The maximum entropy framework

The maximum entropy principle — an example

Suppose we have a random variable X with known states (values of the observations, $x_1, \ldots, x_n$) but unknown probabilities $p_1, \ldots, p_n$; plus some extra constrains, eg. $\langle X \rangle$ is known. We are given the task to attempt to have a good guess for the probabilities.

Let's start with one of the simplest examples: X can take 1, 2 or 3 with unknown probabilities, and $\langle X \rangle = \overline{x}$ is known. Fixing $\langle X \rangle$ does not determine the probabilities, for example for $\overline{x} = 2$ any $(p_1, p_2, p_3) = (\frac{1-p_2}{2}, p_2, \frac{1-p_2}{2})$ satisfies the constraint, including eg. $(0, 1, 0)$ or $(\frac{1}{2}, 0, \frac{1}{2})$ or $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Which one is the "best"? According to the *maximum entropy principle*, the best guess is the one which maximises the information entropy under the given constraints.

To calculate this solution, we need to find the maximum of $H(p_1, p_2, p_3)$ as a function of p_1, p_2, p_3, under two constraints: $\langle X \rangle = 1p_1 + 2p_2 + 3p_3 = \overline{x}$ and $p_1 + p_2 + p_3 = 1$. We use the method of Lagrange multipliers: first calculate the unconditional maximum of the original function plus the constraints added with some multiplying factors (the Lagrange multipliers), which give the probabilities in a functional form with the Lagrange multipliers as parameters.

$$\begin{aligned} 0 &= d\left[H(p_1, p_2, p_3) - \lambda \left(\sum_{i=1}^{3} i p_i - \overline{x} \right) - \mu \left(\sum_{i=1}^{3} p_i - 1 \right) \right] \\ &= d\left[-\sum_{i=1}^{3} p_i \log p_i - \lambda \sum_{i=1}^{3} i p_i - \mu \sum_{i=1}^{3} p_i \right] \\ &= \sum_{i=1}^{3} \big\{ -\log p_i - 1 - \lambda i - \mu \big\} dp_i = 0 \end{aligned}$$

Since this has to hold for any dp_i, the curly brackets need to be zero:

$$-\log(p_i) - 1 - \lambda i - \mu = 0\,, \qquad\qquad i = 1, 2, 3$$

which with the notation $\lambda_0 = \mu + 1$ gives

$$p_i = e^{-\lambda_0 - \lambda i}\,.$$

Now we set the Lagrange multipliers by requiring the constraints to be satisfied. The constraint on the sum of probabilities give

$$1 = \sum_{i=1}^{3} p_i = e^{-\lambda_0} \sum_{i=1}^{3} e^{-\lambda i} \qquad \Rightarrow \qquad e^{-\lambda_0} = \frac{1}{e^{-\lambda} + e^{-2\lambda} + e^{-3\lambda}}$$

so

$$p_i = \frac{e^{-\lambda i}}{e^{-\lambda} + e^{-2\lambda} + e^{-3\lambda}} = \frac{e^{\lambda(1-i)}}{1 + e^{-\lambda} + e^{-2\lambda}}$$

The other constraint, $\langle X \rangle = \overline{x}$ gives

$$\overline{x} = \sum_{i=1}^{3} i p_i = \frac{1 + 2e^{-\lambda} + 3e^{-2\lambda}}{1 + e^{-\lambda} + e^{-2\lambda}} \tag{4}$$

Multiplying the equation with the denominator gives a second degree equation for $e^{-\lambda}$:

$$(\overline{x} - 3)\left(e^{-\lambda}\right)^2 + (\overline{x} - 2)e^{-\lambda} + \overline{x} - 1 = 0$$

which has the solution

$$e^{-\lambda} = \frac{-(\overline{x} - 2) \pm \sqrt{(\overline{x} - 2)^2 - 4(\overline{x} - 3)(\overline{x} - 1)}}{2(\overline{x} - 2)} = \frac{2 - \overline{x} \pm \sqrt{4 - 3(\overline{x} - 2)^2}}{2(\overline{x} - 3)}$$

Now if we rewrite (4) as

$$\overline{x} = \frac{e^{\lambda} + 2 + 3e^{-\lambda}}{e^{\lambda} + 1 + e^{-\lambda}} = 1 + \frac{1 + 2e^{-\lambda}}{e^{\lambda} + 1 + e^{-\lambda}}$$

then p_2 becomes

$$\begin{aligned} p_2 &= \frac{1}{e^{-\lambda} + 1 + e^{\lambda}} = \frac{\overline{x} - 1}{1 + 2e^{-\lambda}} = \frac{\overline{x} - 1}{1 + \frac{1}{\overline{x}-3}\left(2 - \overline{x} \pm \sqrt{4 - 3(\overline{x} - 2)^2}\right)} \\ &= \frac{(\overline{x} - 1)(\overline{x} - 3)}{-1 \pm \sqrt{4 - 3(\overline{x} - 2)^2}} = \frac{(\overline{x} - 1)(\overline{x} - 3)(-1 \mp \sqrt{4 - 3(\overline{x} - 2)^2}}{1 - (4 - 3(\overline{x} - 2)^2)} \\ &= \frac{-1 + \sqrt{4 - 3(\overline{x} - 2)^2}}{3} \end{aligned}$$

In the last step we had to keep the $+$ sign as only this root gives non-negative p_2. Finally the other probabilities become

$$p_1 = \frac{3 - \overline{x} - p_2}{2} \ , \qquad p_3 = \frac{\overline{x} - 1 - p_2}{2}$$

This solution has the right behaviour in the limiting cases: when $\overline{x} = 1$, the probabilities $(p_1, p_2, p_3) = (1, 0, 0)$; and when $\overline{x} = 3$, they are $(0, 0, 1)$. For $\overline{x} = 2$, the solution is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. The maximum entropy solution assigns zero probabilities only when no other possibilities are allowed. This is a very desirable property: it would be a sure failure to propose that a certain state has zero probability, and then find out that a given observation happened to yield that state. The Maximum Entropy solution is guaranteed not to fail there.

Maximum entropy principle — general form

After having this worked out example, we state the maximum entropy principle in a more general form. Suppose we have a random variable X taking (known) values $x_1, \ldots, x_n$ with unknown probabilities $p_1, \ldots, p_n$. In addition, we have m constraint functions $f_k(x)$ with $1 \leq k \leq m < n$, where

$$\langle f_k(X) \rangle = F_k \ ,$$

the F_ks are fixed. Then the maximum entropy principle assigns probabilities in such a way that maximises the information entropy of X under the above constraints. This is the "best guess" in the absence of any further knowledge about the random variable. Since any extra assumption would bring a reduction in uncertainty (see mutual information), we explicitly deny those extra assumptions by maximising the uncertainty.

In the following we calculate various properties of the maximum entropy solution. This may sound dry, but has the advantage that these abstract results can be very easily applied later for concrete examples.

To obtain a formal solution we proceed in a similar way as in the example, maximise the information entropy using Lagrange multipliers:

$$\begin{aligned} 0 &= d\left[H(p_1, \ldots, p_n) - \sum_{k=1}^{m} \lambda_k \left(\sum_{i=1}^{n} f_k(x_i) p_i - F_k\right) - \underbrace{\mu}_{\lambda_0 - 1}\left(\sum_{i=1}^{n} p_i - 1\right)\right] \\ &= \sum_{i=1}^{n} \left\{ -\log(p_i) - 1 - \sum_{k=1}^{m} \lambda_k f_k(x_i) - (\lambda_0 - 1) \right\} dp_i \end{aligned}$$

Since this is zero for any dp_i, all n braces have to be zero, giving

$$p_i = \exp\left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k f_k(x_i)\right) \tag{5}$$

Then all the Lagrange multipliers $(\lambda_0, \lambda_1, \ldots \lambda_m)$ are fixed by writing back into the constraints. The sum of probabilities give

$$1 = \sum_{i=1}^{n} p_i = e^{-\lambda_0} \sum_{i=1}^{n} \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(x_i) \right)$$

The sum after $e^{-\lambda_0}$ appears frequently, so it is useful to consider it separately: we will call it *partition function*

$$Z(\lambda_1, \ldots, \lambda_m) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(x_i) \right) \tag{6}$$

With this notation

$$e^{-\lambda_0} = \frac{1}{Z(\lambda_1, \ldots, \lambda_m)}\,, \qquad \lambda_0 = \log Z(\lambda_1, \ldots, \lambda_m) \tag{7}$$

The other constraints are

$$\begin{aligned} F_k = \sum_{i=1}^{n} f_k(x_i) p_i = e^{-\lambda_0} \sum_{i=1}^{n} f_k(x_i) \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(x_i) \right) = -\frac{1}{Z} \frac{\partial Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k} \\ = -\frac{\partial \log Z(\lambda_1, \ldots, \lambda_m)}{\partial \lambda_k}\,, \end{aligned} \tag{8}$$

which is m implicit equations, just enough to determine in principle the m unknowns λ_k. Using (7) then the probabilities (5) are then fully determined:

$$p_i = \frac{1}{Z(\lambda_1, \ldots, \lambda_m)} \exp\left(-\sum_{k=1}^{m} \lambda_k f_k(x_i) \right) \tag{9}$$

Unlike the simple example we had with three states, in practice it is often not possible to calculate the λ_ks explicitly as a function of F_ks, but as we see later this does not prevent us obtaining lots of useful results.

Consider now the value of the maximised information entropy. It is no longer function of the probabilities, but instead of the constraint values F_k, and to reflect this we change notation to S:

$$\begin{aligned} S(F_1, \ldots, F_m) = H(\underbrace{p_1, \ldots, p_n}_{\text{from (9)}}) = -\sum_{i=1}^{n} p_i \log(p_i) = -\sum_{i=1}^{n} p_i \left(-\lambda_0 - \sum_{k=1}^{m} \lambda_k f_k(x_i) \right) \\ = \lambda_0 + \sum_{k=1}^{m} \lambda_k \sum_{i=1}^{n} f_k(x_i) p_i = \log Z(\lambda_1, \ldots, \lambda_m) + \sum_{k=1}^{m} \lambda_k F_k \end{aligned} \tag{10}$$

Now calculate the partial derivatives of S w.r.t. the F_ks, being careful about what is kept constant in the partial derivatives[1]:

$$\left.\frac{\partial S}{\partial F_k}\right|_{\{F\}} = \sum_{\ell=1}^{m} \underbrace{\left.\frac{\partial \log Z}{\partial \lambda_\ell}\right|_{\{\lambda\}}}_{F_\ell} \left.\frac{\partial \lambda_\ell}{\partial F_k}\right|_{\{F\}} + \sum_{\ell=1}^{m} \left.\frac{\partial \lambda_\ell}{\partial F_k}\right|_{\{F\}} F_\ell + \lambda_k = \lambda_k \tag{11}$$

Here either $S(F_1, \ldots, F_m)$ or $\log Z(\lambda_1, \ldots, \lambda_m)$ give a full description of the system, as the other can be calculated using (10), and there is a symmetric relation between their partial derivatives: (8) and (11). We look at this kind of relation between two functions more closely below.

[1] In thermodynamics and statistical physics functions of many variables are used extensively, and the notation is not always clear on what the free variables are. When taking partial derivatives, it is essential to be clear on what is kept constant; therefore it is often shown at the bottom of the vertical bar after the partial differential. Eg. the notation $\{\lambda\}$ means all λ_js are kept fixed except the one we differentiate with.

Legendre transform

Consider a convex function $f(x)$, and define the following function

$$f^*(p) \stackrel{\text{def}}{=} \max_x \big(px - f(x)\big) \tag{12}$$

We call this[2] the *Legendre tranform* of $f(x)$. If f is differentiable as well, we can calculate the maximum as

$$0 = \frac{d}{dx}(px - f(x)) = p - \frac{df(x)}{dx}$$

Its solution for x depends on p, which we call $x(p)$:

$$\left.\frac{df(x)}{dx}\right|_{x=x(p)} = p$$

which plugged into (12) gives

$$f^*(p) = px(p) - f\big(x(p)\big)$$

Now let's calculate the Legendre transform of f^*:

$$(f^*)^*(y) = \max_p \big(yp - f^*(p)\big)$$

Again, if f^* is differentiable then

$$\left.\frac{df^*(p)}{dp}\right|_{p=p(y)} = y$$

However,

$$\frac{df^*(p)}{dp} = \frac{px(p) - f(x(p))}{dp} = x(p) + p\frac{dx(p)}{dp} - \underbrace{\left.\frac{df(x)}{dx}\right|_{x(p)}}_{p}\frac{dx(p)}{dp} = x(p)$$

so

$$y = \left.\frac{df^*(p)}{dp}\right|_{p=p(y)} = x(p(y))$$

thus

$$f^{**}(y) = yp(y) - f^*(p(y)) = yp(y) - p(y)x(p(y)) + f(x(p(y))) = f(y)$$

Thus the function $f^{**}(\cdot)$ and $f(\cdot)$ are equal, or in other way to say the Legendre transform is its own inverse.

The Legendre transform can be easily generalised to concave functions: in the definition $\max$ needs to be replaced by $\min$.

The other generalisation is functions of multiple variables: the Legendre transform of $f(x_1, \ldots, x_m)$ is

$$f^*(p_1, \ldots, p_k) = \sum_{k=1}^{m} x_k p_k - f(x_1, \ldots, x_m)\,, \qquad \text{where } p_k = \frac{\partial f}{\partial x_k}$$

In the previous section we have seen that $S(F_1, \ldots, F_m)$ and $-\log Z(\lambda_1, \ldots, \lambda_m)$ are Legendre transforms of each other, either one of them provides a full description of the system. The only remaining bit is to show that $-\log Z$ is indeed either convex or concave so that the Legendre transform is defined.

[2]The Legendre transform is often defined with a sign difference: $f^*(p) = \max(f(x) - px)$. The advantage of our notation is that the inverse, as we soon see, is completely symmetric.