This white paper is about machine learning problems that consider gender, race or similar variables as predictors in a model. For example, race may be an important predictor for income level, at least in the USA. This paper doesn't deal with the definition of race categories or how to 'measure' race (or gender). The latter questions are contentious topics in social science and there is no scientific consensus answer. This paper considers what you would do with categorical data, such as gender or race, and in particular how do you measure "diversity." Different fields (e.g. physics or evolutionary biology) propose multiple "right" answers to choose from. But a too-simple measure, such as variance, is not a theoretically sound diversity measure.

Information entropy as a measure of diversity

G. R. Harp, 01-Mar-2020

Statement of the problem and definition of entropy

In one project, we were trying to come up with a practical definition of racial diversity and gender diversity in a human population. One possibility is to simply count the number of different races that appear in the group. But a group that is 99% one category and 1% of a different category is not as diverse as a simple count would show. We also discussed using A) the variance or B) the 'information entropy' as a measure of diversity. Here we make a comparison of the latter two choices.

One suggestion is to use the variance as a measure of diversity. This may be sufficient for some simple comparisons, but variance has a couple of drawbacks. Variance is an extrinsic variable, meaning it has units. This makes it difficult to compare the variance (diversity) of two different measures. Suppose you have one column which is race (categories = 1, 2, 3, and 4), and a second column which is gender (categories = 1, 2). If you compute the variances of these two columns, then it isn't clear how can you compare the diversity of race with the diversity of gender?

Another proposal is to use information entropy (Shannon entropy) as the diversity measure.

*The **information entropy**, often just **entropy**, is a basic quantity in information theory associated to any random variable, which can be interpreted as the average level of "information", "surprise", or "uncertainty" inherent in the variable's possible outcomes.*

The quote is from Wikipedia's definition of information entropy, and gives you some idea of what entropy means. An example helps: Suppose you have a classroom with 29 children of various races and then add another child from the same community. How can you predict the race of the 30th child before they arrive? How likely is it that you will guess wrongly, assuming the theoretically best guess? If your guess is wrong, then you are 'surprised' by the result. That's entropy.

Please go to the Wikipedia page for a description of the formula $\left(-\sum p_i \ln(p_i)\right.$, where p_i is the probability of category $\left. i \right)$ and some explanation. If you haven't thought of it before, entropy can be a difficult concept to get your head around, so I refer you to the explanations on Wikipedia to get started.

An illuminating example

Instead, let's consider an example. You have two cohorts of 100 people and some extrinsic categorical variable X, whatever that may be. For convenience, we assume that X values may fall into one of 100 categories, labeled i, where $0 < i \leq 100$.

In the first cohort, ninety-nine people have $X = 1$ and only one person has $X = 100$. Intuitively we think this sample has very low diversity (whatever that means). A simple calculation shows that X has a variance of 97 units.

In the second cohort, the X follows a normal distribution with values from 1-100, like before. Intuitively, this is a highly diverse sample. However, by coincidence this cohort has a variance $\mathrm{var}(X) = 97$. (In case you think this is impossible, I have prepared two samples with exactly these distributions, shown below.)

As an intuitive measure of diversity the variance is disappointing, since it doesn't distinguish between these two very different groups. What about the entropy? In Figure 1 there is a screen shot of a Jupyter notebook that calculates entropy using the `scipy.stats.entropy()` function in python. (A similar function exists in all mainstream statistics packages, including R.)

```
In [50]: import numpy as np
         import pandas as pd
         from scipy.stats import entropy

In [116]: # each dataset has 100 values
          highDiversity = np.array([18, 24, 37, 24, 18, 27, 24, 31, 34, 28, 32, 2, 10, 15, 27, 19, 37, 22, 19, 11, 38, 22, 34, 39, 17, 35,
                                    16, 18, 23, 11, 16, 37, 30, 37, 20, 23, 6, 23, 43, 28, 20, 31, 20, 21, 6, 17, 23, 8, 16, 31, 21, 19, 21, 23,
                                    42, 32, 28, 3, 20, 17, 23, 42, 25, 36, 30, 23, 3, 26, 21, 6, 38, 29, 33, 37, 30, 27, 19, 21, 42, 21, 19, 31, 18,

          lowDiversity = np.array([100, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                                   1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
                                   1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])

          N = len(lowDiversity)
          maxEntropy = np.log(N)

In [117]: unique_elements, counts_elements = np.unique(highDiversity, return_counts=True)
          prob_elements = counts_elements / len(highDiversity)
          highDiversityEntropy = entropy(prob_elements)/maxEntropy

          unique_elements, counts_elements = np.unique(lowDiversity, return_counts=True)
          prob_elements = counts_elements / len(lowDiversity)
          lowDiversityEntropy = entropy(prob_elements)/maxEntropy

In [119]: pd.DataFrame({'DataSet': ['lowDiversity', 'highDiversity'], 'variance': [np.var(lowDiversity), np.var(highDiversity)],
                        'entropy = diversity quotient':[lowDiversityEntropy, highDiversityEntropy]})

Out[119]:
```

	DataSet	variance	entropy = diversity quotient
0	lowDiversity	97.0299	0.012161
1	highDiversity	97.4444	0.734588

Figure 1: A Jupyter notebook that compares the variance and information entropy of two different cohorts, designed to have an intuitive feeling of either high or low diversity

Skipping to the last cell in Figure 1, we compare the variances and normalized entropies for each of the cohorts. I'm inventing a new name for entropy in this case, 'diversity quotient.' The diversity quotient is a dimensionless measure that varies over the range 0-1. Comparing the variances and diversity quotients of the two cohorts, we see that the variances are almost equal while the diversity quotient better captures our intuitive sense of what diversity means.

Normalization

The only detail I haven't explained is how we normalize entropy to a 0-1 scale. This is done by considering the entropy of the sample with maximum diversity. For a cohort with sample size $N = 100$, this implies there is one person in each category. The entropy here is easy to calculate, since every category i has the same probability $p_i = 1/100 \equiv 1/N$. There are N_{cat} different categories, so

$$\begin{aligned} \text{Maximum entropy} &= \sum_i p_i \ln\left(p_i\right) \\ &= -N_{cat}\left(1/N_{cat}\right)\ln(1/N_{cat}) = \ln\left(N_{cat}\right). \end{aligned} \tag{1}$$

In cases where there are more categories than cases, $N_{cat} \geq N$, you use the smaller of the two numbers (sample size). In cases where $N > N_{cat}$ (typical case), and maximum entropy is achieved when there are an equal number of samples in each category. This is true, even if N_{cat} is not an integer multiple of N, since the sample size drops does not appear in Eq. (1).

Knowing the maximum entropy, we form the normalized entropy (aka diversity quotient) as the sample entropy divided by the theoretical maximum entropy.

Summary

Here we present a short analysis that motivates using the information entropy as our diversity measure. Many other diversity measures are found in the literature. For example, a super-simple measure is simply to take the probability value of the dominant class. This is a measure that varies from 0-1, and is also monotonic with our intuition of diversity.

The best reasons to choose the entropy is because it has a well-defined meaning in information theory as "the maximum information content expressible by the sample." Entropy is commonly used in genetics, where researchers want to compute the information carrying capacity of a single gene or piece of DNA. This answers the interesting question of, how much information is required to fully specify a single human being?

By dividing our sample entropy by the maximum entropy we have a metric, which we call 'diversity quotient', which is an intuitive measure of the fractional information content embedded in our sample, compared to the maximum information there could possibly be.

Update: 05-Apr-2020

One of our readers pointed out that in evolutionary biology, a more common measure of species diversity is the Simpson Index. If n_i is the number of species in category i, and the total number of animals spotted is N, then the Simpson index is computed as

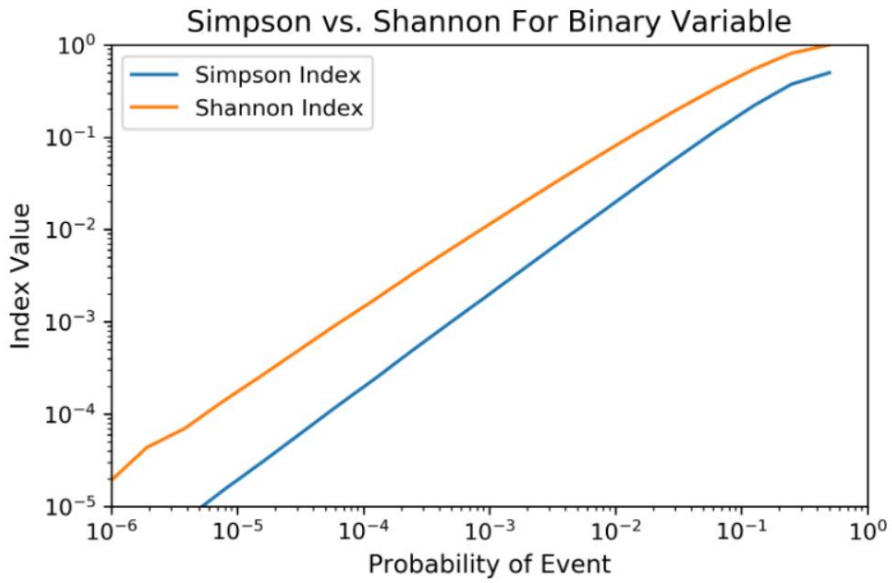
$$\text{Simpson Index} \quad = \quad 1 \; - \left(\frac{\sum_i n_i \left(n_i - 1 \right)}{N \left(N - 1 \right)} \right) \tag{2}$$

The Simpson index has the same desirable properties as entropy (varies 0-1, intuitively consistent). After some discussion, we have a few more remarks.

The Simpson and Shannon indexes will always be monotonically related. That is, any ordering of diversity using one index will agree with ordering by the other index. This is seen if we compare the asymptotic limits of the functional form of the two indices:

$$\text{Shannon} \sim \sum_i p_i \ln\left(p_i \right) \qquad \text{Simpson} \sim \sum_i p_i^2$$

So from one perspective, either index gives the same information, and what changes between them is the interpretation of index values. The Simpson index is generally lower than Shannon, as observed in this plot.



This figure examines a (0 or 1) variable where the probability of "1" varies over a large range. Notice the log-log scaling, where a straight line indicates a power law dependence. It is notable that the Simpson index value is close to the probability value over most of the range. Thus, Simpson is intuitively even more easy to understand than Shannon entropy.

Note: We're plotting the Shannon as defined in my previous post, which is normalized to the number of categories, not the number of samples. This normalization was added by me, and is a bit non-standard.

Another thing to compare are the index values when the probability tends to ½, or the point of maximum diversity. The Shannon index goes to 1, which indicates this is the most diversity you can have with two categories. The Simpson index does not go to 1. Instead, the Simpson measure is compared to the case where the number of categories is the same as the number of samples, and each category has the same number of samples in it. I prefer the Shannon normalization, but both of these are reasonably justified if you keep in mind what it is you're comparing to.

This difference in normalization does have a significant impact when comparing two variables with different numbers of occupied categories. Suppose you have 999,999 samples equally distributed between zeros and ones, and a single sample with a different value, say, 10. Because Shannon is sensitive to the number of occupied categories, this single sample has a large impact. The addition of a third category decreases the entropy from a value of 1 to approximately [log(2)/log(3)]. This suggests that the population could be a lot more diverse

than it is. On the other hand, the Simpson index changes by only a tiny amount, about 10^{-6}. This suggests the diversity hardly changes.

Which of these behaviors is more consistent with intuition? I don't know. If you have three species but only two of them appear frequently, then the diversity is lower than it could be (Shannon). On the other hand, you could argue that the diversity of a large population changes hardly at all if only one animal belongs to a third species. I think both answers are justifiable.

Having said all that, I am now leaning toward using the Simpson index for our models. The Simpson has no physical interpretation, unlike the information entropy. On the other hand, Simpson gives values that are more intuitive to understand. Unless you really care about information content, Simpson is a slightly preferred measure of diversity.