Finding Scientific Topics & Integrating Topics and Syntax

Griffiths, Steyvers, et al

Rebecca Mason
CS2950P Probabilistic Graphical Models
April 21, 2010

Overview

Gibbs Sampler

Paper: Finding Scientific Topics

Paper: Integrating Topics and Syntax

Gibbs Sampler

Gibbs sampler (Geman and Geman 1984) is a special case of Metropolis-Hastings. It is based on the idea that it is easier to consider a sequence of conditional distributions than to obtain the marginal by integration of the joint density.

Example: at step t:

$$x^{(t+1)} \sim p(x_j | x_1^{(t)}, ..., x_{j-1}^{(t)}, x_{j+1}^{(t)}, ..., x_N^{(t)})$$

Approximating the Marginal Distributions

The expectation of any function f of the random variable x is approximated by

$$E[f(x)]_m = \frac{1}{m}\sum_{i=1}^{m} f(x_i)$$

Since

$$p(x) = \int p(x|y)p(y)dy = E_y[p(x|y)$$

one can approximate the marginal density using

$$\hat{p}_m^{(x)} = \frac{1}{m}\sum_{i=1}^{m} p(x|y = y_i)$$

Variants of Gibbs Sampler

- Gibbs Sampler
  1. Draw a conditioned on b,c
  2. Draw b conditioned on a,c
  3. Draw c conditioned on a,b
- Blocked Gibbs Sampler
  1. Draw a,b conditioned on c
  2. Draw c conditioned on a,b
- Collapsed Gibbs Sampler
  1. Draw a conditioned on c
  2. Draw c conditioned on a

Finding Scientific Topics

Thomas Griffiths and Mark Steyvers (2004)

- Uses LDA to model which topics documents address.
- Gibbs sampling for inference
- Example: Applying topic models to images
- Application: identify "hot topics" that are more popular over time
- Application: tagging abstracts

Modeling Documents and Topics

T topics, probability of ith word in given document is

$$P(w_i) = \sum_{j=1}^{T} P(w_i | z_i = j) P(z_i = j)$$

Want to find the posterior distribution over assignments of words to topics

$$P(\mathbf{z}|\mathbf{w}) = \frac{P(\mathbf{w}, \mathbf{z})}{\sum_z P(\mathbf{w}, \mathbf{z})}$$

This distribution cannot be computed directly because the sum in the denominator does not factorize and involves T^n terms, where n is the total number of word items in the corpus.

Using Gibbs Sampling for Inference

To apply Gibbs sampling we need the full conditional distribution.

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

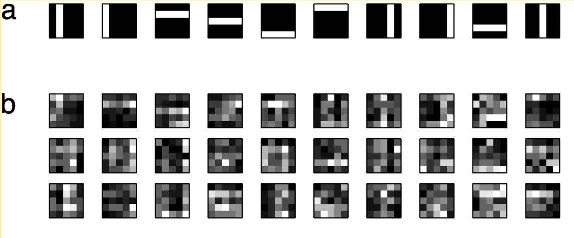
where $n_{-i}^{(\cdot)}$ is a count that does not include the current assignment of z_i

Estimates of ϕ and θ:

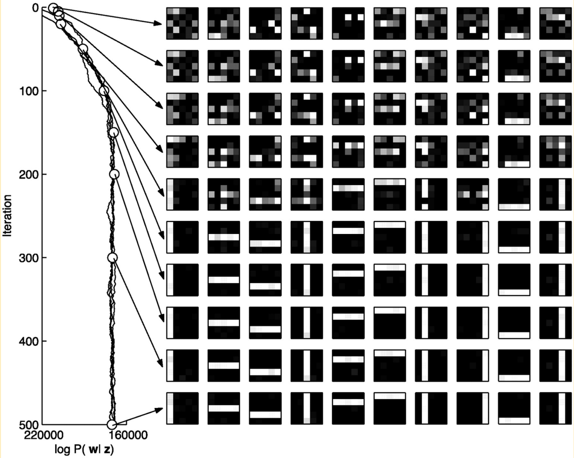
$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + W\beta}$$

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha}$$

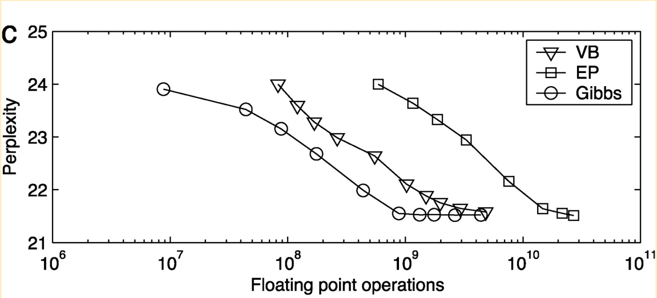
Applying Topic Models to Images



Applying Topic Models to Images

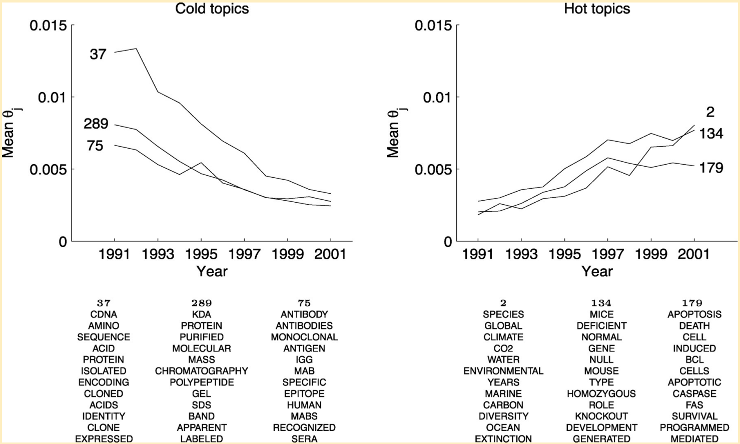


Convergence



Hot and Cold Topics

Conducted linear trend analysis on θ_j to find topics that rose or fell in popularity



Tagging Abstracts

Find words that are important to a topic

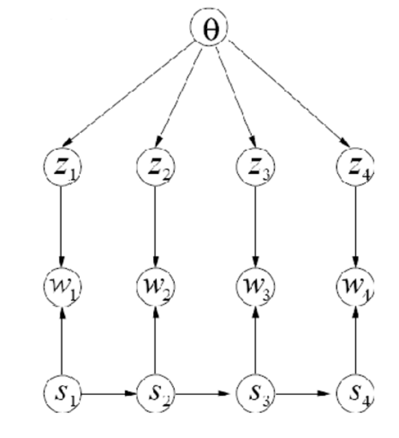
A generalized[3] fundamental[146] theorem[267] of natural[280] selection[280] is derived[233] for populations[280] incorporating[149] both genetic[280] and cultural[280] transmission[25]. The phenotype[3] is determined[17] by an arbitrary[3] number[287] of multiallelic[3] loci[3] with two[271]-factor[60] epistasis[280] and an arbitrary[149] linkage[3] map[3], as well as by cultural[280] transmission[25] from the parents[280]. Generations[280] are discrete[69] but partially[275] overlapping[146], and mating[280] may be nonrandom[280] at either the genotypic[280] or the phenotypic[280] level[199] (or both). I show[25] that cultural[280] transmission[25] has several[173] important[173] implications[17] for the evolution[280] of population[280] fitness[280], most notably[230] that there is a time[72] lag[72] in the response[213] to selection[280] such that the future[287] evolution[280] depends[105] on the past selection[280] history[280] of the population[280].

Integrating Topics and Syntax

Griffiths, Steyvers, Blei, Tenenbaum (2005)

- Presents a generative model that uses short-range syntactic dependencies and long-range semantic dependencies
- Gibbs sampling for inference
- Application: part-of-speech tagging
- Application: document classification

Graphical Model



Semantic states: Generate words from LDA topic model

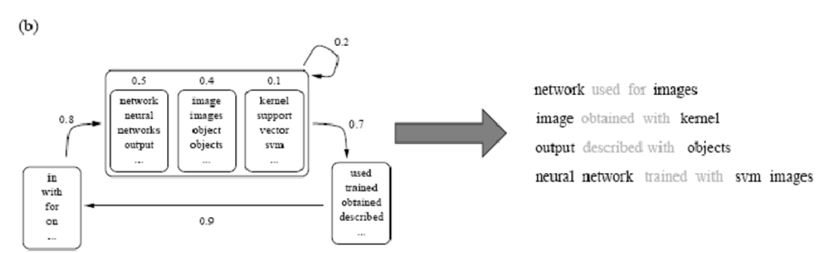
Syntactic states: generate words from HMM

Generating a Document

Sample $\theta^{(d)}$ from a Dirichlet(α) prior

For each word w_i in document d

1. Draw z_i from $\theta^{(d)}$
2. Draw c_i from $\pi^{(c_{i-1}))}$
3. If $c_i = 1$, then draw w_i from $\phi^{(z_i)}$, else draw w_i from $\pi^{(c_i))}$



Example of generating a phrase

Inference

Use Gibbs sampling to iteratively draw a topic assignment z_i and class assignment c_i for each word w_i in the corpus

Each z_i is drawn from:

$$\begin{aligned} P(z_i|\mathbf{z_{-i}}, \mathbf{c}, \mathbf{w}) &\propto P(z_i|\mathbf{z_{-i}})P(w_i|\mathbf{z}, \mathbf{c}, \mathbf{w_{-i}}) \\ &\propto \begin{cases} n_{z_i}^{(d_i)} + \alpha, & \text{if } c_i \neq 1 \\ (n_{z_i}^{(d_i)} + \alpha)\frac{n_{w_i}^{(z_i)}+\beta}{n^{(z_i)}+W\beta}, & \text{if } c_i = 1 \end{cases} \end{aligned}$$

Each c_i is drawn from:

$$\begin{aligned} P(c_i|\mathbf{c_{-i}}, \mathbf{z}, \mathbf{w}) &\propto P(w_i|\mathbf{c}, \mathbf{z}, \mathbf{w_{-i}})P(\mathbf{c}) \\ &\propto \begin{cases} \frac{n_{w_i}^{(c_i)}+\delta}{n^{(c_i)}+W\delta}\frac{(n_{c_i}^{c_{i}-1}+\gamma)(n_{c_i+1}^{c_i}+\gamma)}{n^{c_i}+C\gamma}, & \text{if } c_i \neq 1 \\ \frac{n_{w_i}^{(z_i)}+\beta}{n^{(z_i)}+W\beta}\frac{(n_{c_i}^{c_{i}-1}+\gamma)(n_{c_i+1}^{c_i}+\gamma)}{n^{c_i}+C\gamma}, & \text{if } c_i = 1 \end{cases} \end{aligned}$$

LDA Topics vs HMM-LDA Topics

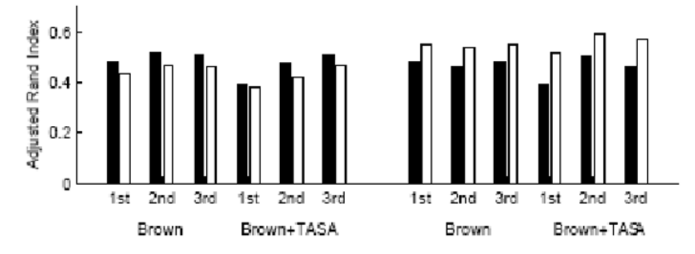
the	the	the	the	the	a	the	the	the
blood	,	,	of	a	the	,	,	,
,	and	and	,	of	of	of	a	a
of	of	of	to	,	,	a	of	in
body	a	in	in	in	in	and	and	game
heart	in	land	and	to	water	in	drink	ball
and	trees	to	classes	picture	is	story	alcohol	and
in	tree	farmers	government	film	and	is	to	team
to	with	for	a	image	matter	to	bottle	to
is	on	farm	state	lens	are	as	in	play
blood	forest	farmers	government	light	water	story	drugs	ball
heart	trees	land	state	eye	matter	stories	drug	game
pressure	forests	crops	federal	lens	molecules	poem	alcohol	team
body	land	farm	public	image	liquid	characters	people	*
lungs	soil	food	local	mirror	particles	poetry	drinking	baseball
oxygen	areas	people	act	eyes	gas	character	person	players
vessels	park	farming	states	glass	solid	author	effects	football
arteries	wildlife	wheat	national	object	substance	poems	marijuana	player
*	area	farms	laws	objects	temperature	life	body	field
breathing	rain	corn	department	lenses	changes	poet	use	basketball
the	in	he	*	be	said	can	time	,
a	for	it	new	have	made	would	way	;
his	to	you	other	see	used	will	years	(
this	on	they	first	make	came	could	day	:
their	with	i	same	do	went	may	part	)
these	at	she	great	know	found	had	number	
your	by	we	good	get	called	must	kind	
her	from	there	small	go		do	place	
my	as	this	little	take		have		
some	into	who	old	find		did		

Top: Topics extracted by LDA model. Middle: Topics from composite model.

Bottom: Classes from composite model

Part-of-Speech Tagging

Black is all tags, and white is 10 top-level tags. Left: HMM. Right: HMM-LDA.



HMM-LDA does slightly worse for all tags, because words that are in the same semantic class will be assigned together, so composite model does not capture all the distinctions.

References

1. Erik Sudderth's CS2950P lecture (April 19th, 2010)
2. web.mit.edu/ wingated/www/introductions/mcmc-gibbs-intro.pdf
3. http://nlpers.blogspot.com/2007/07/collapsed-gibbs.html
4. Finding Scientific Topics
5. Integrating Topics and Syntax