生成模型读书笔记二

2021年1月10日 22:35

1. **最大似然估计Maximum likelihood estimation (MLE)**
   a. 似然函数(likelihood function)
      似然函数，p(x; θ)给出了样本x在以θ为参数的分布下的概率。其实就是我们常说的概率密度函数，不过函数中的参数θ是未知的。
   b. MLE算法
      首先假设一个分布，其中含有未知的参数θ，利用已知的样本数据，反推最有可能产生出这样数据的模型参数。
      于是我们现在有一个分布p(x; θ)，它的参数为θ，假设数据样本都是独立同分布的，把所有样本代进p，因为想要发生这个采样结果的可能性最大，且有样本间独立的前提，观测样本集X发生的概率就是每个xi发生的概率相乘，所以求解θ = argmax$_\theta$ p(X;θ)=arg$\max_\theta \prod_i p(x_i;\theta)$，通常会计算log likelihood把累乘变成求和，即argmax$_\theta$ $\sum \log p(x_i,\theta)$
   c. 两个例子
      i. 离散分布
         有一个重量不均匀的硬币，抛出正面的概率为p，抛出反面的概率为1-p，现在做了N次实验，抛出了a次正面，b次反面，求p。
         这其实就是一个参数为p的伯努利分布，p(正面)=p，p(反面)=1-p，写出最大似然公式有
         $$\begin{aligned}\Theta &= \text{argmax}_\theta \sum \log p(x_i,\theta)\\ &= argmax_\theta\ a\ * logp + b\ * \log(1-p)\end{aligned}$$
         因为要求最大，所以对上式求导，使导数等于0，最后可以求得p=$\frac{a}{a+b}$
      ii. 连续分布
         比如高斯分布，有两个参数均值μ和标准差σ，详细例子就不举了，步骤就是把概率密度函数写出来，代入最大似然公式，对每个参数求导，使导数等于0，求出取得最大值时的参数。
   d. 与贝叶斯公式的联系
      $$f(x|y)=\frac{f(x,y)}{f(y)}=\frac{f(y|x)\,f(x)}{f(y)}$$
      将数据样本(证据)看作贝叶斯公式中的y，将θ看作公式中的x，似然函数就是p(y|x)，最大似然也就是找可以使得p(y|x)最大的x，但是我们不知道p(y|x)的形式(或者说因为θ未知，所以p(y|x)有无数多个)，所以来分析下从贝叶斯公式我们可以怎么求。
      因为数据样本是已知的，所以公式中p(y)确定，然后因为我们对θ没有任何已知的知识，所以假设θ是在整个解空间中均匀分布的，任意一个θ成为解的可能性都是相同的，那么p(x)是常数，这么一来，公式右边的p(x)，p(y)都是常数，就有p(x|y) = p(y|x)*c, 想要求max p(y|x)就相当于求max p(x|y)，也就是上文的p(θ|x)，求最大似然也就等价于求后验分布最大

e. 与交叉熵，KL的关系

Log likelihood的式子：

$$\theta = \arg\max_{\theta} \sum_{i=1}^{n} \log p_\theta(X_i)$$

如果右边的式子除以n，会有什么事情发生呢？

根据数据样本来自于真实分布这个前提，第i个样本X_i在原分布的概率就是1/n，也就是p(X_i)(不考虑X_i的值的情况下，考虑Xi的值的话就更好理解了，比如上文例子中的伯努利分布p，正样本个数a, 负样本个数b，最大似然式子就是a个正样本的概率和b个负样本的概率相加，式子可以写成 $a * \log p_\theta(X^+) + b * \log p_\theta(X^-)$，除以n之后，从采样估计的观点看，a/n就近似是原分布中正样本的概率p，b/n就可以当作1-p，可以自己推一下伯努利分布和高斯分布从最大似然到交叉熵的推导)，所以在除以n之后就可以写成

$\sum p(x_i)\log p_\theta(x_i)$

其中p是真实的数据分布，p_θ是要估计的，相当于当x来自分布p时$\log p_\theta(x)$的期望E[$\log p_\theta(x)$]，这不就是p和p_θ交叉熵差个负号嘛，所以最大化似然其实就是最小化交叉熵。

$$\theta = \arg\max_{\theta} \mathcal{L}(\theta) = \arg\max_{\theta} \mathbb{E}\left[\log p_\theta(X_i)\right]$$

更进一步，因为交叉熵等于原分布的熵加KL散度，原分布的熵是确定的，所以最大化似然⇒最小化交叉熵⇒最小化KL散度，三者都是相通的。

$$\mathrm{H}(p,q) = \mathrm{E}_p[-\log(q)] = \mathrm{H}(p) + D_{\mathrm{KL}}(p \parallel q).$$

不同点在于，最大似然是基于真实分布p(x)已知这个假设(样本数据能够真实完整地表示真实分布，比如最常见的分类任务就是用交叉熵损失，认为样本的统计概率就是真实分布概率)，直接算交叉熵式子就可以了，从而有观点认为KL散度更多地用于当真实分布未知或者只知道部分的情况 (后面会说一下这种观点，因为从最大似然也可以算部分未知的情况，只是迂回一点)。

如果真实分布p确实在我们预设的分布族p_θ里，那么解上述优化问题是可以完美地还原真实分布的。而在实际中，我们所知道的只是一组采样数据，所以也只能是真实分布的一个近似，叫做经验分布，将真实分布的概率都放在采样到的样本点上。

f. 采样估计

假设要计算一个连续分布的均值，而它的积分又比较难算的话

$$\mathbb{E}[x] = \int x p(x) dx$$

有两种方法，一是直接按照积分的定义进行数值估计，在x轴上取有代表性的n个点，x1<x2<...<xn，这个方法比较困难

$$\mathbb{E}[x] \approx \sum_{i=1}^{n} x_i p(x_i)\,(x_i - x_{i-1})$$

二是进行采样估计，从p中采样n个点，两种方法的差别在于，采样估计不需要计算p(x)，因为样本是依概率采样出来的，概率大的x被采样到的次数也多，所以概率已经被

包含在采样操作中了，这也是蒙特卡洛模拟的基础

$$\mathbb{E}[x] \approx \frac{1}{n}\sum_{i=1}^{n} x_i, \quad x_i \sim p(x)$$

从数据推断模型，大多都会使用能写成期望的形式，因为这样就可以使用采样计算了。

2. Maximum A Posterior(MAP)

a. 考虑知道先验概率p(θ)的情况下

这时重新分析贝叶斯公式，如果先验p(x)已知，也就是不同θ成为解的可能性是不同的，p(θ)就不能看成常数，最大化的式子就变成了max p(x|θ)*p(θ)，这就是MAP，给定了观测值后使后验概率最大。

b. MAP的式子

$$\begin{aligned}\arg\max_{\theta} p(x|\theta)\cdot p(\theta) &= \arg\max_{\theta} log\prod_{i=0}^{n} p(x_i|\theta)p(\theta) \\ &= \arg\max_{\theta} \sum_i log(p(x_i|\theta)p(\theta)) \\ &= \arg\max_{\theta} \sum_i log(p(x_i|\theta) + log(p(\theta))\end{aligned}$$

经过推导可以求得上式，最终要最大化的有两项，第一项和MLE一样，第二项就是θ的先验

c. 共轭先验

从贝叶斯公式我们知道，后验分布正比于似然函数和先验分布的乘积。似然函数是我们假设的数据分布，先验也是我们自己定义的，对于先验，选择规则是，选择一个可以使得计算出来的后验拥有与先验分布相同函数形式的，这样的先验叫做似然函数的共轭先验，先验和后验称为共轭分布。共轭先验形式优美，可以极大地简化计算。常见的: 二项分布的共轭先验是Beta分布，多项式分布的共轭先验是dirichlet分布，高斯分布的共轭先验是另一个高斯分布。任何指数族分布来说，都存在一个共轭先验。

d. 与MLE的关系

这个还是有点缕不顺，不太清楚，暂且这么理解：从推断参数这个问题上来说，我们知道数据样本x，求模型参数θ，从贝叶斯公式有

$$P(\theta|x) = \frac{p(x|\theta) * p(\theta)}{p(x)}$$

其实就是在已知证据x的情况下求最有可能的参数θ，即后验分布p(θ|x)最大，MLE和MAP都是在完成这个任务。从贝叶斯公式我们可以知道，在有数据样本(p(x)固定，看作常量)的情况下，后验分布正比于先验和似然函数的乘积，不过在MLE中，问题简化为先验是均匀分布，也可以叫做无信息先验，所以p(θ)也成了常量，求后验分布最大就是在求最大似然，所以这个方法叫最大似然估计，而MAP是一个更一般的情况，先验p(θ)适用范围更广，而不只是均匀分布，最大化后验分布时需要将先验考虑进来，优化公式中右边的式子，所以这个叫最大后验，MLE相当于它的一个特例。