A Role-Reversal in the Log-Log Model

Jason D. M. Rennie
jrennie@gmail.com

May 15, 2005

1 A Change of Parameters

To this point in our discussion of the log-log model, we have largely assumed that the per-word parameter is the additive "constant", b. However, the optimization space for b is non-convex; solving for one b_i per word creates a very hard optimization problem—we may never be able to find the optimal setting of the $\{b_i\}$. So, we instead use a single b for all words and learn one a_i for each word. We can show that if b is fixed, solving for the $\{a_i\}$ is a convex optimization problem [1].

2 Modeling Frequency Counts

Let x_i be the frequency of word i in some document. Our count-based model assigns the following probability to this event:

$$P_i(x_i) = \frac{(x_i + b)^{a_i}}{Z(a_i, b)}, \tag{1}$$

where $Z(a_i, b) = \sum_{x=0}^{\infty}(x + b)^{a_i}$. The negative log-likelihood of a set of documents is simply the product of probabilities over all documents and words,

$$J_{\text{c}} = -\log \prod_j \prod_i P_i(x_{ij}) = \sum_j \sum_i [\log Z(a_i, b) - a_i \log(x_{ij} + b)]. \tag{2}$$

$\log Z$ is a convex function of a_i per our discussion in [1][1]. The second term is linear in a_i, and sums of convex functions are convex, so the entire negative log-likelihood is convex in the $\{a_i\}$.

We learn parameters via Conjugate Gradients, utilizing the objective and gradient. Let n be the number of documents and let $\hat{P}_i$ be the empirical frequency distribution of word i. Then, the partial derivatives of the objective

[1] Note that for $x + b \geq 0$, $(x + b)^{a_i} \equiv \exp(a_i \log(x + b))$.

are

$$\frac{\partial J_c}{\partial a_i} = \sum_j \left[\frac{\sum_{x=0}^{\infty} (x+b)^{a_i} \log(x+b)}{\sum_{x=0}^{\infty} (x+b)^{a_i}} - \log(x_{ij}+b) \right] \tag{3}$$

$$= n \left(E_{x\sim P_i}[\log(x+b)] - E_{x\sim \hat{P}_i}[\log(x+b)] \right), \tag{4}$$

and

$$\frac{\partial J_c}{\partial b} = \sum_j \sum_i \left[\frac{\sum_{x=0}^{\infty} (x+b)^{a_i} \frac{a_i}{x+b}}{\sum_{x=0}^{\infty} (x+b)^{a_i}} - \frac{a_i}{x_{ij}+b} \right] \tag{5}$$

$$= \sum_i n \left(E_{x\sim P_i} \left[\frac{a_i}{x+b} \right] - E_{x\sim \hat{P}_i} \left[\frac{a_i}{x+b} \right] \right). \tag{6}$$

As expected, the partial derivatives are differences between empirical and model expectations.

3 Modeling Frequency Rates

We also look at swapping the roles of the parameters in our model of frequency rates. Our rate-based model assigns the following probability to word i occuring with rate $\frac{x_i}{l}$ in a document with length l:

$$P_i\left(\frac{x_i}{l}\right) = \frac{1}{Z(a_i, b, l)} \int_{\frac{x_i}{l}}^{\frac{x_i+1}{l}} (r+b)^{a_i}\, dr, \tag{7}$$

where $Z(a_i, b, l) = \int_0^{\frac{l+1}{l}} (r+b)^{a_i}\, dr$ is the normalization constant. Working out the integrals and simplifying, we get

$$P_i\left(\frac{x_i}{l}\right) = \frac{\left(\frac{x_i}{l}+b\right)^{a_i+1} - \left(\frac{x_i+1}{l}+b\right)^{a_i+1}}{(b)^{a_i+1} - \left(\frac{l+1}{l}+b\right)^{a_i+1}} \tag{8}$$

Let l_j be the length of the j^{th} document. Define $f_{ij}(x) = \left(\frac{x}{l_j}+b\right)^{a_i+1}$. Again, the negative log-likelihood of a set of documents is simply the product of probabilities over all documents and words,

$$J_{\text{r}} = -\log \prod_j \prod_i P_i\left(\frac{x_{ij}}{l}\right) \tag{9}$$

$$= \sum_j \sum_i \Big(\log\big[f_{ij}(0) - f_{ij}(l_j+1)\big] - \log\big[f_{ij}(x_{ij}) - f_{ij}(x_{ij}+1)\big] \Big). \tag{10}$$

Again, we learn parameters via Conjugate Gradients, utilizing the objective and gradient. First we calculate partial derivatives for f_{ij},

$$\frac{\partial f_{ij}(x)}{\partial a_i} = f_{ij}(x) \log\left(\frac{x}{l_j} + b\right) \tag{11}$$

$$\frac{\partial f_{ij}(x)}{\partial b} = f_{ij}(x) \frac{a_i + 1}{\frac{x}{l_j} + b}. \tag{12}$$

Then, the negative log-likelihood partial derivatives are

$$\frac{\partial J_r}{\partial a_i} = \sum_j \left(\frac{\frac{\partial f_{ij}(0)}{\partial a_i} - \frac{\partial f_{ij}(l+1)}{\partial a_i}}{f_{ij}(0) - f_{ij}(l+1)} - \frac{\frac{\partial f_{ij}(x_{ij})}{\partial a_i} - \frac{\partial f_{ij}(x_{ij}+1)}{\partial a_i}}{f_{ij}(x_{ij}) - f_{ij}(x_{ij}+1)} \right) \tag{13}$$

$$\frac{\partial J_r}{\partial b} = \sum_j \sum_i \left(\frac{\frac{\partial f_{ij}(0)}{\partial b} - \frac{\partial f_{ij}(l+1)}{\partial b}}{f_{ij}(0) - f_{ij}(l+1)} - \frac{\frac{\partial f_{ij}(x_{ij})}{\partial b} - \frac{\partial f_{ij}(x_{ij}+1)}{\partial b}}{f_{ij}(x_{ij}) - f_{ij}(x_{ij}+1)} \right) \tag{14}$$

References

[1] J. D. M. Rennie. A class of convex functions. http://people.csail.mit.edu/~jrennie/writing, May 2005.