Foundations of Graphical Models

David M. Blei

Columbia University
Today’s lecture

- What is this course about?
- Latent Dirichlet allocation: An example of a graphical model
- Other examples of applied probabilistic modeling
- Box’s loop
- What will we cover?
- Prerequisites, requirements, and grades
What is this course about?
Latent Dirichlet Allocation
(An example of a model that I know well)
“are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


Documents exhibit multiple topics.
**Seeking Life’s Bare (Genetic) Necessities**

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, these predictions are not all that far apart, especially in comparison to the 75,000 genes in the human genome, notes Steve Anderson, a geneticist at the University of Utah. Anderson arrived at the 800 number. But coming up with a consensus answer may be more than just a matter of numbers. As more and more genomes are completely sequenced and analyzed, “It may be a way of organizing any newly sequenced genome,” explains Ardy Mushegain, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing any

“...and assignments

**Latent Dirichlet Allocation**
Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life.

One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, says Stephen J. Long, University in St. Andrews, Scotland, who arrived at the 2,000 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arady Mushgian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

---


SCIENCE • VOL. 272 • 24 MAY 1996
**LDA as a graphical model**

- Nodes are random variables; edges indicate dependence.
- Shaded nodes are observed; unshaded nodes are hidden.
- Plates indicate replicated variables.
LDA as a graphical model

- Encodes independence assumptions
- Defines a factorization of the joint distribution
- Connects to algorithms for computing with data
The joint defines a posterior, \( p(\theta, z, \beta \mid w) \).

From a collection of documents, infer

- Per-word topic assignment \( z_{d,n} \)
- Per-document topic proportions \( \theta_d \)
- Per-corpus topic distributions \( \beta_k \)

Then use posterior expectations to perform the task at hand: information retrieval, document similarity, exploration, and others.
• **Data**: The OCR’ed collection of *Science* from 1990–2000
  - 17K documents
  - 11M words
  - 20K unique terms (stop words and rare words removed)

• **Model**: 100-topic LDA model using variational inference.
Seeking Life’s Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here, two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today’s organisms can be sustained with just 256 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn’t be enough.

Although the numbers don’t match precisely, those predictions “are not all that far apart,” especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. “It may be a way of organizing any newly sequenced genome,” explains Arca Mushegan, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an


SCIENCE • VOL. 272 • 24 MAY 1996
human genome dna genetic genes sequence gene molecular sequencing map information genetics mapping project sequences evolution evolutionary species organisms life origin biology groups phylogenetic living diversity group new two common disease host bacteria diseases resistance bacterial new strains control infectious malaria parasite parasites united tuberculosis computer models information data computers system network systems model parallel methods networks software new simulations
Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his exclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his own whims and time constraints, but with the rare advantage of big-studio financial support for all his endeavors.

Kubrick’s films are characterized by a formal visual style and meticulous attention to detail—his later films often have elements of surrealism and expressionism that eschews structured linear narrative. His films are repeatedly described as slow and methodical, and are often perceived as a reflection of his obsessive and perfectionist nature. A recurring theme in his films is man’s inhumanity to man. While often viewed as
| 1 | Game | Second | Season | Team | Coach | Play | Points | Games | Giants | Second | Players |
| 2 | Life | Know | School | Street | Man | Family | Says | House | Children | Night |         |
| 3 | Film | Movie | Show | Life | Television | Films | Director | Man | Story | Says |         |
| 4 | Book | Life | Books | Novel | Story | Man | Author | House | War | Children |         |
| 5 | Wine | Street | Hotel | House | Room | Night | Place | Restaurant | Park | Garden |         |
| 6 | Bush | Campaign | Clinton | Republican | House | Party | Democratic | Political | Democrats | Senator |         |
| 7 | Building | Street | Square | Housing | House | Buildings | Development | Space | Percent | Real |         |
| 8 | Won | Team | Second | Race | Round | Cup | Open | Game | Play | Win |         |
| 9 | Yankees | Game | Mets | Season | Run | League | Baseball | Team | Games | Hit |         |
| 10 | Government | War | Military | Officials | Iraq | Forces | Iraqi | Army | Troops | Soldiers |         |
| 11 | Children | School | Women | Family | Parents | Child | Life | Says | Help | Mother |         |
| 12 | Stock | Percent | Companies | Fund | Market | Bank | Investors | Funds | Financial | Business |         |
| 13 | Church | War | Women | Life | Black | Political | Catholic | Government | Jewish | Pope |         |
| 14 | Art | Museum | Show | Gallery | Works | Artists | Street | Artist | Paintings | Exhibition |         |
| 15 | Police | Yesterday | Man | Officer | Officers | Case | Found | Charged | Street | Shot |         |

Topics found in 1.8M articles from the New York Times
Other examples of applied probabilistic modeling
(from my research group and others)
Communities discovered in a 3.7M node network of U.S. Patents
Neuroscience analysis of 220 million fMRI measurements
Population analysis of 2 billion genetic measurements
Figure 3: Six different styles of art discovered on Etsy. Each column contains high-ranking items from a topic. Note that all items come from unique sellers.

More specifically, the “following” mechanism can be described as follows: Users can “follow” each other on Etsy in the same way that users can follow other users on Twitter. When user A follows user B, user B’s activity (for example: products or shops that user B favorites, or even other users that user B follows) will be shown on user A’s activity feed in the form of story cards (Figure 5). The idea is that a user will want to follow another user who has similar interests, so that it is more likely that user B’s activity will interest user A. Before the deployment of our recommendation system, Etsy users found other users to follow by either 1) knowing the user in person, or 2) stumbling upon them while browsing the site. Thus, the purpose of the user recommendation system was to make the process of finding users with similar interests less arbitrary and more intentional.

4.1.1 Algorithm & Implementation

Once we obtain each user’s interest profile (as described in section 3.2), we conduct a nearest neighbor search across all eligible users on Etsy (i.e. those active users who do not have private settings turned on) to find the top 100 users closest to a user A’s interest profile, which we recommend. These are users, presumably, with the most similar styles and interests. The problem of the nearest neighbor search, of course, is that examining every pair of users to determine the distance between them (the “brute force” approach) is unfeasible due to the large number of users. Therefore, we experimented with two different hashing methods, both of which center around the idea of hashing the interest profiles into buckets, and then computing distances only between users that have similar interests.
Ideological Leanings of Supreme Court Justices

Source Data: Andrew D. Martin and Kevin M. Quinn
http://mqscores.wustl.edu/measures.php

Term Beginning in October

Supreme Court Ideology over time (Martin and Quinn, 2001)
Breaking the Nazi code (Turing and Good, 194?)
Why we like this picture:

- Customized data analysis is important to many fields.
- This pipeline separates assumptions, computation, application.
- It facilitates solving data science problems.
What we need:

- Expressive components from which to build models
- Scalable and generic inference algorithms
- Stretch probabilistic modeling into new areas
Box’s Loop

Make assumptions

Discover patterns

Massive data

Predict & Explore

Criticize model

Revise
What will we cover?
The basics of graphical models

1. Probability: Basic concepts and review
2. Semantics of graphical models
3. D-separation and conditional independence
4. The elimination algorithm
5. Tree propagation and hidden Markov models I
6. Tree propagation and hidden Markov models II
Latent variable models

1. Models, data, and statistical concepts I
2. Models, data, and statistical concepts II
3. Bayesian mixtures of Gaussians and the Gibbs sampler I
4. Bayesian mixtures of Gaussians and the Gibbs sampler II
5. Exponential families, conjugacy, and mixtures of exponential families I
6. Exponential families, conjugacy, and mixtures of exponential families II
7. Mixed-membership, topic models, and variational inference I
8. Mixed-membership, topic models, and variational inference II
9. Matrix factorization and recommendation systems I
10. Matrix factorization and recommendation systems II
Conditional models

1. Regression: Linear and logistic
2. Generalized linear models
3. Hierarchical models, robust models, and empirical Bayes I
4. Hierarchical models, robust models, and empirical Bayes II
Advanced ideas in approximate posterior inference

1. Markov chain Monte Carlo I
2. Markov chain Monte Carlo II
3. Variational inference I
4. Variational inference II
Other topics and summary

1. An brief introduction to Bayesian nonparametrics
2. Summary (and wiggle room)
Some additional discussion

- Programming languages
- Applications
- Box’s loop, again
- Note: We will usually be at the board.
Prerequisites, Requirements, Grades, Etc.

- Office hours: Wednesdays 2:30-4:30, 703 CEPSR
- Prerequisites
  - Probability and Statistics
  - Optimization
  - Programming
- Requirements
  - Weekly paper about the reading (< 1 page)
  - Occasional homework
  - Final project
- Your grade: Mostly the final project