Entropy Calculation

Rahul Soni, Eureka.AI

October 10, 2018

Abstract

This report contains a brief summary of Analysis of Entropy calculation for different probability distributions. Specifically, test cases were designed to provide empirical support for the Entropy Theory.

1 Introduction

Cross entropy is a measure of randomness or disorder in the system. For a given probability distribution associated with a random variable, the binary entropy, E, iscalculated as:

$$E = -\sum_{i=1}^{n} p_i * log(p_i) \tag{1}$$

Since entropy is a measure of randomness, a larger entropy would be higher randomness and thus least amount of information that can be obtained from underlying distribution. In terms of probability distribution, a large entropy would mean that the probability distribution tends towards uniform distribution since all outcomes are equally likely nothing can be inferred with a higher confidence.

We conclude that **The entropy of the system increases when the probability distribution tends towards uniform distrbution.**

2 Experiments

2.1 ZipCode Dataset

ZipCode dataset contains two attributes for location information viz. (i) ZipCode and, (ii) Country. The Binary entropy for each case is as follows:

```
ZipCode: E = 8.22
Country: E = 4.5
```

We note that the entropy is higher for the Zipcode compared to the Country as location. It suggests that the probability distribution of ZipCode should be flat (and tends to uniform) compared to the probability distribution when

Country was selected as location. This indeed is the case as shown in Fig.[1] below.

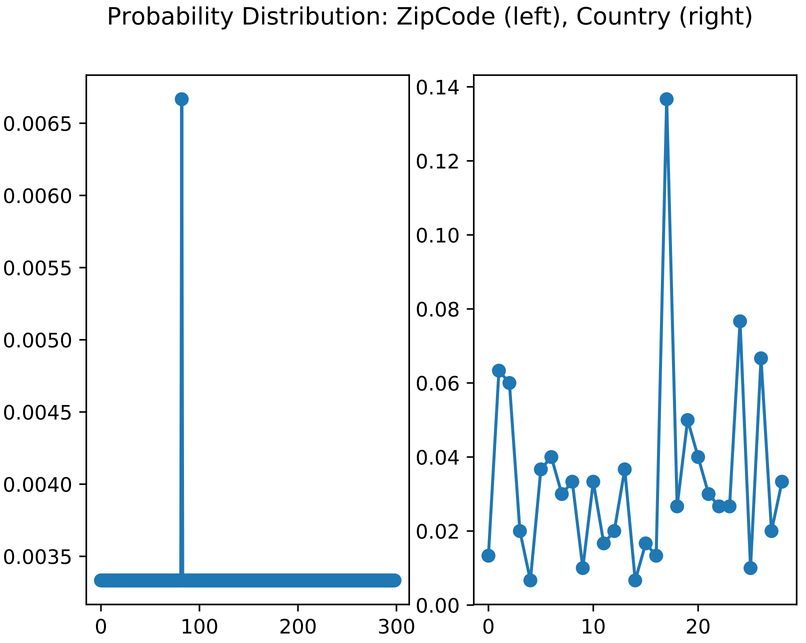

Figure 1: Probability distribution of ZipCode versus Country Code.

There are a total of 300 zip codes containing 299 unique zip codes and only one zipcode with two entries. However, for country code, there are only 29 unique entries with entries repeating with varied probability as seen in the graph.

For ZipCode, 299 probabilities are identical (and $= 1/300$) which tends towards a uniform distribution and explains larger entropy compared to the country code.