PoCS What's The Story?

Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2015
Assignment 5 • code name: The balance of words

Dispersed: Saturday, October 3, 2015.
Due: 5 pm, Monday, October 12, 2015.
Some useful reminders:
Deliverator: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds@uvm.edu
Office hours: 10:30 am to 11:30 am on Mondays, and 2:00 pm to 3:30 pm Wednesdays at Farrell Hall
Course website: http://www.uvm.edu/~pdodds/teaching/courses/2015-08UVM-300

All parts are worth 3 points unless marked otherwise. Please show all your workingses clearly and list the names of others with whom you collaborated.

Graduate students are requested to use LaTeX (or related TeX variant).

Email submission: PDF only! Please name your file as follows (where the number is to be padded by a 0 if less than 10 and names are all lowercase):
`CSYS300assignment[%02d]-[firstname]-[lastname].pdf`

---

1. (6 pts) **Generalized entropy and diversity:**

   For a probability distribution of $i = 1, \ldots, n$ entities with the ith entity having probability of being observed p_i, Shannon's entropy is defined as [2]: $H = -\sum_{i=1}^{n} p_i \ln p_i$. There are other kinds of entropies and we'll explore some aspects of them here.

   Let's use the setting of words in a text (another meaningful framing is abundance of species in an ecology). So we have word i appearing with probability p_i and there are n words.

   Now, a useful quantity associated with any kind of entropy is diversity, D [1]. Given a text T with entropy H, we define D to be the number of words in another hypothetical text T' which (1) has the same entropy, and (2) where all words appear with equal frequency $1/D$. In text T', we have $p_i = 1/D$ for $i = 1, \ldots, D$.

   Diversity is thus a number, and behaves in number-like ways that are more intuitive to grasp than entropy. (Entropy is still the primary thing here.)

   Determine the diversity D in terms of the probabilities $\{p_i\}$ for the following:

(a) Simpson concentration:

$$S = \sum_{i=1}^{n} p_i^2.$$

(b) Gini index:

$$G \equiv 1 - S = 1 - \sum_{i=1}^{n} p_i^2.$$

Please note any connections between diversity for the Simpson and Gini indices.

(c) Shannon's entropy:

$$H = -\sum_{i=1}^{n} p_i \ln p_i.$$

(d) Renyi entropy:

$$H_q^{(\mathrm{R})} = \frac{1}{q-1}\left(-\ln \sum_{i=1}^{n} p_i^q\right),$$

where $q \neq 1$.

(e) The generalized Tsallis entropy:

$$H_q^{(\mathrm{T})} = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right),$$

where $q \neq 1$.

Please note any connections between diversity for Renyi and Tsallis.

(f) Show that in the limit $q \to 1$, the diversity for the Tsallis entropy matches up with that of Shannon's entropy.

2. (3 + 3 points) *Zipfarama via Optimization:*

Complete the Mandelbrotian derivation of Zipf's law by minimizing the function

$$\Psi(p_1, p_2, \ldots, p_n) = F(p_1, p_2, \ldots, p_n) + \lambda G(p_1, p_2, \ldots, p_n)$$

where the 'cost over information' function is

$$F(p_1, p_2, \ldots, p_n) = \frac{C}{H} = \frac{\sum_{i=1}^{n} p_i \ln(i+a)}{-g \sum_{i=1}^{n} p_i \ln p_i}$$

and the constraint function is

$$G(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i - 1 \quad (= 0)$$

to find

$$p_j = e^{-1-\lambda H^2/gC}(j+a)^{-H/gC}.$$

Then use the constraint equation, $\sum_{j=1}^{n} p_j = 1$ to show that

$$p_j = (j+a)^{-\alpha}.$$

where $\alpha = H/gC$.

3 points: When finding λ, find an expression connecting λ, g, C, and H.

Hint: one way may be to substitute the form you find for $\ln p_i$ into H's definition (but do not replace p_i).

Note: We have now allowed the cost factor to be $(j+a)$ rather than $(j+1)$.

3. (3 + 3)

   (a) For $n \to \infty$, use some computation tool (e.g., Matlab, an abacus, but not a clever friend who's really into computers) to determine that $\alpha \simeq 1.73$ for $a = 1$. (Recall: we expect $\alpha < 1$ for $\gamma > 2$)

   (b) For finite n, find an approximate estimate of a in terms of n that yields $\alpha = 1$.

   (Hint: use an integral approximation for the relevant sum.)

   What happens to a as $n \to \infty$?

References

[1] L. Jost. Entropy and diversity. *Oikos*, 113:363–375, 2006. pdf

[2] C. E. Shannon. A mathematical theory of communication. *The Bell System Tech. J.*, 27:379–423,623–656, 1948. pdf