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

Lecture 4: Naive Bayes (the Frequentist approach and the Bayesian approach)

Julia Hockenmaier
juliahmr@illinois.edu
3324 Siebel Center

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

Today's class

The task: text classification (sentiment analysis)

Assign (sentiment) label $L_i \in \{+,-\}$ to a document $\boldsymbol{W}_i=(w_{i1}...w_{iN})$.

$\boldsymbol{W}_1$= *"This is an amazing product: great battery life, amazing features and it's cheap."*
$\boldsymbol{W}_2$= *"How awful. It's buggy, saps power and is way too expensive."*

The data:

A set $\boldsymbol{D}$ of N documents with (or without) labels

The model:

Naive Bayes

Comparing different estimation techniques:

- Supervised MLE
- Unsupervised MLE with EM
- Unsupervised Bayesian Estimation with Gibbs sampling
- Supervised Bayesian Estimation

The model

A Naive Bayes model

The task:

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

$\boldsymbol{W}_1$= *"This is an amazing product: great battery life, amazing features and it's cheap."*
$\boldsymbol{W}_2$= *"How awful. It's buggy, saps power and is way too expensive."*

The model:

- Use Bayes' Rule:
  $L_i = \text{argmax}_L P(L \mid \boldsymbol{W}_i) = \text{argmax}_L P(\boldsymbol{W}_i \mid L)P(L)$
- Assume $\boldsymbol{W}_i$ is a "bag of words":
  $\boldsymbol{W}_1$= *{an:1, and: 1, amazing: 2, battery: 1, cheap: 1, features: 1, great: 1,...}*
  $\boldsymbol{W}_2$= *{awful: 1, and: 1, buggy: 1, expensive: 1,...}*
- $P(\boldsymbol{W}_i \mid L)$ 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: $L \sim Bernoulli(\pi)$

Using this model

The model:

$P(\boldsymbol{W}_i \mid L)$ is a multinomial distribution: $\boldsymbol{W}_i \sim Multinomial(\boldsymbol{\theta}_L)$
$P(L)$ is a Bernoulli distribution: $L \sim Bernoulli(\pi)$

The "frequentist" approach (MLE):

Estimate $\pi, \boldsymbol{\theta}_+, \boldsymbol{\theta}_-$, then:
$P(L_i = + \mid \boldsymbol{W}_i) \propto P(\boldsymbol{W}_i \mid \boldsymbol{\theta}_+)\, \pi$

The Bayesian approach:

Choose priors for $\pi \sim Beta(\alpha,\beta)$,
$\boldsymbol{\theta}_+ \sim Dirichlet(\boldsymbol{\gamma})$, $\boldsymbol{\theta}_- \sim Dirichlet(\boldsymbol{\gamma})$ then compute the following expectation:
$P(L_i = + \mid \boldsymbol{W}_i) \propto \iint P(\boldsymbol{W}_i \mid \boldsymbol{\theta}_+)\, \boldsymbol{\theta}_+ P(\boldsymbol{\theta}_+; \boldsymbol{\gamma})\, P(\pi; \alpha,\beta)\, d\boldsymbol{\theta}_+\, d\pi$

The frequentist approach

Supervised MLE

The data is labeled:

We have a set $\mathbf{D}$ of D documents $\boldsymbol{W}_1 \ldots \boldsymbol{W}_d$ with N words
Each document W_i has N^i words
D^+ documents (subset $\mathbf{D}^+$) have a positive label and N^+ words
D^- documents (subset $\mathbf{D}^-$) have a negative label and N^- words
Each word $w_1 \ldots w_i \ldots w_V$ appears $N^+(w_i)$ times in $\mathbf{D}^+$, $N^-(w_i)$ times in $\mathbf{D}^-$
Each word $w_1 \ldots w_i \ldots w_V$ appears $N^j(w_i)$ times in D^j

MLE: relative frequency estimation

- Labels: $L \sim Bernoulli(\pi)$ with $\pi = D^+/d$
- Words: $\boldsymbol{W}_i \mid + \sim Multinomial(\boldsymbol{\theta}^+)$ with $\boldsymbol{\theta}_i^+ = N^+(w_i)/N^+$
- Words: $\boldsymbol{W}_i \mid - \sim Multinomial(\boldsymbol{\theta}^-)$ with $\boldsymbol{\theta}_i^- = N^-(w_i)/N^-$

Inference with MLE

The inference task:

Given a new document $\boldsymbol{W}_{i+1}$, what is its label L_{i+1}?

Word w_j occurs $N_{i+1}(w_j)$ times in $\boldsymbol{W}_{i+1}$.

$$
\begin{aligned}
P(L = + \mid \mathbf{W}_{i+1}) &\propto P(+)P(\mathbf{W}_{i+1} \mid +) \\
&= \pi \prod_{j=1}^{V} \theta_{+j}^{N_{i+1}(w_j)}
\end{aligned}
$$

Unsupervised MLE

The data is *un*labeled:

We have a set $\mathbf{D}$ of D documents $\boldsymbol{W}_1 \ldots \boldsymbol{W}_d$ with N words
Each document $\boldsymbol{W}_i$ has N^i words
Each word $w_1 \ldots w_i \ldots w_V$ appears $N^j(w_i)$ times in $\boldsymbol{W}_j$

EM algorithm: "expected rel. freq. estimation"

Initialization: pick initial $\pi^{(0)}, \boldsymbol{\theta}^{+(0)}, \boldsymbol{\theta}^{-(0)}$
Iterate:

- Labels: $L \sim Bernoulli(\pi)$ with $\pi^{(t)} = \langle N_+ \rangle_{(t-1)} / \langle N \rangle_{(t-1)}$
- Words: $\boldsymbol{W}_i \,|+ \sim Multinomial(\boldsymbol{\theta}^+)$ with $\theta_i^{+\,(t)} = \langle N^+(w_i) \rangle_{(t-1)} / \langle W^+ \rangle_{(t-1)}$
- Words: $\boldsymbol{W}_i \,|- \sim Multinomial(\boldsymbol{\theta}^-)$ with $\theta_i^{-\,(t)} = \langle N^-(w_i) \rangle_{(i-1)} / \langle W^- \rangle_{(i-1)}$

The Bayesian approach

The Bayesian approach

We need to compute an integral
$P(L_i = + |\, \boldsymbol{W}_i\,) \propto \iint P(\,\boldsymbol{W}_i \mid \boldsymbol{\theta}_+)\, \boldsymbol{\theta}_+ P(\boldsymbol{\theta}_+; \boldsymbol{\gamma})\, P(\pi; \alpha,\beta)\, d\boldsymbol{\theta}_+\, d\pi$

Case 1: we have labeled data

Case 2: we do not have labeled data

Bayesian: supervised

The data is labeled:

We have a set $\mathbf{D}$ of D documents $\boldsymbol{W}_1 \ldots \boldsymbol{W}_d$ with N words
Each document W_i has N^i words
D^+ documents (subset $\mathbf{D}^+$) have a positive label and N^+ words
D^- documents (subset $\mathbf{D}^-$) have a negative label and N^- words
Each word $w_1 \ldots w_i \ldots w_V$ appears $N^+(w_i)$ times in $\mathbf{D}^+$, $N^-(w_i)$ times in $\mathbf{D}^-$

Bayesian estimation

- $P(\,+) = (D^+ + \alpha)/(D + \alpha + \beta)$
- $P(w_i\,|+) = (N^+(w_i) + \gamma_i)/(N^+(w_i) + \gamma_0)$
- $P(\boldsymbol{W}_i|+) = \prod P(w_j\,|+)^{Ni(wj)}$

Bayesian: unsupervised

We need to approximate the integral/expectation:
$P(L_i = + \mid \boldsymbol{W}_i) \propto \iint P(\boldsymbol{W}_i \mid \boldsymbol{\theta}_+)\, \boldsymbol{\theta}_+ P(\boldsymbol{\theta}_+; \boldsymbol{\gamma})\, P(\pi; \alpha,\beta)\, d\boldsymbol{\theta}_+\, d\pi$

We can approximate the expectation of $f(x)$ by sampling a finite number of points $x_1 \ldots x_N$ according to $p(x)$, evaluating $f(x_i)$ for each of them, and computing the average.

How can we sample according to $p(x)$?

For us $p(x) = p(D, \boldsymbol{L}, \pi, \theta+, \theta-; \alpha, \beta, \boldsymbol{\gamma})$

Markov Chain Monte Carlo

If we had discrete distribution $p(\boldsymbol{x}) = p(x_1, \ldots, x_k)$, $p(\boldsymbol{x})$ has only a finite number of outcomes.

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

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

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

Gibbs sampling

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

That is, we will go from state $\boldsymbol{x} = (x_1, \ldots, x_k)$
to state $\boldsymbol{y} = (y_1, \ldots, y_k)$

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

$$P(Y_i = y_i \mid y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_k) = P(y_1, \ldots, y_{i-1}, y_i, x_{i+1}, \ldots, x_k)/(y_1, \ldots, y_{i-1}, x_{i+1}, \ldots, x_k)$$

Gibbs sampling

For us $p(x) = p(D, \boldsymbol{L}, \pi, \theta+, \theta-; \alpha, \beta, \boldsymbol{\gamma})$

$\pi, \theta+, \theta-$ are real-valued, but they will disappear:

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

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

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 $\boldsymbol{N}_i(w_j)$ 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

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

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

Add its word counts $\boldsymbol{N}_i(w_j)$ to word counts for class x'

Final estimate:

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

Why we don't need to estimate π

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