Using a Log-Log Distribution to Model Term Frequency Rates

Jason D. M. Rennie
jrennie@gmail.com

May 8, 2005*

In [1], we outlined a basic architecture for learning a heavy-tailed distribution for term frequencies. Although this model can deal with documents of different lengths, its design is far from optimal in this respect since document length is not captured as part of the model. Here we discuss an alternate model that utilizes the log-log distribution, but uses it to model frequency rates rather than raw frequency values.

Recall the log-log distribution,

$$P(x) = \frac{(x + e^\beta)^{-e^\alpha - 2}}{Z(\alpha, \beta)}. \tag{1}$$

Earlier, we limited $x \in \{0, 1, 2, \dots\}$. Here, we consider the frequency rate of a word in the context of a document of a certain length. Let l be the length of the document. Then, the rate, $r = \frac{x}{l}$ can take on values $r \in \{0, \frac{1}{l}, \frac{2}{l}, \dots, 1\}$. We say that the chance that r takes on the value $r = \frac{x}{l}$ ($x \in \{0, \dots, l\}$) is

$$P\left(\frac{x}{l}\right) = \frac{1}{Z(\alpha, \beta, l)} \int_{\frac{x}{l}}^{\frac{x+1}{l}} (r + e^\beta)^{-e^\alpha - 2} dr, \tag{2}$$

where $Z(\alpha, \beta, l) = \int_0^{\frac{l+1}{l}} (r + e^\beta)^{-e^\alpha - 2} dr$. Performing the integration and simplifying, we get

$$P\left(\frac{x}{l}; \alpha, \beta\right) = \frac{\left(\frac{x}{l} + e^\beta\right)^{-e^\alpha - 1} - \left(\frac{x+1}{l} + e^\beta\right)^{-e^\alpha - 1}}{\left(e^\beta\right)^{-e^\alpha - 1} - \left(\frac{l+1}{l} + e^\beta\right)^{-e^\alpha - 1}}. \tag{3}$$

As before, we learn a single α and a separate β_i for each word and select parameters by maximizing likelihood of the data. Define $f_i(x, l) = \left(\frac{x}{l} + e^{\beta_i}\right)^{-e^\alpha - 1}$.

*Updated May 15, 2005

The negative log-likelihood for a single document is

$$J = -\log P(D) = -\sum_i \log P\left(\frac{x}{l}; \alpha, \beta_i\right) = \sum_i \Big(\log[f_i(0,l) - f_i(l+1,l)] - \log[f_i(x_i,l) - f_i(x_i+1,l)] \Big). \quad (4)$$

We learn parameters for the model by minimizing the negative log-likelihood via Conjugate Gradients (CG). For CG, we need to be able to calculate the gradient. Note that

$$\frac{\partial f_i(x,l)}{\partial \alpha} = -f_i(x,l) e^{\alpha} \log\left(\frac{x}{l} + e^{\beta_i}\right), \quad (5)$$

$$\frac{\partial f_i(x,l)}{\partial \beta_i} = -f_i(x,l) \frac{e^{\alpha}+1}{\frac{x}{l} + e^{\beta_i}} e^{\beta_i}. \quad (6)$$

The partial derivatives are

$$\frac{\partial J}{\partial \alpha} = \sum_i \left(\frac{\frac{\partial f_i(0,l)}{\partial \alpha} - \frac{\partial f_i(l+1,l)}{\partial \alpha}}{f_i(0,l) - f_i(l+1,l)} - \frac{\frac{\partial f_i(x_i,l)}{\partial \alpha} - \frac{\partial f_i(x_i+1,l)}{\partial \alpha}}{f_i(x_i,l) - f_i(x_i+1,l)} \right), \quad (7)$$

$$\frac{\partial J}{\partial \beta_i} = \frac{\frac{\partial f_i(0,l)}{\partial \beta_i} - \frac{\partial f_i(l+1,l)}{\partial \beta_i}}{f_i(0,l) - f_i(l+1,l)} - \frac{\frac{\partial f_i(x_i,l)}{\partial \beta_i} - \frac{\partial f_i(x_i+1,l)}{\partial \beta_i}}{f_i(x_i,l) - f_i(x_i+1,l)}. \quad (8)$$

References

[1] J. D. M. Rennie. Learning a log-log term frequency model. http://people.csail.mit.edu/∼jrennie/writing, May 2005.