Summary of “Producing Power-Law Distributions and Damping Word Frequencies with Two-Stage Language Models”

Matt Dickenson `mcd31`
STA571/CS590.01 Due: 13 January, 2014

This paper introduces a two-stage framework that can produce power law distributions, which often appear when analyzing the frequencies of word tokens in natural languages. The generator (first stage) is generic and can be any “standard” probabilistic model. Then, the adaptor (second stage) transforms the frequencies so that they resemble observed word frequencies in natural languages. The role of the adaptor is to “damp” the frequencies to improve estimation of the generator parameters.

The paper explores two adaptors in particular–the Chinese restaurant process (CRP) and the Pitman-Yor generalization (PYCRP). Both of these are “rich-get-richer” processes used in nonparametric Bayesian statistics. The paper demonstrates that a TwoStage(CRP(α), P_φ) model is equivalent to a (Dirichlet-Process) DP(α, P_φ) model, and a TwoStage(PYCRP(a, b), P_φ) model is equivalent to a (Pitman-Yor) PYP(a, b, P_φ) (where P_φ has infinite support in both cases).

Both DP and PYP can be understood as the results of a stick-breaking process, where the DP has one parameter and the PYP has two. One important difference in the two processes is that the CRP (analogous to DP) treats each word type independently, ignoring dependencies in the generator, whereas the PYCRP does not. Another comparison is how each adaptor damps the frequencies: the CRP can be used to estimate the log transformed frequencies, while the PYCRP helps to estimate the inverse-power transformed frequencies.

Overall, the PYP is more flexible than the DP. In particular, it is more flexible in the tails. PYP is preferred over the DP when the word frequencies appear (or are expected) to have a power law distribution. The DP would be more appropriate when the word frequencies exhibit an exponential distribution. Because natural language word frequencies appear to be distributed according to a power law, the PYCRP is a useful adaptor. Using the framework proposed by the authors, word frequencies can be damped to help estimate parameters of the generator function.