Information entropy and Kullback-Leibler divergence

Richard Torkar

September 2020

We need to compare disparate statistical models and tell which one is better than the other (relatively speaking) concerning out of sample predictions. The main question we need to answer is:

> How much is our uncertainty reduced by learning the outcome?

The question is answered by measuring the *decrease* in uncertainty. The *decrease* in uncertainty is *information*. How do we quantify uncertainty in a probability distribution?

- Should be continuous! We don't want a small change in any of the probabilities to result in a massive change in uncertainty.
- Uncertainty should *increase*(!) as the number of possible events increases, e.g., 50/50 rain/sun vs. one out of every three days rain, shines, hails. In the latter example there is more uncertainty.
- Uncertainty should be additive! In short, should be the sum of the separate uncertainties.

Only one function satisfies the above desiderata, i.e., information entropy,[1]

$$H(p) = -\sum_{i=1}^{n} p_i \log(p_i)$$

The above states that the uncertainty in a probability distribution is the *average log probability* of an event.

[1] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 7 1948. DOI: 10.1002/j.1538-7305.1948.tb01338.x

As an example, consider a software under test (SUT). When we run our test suite we find that 30% of the test cases fail (find a bug), and 70% pass. The information entropy is hence:

```
> p <- c(0.3, 0.7)
> -sum(p*log(p))
[1] 0.6108643
```

Consider now that we re-run the test suite a month later and 20% of the test cases fail, i.e., $p = \{0.2, 0.8\}$.

```
> p <- c(0.2, 0.8)
> -sum(p*log(p))
[1] 0.5004024
```

We have now quantified our uncertainty. It has decreased!

WE NOW HAVE A WAY TO QUANTIFY UNCERTAINTY. Next we want to answer the question:

How far away is a model from the target?

or in other words: What is the additional uncertainty induced by using probabilities from one distribution to describe another distribution? This is called *Kullback-Leibler divergence*, and is key for making model comparisons.[2] Here we not only use p for events, but we also introduce q.

Let's assume that we have a distribution of events $p_1 = 0.3$ and $p_2 = 0.7$. We believe that these events happened with probabilities $q_1 = 0.25$ and $q_2 = 0.75$. How much uncertainty have we introduced by using $\{q_1, q_2\}$ to approximate $\{p_1, p_2\}$? To calculate this uncertainty we use the Kullback-Leibler divergence:

$$\mathrm{D_{KL}} = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$$

i.e., the *average difference in log probability* between the target (p) and our model (q).

AS AN EXAMPLE, let's travel from from Earth to Mars, and vice versa. If we start by traveling from Earth to Mars, we know that Earth has 70% water and 30% land, and when we land we find out that Mars is consisting of 1% water and 99% land. If we first calculate our $\mathrm{D_{K-L}}$ divergence going from Earth to Mars we see that:

```
> q <- c(0.7, 0.3) # Our model
> p <- c(0.01, 0.99) # The true ratio on Mars
> sum(p * log(p/q))
[1] 1.139498
```

i.e., $\mathrm{D_{E\to M}} = \mathrm{D_{KL}}(p,q) = 1.14$.

Let's now look at what happens when we go from Mars to Earth instead (i.e., we know the ratio for Mars but we don't know the ratio for Earth):

```
> # Our model is now Mars' ratio since we're standing on Mars
> q <- c(0.01, 0.99)
> # When we land on Earth we find out that there's 70% water
> p <- c(0.7, 0.3)
> sum(p * log(p/q))
[1] 2.61577
```

i.e., $\mathrm{D_{M\to E}} = \mathrm{D_{KL}}(p,q) = 2.62$.

The divergence is more than double when we go from Mars to Earth! This is a feature and not a bug.

[2] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 03 1951. DOI: 10.1214/aoms/1177729694

Harold Jeffreys had used this measure already in the development of Bayesian statistics.

If we use a distribution with high entropy to approximate an unknown true distribution of events, we will reduce the distance to the truth and therefore the error.