CS598JHM: **Advanced NLP** (Spring '10)

Lecture 5: More on Gibbs sampling

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

http://www.cs.uiuc.edu/class/sp10/cs598jhm

Naive Bayes for text classification

The task:

Assign (sentiment) label $L_i \in \{+,-\}$ to a document $\boldsymbol{W}_i$.

The model:

- $L_i = argmax_L\ P(L \mid \boldsymbol{W}_i) = argmax_L\ P(\boldsymbol{W}_i \mid L)P(L)$
- $P(\boldsymbol{W}_i \mid L)$ is a multinomial distribution: $\boldsymbol{W}_i \sim Multinomial(\boldsymbol{\theta}_L)$
  We have a vocabulary of V words. Thus: $\boldsymbol{\theta}_L = (\theta_1, ..., \theta_V)$
- $P(L)$ is a Bernoulli distribution: $L \sim Bernoulli(\pi)$

Estimation

- Labels: $L \sim Bernoulli(\pi)$
- Words: $\boldsymbol{W}_i \mid L \sim Multinomial(\boldsymbol{\theta}^L)$

	Supervised	**Unsupervised**
Freq.	**Relative frequency estimation** - Labels: $\pi = D^+/d$ - Words: $\boldsymbol{\theta}_i^+ = N^+(w_i)/N^+$	**Expectation Maximization:** At each iteration t: - Labels: $\pi^{(t)} = E[D]_{(t-1)}/d$ - Words: $\boldsymbol{\theta}_i^+ = E[N^+(w_i)]_{(t-1)}/E[N^+(w_i)]_{(t-1)}$
Bayes	**With priors:** - Labels: $\pi = (D^+ + \alpha)/(D + \alpha + \beta)$ - Words: $\boldsymbol{\theta}_i^+ = (N^+(w_i) + \gamma_i)/(N^+(w) + \gamma_0)$	**Gibbs sampling**

Approximating expectations

$$
\begin{aligned}
E[f(x)] &= \int_0^1 f(x)p(x)dx \\
&= \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) \\
&\quad \text{for } x^{(1)}...x^{(i)}...x^{(N)} \text{ drawn from } p(x) \\
&\approx \frac{1}{T}\sum_{i=1}^{T} f(x^{(i)}) \\
&\quad \text{for } x^{(1)}...x^{(i)}...x^{(T)} \text{ drawn from } p(x)
\end{aligned}
$$

Markov Chain Monte Carlo

A multivariate distribution $p(\mathbf{x}) = p(x_1,\ldots,x_k)$ with discrete x_i has only a finite number of possible outcomes.

Markov Chain Monte Carlo methods construct a Markov chain whose states are the outcomes of $p(\mathbf{x})$.

The probability of visiting state $\mathbf{x}_j$ is $p(\mathbf{x}_j)$

We sample from $p(\mathbf{x})$ by visiting a sequence of states from this Markov chain.

Gibbs sampling

We visit states according to transition probabilities $P(\mathbf{y}|\,\mathbf{x})$

We go from state $\mathbf{x} = (x_1,\ldots,x_k)$ to state $\mathbf{y} = (y_1,\ldots,y_k)$

We get from $\mathbf{x} = (x_1,\ldots,x_k)$ to $\mathbf{y} = (y_1,\ldots,y_k)$ in k steps:

$(x_1, x_2, \ldots, x_i, \ldots, x_{k-1}, x_k) = \mathbf{x} = \mathbf{x}^{(t)}$
$(y_1, x_2, \ldots, x_i, \ldots, x_{k-1}, x_k)$
$(y_1, y_2, \ldots, x_i, \ldots, x_{k-1}, x_k)$
$(y_1, y_2, \ldots, x_i, \ldots, x_{k-1}, x_k)$
$(y_1, y_2, \ldots, y_i, \ldots, x_{k-1}, x_k)$
$(y_1, y_2, \ldots, y_i, \ldots, x_{k-1}, x_k)$
$(y_1, y_2, \ldots, y_i, \ldots, y_{k-1}, x_k)$
$(y_1, y_2, \ldots, y_i, \ldots, y_{k-1}, y_k) = \mathbf{y} = \mathbf{x}^{(t+1)}$

Gibbs sampling

Individiual steps:

For $i = 1\ldots k$:
pick $\mathbf{y}_i$ by sampling from $P(Y_i \mid y_1,\ldots,y_{i-1}, x_{i+1},\ldots,x_k)$

$$P(Y_i = \mathbf{y}_i \mid y_1,\ldots,y_{i-1}, x_{i+1},\ldots,x_k) = \frac{P(y_1,\ldots,y_{i-1}, \mathbf{y}_i, x_{i+1},\ldots,x_k)}{P(y_1,\ldots,y_{i-1}, x_{i+1},\ldots,x_k)}$$

Gibbs sampling

Our states:
One label assignment $L_1,\ldots,L_N$ to each of our N documents
$\mathbf{x} = (L_1,\ldots,L_N)$

Our transitions:
We go from one label assignment $\mathbf{x} = (+,+,-,+,-\ldots+)$
to another $\mathbf{y} = (-,+,+,+,\ldots,+)$

Our intermediate steps:
We generate label Y_i conditioned on $Y_1\ldots Y_{i-1}$ and $X_{i+1}\ldots X_N$
Call label assignment $Y_1\ldots Y_{i-1}, X_{i+1}\ldots X_N$ $\mathbf{L}^{(-i)}$
We need to compute $P(Y_i \mid \mathbf{D}\,\mathbf{L}^{(-i)}, \alpha, \beta, \gamma)$

Gibbs sampling

$$\underbrace{P(L_j = +|\mathbf{D}, \mathbf{L}^{(-\mathbf{j})}; \alpha, \beta, \boldsymbol{\gamma})}_{\text{prob. that } \mathrm{D_j} \text{ is pos. review}}$$

$$\propto \underbrace{P(\mathbf{W_j}|+, D_+^{(-j)}; \boldsymbol{\gamma})}_{\text{pos. review generates } \mathrm{D_j}} \underbrace{P(L_j = +|\mathbf{L}^{(-\mathbf{j})}; \alpha, \beta)}_{\text{prob. of pos. review}}$$

$$P(L_j = + \mid \mathbf{L}^{(-\mathbf{j})}; \alpha, \beta) = \frac{\alpha + N_+^{(-j)}}{\alpha + \beta + N - 1}$$

$$P(w_k = y|D_+^{(-j)}; \boldsymbol{\gamma}) = \frac{N_{D_x^{(-j)}}(y) + \gamma_y}{\gamma_0 + N_{D_x^{(-j)}}}$$

NB: $\pi, \theta+, \theta-$ disappear (are integrated out)

Why we don't need to estimate π

$$\begin{aligned}
P(L_j = + \mid \mathbf{L}^{(-\mathbf{j})}; \alpha, \beta) &= \int P(L_j = +|\pi) P(\pi|\mathbf{L}^{(-\mathbf{j})}; \alpha, \beta) d\pi \\
&= \int \pi P(\pi|\mathbf{L}^{(-\mathbf{j})}; \alpha, \beta) d\pi \\
&= \int \pi \frac{\Gamma(\alpha+\beta+N-1)}{\Gamma(\alpha+N_+^{(-j)})\Gamma(\beta+N_-^{(-j)})} \pi^{\alpha+N_+^{(-j)}-1}(1-\pi)^{\beta+N_-^{(-j)}-1} d\pi \\
&= \frac{\Gamma(\alpha+\beta+N-1)}{\Gamma(\alpha+N_+^{(-j)})\Gamma(\beta+N_-^{(-j)})} \int \pi^{\alpha+N_+^{(-j)}}(1-\pi)^{\beta+N_-^{(-j)}-1} d\pi \\
&= \frac{\Gamma(\alpha+\beta+N-1)}{\Gamma(\alpha+N_+^{(-j)})\Gamma(\beta+N_-^{(-j)})} \frac{\Gamma(\alpha+N_+^{(-j)}+1)\Gamma(\beta+N_-^{(-j)})}{\Gamma(\alpha+\beta+N)} \\
&= \frac{\Gamma(\alpha+\beta+N-1)}{\Gamma(\alpha+N_+^{(-j)})} \frac{\Gamma(\alpha+N_+^{(-j)}+1)}{\Gamma(\alpha+\beta+N)} \\
&= \frac{\Gamma(\alpha+\beta+N-1)}{\Gamma(\alpha+N_+^{(-j)})} \frac{\Gamma(\alpha+N_+^{(-j)}+1)}{(\alpha+\beta+N-1)\Gamma(\alpha+\beta+N-1)} \\
&= \frac{\Gamma(\alpha+N_+^{(-j)}+1)}{\Gamma(\alpha+N_+^{(-j)})(\alpha+\beta+N-1)} \\
&= \frac{(\alpha+N_+^{(-j)})\Gamma(\alpha+N_+^{(-j)})}{\Gamma(\alpha+N_+^{(-j)})(\alpha+\beta+N-1)} \\
&= \frac{\alpha+N_+^{(-j)}}{\alpha+\beta+N-1}
\end{aligned}$$

The Gibbs sampler

Initialize:

Define priors α, β, γ.

Assign initial labels $\mathbf{L}^{(0)}$ to documents

Iterate:

For each iteration $t = 1 \ldots T$:

For every document $\boldsymbol{W}_i$ (with current label $x = L_i^{(t-1)}$)

(Temporarily) remove its word counts $N_i(w_j)$ from its class x:

$N_{x\backslash i}^{(t-1)}(w_j) = N_x^{(t-1)}(w_j) - N_i^{(t-1)}(w_j)$

(Temporarily) remove $\boldsymbol{W}_i$ from the documents in its class x:

$D_{x\backslash i}^{(t-1)} = D_x^{(t-1)} - 1$

Assign a new label $x' = L_i^{(t-1)}$ to $\boldsymbol{W}_i$ with

$P(L \mid \boldsymbol{W}_i, L_0^{(t)} \ldots L_{i-1}^{(t)}, L_{i+1}^{(t-1)} \ldots L_D^{(t-1)}; \alpha, \beta, \boldsymbol{\gamma})$

Add $\boldsymbol{W}_i$ to the documents in class x'

Add its word counts $N_i(w_j)$ to word counts for class x'

Final estimate:

Use (some of the) snapshots $\mathbf{L}^{(1)} \ldots \mathbf{L}^{(T)}$ to estimate $P(+)$, $P(w_i \mid +)$, $P(w_i \mid -)$

Estimation

- Labels: $L \sim Bernoulli(\pi)$ Words: $\boldsymbol{W}_i \mid L \sim Multinomial(\boldsymbol{\theta}^L)$

	Supervised	**Unsupervised**
Freq.	**Relative frequency estimation** - Labels: $\pi = D^+/d$ - Words: $\boldsymbol{\theta}_i^+ = N^+(w_i)/N^+$	**Expectation Maximization:** At each iteration t: - Labels: $\pi^{(t)} = E[D]_{(t-1)}/d$ - Words: $\boldsymbol{\theta}_i^+ = E[N^+(w_i)]_{(t-1)}/E[N^+(w_i)]_{(t-1)}$
Bayes	**With priors:** - Labels: $\pi = (D^+ + \alpha)/(D + \alpha + \beta)$ - Words: $\boldsymbol{\theta}_i^+ = (N^+(w_i) + \gamma_i)/(N^+(w) + \gamma_0)$	**Gibbs sampling:** For each ministep i at each iteration t: - Labels: $\pi_i = (D^{+(-i)} + \alpha)/(D-1 + \alpha + \beta)$ - Words: $\boldsymbol{\theta}_i^+ = (N^{+(-i)}(w_i) + \gamma_i)/(N^{+(-i)}(w) + \gamma_0)$