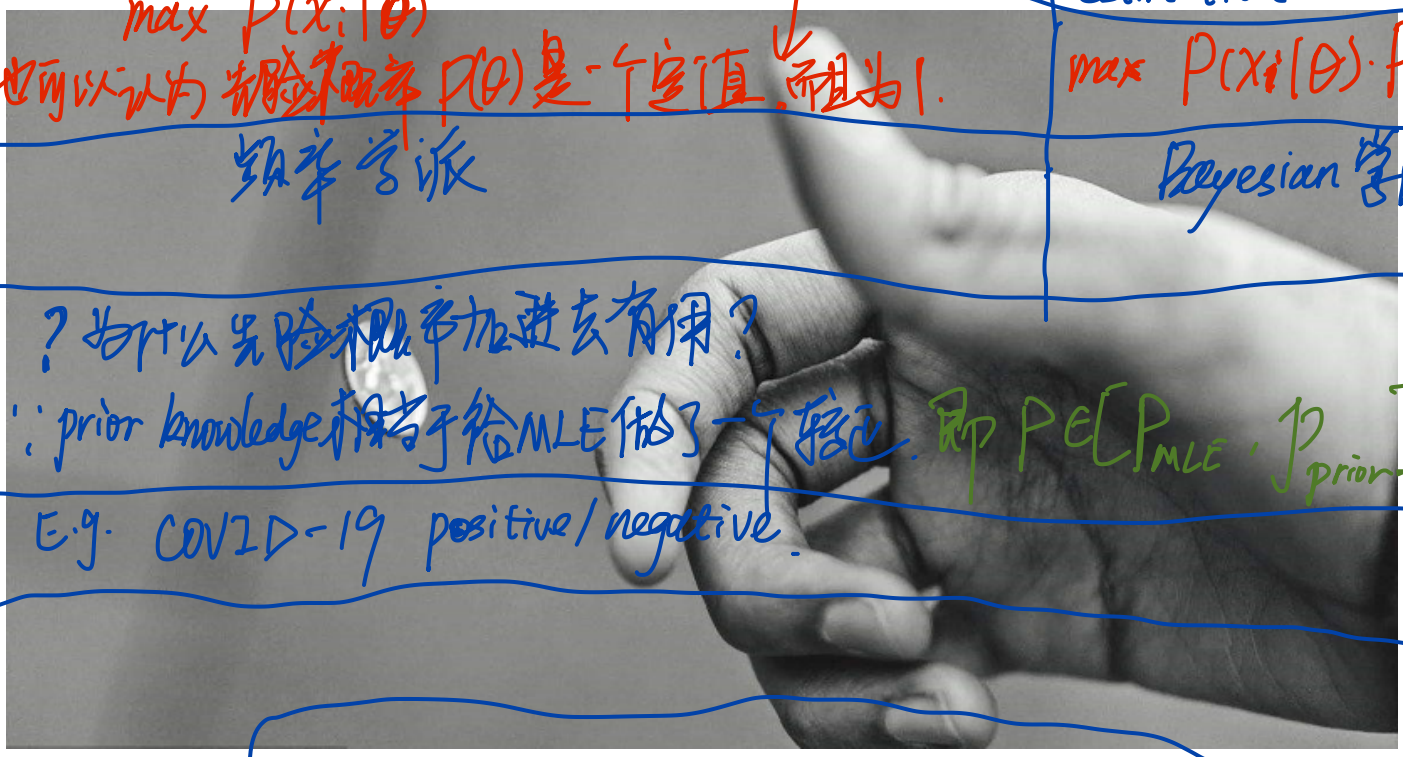
极大似然估计 —— Maximum Likelihood Estimation (MLE)



1 引入

机器学习中，经常会遇到极大似然估计 (Maximum Likelihood Estimation, MLE) 这个名词，它的含义是什么？它能够解决什么问题？我们该如何理解并使用它？本篇就对此进行详细的阐述和回答。

举一个最简单直观的例子，假设投掷硬币，我们每次投掷的结果只有两种：一正一反，古往今来，无数的实验和直觉告诉我们，投硬币这件事情正反两面的概率就是五五分，即正面概率 0.5，反面概率也是 0.5。

然而，我们怎么知道概率是 0.5 的呢？我们凭什么说就是 0.5，不可以是 0.55 或者 0.48 呢？因为有很多人做过这个实验，投多次硬币，比如投100次，大体上正反两面的次数总是都差不多，因此，我们就 **估计** 这个事情(投掷为正面)发生的概率为 0.5。

注意，上述的这个思维推理的过程很直觉化，我们人类很多时候对某件事情的判断其实就是这样：多次经历某件事情，比如买水果，我们每次都在看水果的外观来判断该水果是否新鲜可口，久而久之，我们就会发现，拍着清脆欲裂的西瓜有更大的概率会香甜可口，这其实也就是机器学习的基本世界观：**从经验到规律**。那么换成计算机解决，就是**数据 -> 规则 (Data -> Rule)**。

2 极大似然估计的概念

说了这么多，那么极大似然估计到底是什么呢？再来看看我们刚刚说的抛硬币的例子，其实我们并不知道一枚硬币抛出之后正面朝上的客观概率是多少，因为毕竟我们不是上帝，但是我们还是很想知道这个概率的大小，我们唯一的手段就是，做实验，从实验结果的数据中发现这个事件其中的规律。比如，我们抛掷 100 次，发现正面有 52 次，反面有 48 次。此时，这个结果就给我们判断正面的概率提供了一种依据，现在可能有很多人会立刻说："我知道了，根据这次实验的结果，正面的概率应该是 0.52！"说的没错，**这个论断的思维过程就是概率理论中我们最常看到的一个词："估计"**。

但是，我们是如何估计的？直觉上，100 次中有 52 次正面，因此我们估计正面概率为 0.52，这似乎很简单直观。但是如果从纯粹的数学理论角度去思考，我们应当给出一个完美的解释。为了能够解释好这种估计的方法，数学家提出了极大似然估计。

极大似然估计的哲学内涵就是：我们对某个事件发生的概率未知，但我们做了一些实验，有过一些对这个事件的经历(经验)，那么我们认为，这个事件的概率应该是能够与我们做的实验结果最吻合，当然，前提是我们做的实验次数应当足够多。如果只做一次实验，显然我们就会估计概率为 0 或 1 了。

3 计算

那么，这里的关键问题就是：**我们如何确定正面的概率 p_+ ，使得其能够与**

实际的实验结果吻合度最高？这里的事件只有两种结果，因此要求解的概率有两个，即：p_+ 和 p_- ，且由于

$$p_+ + p_- = 1$$

，所以我们只需要求解 p_+ 这一个概率即可。

将问题推广，对于某一实验，我们可能包含多种情况，其中每个实验结果的概率可记为一个集合 P：

$$P = \{p_1, p_2, p_3, \dots p_n\}$$

其中，

$$\sum_{i=1}^{n} p_i = 1$$

假设我们做了 m 次实验，则该实验结果出现的概率可计算如下：

$$E = \prod_{k=1}^{m} p_k \ , \ p_k \in P$$

例如抛硬币实验，如果我们做了 4 次实验，结果为：正、正、反、反，那么这样的实验出现的概率可计算为：

$$E = p_+ \times p_+ \times p_- \times p_- = p_+^2 \times p_-^2 = p_+^2 \times (1 - p_+)^2$$

现在回到我们在对极大似然估计含义的理解，如何确定正面的概率 p_+ ，使得其能够与实际的实验结果吻合度最高？其实把这句话说的在明白一些，就是：**如何确定正面的概率 p_+ ，使得我们此次实验结果的发生概率尽可能大？**

在转换为数学语言表达就是：求得一组

$$\{p_1, p_2, p_3, \dots p_n\}$$

，使得 E 最大化，数学表达即：

$$p_1^*, p_2^*, p_3^*, \ldots, p_n^* = \arg\max_{p_1, p_2, p_3, \ldots, p_n} E$$

> *这里，我的表达公式可能与标准的统计学书籍中的表达不一致，为的是通过这样的表达，使其内涵更简单地体现出来，完备的数学表达可参考相关的概率统计书籍。*

Eg, 对于我们刚刚做的 4 次抛硬币实验，即可记为：

$$p_+^* = \arg\max_{p_+} p_+^2 \cdot (1 - p_+)^2$$

这个计算似乎我们能搞定，令

$$x = p_+$$

，设：

$$f(x) = x^2(1-x)^2$$

则该问题就转化为求：在 x 取何值时，使得 $f(x)$ 最大?

该求解问题比较简单，由于

$$x \in [0, 1]$$

，且前后两项均为平方项，均为非负数，根据基本不等式：

$$ab \leq \left(\frac{a+b}{2}\right)^2$$

, 当且仅当

$$a = b$$

时取等号，则求得当

$$x^2 = (1-x)^2$$

即：

$$x = 0.5$$

时， $f(x)$ 取得最大值。

因此，在实验结果为“正、正、反、反”时，我们借助极大似然估计法求得

$$p_+ = 0.5$$

时，该实验结果发生的概率最大。

Eg 4 抓豆子实验

刚刚我们用 4 次抛硬币的实验解释了极大似然估计，由于该问题的计算十分简单，我们再举一个抓豆子的例子来具体说明一下极大似然估计的常用计算手法。

假设我们现在有一个麻袋，里面装了很多豆子，且豆子两种：红豆和绿豆，现在我们想知道这两种豆子各自占得比例是多少。显然我们不可能傻乎乎地一颗一颗的去数，懂概率理论的人会这样做：先把袋子中的豆子摇匀，然后随机地抽若干次豆子，记下抽取的豆子中红豆的个数和绿豆的个数，这样就能知道个大概情况了。

这其实也是极大似然估计的一个很实用地运用案例。现在我们假定抽取了 100 次豆子，其中有 70 个是红豆，30 个是绿豆。同样的，设红豆的比例 (或称抓得红豆的概率) 为 x ，则绿豆的比例为

$$1-x$$

，那么我们这次实验结果出现的概率为：

$$E = x^{70} \cdot (1-x)^{30}$$

, 其中

$$x \in [0,1]$$

可以看到，这次我们举的例子中指数项很大，无法像前面的那个例子中简单地求解，那么如何求得使得 E 最大时相应的 x 取值呢？我们可以使用一个数学中使用非常广泛的对数函数来协助解决，这里，我们设：

$$f(x) = \ln\left(x^{70} \cdot (1-x)^{30}\right)$$

我们都知道，函数

$$\ln(x)$$

是定义在

$$(0,+\infty)$$

上的单调增函数，因此加入

$$\ln(x)$$

函数对函数的极值求解并无影响。

根据对数计算公式

$$\log(ab)=\log a+\log b$$

，得：

$$f(x)=\ln(x^{70})+\ln\big((1-x)^{30}\big)$$

再根据对数计算公式

$$\log a^b=b\log a$$

，得：

$$f(x)=70\ln(x)+30\ln(1-x)$$

这样借助对数函数的数学变换，求解问题一下子变得简单很多。

现在，要求 $f(x)$ 的极值，我们可以先对其求导，借助

$$\ln'(x)=\frac{1}{x}$$

这一求导性质，得：

$$f'(x)=\frac{70}{x}-\frac{30}{1-x}$$

仔细观察可以发现，

$$x\in[0,1]$$

, 当 x 很小时

$$\frac{70}{x} > \frac{30}{1-x}$$

, 此时

$$f'(x) > 0$$

, 当 x 逐渐变大时，将会存在

$$\frac{70}{x} < \frac{30}{1-x}$$

, 即

$$f'(x) < 0$$

，因此，函数 $f(x)$ 是一个先上升后下降的函数形式，故当

$$f'(x) = 0$$

时， $f(x)$ 取得最大值。

令:

$$f'(x) = \frac{70}{x} - \frac{30}{1-x} = 0$$

解得:

$$70(1-x) = 30x$$

⇒

$$x = \frac{7}{10}$$

至此，我们利用极大似然估计求得了红豆的比例最可能为 0.7。原来一个我们平时经常直觉上判断出概率值的思维过程包含了这样一个完整的数学求证推理的过程。深入思考的人可能会发现，实际真实的红豆比例可能并非 0.7，的确，我们只能说很可能在 0.7 左右，因为极大似然估计方法本质上也是一种“估计”，既然叫做估计，肯定会存在偏差，但**该中估计策略的基本世界观应该就是我们目前最能够直接认可的一种方式，那就是通过历史经验总**

结出规律啊。 而MAP (Maximum A Posteriori estimation) 能借助人们对 prior knowledge/probability 减小统计出来的误差，很能靠谱这个或 base rate.