Jaeho, Chang

Date: May 11, 2020

G-test, Information Entropy, p-value

==============================

Suppose that we have an entire dataset of size $d : D_1, D_2, \cdots, D_d$. Through tokenization, we got tokens for each data as t_{ij} where $j = 1, \cdots, d$ and $i = 1, \cdots, n_j$.

Then, there are T unique tokens $t_1^*, t_2^*, \cdots, t_T^*$ and their counts $O_1, O_2, \cdots, O_T$. Each count O_k is defined as $\sum_{j=1}^d \sum_{i=1}^{n_j} 1(t_k^* = t_{ij})$ and denote N by the sum of counts $O_1, O_2, \cdots, O_T$.

Token_Count	Token_Value	Token_Type
O_1^{plain}	t_1^*	plain
O_1^{sqli}	t_1^*	sqli
O_2^{plain}	t_2^*	plain
$\vdots$	$\vdots$	$\vdots$

If we convert this dataset to the crosstable, we get:

token\type	plain	sqli	Total
t_1^*	O_1^{plain}	O_1^{sqli}	O_1
t_2^*	O_2^{plain}	O_2^{sqli}	O_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_T^*	O_T^{plain}	O_T^{sqli}	O_T
Total	N^{plain}	N^{sqli}	N

$\textbf{G-test score}$ of a token t_k^* with label $j \in \{\text{plain, sqli}\}$ is defined as:

$$G_{kj} = 2 \cdot O_{kj} \ln\left(\frac{O_{kj}}{E_{kj}}\right)$$

where E_{kj} is the expected count under multinomial distribution, that is, calculated by

$$E_{kj} = N_j \cdot \frac{O_k}{N}$$

by the Maximum Likelihood estimation. Here, note that $\sum_{k=1}^T E_{kj} = N_j$.

$\textbf{Shannon Entropy}$ of a token t_k^* is calculated as:

$$H(t_k^*) = -\sum_{j \in \{\text{plain, sqli}\}} \hat{\pi}_{kj} \log_2 \hat{\pi}_{kj}$$

where $\hat{\pi}_{kj} = O_{kj}/O_k$.

It is well known that the asymptotic distribution of the log-likelihood ratio test statistics approaches to χ^2 distribution. Let's denote the log-likelihood at $\theta = \hat{\theta}$ as $l(\hat{\theta})$ where $\hat{\theta}$ is an MLE of $\theta \in \Theta \subset \mathbb{R}^k$. Suppose that n samples were drawn from an identical distribution independently. Then,

$$l(\hat{\theta}) \approx l(\theta_0) + (\hat{\theta} - \theta_0)' \nabla^2 l(\theta_0)(\hat{\theta} - \theta_0)/2$$
$$\therefore\ 2[l(\hat{\theta}) - l(\theta_0)] \approx n(\hat{\theta} - \theta_0)' \nabla^2 l_1(\theta_0)(\hat{\theta} - \theta_0).$$

Since $-\sqrt{n} I_1(\hat{\theta})^{-\frac{1}{2}}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}_k(0, I_k)$ and $\left\{I_1(\hat{\theta})\right\}^{-1} \nabla^2 l_1(\theta_0) \to 1$ as $n \to \infty$,

$$2[l(\hat{\theta}) - l(\theta_0)] \xrightarrow{d} \chi_k^2.$$

Therefore, we can interpret the p-value of the observed G-test defined as

$$G_o = 2 \cdot \sum_{k,j} O_{kj} \ln\left(\frac{O_{kj}}{E_{kj}}\right)$$

as

$$P(|G| > G_o | Multinomial) \approx 2 \cdot \min\left[P(\chi_{T-1}^2 > G_o), P(\chi_{T-1}^2 < G_o)\right].$$

Note that we estimate $T - 1$ parameters because $\sum_{i=1}^T \theta_i = 1$.