Michael I. Jordan

Representation and Inference

University of California, Berkeley
Computer Science and Statistics

http://www.cs.berkeley.edu/~jordan
anytime algorithms

- Laws of large numbers
- Hierarchical models
- Heterogeneous data sources
- Consistency Guarantees

Advantages

Viewing the core problem as one of probabilistic inference has many

- Notions of uncertainty
- Abstraction, resolution
- Objects, relations, states, parameters

Representational Issues

A key component of any decision-making system: situation assessment

Situation Assessment
- Graph separation ≡ factorization

- Marriage between graph theory and probability theory

- Joint distribution function

- Each node is a random variable
\[
\frac{(\theta | \Lambda, \mathcal{H}) d}{(\theta | \Lambda') \mathcal{H}} = (\theta | \Lambda) d
\]

where

\[
\frac{(\theta | \Lambda) d}{(\theta | \Lambda', \mathcal{H}) d} = (\theta', \Lambda | \mathcal{H}) d
\]

Inference: •

\( \Lambda \cap \mathcal{H} = S \) • partition the nodes: •

Probabilistic Inference
Inference Algorithms

• Variational MCMC
• Structured mean field
• Hybrid algorithms

- sum-product algorithm (loopy belief propagation)
- convexity-based algorithms

• Variational algorithms
- Markov chain Monte Carlo (MCMC)
- Importance sampling

• Sampling algorithms
- Exact algorithms
New Directions

- Algorithms into hardware
- Theoretical analysis of inference algorithms
- Minimax methods
- Semiparametric methods
- Identity uncertainty
- Probabilistic relational models
- Latent Dirichlet allocation models
http://www.cs.berkeley.edu/~jordan

University of California, Berkeley
Computer Science and Statistics

David M. Blei, Michael I. Jordan and Andrew Y. Ng

Latent Dirichlet Allocation Models
Conclusions

- Application to document/image modeling

Correspondence Latent Dirichlet allocation models

- Application to Latent Dirichlet models

Variational algorithms

- Application to document modeling

Latent Dirichlet allocation models

- Document/image modeling

- Document modeling

- The information modeling problem

Outline
Probabilistic methods are currently being under-utilized

• Uncertainty is rampant

  • Image annotation
  • Topic hierarchies
  • Collaborative filtering
  • Text summarization
  • Text classification

• Many related problems

  • Next-generation Google: search with inferential capabilities
  • Free-form information needs
  • Large collections of reports, documents, messages, etc.

Finding useful information

Information Retrieval
annual $100,000 donation, too.

a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual
where music and the performing arts are taught, will get $250,000. The Hearst Foundation,
Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School,
building, which will house young artists and provide new public facilities. The Metropolitan
saw Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new
research, education and the social services, “Hearst Foundation President Randolph A. Hearst
Grants an act every bit as important as our traditional areas of support in health, medical
we had a real opportunity to make a mark on the future of the performing arts with these
Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that
The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center,

Probabilistic Modeling of Documents
How to annotate the images?

Continuous-valued feature vector

Each image is segmented into multiple ‘blobs‘ which are represented as a 45-dimensional

Dataset of images and corresponding text (Barnard & Forsyth, 2000)

Sculpture, Statue, Stone

Probabilistic Modeling of Documents/Images
Still, limited data reduction and no structure revealed

- Discrimination
  - combines computational efficiency with a (bare) minimum of
    thoroughly ad hoc, but somewhat of an empirical success story

\[
\frac{\sum \text{times word } i \text{ occurs across all documents}}{\text{times word } i \text{ occurs in the document}} = ?
\]

Represent a document as a list of ratios of counts of word occurrences:

- Choose a vocabulary \( \mathcal{V} \) of words

**tf-idf Approach**
Probabilistic latent semantic indexing

- Clustering
- Synonymy and polysemy
- Anecdotal arguments in its favor
- Retain a small number of singular vectors
- Singular value decomposition of $X$
- Arrange the tf-idf values into a word-by-document matrix $X$
— thus we use multi-level mixture distributions

But there is also an exchangeability assumption for documents

— De Finetti’s theorem tells us how to represent exchangeable distributions—we must use mixture distributions

This is just exchangeability (for words in a document)

Current methods in information retrieval are based on the “bag-of-words” assumption

“Bag-of-words assumption”
Latent Dirichlet Allocation

- Non-conjugacy
- Three-level conditional hierarchical Bayes (parametric empirical Bayes)
- Random mixture of multivariate normals

exchangeability for words and documents

A (simple) hierarchical Bayesian model, based on an assumption of
Repeated sampling of multinomial topic variable within documents
Repeated sampling of Dirichlet document variable within corpus

• A document is represented as a Dirichlet \( \theta \)
• A topic is represented as a multinomial \( z \)
• A word is represented as a multinomial \( w \)

Core model structure:

Latent Dirichlet Allocation (LDA) Model
Related Models

Mixture of uniGrams model

UniGram model
Latent Dirichlet Allocation (LDA) Model
The Dirichlet distribution

- The beta distribution is the special 2-d case of the Dirichlet and illustrates which \( z \) is drawn.
- The Dirichlet distribution is the exponential family distribution that parameterizes the simplex—the multinomial distribution over topics from

The Dirichlet distribution.
A corpus is a Dirichlet distribution on the simplex.

A document is a point in the simplex—a distribution over topics.

Each corner of the simplex corresponds to a topic—a component of the vector.*
Joint probability

\[ P(u_z | m) d(\theta | u_z) d \prod_{n=1}^{u} (\nu_t | \theta) d = (\mathbb{g} | \nu_m z, \theta) d \]

Joint probability for a single document

Topics, respectively

For a given document, let \( m \) and \( z \) denote the (observed) words and (latent)
But variational approximations can be readily developed (Dickey, 1983).

\[
\begin{align*}
(2) & \quad \theta P \left( u_n | \prod_{\gamma=1}^{I=\ell} \prod_{A=1}^{\Pi=\gamma} \prod_{N=1}^{\Pi=\gamma} \right) \left( \prod_{\gamma=1}^{I=\ell} \prod_{A=1}^{\Pi=\gamma} \theta \prod_{N=1}^{\Pi=\gamma} \right) \int \frac{\prod_{\gamma=1}^{I=\ell} \prod_{A=1}^{\Pi=\gamma}}{\prod_{\gamma=1}^{I=\ell} \prod_{A=1}^{\Pi=\gamma}} d \theta = \\
(1) & \quad \theta P \left( \theta | u, \gamma \right) d \left( \gamma | u, \gamma \right) d \left( u, n \right) \prod_{\gamma=1}^{I=\ell} \prod_{A=1}^{\Pi=\gamma} \prod_{N=1}^{\Pi=\gamma} \int \frac{\prod_{\gamma=1}^{I=\ell} \prod_{A=1}^{\Pi=\gamma}}{\prod_{\gamma=1}^{I=\ell} \prod_{A=1}^{\Pi=\gamma}} d \theta = \left( \gamma | \gamma, w \right) d \theta
\end{align*}
\]

The probability of the observables \( w \) is a mixture:

**Posterior Blues**
\[ (x|\theta)d \sim \theta \]

\[ (\theta|x)d \sim x \]
\[
( ( \forall' x | \theta)^b \| (n, x | \theta)^b ) \mathcal{D} = \arg \max_b \ = \ n \\
\{(n, x | \theta)^b\} \text{ indexed by } n \\
\{ (n, x | \theta)^b \} \text{ from a family of simplified distributions} \\
\text{Choose } b \text{ from a family of simplified distributions} \\
( ( \forall' x | \theta)^b \| (x | \theta)^b ) \mathcal{D} = (\forall' b)^b \mathcal{J} \arg \max_b - (x | \theta)^b \mathcal{D} \\
\text{Fact 2:} \\
(\forall' b)^b \mathcal{J} \arg \max_b = (\forall' x | \theta)^b \\
\text{Fact 1:} \\
\text{Why „variational“?} \\
\text{Variational bound for a marginal likelihood (cont.)}
Empirical Bayes using Variational Inference

\[ \max_{\lambda} \mathcal{L}(\lambda, b) \mathcal{J} \]  
\[ \max_{\lambda} \mathcal{L}(\lambda, b) \mathcal{J} \]  
\[ \max_{\lambda} \mathcal{L}(\lambda, b) \mathcal{J} \]  
\[ \max_{\lambda} \mathcal{L}(\lambda, b) \mathcal{J} \]  

A two-phase procedure:

Thus we maximize the lower bound \( \mathcal{J}(\lambda, b) \) instead.

- But it is intractable to compute and/or maximize \( \mathcal{L}(\lambda | x) \) in general.

- An empirical Bayesian wants to maximize \( \mathcal{L}(\lambda | x) \) with respect to \( \lambda \)
So consider a model in which this coupling is removed variationally.

The coupling between $\theta$ and $\theta'$ in the likelihood is the source of the problems.

Latent Dirichlet allocation (LDA) model
where are the variational parameters

\[
(u\phi|u\mathbf{z}) \prod_{N}^{u=1} (\mathcal{H} | \theta)^{u} \mathcal{D} \mathcal{I} = (\phi, \mathcal{N}, m | \mathbf{z}, \theta) \mathcal{B}
\]

Family of variational approximations:

Approximating family
The problem-dependent variables are decoupled under the variational distribution $\theta$ and is apparent.

\[
\theta \mathcal{P} \left( \left( \theta | u_z \right) d (v, u_z | u_m) d \left( \begin{array}{ccc} \mathcal{L} & \mathcal{L} & \mathcal{L} \\ \\
\mathcal{L} & \mathcal{L} & \mathcal{L} \\ \\
\mathcal{L} & \mathcal{L} & \mathcal{L} \\
\end{array} \right) \left( \begin{array}{ccc} \mathcal{L} & \mathcal{L} & \mathcal{L} \\ \\
\mathcal{L} & \mathcal{L} & \mathcal{L} \\ \\
\mathcal{L} & \mathcal{L} & \mathcal{L} \\
\end{array} \right) \right) \int \left( \begin{array}{ccc} \mathcal{L} & \mathcal{L} & \mathcal{L} \\ \\
\mathcal{L} & \mathcal{L} & \mathcal{L} \\ \\
\mathcal{L} & \mathcal{L} & \mathcal{L} \\
\end{array} \right) \left( \begin{array}{ccc} \mathcal{L} & \mathcal{L} & \mathcal{L} \\ \\
\mathcal{L} & \mathcal{L} & \mathcal{L} \\ \\
\mathcal{L} & \mathcal{L} & \mathcal{L} \\
\end{array} \right) \right) = (v, \theta | \mathcal{M}) d
\]

Recover the problem-dependent marginal likelihood:

Marginal Likelihood
the tightest possible bound

And maximize with respect to the variational parameters and to obtain

which is an exercise in exponential family manipulation

\[
( (\phi', \lambda^\theta \mid z, \theta) d \| (\phi', \lambda^\theta \mid z, \theta) b ) D
\]

We now compute the KL divergence:

**Variational Approximation**
Fast approximate inference

approximate posterior for documents and topics

The result is a set of parameters \( \phi \) and \( \lambda \) that can be used to form an

where \( \Phi(x) \) is the digamma function

\[
\left\{ \left( \frac{\lambda}{\phi} \right) \Phi - (\gamma) \Phi \right\} \text{d}x \propto \Phi \\
\Phi \left| \begin{array}{l}
\gamma = u \\
\lambda = u \end{array} \right| = \gamma
\]

We minimize the KL divergence by iterating between the following equations:

Variational inference
\[
\left( \mathcal{P} \mathcal{L} \left. \frac{I = p}{\mathcal{L}} \right) \Phi - \left( \mathcal{P} \mathcal{L} \right) \Phi \left[ \frac{I = p}{\mathcal{L}} \right] + \left( \mathcal{P} \mathcal{L} \right) \Phi \left[ \frac{I = p}{\mathcal{L}} \right]
\]

\[
\mathcal{J} \mathcal{C} = \left( \mathcal{P} \mathcal{L} \right) \Phi \left[ \frac{I = p}{\mathcal{L}} \right]
\]


\begin{align*}
\text{Gradient is:} & \quad \mathcal{P} \mathcal{L} \text{ where the Dirichlet parameters } \varphi \text{ are maximized via Newton-Raphson, where the} \\
\text{The multilogarithm parameters } \phi \text{ can be obtained analytically:} & \quad \varphi
\end{align*}

\[
\mathcal{J} \mathcal{C} = \left( \mathcal{P} \mathcal{L} \right) \Phi \left[ \frac{I = p}{\mathcal{L}} \right]
\]

\text{Recall that we maximize the lower bound with respect to the hyperparameters } \varphi \text{ and } \mathcal{J} \mathcal{C} \text{ (Empirical Bayes).}
• Fit a 100-factor LDA model to this dataset

• 90,150 different terms, 3,189,629 words in corpus

• 15,000 documents from the TREC AP corpus from 1988-1990
a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual where music and the performing arts are taught, will get $250,000. The Hearst Foundation, Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, building, which will house young artists and provide new public facilities. The Metropolitan A. Heart Research, Education and the Social Services. "Hearst Foundation President Randolph A. Heart grants are an act of every bit as important as our traditional areas of support in health, medical that we had a real opportunity to make a mark on the future of the performing arts with these Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, a new document.
<table>
<thead>
<tr>
<th>HAITI</th>
<th>LIFE</th>
<th>CONGRESS</th>
<th>LOVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ELEMENTARY</td>
<td>CARE</td>
<td>GOVERNMENT</td>
<td>ACTRESS</td>
</tr>
<tr>
<td>PRESIDENT</td>
<td>PERCENT</td>
<td>PROGRAMS</td>
<td>THEATER</td>
</tr>
<tr>
<td>STATE</td>
<td>MEN</td>
<td>MONEY</td>
<td>OPERA</td>
</tr>
<tr>
<td>NAMPHY</td>
<td>WELFARE</td>
<td>PLAN</td>
<td>YORK</td>
</tr>
<tr>
<td>MANICAT</td>
<td>FAMILY</td>
<td>STATE</td>
<td>FIRST</td>
</tr>
<tr>
<td>BENNETT</td>
<td>WORK</td>
<td>NEW</td>
<td>ACTOR</td>
</tr>
<tr>
<td>TEACHER</td>
<td>PARENTS</td>
<td>YEAR</td>
<td>BEST</td>
</tr>
<tr>
<td>PUBLIC</td>
<td>WORK</td>
<td>SPENDING</td>
<td>MUSICAL</td>
</tr>
<tr>
<td>HIGH</td>
<td>FAMILIES</td>
<td>FEDERAL</td>
<td>PLAY</td>
</tr>
<tr>
<td>TEACHERS</td>
<td>YEARS</td>
<td>BILLION</td>
<td>MOVIE</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>CHILD</td>
<td>BUDGET</td>
<td>MUSIC</td>
</tr>
<tr>
<td>SCHOOLS</td>
<td>PEOPLE</td>
<td>PROGRAM</td>
<td>SHOW</td>
</tr>
<tr>
<td>STUDENTS</td>
<td>WOMEN</td>
<td>TAX</td>
<td>FILM</td>
</tr>
<tr>
<td>SCHOOL</td>
<td>CHILDREN</td>
<td>MILLION</td>
<td>NEW</td>
</tr>
</tbody>
</table>

“Education” “Children/Family” “Budgets” “Arts”
monotonically decreasing in the likelihood of held-out data.
Prior on the topic distributions

Extended LDA model:

- There are a wide variety of ad hoc methods
- "Smoothing" is a key issue in text modeling
• Variational Empirical Bayes for Inference $\alpha$ and $\gamma$

\[ \dot{d} \land \phi \lor \phi \]

Variational Inference for Inference $\phi$ and $\phi$

Variational approximation of smoothed model
Hierarchical topic model
Each image is segmented into multiple "blobs" which are represented as a 45-dimensional continuous-valued feature vector.

Dataset of images and corresponding text (Barnard & Forsyth, 2000)

Sculpture, Statue, Stone
(Right) The Gaussian-multinomial LDA model for images and captions

\( \theta \) denotes \( N \) (conditionally independent) replicates of \( \theta \)

(Left) The GLM-MIXTURE model of images and captions. The box around

Pairwise models
Variables $\psi_m$ are conditioned on $\psi_m$, the number of image regions. The graphical model representation of the Corr-LDA model. Note that the correspondence LDA model
The models. The overfitting problems in GM-Mixture and Cor-LDA have been corrected.

(Middle) Caption: perplexity for the empirical Bayes smoothed estimates of

(perplexity) Caption: perplexity in Cor-LDA.

(left) The average negative log probability of the held-out test set.

perplexity results
(Bottom) A segmented image and its labeling

Annotatons
Mountain, Sand, Helicopter, Skis, People, Dunes
- 70/30 training/testing split
- Each class has 300 - 1000 documents
- 4 classes of HTML documents: student, course, project, and faculty
dataset

We compared LDA to mixture of unigrams and naive Bayes on the WebKB marginal likelihoods, and maximizing classification by estimating a model for each class c, computing approximate
WEBKB Classification Results
other movies which that user saw.

We hold out a movie from each test user and compute its probability given the
unigrams. We evaluate LDA against the unigram and mixture of

\[
\theta p(\text{seen} \mid \theta) d(\theta \mid z) d(z \mid m) \sum_{\theta} = (m_{\text{seen}}) d
\]

1. Compute the posterior for a given set of movies.
2. Compute a distribution over movies.
3. Output the most likely movie still not seen by the user.

\[
\theta p(\text{seen} \mid \theta) d(\theta \mid z) d(z \mid m) \sum_{\theta} = (m_{\text{seen}}) d
\]

We train on a portion of the EachMovie dataset. Given a new user,

analogous to documents, items are analogous to words.

In collaborative filtering, we try to predict an item that a user will select given the previous

Collaborative Filtering
Predictive perplexity results
Conclusions