最大熵原理的简要介绍

15220152202201

罗雨茵

生活中，我们总会遇到需要排队的时候，却很少遇到队伍长度相差很多的情况，这是因为每个人都会无意识地遵循最大熵原理。假设你去食堂打饭，在不知道阿姨给菜速度时，面对长度相差无几的队伍，你会随机选择一条；比你晚来一步的人则排除你所在的队伍，在剩下的队伍中随机选择一条，这样所有的队伍最终看起来仍是差不多长的。在这个过程中，你由于一无所知，不作未知假设，认为随机排一条队伍排到最快的这一事件是等概率的；而晚来一步的人将你的出现视为已知约束条件，再做了与你同样的思考。

这正是最大熵原理——在已有约束条件下，不再作任何未知假设，将未知事件视同等概率事件处理，这是最有效率的问题解决方式。

在信息论中，我们引入“信息熵”概念，把问题进一步量化，运用最大熵原理寻求解决问题的最优效率。首先，我们需要理解什么是信息，信息即通过传递从未知变成已知的事实。“这周末放假。”大家都知道周末放假，这不是一个信息。“这周末班级组织聚餐。”得知了班级对周末的安排，这才是信息。一个信息的信息量要怎么度量呢？“成绩很好的同学 A 保研北大了。”这看起来顺理成章；“成绩不算拔尖、经历也不算突出的同学 B 拿到了耶鲁的 Offer。”这带给人的冲击就比较大，也就是拥有较大的信息量。前者发生的概率较大，后者发生的概率较小，我们可以认为事件的发生概率与信息量呈反比关系。于是信息论男神克劳德・艾尔伍德・香农（Claude Elwood Shannon ）创造了“信息熵”，借用“熵”在热力学中用以度量体系混乱程度的物理意义，来描述信息的不确定程度，并赋予其数学意义以将一个事件的总信息量进行量化：

$$H = -\sum p(x) log p(x)$$

这里 $p(x)$ 表示事件每个可能性发生的概率，$log p(x)$ 表示每种可能性发生后的信息量，将它们相乘加总的和就是事件发生后总信息量的数学期望

——也就是信息熵。我们可以这样解释信息量的对数表达，多个事件的发生概率相乘即它们同时发生的概率，而多个事件的信息量加总即它们的总信息量。

有了这个公式，我们就能进一步证明为什么最大熵状态下的解决方式是最有效率的。在过去的学习中，效率问题往往就意味着极值问题，而存在约束条件的极值问题可以通过拉格朗日乘数法进行简单求解。在证明之前，我们先明确信息熵的单位。现代计算机科学用 bit 作为信息量的最小单位，因此我们把 log 的底数设为 2，上述式子变化为：

$$H(x) = -\sum_i^n p(x_i) log_2 p(x_i)$$

用拉格朗日乘数法求解 $maxH(x)$：

已知概率总 1 和为 1，$g(p_1, p_2, \ldots, p_n) = \sum_k^n p_k = 1$

设 $f(p_1, p_2, \ldots, p_n) = -\sum_{k=1}^n p_k log_2 p_k$，对于 $\forall k = 1, \ldots, n$，令

$$\partial/(\partial p_k)(f + \theta(g-1)) = 0$$

由此得到

$$\partial/(\partial p_k)(-\sum_{k=1}^n p_k log_2 p_k + \theta(\sum_{k=1}^n p_k - 1)) = 0$$

即

$$-(1/ln2 + log_2 p_k) + \theta = 0$$

可知 $p_1 = p_2 = \cdots = p_n$，$p_k = 1/n$。

即使用均匀分布时可得到最大熵的值。除此之外还有其它证法，本文就不加以赘述了。

参考文献

[1]dog250,2018,不知为不知--信息论和最大熵原则

[2] 扬子落木,2018,图解最大熵原理（The Maximum Entropy Principle）

[3]starINsky_mike,2011,熵最大定理两种理解