Probabilistic Graphical Models

Special Topics in Machine Learning
Brown University CSCI 2950-P, Spring 2013
Tuesdays & Thursdays, 1:00-2:20pm, CIT506

Instructor:  Erik Sudderth
Teaching Assistant:  Jason Pacheco
Learning from Structured Data
Hidden Markov Models (HMMs)

Visual Tracking

\[ p(x, y) = p(x_0) \prod_{t=1}^{T} p(x_t \mid x_{t-1}) p(y_t \mid x_t) \]

“Conditioned on the present, the past and future are statistically independent”
Kinematic Hand Tracking

Kinematic Prior

Structural Prior

Dynamic Prior
Dynamic Bayesian Networks

Murphy,

2002
Nearest-Neighbor Grids

Low Level Vision
- Image denoising
- Stereo
- Optical flow
- Shape from shading
- Superresolution
- Segmentation

$x_s$ → unobserved or hidden variable
$y_s$ → local observation of $x_s$
Wavelet Decompositions

- Bandpass decomposition of images into multiple scales & orientations
- Dense features which simplify statistics of natural images
Hidden Markov Trees

- Hidden *states* model evolution of image patterns across scale and location
Medical Diagnosis

Parameterization: Noisy-OR, logistic regression, generalized linear models...
Low Density Parity Check (LDPC) Code
Sensor localization
Sensor localization
Example Data for a Topic Model

Poisoning by ice cream.

No chemist certainly would suppose that the same poison exists in all samples of ice cream which have produced gastrointestinal symptoms in man. Mineral poisons, copper, lead, arsenic, and mercury, have all been found in ice cream. In some instances these have been used with criminal intent. In other cases their presence has been accidental. Likewise, that vanilla is sometimes the bearer, at least, of the poison, is well known to all chemists. Dr. Bartley's idea that the poisonous properties of the cream which he examined were due to putrid gelatine is certainly a rational theory. The poisonous principle might in this case arise from the decomposition of the gelatine; or with the gelatine there may be introduced into the milk a ferment, the growth of which a poison is produced.

But in the cream which I examined, none of the above sources of the poisoning existed. There were no mineral poisons present. No gelatine of any kind had been used in making the cream. The vanilla used was shown to be not poisonous. This showing was made, not by a chemical analysis, which might not have been conclusive, but by Mr. Nove and I drank of the vanilla extract which was used, and no ill results followed. Still, from this cream we isolated the same poison which I had before found in poisonous cheese (Zeitschrift für physiologische chemie, x, 114).

RNA Editing and the Evolution of Parasites
Larry Simpson and Dmitri A. Maslov

The Lymnaea stagnalis flagellum, together with their inner group of Euglenoids, represent the rice strain of the Listeria species. The organisms comprise microorganisms (I). Within the lichenophylls there are two major groups, the poorly studied hololichenophylls, which consist of both free-living and parasitic algae, and the better known lichenophylls, which are obligate parasites (2).

Perhaps because of the antiquity of the trilobate annelid lineage, these cells possess several unique genetic features (e.g., a complex, circular, non-replicating bacterial genome). This unique arrangement of genes, which is RNA editing, has been identified in some terrestrial invertebrates. The editing process is thought to be an important factor in the evolution of complex genomic structures. RNA editing can alter mRNA stability, translation efficiency, and may be involved in the production of functional proteins. The availability of these proteins is likely due to the high efficiency of RNA editing, which can sometimes exceed 90%.

Chaotic Beetles
Charles Godfrey and Michael Hassell

Evolution has been known to influence the evolutionary history of species. The study of population dynamics has been crucial in understanding the evolution of complex ecosystems. The analysis of chaotic population models, such as the logistic population model, can provide insights into the dynamics of complex ecosystems. The logistic model, which is a simple mathematical model of population growth, can exhibit complex dynamics, including period-doubling bifurcations and chaotic behavior. These findings highlight the importance of understanding the dynamics of complex ecosystems, which can have implications for the management of natural resources.

Some of the recent work on the dynamics of complex systems has been focused on the analysis of time series data. The analysis of time series data can be used to identify underlying patterns and structures in complex systems. These patterns can provide insights into the dynamics of complex ecosystems, which can have implications for the management of natural resources.
Example Output: 4 Topics

<table>
<thead>
<tr>
<th>human</th>
<th>evolution</th>
<th>disease</th>
<th>computer models</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome</td>
<td>evolutionary</td>
<td>host</td>
<td>information</td>
</tr>
<tr>
<td>dna</td>
<td>species</td>
<td>bacteria</td>
<td>data</td>
</tr>
<tr>
<td>genetic</td>
<td>organisms</td>
<td>diseases</td>
<td>computers</td>
</tr>
<tr>
<td>genes</td>
<td>life</td>
<td>resistance</td>
<td>system</td>
</tr>
<tr>
<td>sequence</td>
<td>origin</td>
<td>bacterial</td>
<td>network</td>
</tr>
<tr>
<td>gene</td>
<td>biology</td>
<td>new</td>
<td>systems</td>
</tr>
<tr>
<td>molecular</td>
<td>groups</td>
<td>strains</td>
<td>model</td>
</tr>
<tr>
<td>sequencing</td>
<td>phylogenetic</td>
<td>control</td>
<td>parallel</td>
</tr>
<tr>
<td>map</td>
<td>living</td>
<td>infectious</td>
<td>methods</td>
</tr>
<tr>
<td>information</td>
<td>diversity</td>
<td>malaria</td>
<td>networks</td>
</tr>
<tr>
<td>genetics</td>
<td>group</td>
<td>parasite</td>
<td>software</td>
</tr>
<tr>
<td>mapping</td>
<td>new</td>
<td>parasites</td>
<td>new</td>
</tr>
<tr>
<td>project</td>
<td>two</td>
<td>united</td>
<td>simulations</td>
</tr>
<tr>
<td>sequences</td>
<td>common</td>
<td>tuberculosis</td>
<td></td>
</tr>
</tbody>
</table>

*Columns sorted by probability of word given topic.*

D. Blei, 2008
Every document discusses a mixture of multiple topics.

D. Blei, 2008
LDA: Generative Model

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, notes Svante Andersson, an evolutionary biologist at Stockholm University. "But we are still far from the 800 number. But coming up with a consensus answer may be more than just a 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 Araceli Mushegian, a computational molecular biologist at the National Center for Biotechnology Information in Bethesda, Maryland. Comparing an organism to its neighbors can provide clues about its evolutionary past.


- Cast these intuitions into a generative probabilistic process
- Each document is a random mixture of corpus-wide topics
- Each word is drawn from one of those topics

D. Blei, 2008
LDA: Graphical Model

- Dirichlet parameter
- Per-document topic proportions
- Per-word topic assignment
- Observed word
- Topics
- Topic hyperparameter

D. Blei, 2008
Graphical Models

Directed Bayesian Network

Factor Graph

Undirected Graphical Model
Undirected Graphical Models

An undirected graph $\mathcal{G}$ is defined by

\[ \mathcal{V} \rightarrow \text{set of } N \text{ nodes } \{1, 2, \ldots, N\} \]

\[ \mathcal{E} \rightarrow \text{set of edges } (s, t) \text{ connecting nodes } s, t \in \mathcal{V} \]

Nodes $s \in \mathcal{V}$ are associated with random variables $x_s$

Nodes $s \in \mathcal{V}$ are associated with random variables $x_s$

Graph Separation

Conditional Independence

\[
p(x_A, x_C|x_B) = p(x_A|x_B)p(x_C|x_B)
\]
Inference in Graphical Models

\[ p(x \mid y) = \frac{1}{Z} \prod_{s \in \mathcal{V}} \psi_s(x_s) \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \]

\( y \longrightarrow \) observations (implicitly encoded via compatibilities)

Maximum a Posteriori (MAP) Estimates

\[ \hat{x} = \arg \max_x p(x \mid y) \]

Posterior Marginal Densities

\[ p_t(x_t \mid y) = \sum_{x_{\mathcal{V} \setminus t}} p(x \mid y) \]

- Provide both estimators and confidence measures
- Sufficient statistics for iterative parameter estimation
Why the Partition Function?

\[ Z = \sum_x \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E} \psi_{st}(x_s, x_t) \]

**Statistical Physics**
- Sensitivity of physical systems to external stimuli

**Hierarchical Bayesian Models**
- Marginal likelihood of observed data
- Fundamental in hypothesis testing & model selection

**Cumulant Generating Function**
- For exponential families, derivatives with respect to parameters provide marginal statistics

**PROBLEM:** Computing \( Z \) in general graphs is NP-complete
Exact Inference

MESSAGES: Sum-product or belief propagation algorithm

\[ m_{ts}(x_s) = \alpha \sum_{x_t} \psi_{st}(x_s, x_t) \psi_t(x_t, y) \prod_{u \in \Gamma(t) \setminus s} m_{ut}(x_t) \]

Computational cost:

- Number of nodes: \( N \)
- Discrete states: \( M \)

Belief Prop: \( \mathcal{O}(NM^2) \)

Brute Force: \( \mathcal{O}(M^N) \)
Continuous Variables

\[ m_{ij}(x_j) \propto \int_{x_i} \psi_{j,i}(x_j, x_i) \psi_i(x_i, y) \prod_{k \in \Gamma(i) \setminus j} m_{ki}(x_i) \, dx_i \]

Discrete State Variables

- Messages are *finite vectors*
- Updated via matrix-vector products

Gaussian State Variables

- Messages are *mean & covariance*
- Updated via information Kalman filter

Continuous Non-Gaussian State Variables

- Closed parametric forms unavailable
- Discretization can be *intractable* even with 2 or 3 dimensional states
Variational Inference: An Example

\[ p(x \mid y) = \frac{1}{Z} \prod_{(s,t) \in \mathcal{E}} \psi_{st}(x_s, x_t) \prod_{s \in \mathcal{V}} \psi_s(x_s, y) \]

• Choose a family of approximating distributions which is tractable. The simplest example:

\[ q(x) = \prod_{s \in \mathcal{V}} q_s(x_s) \]

• Define a distance to measure the quality of different approximations. One possibility:

\[ D(q \parallel p) = \sum_x q(x) \log \frac{q(x)}{p(x \mid y)} \]

• Find the approximation minimizing this distance
Advanced Variational Methods

- Exponential families
- Mean field methods: naïve and structured
- Variational EM for parameter estimation
- Loopy belief propagation (BP)
- Bethe and Kikuchi entropies
- Generalized BP, fractional BP
- Convex relaxations and bounds
- MAP estimation and linear programming
- ........
Markov Chain Monte Carlo

At each time point, state $z^{(t)}$ is a configuration of all the variables in the model: parameters, hidden variables, etc.

We design the transition distribution $q(z | z^{(t)})$ so that the chain is irreducible and ergodic, with a unique stationary distribution $p^*(z)$

$$p^*(z) = \int_{z'} q(z | z') p^*(z') \, dz'$$

For learning, the target equilibrium distribution is usually the posterior distribution given data $x$: $p^*(z) = p(z | x)$

Popular recipes: Metropolis-Hastings and Gibbs samplers
Sequential Monte Carlo

Particle Filters, Condensation, Survival of the Fittest,…

- Nonparametric approximation to optimal BP estimates
- Represent messages and posteriors using a set of samples, found by simulation

Sample-based density estimate

Weight by observation likelihood

Resample & propagate by dynamics
Course Evaluation

Homeworks: 60%

- Four equally weighted assignments
- Each assignment available for two weeks before due date
- Combine mathematical derivations, algorithm design, programming, and analysis of real datasets
  - Multiscale models of images, objects, visual scenes
  - Particle filters for localization and tracking
  - Topic models of text document collections
  - ...

Final Project: 40%

- Proposal: 1-3 pages, due on March 22 (5%)
- Presentation: ~10 minutes, on May 7 (10%)
- Conference-style technical report, due on May 13 (25%)
Final Projects

Best case: Application of course material to your own area of research

Key Requirements: Novel use of graphical models

- Identify a family of graphical models suitable for a particular application, try baseline learning algorithms
- Propose, develop, and experimentally test a new type of graphical learning or inference algorithm
- Experimentally compare different models or algorithms on an interesting, novel dataset
- **There will not be a list of projects to choose from.** You must propose your own (with the instructor’s advice). We will include pointers to many research papers with relevant applications.
Changes from Previous Years

• Readings from books & in-depth tutorials, not recent research papers. *More accessible.*
• No reading comments or student presentation of research papers. *Course staff will lecture.*
• Homework assignments require *mathematical derivations and algorithm implementation.*
• Subject matter: *Probabilistic Graphical Models*
  ➢ Fall 2011 topic was *Applied Bayesian Nonparametrics*, may repeat for credit
  ➢ Spring 2010 topics similar. *You are welcome to (officially) audit, but see me about taking for credit.*
Textbook & Readings

An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

- Draft textbook by Michael I. Jordan, available as a printed course reader, more details soon…
- Variational tutorial by Wainwright and Jordan (2008)
- Background chapter of Prof. Sudderth’s thesis
- Tutorial articles on Markov chain Monte Carlo, particle filters
- A few other papers for advanced topics…
Course Prerequisites

• A course in modern statistical machine learning
  ➢ Brown CSCI 1950F: Intro to Machine Learning
  ➢ Brown APMA 1690: Computational Probability and Statistics (also APMA 2690)
  ➢ Possibly other classes or experience...

• Programming experience (Matlab, Java, …)

• Readings will require “mathematical maturity”

• Insufficient background by themselves:
  ➢ Brown CSCI 1410: Introduction to AI
  ➢ Traditional undergrad statistics (APMA 1650/1660)
### Prereq: Intro Machine Learning

#### Supervised Learning
- Classification or categorization
- Regression

#### Unsupervised Learning
- Clustering
- Dimensionality reduction

#### Bayesian and frequentist estimation

#### Model selection, cross-validation, overfitting

#### Expectation-Maximization (EM) algorithm
You will probably want a copy of one of these books…
Naïve Bayes Inference:

\[ p(y, \mathbf{x}) = p(y) \prod_{j=1}^{D} p(x_j | y) \]

Convention: Shaded nodes are observed, open nodes are latent/hidden.

Plates denote replication of random variables.
Supervised Learning

Generative ML or MAP Learning: *Naïve Bayes*

\[
\max_{\pi, \theta} \log p(\pi) + \log p(\theta) + \sum_{i=1}^{N} \left[ \log p(y_i \mid \pi) + \log p(x_i \mid y_i, \theta) \right]
\]

Discriminative ML or MAP Learning: *Logistic regression*

\[
\max_{\theta} \log p(\theta) + \sum_{i=1}^{N} \log p(y_i \mid x_i, \theta)
\]
Learning via Optimization

ML Estimate: \[ \hat{w} = \arg \min_w - \sum_i \log p(y_i | x_i, w) \]

MAP Estimate: \[ \hat{w} = \arg \min_w - \log p(w) - \sum_i \log p(y_i | x_i, w) \]

Gradient vectors:
\[ f : \mathbb{R}^M \rightarrow \mathbb{R} \]
\[ \nabla_w f : \mathbb{R}^M \rightarrow \mathbb{R}^M \]

Hessian matrices:
\[ \nabla^2_w f : \mathbb{R}^M \rightarrow \mathbb{R}^{M \times M} \]
\[ (\nabla_w f(w))_{k,l} = \frac{\partial^2 f(w)}{\partial w_k \partial w_\ell} \]

Optimization of Smooth Functions:
- **Closed form**: Find zero gradient points, check curvature
- **Iterative**: Initialize somewhere, use gradients to take steps towards better (by convention, smaller) values
Unsupervised Learning

Clustering:

$$\max_{\pi, \theta} \log p(\pi) + \log p(\theta) + \sum_{i=1}^{N} \log \left[ \sum_{z_i} p(z_i \mid \pi) p(x_i \mid z_i, \theta) \right]$$

Dimensionality Reduction:

$$\max_{\pi, \theta} \log p(\pi) + \log p(\theta) + \sum_{i=1}^{N} \log \left[ \int_{z_i} p(z_i \mid \pi) p(x_i \mid z_i, \theta) \, dz_i \right]$$

- No notion of training and test data: labels are never observed
- As before, maximize posterior probability of model parameters
- For hidden variables associated with each observation, we marginalize over possible values rather than estimating
  - Fully accounts for uncertainty in these variables
  - There is one hidden variable per observation, so cannot perfectly estimate even with infinite data
- Must use generative model (discriminative degenerates)
Expectation Maximization (EM)

- **Initialization**: Randomly select starting parameters
- **E-Step**: Given parameters, find posterior of hidden data
  - Equivalent to test inference of full posterior distribution
- **M-Step**: Given posterior distributions, find likely parameters
  - Similar to supervised ML/MAP training
- **Iteration**: Alternate E-step & M-step until convergence
Gaussian Mixture Models vs. HMMs

**Mixture Model**

\[
p(z_i \mid \pi, \mu, \Sigma) = \text{Cat}(z_i \mid \pi)
\]

\[
p(x_i \mid z_i, \pi, \mu, \Sigma) = \text{Norm}(x_i \mid \mu_{z_i}, \Sigma_{z_i})
\]

**Hidden Markov Model**

\[
p(z_t \mid \pi, \mu, \Sigma, z_{t-1}, z_{t-2}, \ldots) = \text{Cat}(z_t \mid \pi_{z_{t-1}})
\]

\[
p(x_t \mid z_t, \pi, \mu, \Sigma) = \text{Norm}(x_t \mid \mu_{z_t}, \Sigma_{z_t})
\]

Recover mixture model when all rows of state transition matrix are equal.
Probabilistic PCA & Factor Analysis

- **Both Models:** Data is a linear function of low-dimensional latent coordinates, plus Gaussian noise

\[
p(x_i \mid z_i, \theta) = \mathcal{N}(x_i \mid Wz_i + \mu, \Psi) \quad p(z_i \mid \theta) = \mathcal{N}(z_i \mid 0, I)
\]

\[
p(x_i \mid \theta) = \mathcal{N}(x_i \mid \mu, WW^T + \Psi) \quad \text{low rank covariance parameterization}
\]

- **Factor analysis:** \(\Psi\) is a general diagonal matrix
- **Probabilistic PCA:** \(\Psi = \sigma^2 I\) is a multiple of identity matrix
A Quick Poll
Administration

Registration: E-mail sunderth@cs.brown.edu with
• Your name and CS logon
• Your department, major, and year
• Your background in statistical machine learning
  ➢ If you’ve taken Brown courses, just say which ones
  ➢ Otherwise, a few sentences about your experience

Course webpage: Up now, watch for more information

http://cs.brown.edu/courses/csci2950-p/index.html

Readings for Tuesday:
• Chapter 2 from textbook (available soon)