Entropía, Divergencia de Kullback-Leibler e Información

Irving Gómez Méndez

1 Entropy and surprise

The example and explanation of this section was taken from the video of StatQuest with Josh Starmer (2021)

Entropy is the base of something called Relative Entropy, better known as divergence of Kullback-Leibler and Cross-Entropy. The entropy is used to quantify similarities and differences. To define it, we need first to understand the concept of **surprise**.

Think in three groups of marbles. Group A has 6 black marbles and 1 white marble, group B has 1 black marble and 10 white marbles, and group C has 7 black marbles and 7 white marbles.

If we select a marble from group A and this is black, we would not be very surprised. But we would be surprised if we take a white marble. Similarly, if we take a marble from group B and this is black, we would be surprised. Meanwhile, for group C we would be equally surprised if we take a black marble or a white marble.

This indicates that **surprise is inversely related with the probability**. It is tempting to simply use the inverse of the probability to define the surprise. But, before doing it, imagine the next case. Assume that we have a terrible coin that always shows head when we flip it. If the next time that we flip it we get head, we would not be surprised at all, so the surprise should equal to zero. However, we have that

$$\frac{1}{\mathbb{P}(\text{head})} = \frac{1}{1} = 1 \neq 0.$$

Thus, instead of just taking the inverse of the probability, we use the logarithm of the inverse of the probability,

$$\text{Surprise} = \log\left(\frac{1}{\text{Probability}}\right) = -\log(\text{Probability}).$$

Assume now, that we have a coin that has a probability of 0.9 of showing head and 0.1 of showing tail when we flip it, and we observe the sample (H, H, T). The probability of this sample is

$$P\left\{(H, H, T)\right\} = 0.9 \times 0.9 \times 0.1,$$

and the surprise is

$$\begin{aligned}\text{Surprise} &= \log\left(\frac{1}{0.9 \times 0.9 \times 0.1}\right) \\ &= -\left[\log(0.9) + \log(0.9) + \log(0.1)\right].\end{aligned}$$

That is, the total surprise is simply the sum of the surprises of each one of the tosses.

If we want to estimate the total surprise after flipping the coin 100 times, we estimate hwo many times we would observe head and tail (0.9×100 and 0.1×100, respectively), and we multiply it by the surprise of head and tail, respectively.

$$\text{Total expected surprise} = (0.9 \times 100) \log\left(\frac{1}{0.9}\right) + (0.1 \times 100) \log\left(\frac{1}{0.1}\right).$$

if we divide by the number of tosses, we obtain the surprise of each toss, we define this quantity as the **entropy** of the coin:

$$\begin{aligned}
\text{Entropy} &= \mathbb{E}[\text{Suprise}] \\
&= \frac{(0.9 \times 100) \log\left(\frac{1}{0.9}\right) + (0.1 \times 100) \log\left(\frac{1}{0.1}\right)}{100} \\
&= 0.9 \log\left(\frac{1}{0.9}\right) + 0.1 \log\left(\frac{1}{0.1}\right) \\
&= \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y) \text{Suprise}(y) \\
&= \sum_{y \in \mathcal{Y}} p(y) \log\left(\frac{1}{p(y)}\right) \\
&= -\sum_{y \in \mathcal{Y}} p(y) \log(p(y)).
\end{aligned}$$

The last equality corresponds with the equation of the Entropy that Claude Shannon published for the first time in 1948.

Consider again the example of the marbles, we can calculate the entropy of each group.

Group A: The probability of taking a black marble is 6/7 and the probability of taking a black coin is 1/7. The probability of each type of coin, as well as its associated surprise is presented in the next table.

Group A	Probability	Surprise
Black	6/7	$\log\left(\frac{1}{6/7}\right) \approx 0.15$
White	1/7	$\log\left(\frac{1}{1/7}\right) \approx 1.95$

The entropy of group A is (approximately) 0.41.

Analogously, for the groups B and C, we have the following.

Group B:

Group B	Probability	Surprise
Black	1/11	$\log\left(\frac{1}{1/11}\right) \approx 2.4$
White	10/11	$\log\left(\frac{1}{10/11}\right) \approx 0.1$

The entropy of group B is (approximately) 0.3.

Group C:

Group C	Probability	Surprise
Black	7/14	$\log\left(\frac{1}{7/14}\right) \approx 0.69$
White	7/14	$\log\left(\frac{1}{7/14}\right) \approx 0.69$

the entropy of group C is (approximately) 0.69.

Note that, even while group B has a marble with a large surprise, most of the marbles therein have a low surprise. This results in a low entropy for group B. Moreover, this is the most homogeneous group of the three. Meanwhile, group C which has the same number of marbles of each color is the group with the largest entropy. That is, a more homogeneous group has a lower entropy.

From another perspective, in group C both types of marbles have the same probability to be selected. That is, there is a uniform distribution for the color of the marble. From this observation, we deduce that the discrete distribution of maximum entropy with a finite number of possible values corresponds to the uniform distribution. This idea of maximum entropy, can also be interpreted as a less ordered group or as the group of maximum uncertainty, in the sense that it corresponds with a less homogeneous group.

2 Cross-entropy and divergence of Kullback-Leibler

In general, if Y is a r.v. with density p, we define the entropy of p as

$$H(p) = -\mathbb{E}_{Y\sim p}[\log p(Y)].$$

Assume now, that instead of using p we make the "mistake" of modeling Y with the density q, in this case we define the **cross-entropy** as

$$\begin{aligned} H(p||q) &= -\mathbb{E}_{Y\sim p}[\log q(Y)] \\ &= -\sum_{y\in\mathcal{Y}} p(y)\log q(y). \end{aligned}$$

It can be shown that

$$\min_q H(p||q) = H(p||p) = H(p).$$

That is, the cross-entropy quantifies the "disorder" or uncertainty of p plus what we are adding for making the "mistake" of modeling Y through q instead of using its real density p.

Finally, we define the **divergence of Kullback-Leibler**, $D_{KL}(p||q)$, sometimes called **Relative Entropy**, as the difference between $H(p||q)$ and $H(p)$. That is, $D_{KL}(p||q)$ corresponds with the additional entropy induced by q.

$$\begin{aligned} D_{KL}(p||q) &= H(p||q) - H(p) \\ &= -\sum_{y\in\mathcal{Y}} p(y)\log q(y) + \sum_{y\in\mathcal{Y}} p(y)\log p(y) \\ &= \sum_{y\in\mathcal{Y}} p(y)\log\frac{p(y)}{q(y)}. \end{aligned}$$

The KL divergence satisfies $D_{KL}(p||q) \geq 0$ and $D_{KL}(p||q) = 0 \Leftrightarrow q = p$.

3 Bayesian statistics

The goal of Bayesian statistics is to maximize the KL divergence between the posterior and the prior distribution. That is, using Bayesian analysis we are moving from a distribution of high entropy to a distribution of low entropy. In other words, with Bayesian analysis we move from one state of high uncertainty with a huge potential to learn from it to a state of low uncertainty from which we cannot learn too much more.

4 Relation between Kullback-Leibler divergence and maximum likelihood

Assume that Y is a r.v. with distribution F whose density is f, both of them unknown. In our ignorance, we decide to model Y through a distribution G which belongs to the family of distributions $\mathcal{G}$, whose density is g. Note that F might belong or not to $\mathcal{G}$. Let be

$$\begin{aligned} g^\star &\in \arg\min_{g\in\mathcal{G}} D_{KL}(f||g) \\ &= \arg\min_{g\in\mathcal{G}} \mathbb{E}_{Y\sim F}[\log f(Y)] - \mathbb{E}_{Y\sim F}[\log g(Y)] \\ &= \arg\max_{g\in\mathcal{G}} \mathbb{E}_{Y\sim F}[\log g(Y)]. \end{aligned}$$

If we count with a sample $Y_1, \ldots, Y_n \overset{iid}{\sim} Y$, then we can approximate F through the empirical distribution function F_n, for n sufficiently large. Thus,

$$\begin{aligned} g^\star &= \arg\max_{g\in\mathcal{G}} \mathbb{E}_{Y\sim F}[\log g(Y)] \\ &\simeq \arg\max_{g\in\mathcal{G}} \mathbb{E}_{Y\sim F_n}[\log g(Y)] \\ &= \arg\max_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}[\log g(Y_i)] \\ &= \arg\max_{g\in\mathcal{G}} \sum_{i=1}^{n}[\log g(Y_i)] \\ &= \arg\max_{g\in\mathcal{G}} \prod_{i=1}^{n} g(Y_i). \end{aligned}$$

Let be $\hat{g}$ the estimator of maximum likelihood when we model Y through g,

$$\hat{g} = \arg\max_{g\in\mathcal{G}} \prod_{i=1}^{n} g(Y_i).$$

That is, when we use maximum likelihood we attempt to minimize the KL divergence between the real distribution and our model.

Note that $\hat{g} \to g^\star$ when $n \to \infty$. The difference between $g^\star$ and $\hat{g}$ is called the **estimation error**. On the other hand, the difference between f and $g^\star$ is called the **approximation error**.

5 Entropy and information

The goal of **Theory of Information** is to measure the amount of information that contains a message.

Assume that we have an alphabet of 2 symbols, for example x and y, this alphabet has one bit of information. The ASCII code has $2^8 = 256$ characters, which means that it requires 8 bits of information (1 byte) to write any character of the code.

Now, consider the case of coin that has head on both of its sides, that is $\mathbb{P}(\text{head}) = 1$, observing a head in a toss gives us no information at all. On the other hand, if the probability of an event is low and such event occurs, it would give us a lot of information.

Accordingly to Chaude Shannon, a measurement for the possible amount of information in a message must have the following properties:

- It must be continuous. That is, small changes in the probabilities of the events should not make large changes in the uncertainty.
- It must increase with the number of possible events.
- It must be additive.

The only function that satisfies these three conditions is the entropy. Assume that $\mathcal{Y}$ is the alphabet used to write a message, and assume that $y \in \mathcal{Y}$, we define the amount of information in y as

$$I(y) = -\log_2(p(y)),$$

the base 2 in the logarithm is just a convention due for measuring information in bits. The entropy of a message Y is defined as

$$H(Y) = -\sum_{y \in \mathcal{Y}} p(y) \log_2(p(y)).$$

The less the entropy of a message, the larger the potential that it has information.

Note that if $p(y_i) = 1$ and $p(y_j) = 0$ for all $j \neq i$, then $H(Y) = 0$. On the other hand, if $p(y_1) = \cdots = p(y_n)$, then

$$H(Y) = -\log_2(1/n) = \log_2(n).$$

That is, a uniform variable with n possible values has $\log_2(n)$ bits of information. To understand this, consider the following example.

Assume that you have a dice with 16 sides, then we need $4 = \log_2(16)$ binary questions to guess the number when the dice is tossed. To show this, assume that we toss our dice and we observe the number 2. We can make the following questions to guess the number:

1. The number is bigger or equal to 9? -No.
2. The number is bigger or equal to 5? -No.
3. The number is bigger or equal to 3? -No.
4. The number is 2? -Yes.
5. Then, the number is 2.

Of course, this is not the only strategy, we could ask one number at a time, in the following way:

1. The number is 1? -No.
2. The number is 2? -Yes.
3. Then, the number is 2.

With this strategy, we arrived to the result with just 2 questions. However, if the number was, for example, 13, than we would need 13 questions to know the result. Meanwhile, if we use the first strategy, the number of questions needed is still 4, as follows:

1. The number is bigger or equal to 9? -Yes.
2. The number is bigger or equal to 13? -Yes.
3. The number is bigger or equal to 15? -No.
4. The number is 14? -No.
5. Then, the number is 13.

References

StatQuest with Josh Starmer (2021). *Entropy (for data science) Clearly Explained!!!* Youtube. URL: `https://www.youtube.com/watch?v=YtebGVx-Fxw`.