Documentation

Jerry Wang

Contents

1 Introduction

1.1 Entropy - Surprise Factor

First we would like to define the concept of the "surprise" of an event E occurring and define entropy based off that. We deem this important since entropy plays a large role in many of the metric developed and implemented. As such, it would be helpful to define a consistent mathematical framework as well as develop intuition for these definitions at a precise mathematical level.

The following methods and definitions are based on Ross's A First Course in Probability [1]. Mathematically, it makes sense that the surprise invoked by an event E occurring should be a function of the probability of event E itself, which we will denote p. Thus we define $S(p)$ as the surprise invoked by an event with a probability of p. Now we will begin by stating some axioms for this definition of surprise.

Axiom 1

$$S(1) = 0$$

Intuitively, that just means we should feel no surprise when a event with probability 1 occurs. That is, it is not surprising at all for a sure event to occur.

Axiom 2

$$p < q \implies S(p) < S(q)$$

That is, $S(p)$ is a strictly decreasing function of p.

Axiom 3

$$S(p) \text{ is a continuous function with respect to } p$$

Axiom 4

$$S(pq) = S(p) + S(q)$$

The intuition of this axiom is given when we consider independent events E_1 and E_2 with probabilities of occurring with p and q respectively. Since E_1 and E_2 are independent, $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2) = pq$. Thus the surprise invoked by $E_1 \cap E_2$ should be defined as $S(pq)$. But now, let us consider that we are first told that event E_1 occurred, and then afterwards event E_2 occur. We know the total surprise invoked is simply the surprised invoked by $E_1 \cap E_2$ which is $S(pq)$, and since we know that $S(p)$ is the surprise invoked by event E_1 alone, it follows that $S(pq) - S(p)$ should represent the initial surprise invoked by E_2. Due to independence, we know the probability of E_2 is still q and thus the surprise of event E_2 should remain $S(q)$. Thus we have that $S(pq) - S(p) = S(q)$ or that $S(pq) = S(p) + S(q)$.

From the axioms, we can prove that

$$S(p) = -C \log_2 p$$

where $C > 0$. From Axiom 4, we have that

$$S(p^2) = S(p) + S(p) + 2S(p)$$

From induction, we also have that for $m \in \mathbb{Z} > 0$

$$S(p^m) = mS(p) \tag{1}$$

Also note that for $n \in \mathbb{Z} > 0$

$$S(p) = S\left((p^{\frac{1}{n}})^n\right) = nS(p^{\frac{1}{n}})$$

which implies that

$$S(p^{\frac{1}{n}}) = \frac{S(p)}{n} \tag{2}$$

Combining equations 1 and 2, we can define

$$S(p^x) = xS(p) \tag{3}$$

for $x \in \mathbb{Q} > 0$. Now from Axiom 3, we can define equation 3 $\forall x \in \mathbb{Q} > 0$.

Now let us take $x = -\log_2 p$ which implies $p = \left(\frac{1}{2}\right)^x$ which allows for the following relation.

$$S(p) = S\left((\frac{1}{2})^x\right) = xS\left(\frac{1}{2}\right) = -S\left(\frac{1}{2}\right) \cdot \log_2 p$$

Now notice $S\left(\frac{1}{2}\right) = C$ is simply a constant (that is non-zero thanks to Axiom 1 and Axiom 2).

Thus we have shown that

$$S(p) = -C\log_2 p \tag{4}$$

follows from the axioms. Lastly, note that it is standard to let $C = 1$.

2 Mathematical Metrics

2.1 Shannon Entropy Metric

Now we aim to quantify the expected amount of surprise incurred by a random variable X. Since $\log_2 p_i$ is the value of the surprise invoked when an event (let's say E_i) with probability p_i occurs, the expected value of $S(p_i)$ for any outcome of E_i is $p_i \cdot \log_2 p_i$. Now summing over all expectations, we have the expected surprise of a random variable X which takes on n values with probability $p_1, p_2, ...p_n$ is

$$H(x) = \sum_{i=1}^{n} p_i \log_2 p_i$$

For us, we take the random variable x to be all the characters in the password.

2.1.1 Shannon Entropy Ratio

It can be proven using Lagrange Multipliers that the distribution that maximizes entropy is a discrete uniform distribution of probabilities.

We will use the ratio of the Shannon Entropy to the maximum possible Shannon Entropy as our metric to analyze how varied the characters are in the password.

$$R = \frac{H(x,n)}{H_{\max}(n)}$$

2.2 Password Entropy Metric

Password Entropy is defined as

$$E = \log_2(R^L) = L\log_2(R)$$

where R is the character pool and L is the length of the password. This will be used to measure the total amount of possible passwords of the given length. Under the assumption that passwords are guessed randomly, the password entropy corresponds to the amount of total passwords that are possible. For this model, R is based on the amount of character sets (i.e: lowercase letters, capital letters, special symbols, etc...).

2.3 Huffman Encoding

Finally, a metric needs to be used to consider the rarity of letters used in the password. The Huffman encoding is often used for this, which is an algorithm for converting letters into strings of binary digits such that more common letters are assigned shorter strings. For example, the string “Hello world” could be encoded to:

Letter	code
l	000
o	001
e	01
t	100
r	101
h	11

The length of the Huffman encoding could be considered a measure of surprise of each individual character treated independently - in particular, if the following formula is used, where w is the string inputted, and n is its length, $\mathcal{H}(w,n)$ satisfies the surprise axioms. Certainly, $\mathcal{H}(a,n) = 0$ if a only consists of a single character, because no information need be given to fully specify the string and thus the Huffman encoding is an empty string, and if only a single character exists then a string containing that character will have probability 1. A useful feature of the Huffman encoding is that, if any individual character is rarer than any other, the length of its Huffman encoding will be longer, which means that Axiom 2 is satisfied. Since the probabilities of letters are discrete, axiom 3 is trivially satisfied. Finally, for any two words w_1 and w_2 (with lengths n_1 and n_2), the probability of both of them occurring in a string of length $n_1 + n_2$ is equal to each individual probability multiplied. However, the Huffman encoding for each will be simply the length of the two strings concatenated. So, $\mathcal{H}(w_1w_2, n_1+n_2) = \mathcal{H}(w_1,n_1) + \mathcal{H}(w_2,n_2)$

To measure the password strength, we created a Huffman encoding tree based on the NCSC list of 100,000 most commonly used passwords. Then, for each inputted password, we measured the length of its Huffman encoding. So, passwords that contain more unexpected characters will have a higher value from the Huffman encoding. For the application of our website, we used a slightly modified version of length so that reasonable passwords are scaled between a value of 0 and 1. Here $\mathcal{H}_{min}(n)$ represents the Huffman encoding if only the minimum-value characters are used and $\mathcal{H}_{ascii}(n)$ represents the length of the default binary encoding for ascii characters, which is equal to $8n$. Such a value is used as our "maximum" value. So, if the Huffman encoding of the password based on the tree generated from common passwords improves on the ascii encoding, then the password's score will be low.

$$\mathcal{H}_{mod}(w,n) = \frac{\mathcal{H}(w,n) - \mathcal{H}_{min}(n)}{\mathcal{H}_{ascii}(n) - \mathcal{H}_{min}(n)}$$

3 Commonality Metric

Purely mathematical models based on information theory alone may not be able fully quantify the strength of a password. Human tendencies about known popular, and thus vulnerable passwords, must also be considered. Sometimes passwords may score highly on many of the mathematical metrics, but in reality, may be poor passwords due to how prevalent they are. Consider for example passwords like `qazwsx#123` which scores quite highly among all the mathematical metrics, but is probably in reality a poor password due to how similar they are to known passwords.

3.1 Commonality Through Sequence Alignment

To quantify a metric for commonality, a Dynamic Programming approach for the Sequence Alignment Problem as implemented. It scans through a list of 5000 known common passwords and seeks similarities between the given passwords and the common passwords.

4 Combining of the Metrics

The metrics described above were combined via a weighted geometric mean. A geometric mean was chosen to better reflect an accurate combination of the metrics. This choice was made so that the overall password metric would be sensitive to low values in any of the metrics, as a password that scores extremely poorly in any given metric should not score highly overall. A weighted geometric mean is given as follows

$$\bar{x} = \left(\prod_{i=1}^{n} x_i^{w_i} \right)^{1 / \sum_{i=1}^{n} w_i}$$

For our specific model, the password entropy is weighted higher compared to the other metrics.

References

[1] Ross, Sheldon M. 2019. A First Course in Probability. 10th ed. Pearson.