

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 W_i .

The model:

- $L_i = \operatorname{argmax}_L P(L | W_i) = \operatorname{argmax}_L P(W_i | L)P(L)$
- $P(W_i | L)$ is a multinomial distribution: $W_i \sim \text{Multinomial}(\theta_L)$
We have a vocabulary of V words. Thus: $\theta_L = (\theta_1, \dots, \theta_V)$
- $P(L)$ is a Bernoulli distribution: $L \sim \text{Bernoulli}(\pi)$

Estimation

- Labels: $L \sim \text{Bernoulli}(\pi)$
- Words: $W_i | L \sim \text{Multinomial}(\theta^L)$

	Supervised	Unsupervised
Freq.	<p>Relative frequency estimation</p> <ul style="list-style-type: none"> - Labels: $\pi = D^+ / d$ - Words: $\theta_i^+ = N^+(w_i) / N^+$ 	<p>Expectation Maximization:</p> <p>At each iteration t:</p> <ul style="list-style-type: none"> - Labels: $\pi^{(t)} = E[D]_{(t-1)} / d$ - Words: $\theta_i^+ = E[N^+(w_i)]_{(t-1)} / E[N^+(w_i)]_{(t-1)}$
Bayes	<p>With priors:</p> <ul style="list-style-type: none"> - Labels: $\pi = (D^+ + \alpha) / (D + \alpha + \beta)$ - Words: $\theta_i^+ = (N^+(w_i) + \gamma_i) / (N^+(w) + \gamma_0)$ 	<p>Gibbs sampling</p>

Approximating expectations

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

Markov Chain Monte Carlo

A multivariate distribution $p(\mathbf{x}) = p(x_1, \dots, x_k)$ 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 x_j is $p(x_j)$

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 $\mathbf{x} = (x_1, \dots, x_k)$ to state $\mathbf{y} = (y_1, \dots, y_k)$

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

$$(x_1, x_2, \dots, x_i, \dots, x_{k-1}, x_k) = \mathbf{x} = \mathbf{x}^{(t)}$$

$$(y_1, x_2, \dots, x_i, \dots, x_{k-1}, x_k)$$

$$(y_1, y_2, \dots, x_i, \dots, x_{k-1}, x_k)$$

$$(y_1, y_2, \dots, x_i, \dots, x_{k-1}, x_k)$$

$$(y_1, y_2, \dots, y_i, \dots, x_{k-1}, x_k)$$

$$(y_1, y_2, \dots, y_i, \dots, x_{k-1}, x_k)$$

$$(y_1, y_2, \dots, y_i, \dots, y_{k-1}, x_k)$$

$$(y_1, y_2, \dots, y_i, \dots, y_{k-1}, y_k) = \mathbf{y} = \mathbf{x}^{(t+1)}$$

Gibbs sampling

Individual steps:

For $i = 1 \dots k$:

pick y_i by sampling from $P(Y_i | y_1, \dots, y_{i-1}, x_{i+1}, \dots, x_k)$

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

Gibbs sampling

Our states:

One label assignment L_1, \dots, L_N to each of our N documents

$$\mathbf{x} = (L_1, \dots, L_N)$$

Our transitions:

We go from one label assignment $\mathbf{x} = (+, +, -, +, - \dots +)$

to another $\mathbf{y} = (-, +, +, +, \dots, +)$

Our intermediate steps:

We generate label Y_i conditioned on $Y_1 \dots Y_{i-1}$ and $X_{i+1} \dots X_N$

Call label assignment $Y_1 \dots Y_{i-1}, X_{i+1} \dots X_N$ $\mathbf{L}^{(-i)}$

We need to compute $P(Y_i | \mathbf{D} \mathbf{L}^{(-i)}, \alpha, \beta, \gamma)$

Gibbs sampling

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

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

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

NB: π, θ_+, θ_- disappear (are integrated out)

Why we don't need to estimate π

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

The Gibbs sampler

Initialize:

Define priors α, β, γ .

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

Iterate:

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

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

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

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

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

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

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

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

Add \mathcal{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 $\mathbf{L}^{(1)} \dots \mathbf{L}^{(T)}$ to estimate $P(+), P(w_i | +), P(w_i | -)$

Estimation

- Labels: $L \sim \text{Bernoulli}(\pi)$ Words: $W_i | L \sim \text{Multinomial}(\theta^L)$

	Supervised	Unsupervised
Freq.	<p>Relative frequency estimation</p> <ul style="list-style-type: none"> - Labels: $\pi = D^+ / d$ - Words: $\theta_i^+ = N^+(w_i) / N^+$ 	<p>Expectation Maximization:</p> <p>At each iteration t:</p> <ul style="list-style-type: none"> - Labels: $\pi^{(t)} = E[D]_{(t-1)} / d$ - Words: $\theta_i^+ = E[N^+(w_i)]_{(t-1)} / E[N^+(w)]_{(t-1)}$
Bayes	<p>With priors:</p> <ul style="list-style-type: none"> - Labels: $\pi = (D^+ + \alpha) / (D + \alpha + \beta)$ - Words: $\theta_i^+ = (N^+(w_i) + \gamma_i) / (N^+(w) + \gamma_0)$ 	<p>Gibbs sampling:</p> <p>For each ministep i at each iteration t:</p> <ul style="list-style-type: none"> - Labels: $\pi_i = (D^{+(-i)} + \alpha) / (D - I + \alpha + \beta)$ - Words: $\theta_i^+ = (N^{+(-i)}(w_i) + \gamma_i) / (N^{+(-i)}(w) + \gamma_0)$