Learning a Log-Log Term Frequency Model

Jason D. M. Rennie
jrennie@gmail.com

May 2, 2005*

1 The Log-Log Model Revisited

Earlier [1], we defined our term frequency distribution as

$$P(x) = \frac{(x+b)^a}{Z(a,b)}, \tag{1}$$

for $x \in \{0, 1, 2, \dots\}$, where the normalization constant is $Z(a,b) = \sum_{x=0}^{\infty}(x+b)^a$. Note that the noramlization constant is finite only when $a < -1$ and $b > 0$. We also note that $Z(-a, 1) = \zeta(a)$ is the Riemann Zeta Function. Generally speaking, we cannot compute the normalization constant exactly. However, an excellent approximation can be achieved using partial sums. Let $S_n = \sum_{x=0}^{n}(x+b)^a$. We fit the function $f(x) = S_{1/x}$ for $x \in \{1/1, 1/2, \dots, 1/n, \dots\}$, and use the fit at $x = 0$ as our approximation for Z.

As noted, P may not be a distribution if $a \geq -1$ or $b \leq 0$. Additionally, the expectation of P is not finite if $a \geq -2$. Without constraints or reparameterization, our optimization routine may try to evaluate such out-of-bounds values. We choose to reparameterize so that any finite values of the parameters yield a distribution with finite expectation. We reparameterize b as e^β and a as $-e^\alpha - 2$. Thus, we have $P(x) = (x + e^\beta)^{-e^\alpha - 2}/Z(\alpha, \beta)$ and $Z(\alpha, \beta) = \sum_x (x + e^\beta)^{-e^\alpha - 2}$.

Now, consider the problem of fitting the distribution given a set of documents. Let the distribution for word i be $P_i(x) \propto (x + e^{\beta_i})^{-e^\alpha - 2}$. We learn a separate β_i for each word and a single α for the entire data set. Let x_{ij} be the frequency of word i in document j. Let n be the number of documents; let d be the vocabulary size. Then, the data negative log-likelihood is

$$J = -\log P(D) = \sum_{i=1}^{d} \sum_{j=1}^{n} (e^\alpha + 2) \log(x_{ij} + e^{\beta_i}) + \sum_{i=1}^{d} n \log\left(\sum_{x=0}^{\infty} (x + e^{\beta_i})^{-e^\alpha - 2}\right). \tag{2}$$

*Update June 16, 2005

We can use gradient-descent-type techniques to learn the parameters. Note that when we are assessing the probability of frequencies in a given document, there is no need for us to be concerned with document length. Our document likelihood is a joint probability that implicitly includes the event of the documents having certain lengths. It is computationally difficult to calculate the chance of observing a document of a given length. However, we find no need to calculate this probability. Note that for other models, such as the unigram/multinomial, it is also computationally difficult to calculate this marginal probability.

2 Optimization

Here we discuss the task of learning parameters of our log-log term frequency model. As mentioned earlier, we use gradient-descent-type techniques. In particular, we focus on techniques that require only objective and gradient information. We have already discussed the objective (2) and the approximation necessary to calculate the normalization constant. Here we calculate the gradient. Define $\hat{P}_i$ as the empirical distribution of frequencies for word i. Recall that n is the number of documents. The gradient takes the usual form, the difference between the expectations of the estimated and empirical distributions.

$$Z = \sum_{x=0}^{\infty} (x + e^{\beta_i})^{-e^{\alpha}-2} = \sum_{x=0}^{\infty} \exp\left[-(e^{\alpha}+2)\log(x+e^{\beta_i})\right] \tag{3}$$

$$\frac{\partial Z}{\partial \alpha} = -\sum_{x=0}^{\infty} (x + e^{\beta_i})^{-e^{\alpha}-2} e^{\alpha} \log(x + e^{\beta_i}) \tag{4}$$

$$\frac{\partial Z}{\partial \beta_i} = -\sum_{x=0}^{\infty} (x + e^{\beta_i})^{-e^{\alpha}-2} \frac{(e^{\alpha}+2)e^{\beta_i}}{x + e^{\beta_i}} \tag{5}$$

$$\frac{\partial J}{\partial \alpha} = \sum_i \sum_j e^{\alpha} \log(x_{ij} + e^{\beta_i}) - n \sum_i \frac{\sum_x (x + e^{\beta_i})^{-e^{\alpha}-2} e^{\alpha} \log(x + e^{\beta_i})}{\sum_{x'} (x' + e^{\beta_i})^{-e^{\alpha}-2}} \tag{6}$$

$$= ne^{\alpha} \sum_i \left(E_{\hat{P}_i(x)}[\log(x + e^{\beta_i})] - E_{P_i(x)}\left[\log(x + e^{\beta_i})\right]\right) \tag{7}$$

$$\frac{\partial J}{\partial \beta_i} = \sum_j \frac{(e^{\alpha}+2)e^{\beta_i}}{x_{ij} + e^{\beta_i}} - n \frac{\sum_x (x + e^{\beta_i})^{-e^{\alpha}-2} \frac{(e^{\alpha}+2)e^{\beta_i}}{x+e^{\beta_i}}}{\sum_{x'} (x' + e^{\beta_i})^{-e^{\alpha}-2}} \tag{8}$$

$$= n(e^{\alpha}+2) \left(E_{\hat{P}_i(x)}\left[\frac{e^{\beta_i}}{x + e^{\beta_i}}\right] - E_{P_i(x)}\left[\frac{e^{\beta_i}}{x + e^{\beta_i}}\right]\right) \tag{9}$$

Again, we must use approximations to compute these sums. We can use the partial-sum/regression technique discussed earlier.

3 A Length-Conditional Version

So far, we have described a distribution that is not conditioned on the length of the docuemnt. However, it is common for models to condition on document length. The unigram model is a prime example. So, here we describe a length-conditioned version of the log-log term frequency model.

Conditioning on length involes only a minor change in the math. Instead of summing to infinity for the normalization constant, we sum to the length of the document. Since the log-log model is heavy-tailed, this might yield a non-trivial improvement in data likelihood. We use parameters a and $\{b_i\}$ (avoiding the messy reparameterizations); our data negative log-likelihood is

$$J = \sum_{i=1}^{d} \sum_{j=1}^{n} \log \left(\sum_{x=0}^{l_j} (x+b_i)^a \right) - \sum_{i=1}^{d} \sum_{j=1}^{n} a \log(x_{ij} + b_i), \tag{10}$$

where $l_j = \sum_i x_{ij}$ is the length of document j. Note that the normailization term now depends on the document.

The partial derivatives are

$$\begin{aligned} \frac{\partial J}{\partial a} &= \sum_{i=1}^{d} \sum_{j=1}^{n} \frac{\sum_{x=0}^{l_j} (x+b_i)^a \log(x+b_i)}{\sum_{x=0}^{l_j} (x+b_i)^a} - \sum_{i=1}^{d} \sum_{j=1}^{n} \log(x_{ij} + b_i) && (11) \\ &= \sum_{i=1}^{d} \sum_{j=1}^{n} E_{P_{ij}}[\log(x+b_i)] - E_{\hat{P}_{ij}}[\log(x+b_i)], && (12) \end{aligned}$$

$$\begin{aligned} \frac{\partial J}{\partial b_i} &= \sum_{j=1}^{n} \frac{\sum_{x=0}^{l_j} (x+b_i)^a \frac{a}{x+b_i}}{\sum_{x=0}^{l_j} (x+b_i)^a} - \sum_{j=1}^{n} \frac{a}{x_{ij} + b_i} && (13) \\ &= \sum_{j=1}^{n} E_{P_{ij}} \left[\frac{a}{x+b_i} \right] - E_{\hat{P}_{ij}} \left[\frac{a}{x+b_i} \right], && (14) \end{aligned}$$

References

[1] J. D. M. Rennie. A better model for term frequencies. http://people.csail.mit.edu/~jrennie/writing, April 2005.