CSYS 300 PoCS Assignment 5

Thomas Waters

October 5th 2018

Code is located at https://github.com/Evelios/PoCS_Assignment_05

1 Problem 1

This problem is about evaluating a diversity index D, which is easier to intuit about than the entropy number. This diversity index is supposed to represent a different hypothetical text T' which shares the same entropy H as the origional text T. The text T' has all the words appearing in equal frequency $1/D$, so there is $p_i = 1/D$ for $i = 1, ..., D$. This means that we set $H = H'$, substituting the p_i from T with $p'_i = 1/D$ in H' from T'. Also remembering that the sum of the text is from $i = 1$ to D.

Simpson Concentration

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

$$S = S' \tag{2}$$

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{D} (1/D)^2 \tag{3}$$

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{D} (1/D)^2 \tag{4}$$

$$\sum_{i=1}^{n} p_i^2 = D(1/D)^2 \tag{5}$$

$$\sum_{i=1}^{n} p_i^2 = (1/D) \tag{6}$$

$$D = \left(\sum_{i=1}^{n} p_i^2 \right)^{-1} = S^{-1} \tag{7}$$

Gini Index

$$G \equiv 1 - S = 1 - \sum_{i=1}^{n} p_i^2 \tag{8}$$

$$G = G' \tag{9}$$

$$1 - \sum_{i=1}^{n} p_i^2 = 1 - \sum_{i=1}^{n} (1/D)^2 \tag{10}$$

$$\sum_{i=1}^{n} p_i^2 = \sum_{i=1}^{D} (1/D)^2 \tag{11}$$

$$\sum_{i=1}^{n} p_i^2 = D(1/D)^2 \tag{12}$$

$$\sum_{i=1}^{n} p_i^2 = (1/D) \tag{13}$$

$$D = \left(\sum_{i=1}^{n} p_i^2 \right)^{-1} = S^{-1} \tag{14}$$

The diversity index for the Simpson Concentration and the Gini Index come out to be the same! The diversity index for the Gini Index can even be written in terms of the Simpson Concentration.

Shannon's Entropy

$$H = -\sum_{i=1}^{n} p_i \ln p_i \tag{15}$$

$$H = H' \tag{16}$$

$$H = -\sum_{i=1}^{D} (1/D) \ln(1/D) \tag{17}$$

$$H = -D(1/D) \ln(1/D) \tag{18}$$

$$H = -\ln(1/D) = \ln D \tag{19}$$

$$D = e^H = e^{-\sum_{i=1}^{n} p_i \ln p_i} \tag{20}$$

Renya Entropy, $q \neq 1$

$$H_q^{(R)} = \frac{1}{q-1}\left(-\ln \sum_{i=1}^{n} p_i^q\right) \tag{21}$$

$$H_q^{(R)} = (H_q^{(R)})'$$

$$\frac{1}{q-1}\left(-\ln \sum_{i=1}^{n} p_i^q\right) = \frac{1}{q-1}\left(-\ln \sum_{i=1}^{D} (1/D)^q\right) \tag{22}$$

$$-\ln \sum_{i=1}^{n} p_i^q = -\ln \sum_{i=1}^{D} (1/D)^q \tag{23}$$

$$\sum_{i=1}^{n} p_i^q = \sum_{i=1}^{D} (1/D)^q \tag{24}$$

$$\sum_{i=1}^{n} p_i^q = D(D)^{-q} \tag{25}$$

$$\sum_{i=1}^{n} p_i^q = D^{q+1} \tag{26}$$

$$D = \left(\sum_{i=1}^{n} p_i^q\right)^{\frac{1}{q+1}} \tag{27}$$

Generalized Tsallis Entropy, $q \neq 1$

$$H_q^{(T)} = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right) \tag{28}$$

$$H_q^{(T)} = (H_q^{(T)})'$$

$$\frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{D} (1/D)^q\right) \tag{29}$$

$$1 - \sum_{i=1}^{n} p_i^q = 1 - \sum_{i=1}^{D} (1/D)^q \tag{30}$$

$$\sum_{i=1}^{n} p_i^q = \sum_{i=1}^{D} (1/D)^q \tag{31}$$

$$\sum_{i=1}^{n} p_i^q = \sum_{i=1}^{D} (1/D)^q \tag{32}$$

$$\sum_{i=1}^{n} p_i^q = D(D)^{-q} \tag{33}$$

$$\sum_{i=1}^{n} p_i^q = D^{q+1} \tag{34}$$

$$D = \left(\sum_{i=1}^{n} p_i^q \right)^{\frac{1}{q+1}} \tag{35}$$

We see that the Generalized Tallis Entropy and the Renya Entropy come out to be the same!

2 Problem 2

Mandelbrotian derivation of Zipf's law by minimizing the function

$$\Psi(p_1, p_2, ..., p_n) = F(p_1, p_2, ..., p_n) + \lambda G(p_1, p_2, ..., p_n) \tag{36}$$

The cost of information is,

$$F(p_1, p_2, ..., p_n) = \frac{C}{H} = \frac{\sum_{i=1}^{n} p_i ln(i+a)}{-g \sum_{i=1}^{n} p_i \ln p_i} \tag{37}$$

to find

$$p_j = e^{-1-\lambda H^2/gC} (j+a)^{-H/gC} \tag{38}$$

Then use the equation

$$\sum_{j=1}^{n} p_j = 1 \tag{39}$$

To find

$$\alpha = H/gC \tag{40}$$

$$p_j = (j+a)^{\alpha} \tag{41}$$

3 Problem 3

4 Problem 4