MLE与MAP

1. MLE for Discrete Random Variable

假设我们有训练集X，设其为离散随机变量，且其可能的取值为$\{x_1, ..., x_n\}$。我们另有模型$f(x;\theta)$，且$P(x \mid \theta) = f(x;\theta)$，同时我们假设样本是$i.i.d.$的。MLE是一个用来估计参数$\theta$取值的方法，它旨在寻找一个最优的$\theta_{MLE}$，使得$P(X \mid \theta_{MLE})$最大化。也就是说，在$\theta_{MLE}$参数化下，训练集$X$出现的概率会是最大的。

具体来说，当X为离散随机变量时，我们有似然函数：

$$
\begin{aligned}
L(\theta) &= P(X \mid \theta) \\
&= P(x_1 \mid \theta) \cdot P(x_2 \mid \theta) \cdots P(x_n \mid \theta) \\
&= \prod_{i=1}^{n} P(x_i \mid \theta).
\end{aligned} \tag{1}
$$

显然，

$$
\begin{aligned}
\theta_{MLE} &= \operatorname{argmax}_\theta L(\theta) \\
&= \operatorname{argmax}_\theta \log L(\theta) \\
&= \operatorname{argmin}_\theta - \log L(\theta) \\
&= \operatorname{argmin}_\theta - \sum_{i=1}^{n} \log P(x_i \mid \theta).
\end{aligned} \tag{2}
$$

当X为连续随机变量时，这篇blog用一个很好的角度将离散/连续两种情况联系在了一起。

这里提一句，为什么我们将$L(\theta) = P(X \mid \theta)$叫作Likelihood，而不是Probability呢？其实两个不同的叫法只是为了区分出哪一个是$P(X \mid \theta)$中的变量。

1. 给定一个$P(X \mid \theta)$，如果我们称之为Probability，那么我们可以推出，其中θ已知，而X是变量，$P(X \mid \theta)$是一个关于X的函数。在这个称呼下，我们感兴趣的问题是：给定θ后，对任意的X，得到其出现的概率。
2. 给定一个$P(X \mid \theta)$，如果我们称之为Likelihood，那么我们可以推出，其中X已知，而θ是变量，$P(X \mid \theta)$是一个关于θ的函数。在这个称呼下，我们感兴趣的问题是：给定X，寻找一个θ，使得在该θ下，X出现的概率最大。

2. MAP for Discrete Random Variable

MLE通过优化参数θ，使得训练集X出现的概率最大化。但有些情况下我们会更加希望加入人工先验知识。

举个例子，如果我们掷硬币10次，结果全都是正面。我们若认为该实验是一个以θ参数化的伯努利分布，那么MLE就会得出$\theta = 1$的结论。这种估计有可能是过激的。有可能这个硬币本身是均匀的，只是恰好这10次实验全是正面而已。如果能在估计中加入先验知识，那么我们估计的结果可能会更鲁棒。

MAP则是一种在估计中嵌入先验知识的方法。具体来说，若我们假设θ也是一个随机变量，同时它服从某个先验分布，我们可以令$\theta_{MAP} = \operatorname{argmax}_\theta P(\theta \mid X)$。也就是说，给定$\theta$的先验分布和$X$后，我们希望找到在这个情况下概率最大的取值$\theta_{MAP}$。

显然，在$P(\theta \mid X)$中，X是已观测到的随机变量，θ是未观测到的随机变量，因此我们无法直接对其进行计算。所以我们这里应用贝叶斯公式来规避$P(\theta \mid X)$的计算，具体来说，

$$
\begin{aligned}
\theta_{MAP} &= \operatorname{argmax}_\theta P(\theta \mid X) \\
&= \operatorname{argmax}_\theta \frac{P(X \mid \theta)P(\theta)}{P(X)} && \text{贝叶斯公} \\
&= \operatorname{argmax}_\theta P(X \mid \theta)P(\theta) && P(X)\text{与}\theta\text{无} \\
&= \operatorname{argmax}_\theta P(\theta)\prod_{i=1}^{n} P(x_i \mid \theta) \\
&= \operatorname{argmax}_\theta \log\left(P(\theta)\prod_{i=1}^{n} P(x_i \mid \theta)\right) \\
&= \operatorname{argmax}_\theta \left(\log\left(P(\theta)\right) + \sum_{i=1}^{n} \log\left(P(x_i \mid \theta)\right)\right) \\
&= \operatorname{argmin}_\theta \left(-\log\left(P(\theta)\right) - \sum_{i=1}^{n} \log\left(P(x_i \mid \theta)\right)\right).
\end{aligned}
$$

因此，我们可以通过预设θ的先验分布$P(\theta)$，来嵌入先验知识。

比如，对于上面掷硬币的例子，我们可以假设θ服从$Bernoulli$分布的共轭先验分布$Beta$分布，即$\theta \sim Beta(\alpha, \beta)$，从而使得估计更加鲁棒。当然这里也可以选择其它的先验分布，只是如果我们采用共轭先验分布的话，可以使得后验分布$P(\theta \mid X)$与先验分布$P(\theta)$有着相同的数学形式，在计算和理解上带来方便。

MAP中两个值得注意的地方：

1. 若假设先验分布θ服从均匀分布，则$P(\theta)$为常数。此时θ_{MAP}= θ_{MLE}，即MAP等价于MLE。

2. 从机器学习的角度来看，MAP可以看作是MLE增加了一个关于参数的先验分布的正则项$\log\left(P(\theta)\right)$。

References

[1] Link1

[2] Link2

[3] Link3

[4] Link4