\text{tb}} - k_{i-1,\delta}^{\text{tb}} + k_{i-1,\delta}^{\text{tb}}) \]

© Eric Xing @ CMU, ACL Tutorial 2012
Sampling Topic Parameters

- $V|\phi \sim \text{Multi}(\text{Logistic}(\phi))$
- Linear-State space model with non-Gaussian emission
- Use Laplace approximation inside the Forward-Backward algorithm
- Use the resulting distribution as a proposal

Experiments

- **Simulated** data
  - Simulated 20 epochs with 100 data points in each epoch
- Timeline of the NIPS conference
  - 13 years
  - 1740 documents
  - 950 words per document
  - ~3500 vocabulary
Simulation Experiment

Sample Documents:

Ground Truth

Recovered

© Eric Xing @ CMU, ACL Tutorial 2012

© Eric Xing @ CMU, ACL Tutorial 2012
Analyzing the NIPS Corpus

Start state

(a) Posterior sample

(b) Representative KL (f, t)

(c) # Alive Topic

© Eric Xing @ CMU, ACL Tutorial 2012

SOM
speech
Neuro
science
NN
Classification
Methods
Control
Prob.
Models
image

1987
1990
1991
1994
1995
1996

boosting
ICA
Memory
Kernels

ICA
PM
control
speech
Mixtures
ICA
Kernels

Classification

© Eric Xing @ CMU, ACL Tutorial 2012
- Support Vector Method for Function Approximation, Regression Estimation, and Signal Processing, V. Vapnik, S. E. Golowich and A. Smola
- Support Vector Machines, H. Drucker, C. Burges, L. Kaufman, A. Smola and V. Vapnik
- Improving the Accuracy and Speed of Support Vector Machines, C. Burges and B. Scholkopf

- From Regularization Operators to Support Vector Kernels, A. Smola and B. Scholkopf
- Prior Knowledge in Support Vector Kernels, B. Scholkopf, P. Simard, A. Smola and V. Vapnik
- Uniqueness of the SVM Solution, C. Burges and D. Crisp
- An Improved Decomposition Algorithm for Regression Support Vector Machines, P. Laskov
The Big Picture

LDA

Dynamic LDA

HDPM

Infinite Dynamic Topic Models

Quantitative Analysis

© Eric Xing @ CMU, ACL Tutorial 2012
Hyper-parameter Sensitivity

Varying Variance of base measure

Hyper-parameter Sensitivity

topic evolution variance

© Eric Xing @ CMU, ACL Tutorial 2012
Hyper-parameter Sensitivity

\[ m_{k,t} = \sum_{\delta=1}^{\Delta} \exp -\delta m_{k,t-\delta} \]

Global Menu \( T=3 \)

Conclusions and Future Work

- Infinite Dynamic Topic Model
  - Evolve all topical aspects
- Application to other data type
  - Community discovery in social media
- Alternative inference algorithms
  - Particle filters
  - Collapsed Variational Inference
9: Other apps (Optional)

I. Machine translation

B. Zhao and E.P Xing, ACL 2006
The economy and trade relations between Russia and Tianjin develop steadily.
BiTAM: From monolingual to bilingual topic models  

- Monolingual space, a unigram LM $p(w|z)$
  - A topic corresponding to a point in the word simplex.
  - AdMixture of unigrams (Blei, et al. 2003)
- Bilingual space, a translation lexicon $p(f|e, z)$
  - Given a topic $z$, a word usually has limited translations.
  - Topic-specific translation lexicons are sharper
  - Each topic is a point in the conditional simplex
  - AdMixture of topic-specific translation lexicons (Zhao & Xing, ACL/Coling 2006)

Example
- A Chinese word “club”, the translations can be:

<table>
<thead>
<tr>
<th></th>
<th>ogre</th>
<th>war</th>
<th>socialize</th>
<th>interests</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4</td>
<td>0.5</td>
<td>0.0</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>0.1</td>
<td>0.5</td>
<td>0.4</td>
<td></td>
</tr>
</tbody>
</table>

BiTAM: A Generative Process

- Sample topic weights $\theta$ from a Dirichlet($\alpha$)
- Sample a topic $z$ from multinomial ($\theta$)
- For each word $f$ in the sentence $\vec{f}$
  - Sample an alignment $\hat{a}$ from an alignment model
  - Generate $f$ with word $\vec{e}_a$ in a topic-specific lexicon
BiTAM Model-1

- Graphical Model (a language to encode dependencies)

\[ p(F \mid A, E, \alpha, B) = \int \prod_{n=1}^N p(z_n \mid \theta) p(f_n \mid a_n, e_n, B, \omega) d\theta \]

GMF Inference

Approximate the Integral

Approximate the Posterior

\[ q(\theta, z, a) = q(\theta \mid \gamma) \prod_{n=1}^N q(z_n \mid \phi_n) \prod_{j=1}^J q(a_{n,j} \mid \lambda_{nj}) \]

\[ \text{argmin}_{\gamma, \phi, \lambda} KL(p(\theta, z, a), q(\theta, z, a)) \]

Optimization Problem
An upgrade path for BiTAMs

HMM for Alignment

Word-pair level topics

Sent-pair level topics

Word-Pair & HMM

Experiments

- Training data
  - Small: Treebank 316 doc-pairs (133K English words)
  - Large: FBIS-Beijing, Sinorama, XinHuaNews, (15M English words).

<table>
<thead>
<tr>
<th>Train</th>
<th>#Doc.</th>
<th>#Sent.</th>
<th>#Tokens English</th>
<th>#Tokens Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treebank</td>
<td>316</td>
<td>4172</td>
<td>133K</td>
<td>105K</td>
</tr>
<tr>
<td>FBIS.BJ</td>
<td>6,111</td>
<td>105K</td>
<td>4.18M</td>
<td>3.54M</td>
</tr>
<tr>
<td>Sinorama</td>
<td>2,373</td>
<td>103K</td>
<td>3.81M</td>
<td>3.60M</td>
</tr>
<tr>
<td>XinHua</td>
<td>19,140</td>
<td>115K</td>
<td>3.85M</td>
<td>3.93M</td>
</tr>
<tr>
<td>FOUO</td>
<td>15,478</td>
<td>368K</td>
<td>13.14M</td>
<td>11.93M</td>
</tr>
<tr>
<td>Test</td>
<td>95</td>
<td>627</td>
<td>25,500</td>
<td>19,726</td>
</tr>
</tbody>
</table>

- Word Alignment Accuracy & Translation Quality
  - F-measure
  - BLEU
Model Selection

- Choosing num-topics $K$
  - 10-fold cross-validation
  - Number of topics is set to be 50 for 23 million words corpus

![Graph showing log(likelihood) over different num topics]

Topics

| T1 | Teams, sports, disabled, games members, people, cause, water, national, handicapped |
| T2 | Shenzhen, singapore, hongkong, stock, national, investment, yuan, options, million, dollar |
| T3 | Chongqing, company, takeover, shenzhen, tianjin, city, national, government, project, companies |
| T4 | Hongkong, trade, export, import, foreign, tech., high, 1998, year, technology |
| T5 | House, construction, government, employee, living, provinces, macau, anhui, yuan |
| T6 | Gas, company, energy, usa, russia, france, chongqing, resource, china, economy, oil |

© Eric Xing @ CMU, ACL Tutorial 2012
HM-BiTAM versus others

Translation Evaluations
II. Exploring and deciphering social networks

<table>
<thead>
<tr>
<th>Systems</th>
<th>1-gram</th>
<th>2-gram</th>
<th>3-gram</th>
<th>4-gram</th>
<th>BLEU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiero Sys.</td>
<td>73.92</td>
<td>40.57</td>
<td>23.21</td>
<td>13.84</td>
<td>30.70</td>
</tr>
<tr>
<td>Gale Sys.</td>
<td>75.63</td>
<td>42.71</td>
<td>25.00</td>
<td>14.30</td>
<td>32.78</td>
</tr>
<tr>
<td>HM-BiTAM</td>
<td>76.77</td>
<td>42.99</td>
<td>25.42</td>
<td>14.04</td>
<td>33.19</td>
</tr>
<tr>
<td>Ground Truth</td>
<td>76.10</td>
<td>43.85</td>
<td>26.70</td>
<td>15.73</td>
<td>34.17</td>
</tr>
</tbody>
</table>
Dynamic network tomography

- How to model dynamics in a simplex?

Evolving networks
Dynamic MMSB (dMMSB) [Xing, Fu, and Song, AOAS 2009]

<table>
<thead>
<tr>
<th>Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hidden role prior</td>
</tr>
<tr>
<td>Observed interactions</td>
</tr>
<tr>
<td>Role compatibility matrix</td>
</tr>
</tbody>
</table>

Dynamic Mixture of MMSB (dM^3SB) [Ho, Le, and Xing, submitted 2010]

<table>
<thead>
<tr>
<th>Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-varying Role Prior</td>
</tr>
<tr>
<td>Cluster Selection Prior</td>
</tr>
<tr>
<td>Time-varying Network Model</td>
</tr>
</tbody>
</table>

© Eric Xing @ CMU, ACL Tutorial 2012
Algorithm: Generalized Mean Field  
(xing et al. 2004)

Approximate the joint posterior
\[ p(\{\tilde{Z}^\alpha, \tilde{\pi}^\alpha, \mu^\alpha, B^\alpha\}_{\alpha=1}^T | \Theta, \{G^\alpha\}_{\alpha=1}^T) \]
where \( \Theta \) denotes the model parameters, by a factored approximate distribution:
\[
q\left(\{\tilde{Z}^\alpha, \tilde{\pi}^\alpha, \mu^\alpha, B^\alpha\}_{\alpha=1}^T\right) = q_1(\{\tilde{Z}^\alpha, \tilde{\pi}^\alpha\}_{\alpha=1}^T) \times q_2(\{\mu^\alpha\}_{\alpha=1}^T) \times q_3(\{B^\alpha\}_{\alpha=1}^T).
\]

- Inference via variational EM
  - Generalized mean field
  - Laplace approximation
  - Kalman filter & RTS smoother

---

dMMSB vs. MMSB

![Graph showing comparison between dMMSB and MMSB]
dM^3SB vs. dMMSB

Goodness of fit

US Senator voting data
Average held-out marginal log-likelihood over 10 random hold-outs
10,000 samples taken per hold-out marginal log-likelihood
Case Study 1: Sampson’s Monk Network

- Dataset Description
  - 18 monks (junior members in a monastery)
  - Liking relations recorded
  - 3 time-points in one year period
  - Timing: before a major conflict outbreak

- Recall static analysis:

Sampson’s Monk Network: role trajectories

- The trajectories of the varying role-vectors over time

© Eric Xing @ CMU, ACL Tutorial 2012
Case Study 2: The 109th congress

March 2005 | January 2006 | August 2006

US senator voting records
100 senators, 109th Congress (Jan 2005 – Dec 2006) in 8 epochs

Senate Network: role trajectories

Voting data preprocessed into a network graph using (Kolar et al., 2008)

Colored bars: Estimated latent space vector
Numbers under bars: Estimated cluster
Letters beside actor index: Political party and State

Role Compatibility Matrix B
Role 1 = Passive, 2/4 = Democratic clique, 3 = Republican clique
Senate Network: role trajectories

Jon Corzine’s seat (#28, Democrat, New Jersey) was taken over by Bob Menendez from t=5 onwards. Corzine was especially left-wing, so much that his views did not align with the majority of Democrats (t=1 to 4).

Once Menendez took over, the latent space vector for senator #28 shifted towards role 4, corresponding to the main Democratic voting clique.

Ben Nelson (#75) is a right-wing Democrat (Nebraska), whose views are more consistent with the Republican party.

Observe that as the 109th Congress proceeds into 2006, Nelson’s latent space vector includes more of role 3, corresponding to the main Republican voting clique.

This coincides with Nelson’s re-election as the Senator from Nebraska in late 2006, during which a high proportion of Republicans voted for him.

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
More research questions we ask:

- **Event detection**
  - Emergence/disappearance/evolution of perspective, bias, object, theme, etc.

- **Automated summary**
  - Temporally-connected storylines in the news
  - Describe a scene or arbitrary image
  - From keyword or class-label to story
  - Hierarchical categorization of news, images, videos

- **Semantic-based browsing and search**
  - Ranking/matching based on topic/perspective
  - Can we Google image using image? Where is this place on earth?
  - Video retrieval based on story

- **Prediction**
  - Is there going to be a war? When?
  - Can we predict the economy or stock from traditional or internet news?

- **Doing all these with Facebook, MySpace, or Twitter**