How to Describe Variety of a Probability Distribution: A Possible Answer to Yager's Question

Vladik Kreinovich
[2]Department of Computer Science
University of Texas at El Paso
El Paso, Texas 79968, USA
vladik@utep.edu

1. Variety: an Intuitive Notion

- For probability distributions, we have an intuitive understand that some probability distributions are "more random" than the others.
- This intuitive notion of degree of randomness is captured by the formal definition of an *entropy* of a probability distribution.
- Entropy can be defined as an average number of binary ("yes"-"no") questions that one needs to ask to determine the exact alternative.
- When an alternative i appears with probability p_i, this average number of questions is described by Shannon's formula:

$$S = -\sum_{i=1}^{n} p_i \cdot \log_2(p_i).$$

2. Variety: an Intuitive Notion (cont-d)

- For a continuous probability distribution with a probability density $\rho(x)$, we can similarly ask:
  - how many binary questions are needed, on average,
  - to determine x with a given accuracy ε.
- Asymptotically, when $\varepsilon \to 0$, this number of questions can be described as $S - \log_2(2\varepsilon)$, where

$$S = -\int \rho(x) \cdot \log(\rho(x))\, dx.$$

- For discrete case, entropy progresses:
  - from its smallest possible value 0 – when we have a deterministic case in which one alternative i occurs with probability 1 ($p_i = 1$),
  - to the largest possible value which is attained at a uniform distribution $p_1 = \ldots = p_n = \dfrac{1}{n}$.

3. Variety: an Intuitive Notion (cont-d)

- Intuitively:
  - both in the deterministic case and in the uniform distribution case, there is not much variety in the distribution, while
  - in the intermediate cases, when we have several different values p_i, there is a strong variety.
- Entropy does not seem to capture this notion of variety.
- So, Ron Yager asked a question: How can we describe this intuitive notion of variety in precise terms?
- In this paper, we provide a possible answer to this question.

4. Main idea behind our approach

- The value of entropy only depends on the values of the probability.
- It does not depend on which alternatives have different probabilities:
  - if we apply a permutation $\pi : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ to the alternatives,
  - then the resulting probability distribution $p'_i \stackrel{\text{def}}{=} p_{\pi(i)}$ will have the same entropy as the original probability distribution p_i.
- So, when analyzing related properties of randomness, we can assume that:
  - we only know the values $p_1, \ldots, p_n$,
  - but we do not know which alternative has which probability.
- In general, because of the possible permutations, we can have different distributions with the same set of values $\{p_1, \ldots, p_n\}$.

5. Main idea behind our approach (cont-d)

- In other words:
  - once we fix the set of values $\{p_1, \ldots, p_n\}$,
  - we get, in general, not a single distribution but rather a *variety* of different distributions.
- Let us see how big this variety is in different cases.
- Let us first consider the deterministic case, in which all the values p_i are equal to 0 except for one value which is equal to 1.
- In this case, we have $N = n$ possible probability distributions:
  - the first one in which alternative 1 occurs with probability 1,
  - the second one in which alternative 2 occurs with probability 1, etc.
- In the case of a uniform distribution, all the values of p_i are equal.
- So, no matter what permutations we apply, we end up with the exact same uniform distribution.

6. Main idea behind our approach (cont-d)

- Thus, in this case, the variety consists of a single probability distribution: $N = 1$.
- In the generic case, when all n probabilities p_i are different:
  - we get as many probability distributions as we have permutations, i.e,
  - we get $N = n!$ different distributions.
- We see that in this sense:
  - deterministic and uniform cases indeed have low variety, while
  - the generic case has a much larger variety.
- It is therefore reasonable to consider the corresponding value N when formalizing the intuitive notion of variety.

7. How to measure variety: case of discrete distributions

- In line with the above definition of entropy, it is reasonable to describe the variety as:
  - the smallest number of binary questions
  - which are needed to uniquely determine the actual distribution.
- After each binary question, we can have 2 possible answers.
- So, if we ask q binary questions, then, in principle, we can have 2^q possible results.
- Thus:
  - if we know that our (unknown) distribution is one of N distributions, and
  - we want to uniquely pinpoint the distribution after all these questions,
  - then we must have $2^q \geq N$.

8. How to measure variety for discrete distributions (cont-d)

- In this case, the smallest number of questions is the smallest integer q that is $\geq \log_2(N)$.
- Thus, $\log_2(N)$ is the natural measure of variety in the discrete case.
- For the uniform case, $N = 1$, so the variety is $\log_2(1) = 0$.
- In the deterministic case, we have $q = \log_2(n)$.
- In the generic case, we have $q = \log(n!)$.
- To simplify computations, we can use the well-known Stirling formula $n! \sim (n/e)^n \cdot \sqrt{2\pi \cdot n}$, hence $q = \log_2(n!) \approx n \cdot \log_2(n)$.
- It is worth mentioning that:
  - since the variety only depends on the set of probability values $\{p_1, \ldots, p_n\}$ and not on their order,
  - we can, without losing generality, assume that the values p_i are listed in increasing order $p_1 \leq p_2 \leq \ldots \leq p_n$.

9. How to measure variety: case of continuous distributions

- Without losing generality, we can similarly assume that the probability density $\rho(x)$ is an increasing function of x.
- Similarly to entropy, a natural way to go from the discrete case to the continuous case is to take into account that in reality;
  - we can only determine both the value of the variable x and the probability p
  - with a certain accuracy.
- Let us fix the accuracy ε of measuring x.
- Then, within this accuracy, we have only finitely many possible values of x.
- A value x_i covers the whole interval $[x_i - \varepsilon, x_i + \varepsilon]$.

10. How to measure variety: case of continuous distributions (cont-d)

- So we only need values:
  - x_0 – which covers $[x_0 - \varepsilon, x_0 + \varepsilon]$,
  - $x_1 = x_0 + 2\varepsilon$ – which covers $[x_1 - \varepsilon, x_1 + \varepsilon] = [x_0 + \varepsilon, x_0 + 3\varepsilon]$,
  - $x_2 = x_0 + 4\varepsilon$, etc.
- As a result, we get a discrete problem which we already know how to handle.
- When the accuracy ε tends to 0, the discrete problem tends to the original continuous one.
- For *entropy*, it was sufficient to take into account that x cannot be measured exactly.

11. How to measure variety: case of continuous distributions (cont-d)

- For *variety*:
  - we need to distinguish between different and equal values of probability p_i,
  - so, we must also take into account that the probabilities can only be measured with a certain accuracy.
- Let us fix the accuracy ε with which we measure x, and the accuracy δ with which we measure probability.
- Once we fix ε, we get values

$$x_0, x_1 = x_0 + 2\varepsilon, x_2 = x_0 + 4\varepsilon, \ldots, x_i = x_0 + i \cdot (2\varepsilon), \ldots$$

- Each of these values x_i covers an interval $[x_i - \varepsilon, x_i + \varepsilon]$.
- For the probability distribution with the density $\rho(x)$, the probability p_i of x_i is equal to $p_i = \int_{x_i-\varepsilon}^{x_i+\varepsilon} \rho(x)\, dx \approx \rho(x_i) \cdot (2\varepsilon)$.

12. How to measure variety: case of continuous distributions (cont-d)

- We can only determine probabilities with accuracy δ.
- This means, in effect, that we divide the interval $[0,1]$ of possible values of probability into intervals:
  - $\mathbf{p}_0 = [0, 2\delta]$ (probabilities which are approximately equal to δ),
  - $\mathbf{p}_1 = [2\delta, 4\delta]$ (probabilities which are approximately equal to 3δ),

    $\ldots,$
  - $\mathbf{p}_j = [j \cdot (2\delta), (j+1) \cdot (2\delta)]$ (probabilities which are approximately equal to $(j + 1/2) \cdot (2\delta)$), $\ldots,$
  - $[1 - 2\delta, 1]$ (probabilities which are approximately equal to $1 - \delta$),
- We consider events p_i for which the probabilities fall into the same probability interval.
- Let n_j denote the number of events for which the corresponding probability $p_i \approx \rho(x_i) \cdot (2\varepsilon)$ falls within the j-th probability interval $\mathbf{p}_j$.

13. How to measure variety: case of continuous distributions (cont-d)

- Then, the number of possible permutations is equal to the number of ways:
  - to subdivide the overall number of $n = n_1 + n_2 + \ldots$ values
  - into groups of n_1, n_2, etc.
- The total number C_1 of ways to choose n_1 elements out of n is well-known in combinatorics, and is equal to

$$\binom{n}{n_1} = \frac{n!}{(n_1)! \cdot (n - n_1)!}.$$

- After we choose these n_1 elements, we have a problem in choosing n_2 out of the remaining $n - n_1$ elements.
- So for every selection of n_1 elements we have $C_2 = \binom{n - n_1}{n_2}$ possibilities to choose these n_2 elements.

14. How to measure variety: case of continuous distributions (cont-d)

- Therefore, in order to get the total number of selections of n_1 elements and n_2 elements, we must multiply C_2 by C_1.
- Adding selections of n_3, n_4, ..., we get finally the following formula:

$$N = C_1 \cdot C_2 \cdot \ldots \cdot C_{n-1} =$$

$$\frac{n!}{n_1! \cdot (n-n_1)!} \cdot \frac{(n-n_1)!}{n_2! \cdot (n-n_1-n_2)!} \cdot \ldots = \frac{n!}{n_1! \cdot n_2! \cdot \ldots}$$

- Thus, the resulting degree of variety q is equal to

$$q = \log_2(N) = \log_2(n!) - \log_2(n_1!) - \log_2(n_2!) - \ldots$$

- Since $\log_2(n!) \approx n \cdot \log(n)$, we conclude that

$$q = n \cdot \log_2(n) - n_1 \cdot \log_2(n_1) - n_2 \cdot \log_2(n_2) - \ldots$$

15. How to measure variety: case of continuous distributions (cont-d)

- Here the total number of points $n \approx L/(2\varepsilon)$:
  - only depends on the width L of the interval on which the probability distribution is located but
  - not on the distribution itself.
- How big are the values n_j?
- By definition, n_j is the number of values x_i for which

$$j \cdot (2\delta) \le p_i = \rho(x_i) \cdot (2\varepsilon) \le (j+1) \cdot (2\delta), \text{ i.e.,}$$

$$j \cdot \frac{\delta}{\varepsilon} \le \rho(x_i) \le (j+1) \cdot \frac{\delta}{\varepsilon}.$$

- Since $\rho(x)$ is an increasing function of x, this is equivalent to

$$x^{(j)} \le x_i \le x^{(j+1)}, \text{ where } x^{(j)} \stackrel{\text{def}}{=} \rho^{-1}\left(j \cdot \frac{\delta}{\varepsilon}\right).$$

16. How to measure variety: case of continuous distributions (cont-d)

- Here, ρ^{-1} denotes the inverse function to $\rho(x)$, so $\rho(x^{(j)}) = j \cdot \dfrac{\delta}{\varepsilon}$.
- The difference $\Delta x^{(j)} \stackrel{\text{def}}{=} x^{(j+1)} - x^{(j)}$ can be determined from the fact that asymptotically,
$$\rho(x^{(j+1)}) = \rho(x^{(j)} + \Delta x^{(j)}) \approx \rho(x^{(j)}) + \rho'(x^{(j)}) \cdot \Delta x^{(j)}.$$
- Here $\rho'(x)$ denotes the derivative of the density function.
- We know that $\rho(x^{(j)}) = j \cdot \dfrac{\delta}{\varepsilon}$ and $\rho(x^{(j+1)}) = (j+1) \cdot \dfrac{\delta}{\varepsilon}$.
- So, we conclude that $\Delta x^{(j)} \approx \dfrac{\delta}{\varepsilon} \cdot \dfrac{1}{\rho'(x^{(j)})}$.
- On this interval, we have $n_j \approx \Delta x^{(j)}/(2\varepsilon)$ values x_i.
- So $n_j \approx \dfrac{\delta}{2\varepsilon^2} \cdot \dfrac{1}{\rho'(x^{(j)})}$.

17. How to measure variety: case of continuous distributions (cont-d)

- Hence, $q = n \cdot \log_2(n) - \sum n_j \cdot \log_2(n_j)$ can be described as $q = n \cdot \log_2(n) + \sum a(x^j)$, where

$$a(x^{(j)}) \stackrel{\text{def}}{=} -\frac{\delta}{2\varepsilon^2} \cdot \frac{1}{\rho'(x^{(j)})} \cdot \log_2 \left(\frac{\delta}{2\varepsilon^2} \cdot \frac{1}{\rho'(x^{(j)})} \right).$$

- When accuracies tend to 0, this sum gets close to an integral.
- For every function $f(x)$, the integral is approximately equal to its integral sum $\int f(x)\,dx \approx \sum f(x^{(j)}) \cdot \Delta x^{(j)}$.
- The smaller ε and δ, the closer the integral sum to the integral.
- So, we conclude that the sum $\sum a(x^{(j)})$ can be approximately described as $\sum b(x^{(j)}) \cdot \Delta x^{(j)} \approx \int b(x)\,dx$, where $b(x^{(j)}) \stackrel{\text{def}}{=} \dfrac{a(x^{(j)})}{\Delta x^{(j)}}$.
- So $b(x^{(j)}) = -\dfrac{1}{2\varepsilon} \cdot \log_2 \left(\dfrac{\delta}{2\varepsilon^2} \cdot \dfrac{1}{\rho'(x^{(j)})} \right)$.

18. How to measure variety: case of continuous distributions (cont-d)

- Hence $q \approx n \cdot \log_2(n) - \int \frac{1}{2\varepsilon} \cdot \log_2\left(\frac{\delta}{2\varepsilon^2} \cdot \frac{1}{\rho'(x)}\right)\,dx.$
- Since the logarithm of the product is equal to the sum of the logarithms, we can see that

$$q = n \cdot \log_2(n) - \frac{1}{2\varepsilon} \cdot \left(\int \log_2\left(\frac{\delta}{2\varepsilon^2}\right)\,dx + \int \log_2\left(\frac{1}{\rho'(x)}\right)\,dx\right).$$

- Thus, asymptotically, the value q can be determined once we know the value $Q \stackrel{\text{def}}{=} \int \log_2(\rho'(x))\,dx$.
- In the general case, the function $\rho(x)$ is not necessarily increasing, it can be decreasing as well, so $Q \stackrel{\text{def}}{=} \int \log_2(|\rho'(x)|)\,dx$.

19. Conclusions

- For a continuous probability distribution, the above measure of variety can be computed as follows: $Q = \int \log_2(|\rho'(x)|)\,dx$.
- For the (almost) deterministic case, when $\rho(x) \approx \frac{1}{\varepsilon}$ on a narrow interval of width ε, we have $\rho'(x) \approx \frac{\rho(x)}{\varepsilon} \approx \frac{1}{\varepsilon^2}$.
- So $Q \approx \varepsilon \cdot \log_2(\varepsilon^{-2}) \approx 0$.
- For a uniform distribution $\rho(x) = \text{const}$, we have $\rho'(x) = 0$, hence $Q = -\infty$.
- For non-uniform distributions in which $|\rho'(x)| > 0$, as expected, we get higher variety.

20. Future work

- In this talk, we analyzed the case of probabilistic uncertainty.
- It is desirable to extend the corresponding definition and analysis to other types of uncertainty, e.g., fuzzy, interval-valued fuzzy.
- It is also desirable to compare this approach with alternative approaches of defining variety.
- For example, it is desirable to compare with a definition from (Fuller at al. 2003) for OWA operators.

21. Acknowledgments

- This work was supported in part by the National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and
  - HRD-1834620 and HRD-2034030 (CAHSI Includes).
- It was also supported by the AT&T Fellowship in Information Technology.
- It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478.