LN5. 概率估计 (MLE 和 MAP)

李钦宾

先进智能计算与系统团队

邮箱： *qinbin@hust.edu.cn*

2025 年 3 月

目录

目录

贝叶斯最优分类器

主要思想:

前面提到，如果知道 P(X,Y)，则可以使用贝叶斯最优分类器来预测 x 最有可能的标签，形式化为 $\arg\max_y P(Y \mid X)$。那么，是否可以直接从训练数据中估计 P(X,Y) 呢？

如果上述方法能有一个很好的近似，就可以在实践中使用贝叶斯最优分类器来估计 P(X,Y)。

许多监督学习可以被看作是估计 P(X,Y)。一般来说，它们可分为两类:

- 当估计 $P(X,Y) = P(X|Y)P(Y)$ 时，称为**生成学习** (**Generative learning**)。
- 当直接估计 $P(Y|X)$ 时，称为**判别学习** (**Discriminative learning**)。

那么如何估计样本的概率分布呢?

目录

简单场景：抛硬币的例子

问题

假设你发现了一枚古老而珍贵的硬币。你可能会问，“我若投掷这枚硬币，正面朝上的概率是多少?”投掷 $n = 10$ 次，得到如下的结果序列：$D = \{H, T, T, T, H, H, T, T, T, T\}$。基于这些样本，你如何估计 $P(H)$?

我们观察到 n_H=3 个正面，n_T=7 个反面。所以，凭直觉，我们有：

$$P(H) \approx \frac{n_H}{n_H + n_T} = \frac{3}{10} = 0.3$$

那么，能形式化地给出推导吗?

最大似然估计 (Maximum Likelihood Estimation, MLE)

MLE

刚才提到的估计器是最大似然估计 (Maximum Likelihood Estimation, MLE)。对于 MLE，通常分为两个步骤:

1） 对于数据采样于**什么样的分布**进行明确的建模假设。

2） 设置该分布的**参数**，使观察到的数据尽可能拟合。

回到抛硬币的例子:
关于抛硬币的一个天然假设是，所观察到结果的分布是二项分布 (Binomial distribution)。二项分布有两个参数，n 和 θ。它捕获了 n 个独立的伯努利随机事件的分布 (Bernoulli distribution)，其输出为正的概率为 θ。

在上面的例子中，n 是抛硬币的次数，设 θ 为硬币正面朝上的概率 (例如 $P(H) = \theta$)。
形式化地，二项分布被定义为:

$$P(D;\theta) = \binom{n_H + n_T}{n_H} \theta^{n_H}(1-\theta)^{n_T}$$

它计算了我们恰好观察到 n_H 个正面、n_T 个反面的概率。
(硬币被投掷 $n = n_H + n_T$ 次，每次得到正面的概率是 θ)

MLE 原理

找到 $\hat{\theta}$ 使数据出现的似然最大，$P(D;\theta)$:

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\ P(D;\theta)$$

如何求解该最大化问题

两步:
1. 代入分布的所有项，取函数的对数
2. 计算其导数，并使之等于零

取似然的 $\log$(通常称为对数似然) 不会改变它的最大值 (因为对数是单调函数，而似然为正)，但会将所有的乘积转化为求和，在求导时更容易处理。

求导等于零是求极值点的标准方法 (准确地说，应该通过验证二阶导数为负来验证它真的是极大值而不是极小值)

MLE 原理 (对数似然，log-likelihood)

回到二项分布，现在可以代入定义并计算对数似然:

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \underset{\theta}{\operatorname{argmax}}\ P(D;\theta) \\
&= \underset{\theta}{\operatorname{argmax}} \binom{n_H+n_T}{n_H} \theta^{n_H}(1-\theta)^{n_T} \\
&= \underset{\theta}{\operatorname{argmax}} \log \binom{n_H+n_T}{n_H} + n_H \cdot \log(\theta) + n_T \cdot \log(1-\theta) \\
&= \underset{\theta}{\operatorname{argmax}}\ n_H \cdot \log(\theta) + n_T \cdot \log(1-\theta)
\end{aligned}
$$

MLE 原理（log-likelihood）

然后我们就可以通过求导并使它等于 0 来求出 θ。其结果是：

$$\frac{n_H}{\theta} = \frac{n_T}{1-\theta} \Longrightarrow n_H - n_H\theta = n_T\theta \Longrightarrow \theta = \frac{n_H}{n_H + n_T}$$

可以发现，$\theta \in [0,1]$。

- MLE 给出了所观察到数据的合理解释。
- 若 n 很大，且你的模型/分布是正确的 (即 H 包含真实的模型)，那么 MLE 会找到真实的参数。
- 若 n 很小，MLE 会过度拟合数据。当 n 很大时，它工作得很好。
- 如果你没有正确的模型 (且 n 很小)，那么 MLE 可能是非常错误的!

例如，假设你观察到：H, H, H, H, H。$\hat{\theta}_{MLE}$ 是多少?

目录

简单场景：有先验知识的抛硬币

假设你有一种直觉的猜测，θ 应该近似等于 0.5。但你的样本量很小，所以你并不太相信从样本得到的这个估计。

简单修正

加上 $2m$ 次的假想抛掷，其结果应是 θ'(例如，$\theta' = 0.5$)。添加 m 次正面和 m 次反面到你的数据。

$$\hat{\theta} = \frac{n_H + \mathbf{m}}{n_H + n_T + \mathbf{2m}}$$

对于较大的 n，这是一个微不足道的变化。
对于很小的 n，它包含了你关于 θ 应该是什么的“先验信念”。

我们能给出形式化的推导吗？

贝叶斯方法

将 θ 建模为一个随机变量，从分布 P(θ) 中采样。
注意 θ 不是与样本空间中的事件相关的随机变量。

在频率派统计中，这是不允许的。
在贝叶斯统计中，这是允许的，可以指定一个先验 $P(\theta)$ 来定义你认为 θ 可能会取什么值。

由贝叶斯法则，可以得出:

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{P(D)}$$

其中:

- P(θ) 是参数 θ 的先验分布（在我们看到任何数据之前）
- $P(D|\theta)$ 是给定参数 θ 时数据 D 的似然
- $P(\theta|D)$ 是我们观察到数据 D 后，参数 θ 的后验分布

贝叶斯方法

先验 P(θ) 的一个天然选择是 Beta 分布:

$$P(\theta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}$$

其中 $B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ 是归一化常数 (为了确保所有概率之和为 1)

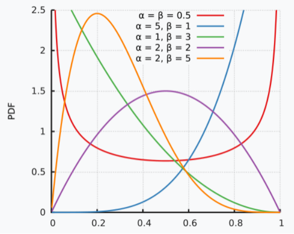


图: Probability density function

注意，这里只需要二元随机变量 θ 的分布。
二项分布的共轭先验是 Beta 分布。通俗来讲，即 Beta 分布描述了二项分布中 θ 取值的可能性。
(Beta 分布的多元泛化是狄利克雷 (Dirichlet) 分布。
多项分布的共轭先验是狄利克雷分布。)

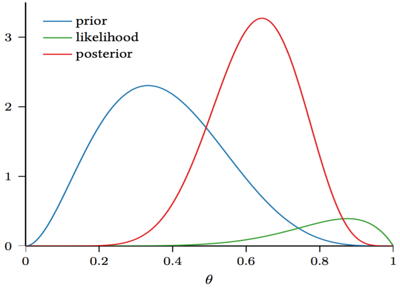
贝叶斯方法

为什么 Beta 分布很适合?

- 它对概率 $\theta \in [0,1]$ 进行建模
- 它与二项分布属于同一个分布族 (共轭先验)，从数学上可以得到很好的结果:

$$P(\theta \mid D) \propto P(D \mid \theta)P(\theta) \propto \theta^{n_H+\alpha-1}(1-\theta)^{n_T+\beta-1}$$

到目前为止，我们得到了 θ 的分布。接下来，我们如何得到 θ 的估计值?



最大后验估计 (Maximum a Posteriori Probability Estimation, MAP)

我们可以选择 $\hat{\theta}$ 作为给定数据的最有可能的 θ。

MAP 原理：寻找一个 $\hat{\theta}$ 使后验分布 $P(\theta \mid D)$ 最大化

$$\begin{aligned}\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}}\ P(\theta \mid D) \\ &= \underset{\theta}{\operatorname{argmax}}\ \log P(D \mid \theta) + \log P(\theta)\end{aligned}$$

最大后验概率估计 (MAP)

对于我们的抛硬币场景，可以得到:

$$
\begin{aligned}
\hat{\theta}_{MAP} &= \underset{\theta}{\operatorname{argmax}}\ P(\theta|D) \\
&= \underset{\theta}{\operatorname{argmax}}\ \frac{P(D|\theta)P(\theta)}{P(D)} && \text{(By Bayes rule)} \\
&= \underset{\theta}{\operatorname{argmax}}\ \log(P(D|\theta)) + \log(P(\theta)) \\
&= \underset{\theta}{\operatorname{argmax}}\ n_H \cdot \log(\theta) + n_T \cdot \log(1-\theta) + (\alpha - 1) \cdot \log(\theta) + (\beta - 1) \cdot \log(1-\theta) \\
&= \underset{\theta}{\operatorname{argmax}}\ (n_H + \alpha - 1) \cdot \log(\theta) + (n_T + \beta - 1) \cdot \log(1-\theta) \\
&\Longrightarrow \hat{\theta}_{MAP} = \frac{n_H + \alpha - 1}{n_H + n_T + \beta + \alpha - 2}
\end{aligned}
$$

最大后验估计 (MAP)

讨论

- MAP 估计与有 $\alpha - 1$ 个虚拟正面和 $\beta - 1$ 个虚拟反面的 MLE 相同。
- 当 $n \to \infty$ 时，$\hat{\theta}_{MAP} \to \hat{\theta}_{MLE}$。因为 $\alpha - 1$ 和 $\beta - 1$ 与非常大的 n_H, n_T 相比变得无关紧要。
- 如果有一个准确的先验 (并且在数学上是可处理的)，则 MAP 是一个很好的估计器。
- 如果 n 很小，而先验是错误的，MAP 可能是非常错误的!

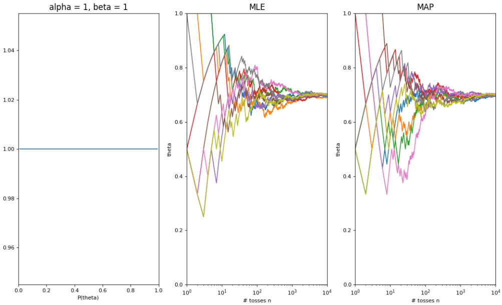
MLE 和 MAP 演示

MLE & MAP 总结

- $\hat{\theta}_{MLE} = \frac{n_H}{n_H + n_T}$, $\hat{\theta}_{MAP} = \frac{n_H + \alpha - 1}{n_H + n_T + \beta + \alpha - 2}$
- MAP 估计与有 $\alpha - 1$ 个虚拟正面和 $\beta - 1$ 个虚拟反面的 MLE 相同

MLE & MAP Demo

- $\alpha = 1, \beta = 1,\ \theta = 0.7$



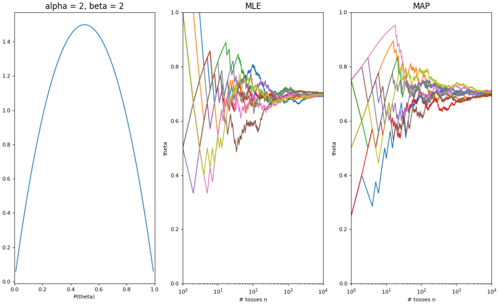
MLE 和 MAP 演示

MLE & MAP 总结

- $\hat{\theta}_{MLE} = \frac{n_H}{n_H+n_T}$, $\quad \hat{\theta}_{MAP} = \frac{n_H+\alpha-1}{n_H+n_T+\beta+\alpha-2}$
- MAP 估计与有 $\alpha-1$ 个虚拟正面和 $\beta-1$ 个虚拟反面的 MLE 相同

MLE & MAP Demo

- $\alpha=2, \beta=2, \theta=0.7$



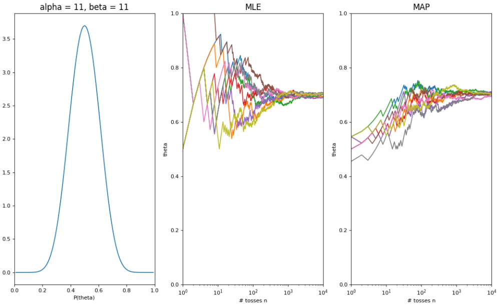
MLE 和 MAP 演示

MLE & MAP 总结

- $\hat{\theta}_{MLE} = \frac{n_H}{n_H+n_T}$, $\hat{\theta}_{MAP} = \frac{n_H+\alpha-1}{n_H+n_T+\beta+\alpha-2}$
- MAP 估计与有 $\alpha-1$ 个虚拟正面和 $\beta-1$ 个虚拟反面的 MLE 相同

MLE & MAP Demo

- $\alpha = 11, \beta = 11, \theta = 0.7$

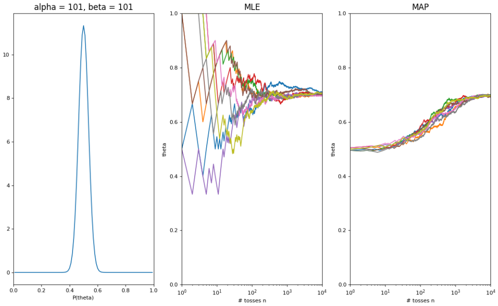


MLE & MAP 总结

- $\hat{\theta}_{MLE} = \frac{n_H}{n_H+n_T}$, $\hat{\theta}_{MAP} = \frac{n_H+\alpha-1}{n_H+n_T+\beta+\alpha-2}$
- MAP 估计与有 $\alpha-1$ 个虚拟正面和 $\beta-1$ 个虚拟反面的 MLE 相同

MLE & MAP Demo

- $\alpha = 101, \beta = 101, \theta = 0.7$



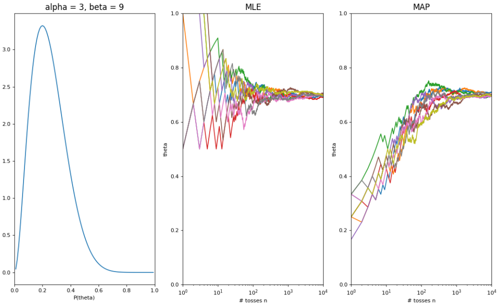
MLE 和 MAP 演示

MLE & MAP 总结

- $\hat{\theta}_{MLE} = \frac{n_H}{n_H+n_T}$, $\quad \hat{\theta}_{MAP} = \frac{n_H+\alpha-1}{n_H+n_T+\beta+\alpha-2}$
- MAP 估计与有 $\alpha - 1$ 个虚拟正面和 $\beta - 1$ 个虚拟反面的 MLE 相同

MLE & MAP Demo

- $\alpha = 3, \beta = 9, \theta = 0.7$



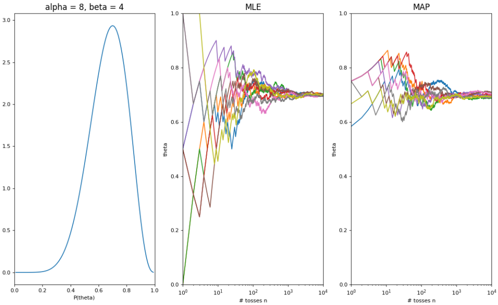
MLE 和 MAP 演示

MLE & MAP 总结

- $\hat{\theta}_{MLE} = \frac{n_H}{n_H+n_T}$, $\hat{\theta}_{MAP} = \frac{n_H+\alpha-1}{n_H+n_T+\beta+\alpha-2}$
- MAP 估计与有 $\alpha-1$ 个虚拟正面和 $\beta-1$ 个虚拟反面的 MLE 相同

MLE & MAP Demo

- $\alpha = 8, \beta = 4,\ \theta = 0.7$



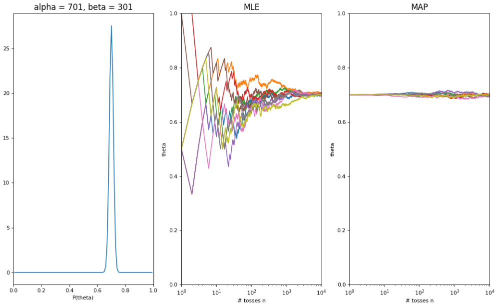
MLE 和 MAP 演示

MLE & MAP 总结

- $\hat{\theta}_{MLE} = \frac{n_H}{n_H+n_T}$, $\hat{\theta}_{MAP} = \frac{n_H+\alpha-1}{n_H+n_T+\beta+\alpha-2}$
- MAP 估计与有 $\alpha - 1$ 个虚拟正面和 $\beta - 1$ 个虚拟反面的 MLE 相同

MLE & MAP Demo

- $\alpha = 701, \beta = 301$, $\theta = 0.7$



目录

用贝叶斯法预测

注意 MAP 只是获得 estimator 的一种方法。在 $P(\theta \mid D)$ 中有更多的信息，简单地按此模式计算而丢弃所有其他信息似乎是不对的。

全贝叶斯法是直接使用后验预测分布对具有特征 X 的测试样本的标签 Y 进行预测:

$$P(Y \mid D, X) = \int_{\theta} P(Y, \theta \mid D, X) d\theta = \int_{\theta} P(Y \mid \theta, D, X) P(\theta|D) d\theta$$

遗憾的是，上述问题在闭式下通常难以解决，采样技术 (如蒙特卡罗 Monte Carlo 近似) 被用来对分布进行近似。

不过，在“高斯过程”对应的设定下，上式是可以求解的。我们后续有时间会讲到。

用贝叶斯法预测

另一个例外是抛硬币的例子。要在上述抛硬币的示例中使用 θ 进行预测，我们可以使用

$$\begin{aligned} P(heads \mid D) &= \int_\theta P(heads, \theta \mid D) d\theta \\ &= \int_\theta P(heads \mid \theta, D) P(\theta \mid D) d\theta \\ &= \int_\theta \theta P(\theta \mid D) d\theta \\ &= E[\theta|D] \\ &= \frac{n_H + \alpha}{n_H + \alpha + n_T + \beta} \end{aligned}$$

(链式法则: $P(A, B|C) = P(A|B, C)P(B|C)$.)

在这里，我们使用了定义 $P(heads \mid D, \theta) = P(heads \mid \theta) = \theta$ 的事实 (这只是因为我们假设数据来自二项分布。一般来说这是不成立的)。

目录

在监督学习中，我们有训练数据 D。使用这些数据来训练一个模型，该模型由参数 θ 表示。然后利用该模型，对测试点 x_t 进行预测。

- **MLE 预测**:
  $P(y|x_t;\theta)$: $\theta = \mathrm{argmax}_\theta\, P(D;\theta)$。这里 θ 纯粹是一个模型参数。
- **MAP 预测**:
  $P(y|x_t;\theta)$: $\theta = \mathrm{argmax}_\theta\, P(\theta|D) \propto P(D \mid \theta)P(\theta)$。这里 θ 是一个随机变量。
- **全贝叶斯法**预测: $P(y|x_t, D) = \int_\theta P(y|\theta)P(\theta|D)d\theta$. 这里 θ 被用于积分。即我们的预测考虑了所有可能的模型。

注意，它们之间的差异是很微妙的：
在 MLE 中，我们最大化 $\log[P(D;\theta)]$，
在 MAP 中，我们最大化 $\log\left[P(D|\theta)\right] + \log\left[P(\theta)\right]$.
所以本质上在 MAP 中，我们只是在优化中添加了 $\log\left[P(\theta)\right]$。这一项与数据无关，如果参数 θ 偏离合理值过大，就会受到惩罚。

我们稍后将把它作为正则项 (regulation) 的一种形式进行讨论，其中 $\log\left[P(\theta)\right]$ 将被解释为分类器复杂性的度量。

The End