Inference and Representation: CTMs, Gibbs Sampling & Markov Chains

Rahul G. Krishnan

New York University

Lab 6, October 12, 2016

Outline

Modeling Correlations in the Simplex

- LDA assumes that per-document topics are drawn from a Dirichlet distribution
- $\theta_d \sim \text{Dir}(\alpha)$
- Instead we will use the logistic normal distribution:
- $\eta \sim \mathcal{N}(\mu; \Sigma) \qquad \theta_d^{(i)} = \frac{\exp(\eta_i)}{\sum_j \exp(\eta_j)}$
- The covariance structure Σ determines how related topics are

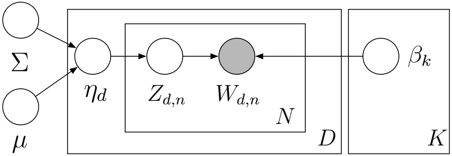
Correlated Topic Model (Blei and Lafferty, 2006)

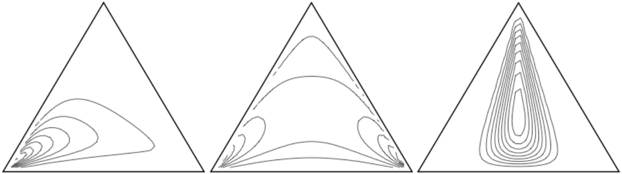
- Goal: model relationships of topics (e.g., “a document about genetics is more likely to be about disease than x-ray astronomy”)
- Training data: corpora of documents, just like for LDA
  - 16,351 OCR articles from *Science*
- Why do this?
- How to do this?

Correlated Topic Model (Blei and Lafferty, 2006)

Replaces Dirichlet prior for θ with the *logistic Normal* distribution:

1. $\eta \mid \{\mu, \Sigma\} \sim \mathcal{N}(\mu, \Sigma)$
2. $\theta_t = \exp \eta_t / \sum_{t'} \exp \eta_{t'}$

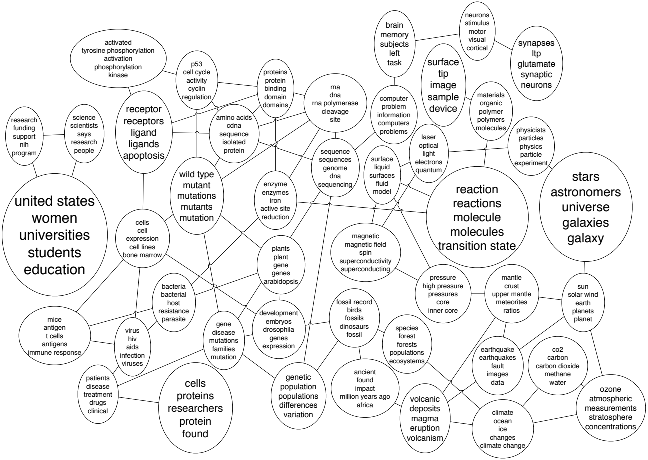




Inverse Covariance Matrix

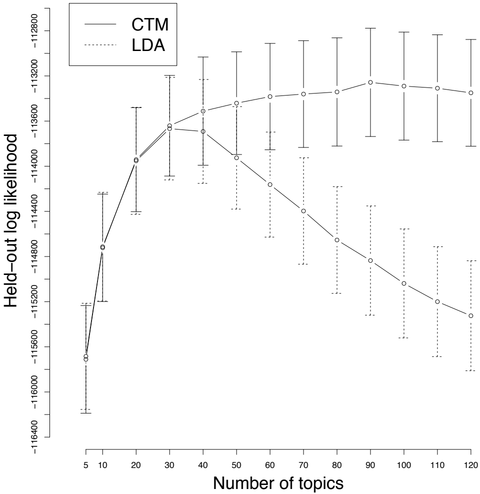
- If we learn Σ, the covariance matrix in the prior then we can construct a graph of relationships between topics
- Recall, $p(\eta_d) \propto (\eta_d - \mu)^T \Sigma^{-1} (\eta_d - \mu)^T$
- The non-zero entries (and off-diagonal) entries in the matrix Σ^{-1} represent relationships between two topics
- Form a graph $\mathcal{G} = (V, E)$ where V are the rows/columns of Σ^{-1} corresponding to topics
- The edges E are the set (ij) $s.t.$ $\Sigma^{-1}_{ij} \neq 0$

Visualizing the inverse covariance matrix



http://www.cs.cmu.edu/~lemur/science/

Held-out likelihood as a function of number of topics



Recap

- Recall that inference in Markov Random Fields is hard.
- Loopy BP was one algorithm that gave us a way to estimate marginal probabilities in an MRF
- Lets use Gibbs sampling to perform approximate inference

Ising Model

- Under an pairwise Ising model with only edge potentials, we have:
- $p(X_1, \ldots, X_N) \propto \prod_{(ij) \in E} \psi_{ij}(x_i, x_j)$
- Instead of a table of potentials like we familiar with,
- Ising Model Edge Potential: $\psi_{ij}(x_i, x_j) = \exp(J x_i x_j)$
- J is a number that denotes the edge strength
- $x_i \in \{-1, 1\}$

Gibbs Sampling

- Gibbs sampling proceeds by iteratively sampling from $p(X_i|X_{\neg i})$
- $p(X_i|X_{\neg i}) \propto \frac{p(X_1,\ldots,X_N)}{p(X_{\neg i})}$
- Since we condition on the Markov Blanket ($N(i)$, the neighbors), for a node,
- $p(X_i = x_i|X_{\neg i}) \propto \prod_{j \in N(i)} \psi_{ij}(x_i, x_j)$
- i.e it suffices to consider the set of edge potentials corresponding to the edges between i and $N(i)$

Deriving the Updates

Assuming $\exists$ assignments $X_j = x_j \in N(i)$

$$p(x_i = +1|x_{\neg i}) = \frac{\prod_{j\in N(i)} \psi_{ij}(X_i = +1, x_j = x_j)}{\prod_{j\in N(i)} \psi_{ij}(x_i = +1, x_j) + \prod_{j\in N(i)} \psi_{ij}(x_i = -1, x_j)}$$

$$p(x_i = +1|x_{\neg i}) = \frac{\prod_{j\in N(i)} \exp(Jx_j)}{\prod_{j\in N(i)} \exp(Jx_j) + \prod_{j\in N(i)} \exp(-Jx_j)}$$

$$p(x_i = +1|x_{\neg i}) = \frac{\exp(J\sum_{j\in N(i)} x_j)}{\exp(J\sum_{j\in N(i)} x_j) + \exp(-J\sum_{j\in N(i)} x_j)}$$

Derivation

$$p(x_i = +1|x_{\neg i}) = \frac{\exp(J\sum_{j\in N(i)} x_j)}{\exp(J\sum_{j\in N(i)} x_j) + \exp(-J\sum_{j\in N(i)} x_j)}$$

Denote: $\eta_i := \sum_{j\in N(i)} x_j$

$$p(x_i = +1|x_{\neg i}) = \frac{\exp(J\eta_i)}{\exp(J\eta_i) + \exp(-J\eta_i)}$$

Divide numerator and denominator by $\exp(J\eta_i)$:

$$p(x_i = +1|x_{\neg i}) = \frac{1}{1 + \exp(-2J\eta_i)} = \text{sig}(2J\eta_i)$$

Procedure

1. Start with a random assignments x_i for all random variables
2. For a random variable $X_1, \ldots, X_N$, sample an assignment $\hat{x}_i \sim \mathrm{sigm}(2J\eta_i)$
3. Set $x_i = \hat{x}_i$ and continue sampling other random variables
4. Repeat (2)

Recap (from lecture)

- We're interested in evaluating $\mathbb{E}_{p(x)}\left[f(x)\right]$
- If we can sample from $p(x)$
- Then, we can evaluate $\mathbb{E}_{p(x)}\left[f(x)\right] = \frac{1}{N}\sum_{i=1}^{N} f(\hat{x}_i) \qquad \hat{x}_i \sim p(x)$
- How do we sample *independantly* from $p(x)$?
- Enter Markov Chains.

What is it?

- A Markov Chain is a stochastic process that operates sequentially, going from one state to the next
- $x_0 \rightarrow x_1 \rightarrow x_2 \rightarrow \ldots x_k \rightarrow \ldots x_{k+1} \rightarrow \ldots, x_K$
- *State space:* The set of all values of the random variable
- *Index/Time:* k, the step of the Markov Chain we are currently at
- *Markov Property:* $p(x_{k+1}|x_k, x_{<k}) = p(x_{k+1}|x_k)$
- *Transition Operator:* $\mathcal{T}(x_k \rightarrow x_{k+1})$
- *Time Homogenous:* A MC is time homogenous if the $\mathcal{T}$ stays constant across time

Visualization

- See a visualization of Markov Chains `http://setosa.io/ev/markov-chains/`
- Each state in the visualization A, B etc. corresponds to a different state of the random variable
- We want a way to sample from $p(x)$
- We're going to build a Markov Chain to visit states of the random variables in a manner such that the number of times A is visited is proportional to $p(x = A)$
- We use a Markov Chain to form a Monte Carlo approximation to $\mathbb{E}_{p(x)}[f(x)]$. Hence the name Markov Chain Monte Carlo

Properties of Markov Chains

- *Irreducibility:* For every state of the Markov chain, there is a positive probability of visiting all other states
- *Aperiodicity:* The chain should not be cyclic. i.e the greatest common denominator of the set of time indices that we visit any state x_i should be one
- *Detailed Balance:* $\pi(x)\mathcal{T}(x \rightarrow x') = \pi(x')\mathcal{T}(x' \rightarrow x)$.

Why are these properties relevant?

- *Limiting distribution:* A distribution π where if we start in any initial distribution π_0, we eventually converge to π by running the Markov Chain
- *Stationary Distribution:* A distribution π such that if we start the chain in π, we stay in π
- *Proposition:* If the Markov Chain is irreducible, aperiodic and satisfies detailed balance, then there exists a limiting distribution.
- *Proposition:* If the limiting distribution exists, it must be the stationary distribution
- If we run the Markov Chain we construct for long enough, we will be sampling from the stationary distribution.

Tying it all together

- We have some complex, high dimensional, probability distribution $\pi(x)$ of interest
- Direct sampling not possible
- Use MCMC:
  - Construct a Markov Chain $\{X_i\}_{i=1}^{\infty}$ such that $\lim_{i \to \infty} p(X_i = x) = \pi(x)$
  - Simulate this and for large i, take samples $\{x_i, \ldots, x_{i+m}$. These are samples from $\pi(x)$
  - Use these samples to form your Monte Carlo estimate

Coming Up

- Key Question: How do we move from one state to the next? What $\mathcal{T}$ do we use? How do we guarantee that the Markov Chain that we construct will have the right stationary distribution?
- In upcoming lectures you will learn:
  1. Metropolis Hasting : How to construct a valid transition kernel
  2. Gibbs Sampling : Sampling from the conditional distribution as one way to perform Metroplis Hastings.