最大似然估计与最大后验估计

0. 参考资料

主要参考如下：

周志华《机器学习》第七章

CSDN@nebulaf91 详解最大似然估计（MLE）、最大后验概率估计（MAP），以及贝叶斯公式的理解

CSDN@段子手实习生 最大似然估计和最大后验估计（转）

1. 似然函数

在统计学中，似然函数和概率函数是两个不同的概念。

对于这个函数：

$$P(x|\theta)$$

x表示一个具体的数据，θ表示模型的参数。

如果θ是已知确定的，x是变量，这个函数就叫做概率函数，它描述在给定θ的条件下，对于不同的样本点x，其出现概率是多少。

如果x是已知确定的，θ是变量，这个函数叫做似然函数，它描述对于不同的模型参数，出现x这个样本点的概率是多少。

2. 最大似然估计

最大似然估计（maximum likelihood estimation，MLE）是频率学派使用的参数估计方法。

回顾一下贝叶斯公式：

$$p(\theta|X) = \frac{p(X|\theta) \cdot p(\theta)}{p(X)}$$

它代表的意义是：

$$posterior = \frac{likelihood \times prior}{evidence}$$

顾名思义，最大似然就是要用似然函数取到最大值时的参数值作为估计值，似然函数可以写做：

$$L(\theta|X) = p(X|\theta) = \prod_{x \in X} p(X = x|\theta)$$

连乘操作容易造成数值的下溢，通常对似然函数取对数，即最大化对数似然。于是最大似然估计问题可以写成：

$$\hat{\theta}_{ML} = \text{argmax}_\theta \, L(\theta|X) = \text{argmax}_\theta \sum_{x \in X} \log p(x|\theta)$$

3. 最大后验估计

最大后验估计（maximum a posterior estimation，MAP）与最大似然估计相似，不同点在于加入了先验分布$P(\theta)$。MAP是贝叶斯学派常用的参数估计方法。

根据贝叶斯公式有：

$$\begin{aligned}\hat{\theta}_{MAP} &= \operatorname{argmax}_\theta \frac{p(X|\theta)p(\theta)}{p(X)} \\ &= \operatorname{argmax}_\theta p(X|\theta)p(\theta) \\ &= \operatorname{argmaxo}\{L(\theta|X) + \log p(\theta)\} \\ &= \operatorname{argmax}_\theta \left\{\sum_{x\in X} \log p(x|\theta) + \log p(\theta)\right\}\end{aligned}$$

注意这里$p(X)$与参数θ无关，因此等价于要使分子最大。直观上，MAP求得的参数θ不仅要使似然函数大，θ自己出现的先验概率也要大（有点类似正则化的思想）。

先验分布的参数称为超参数，即

$$p(\theta) = p(\theta|\alpha)$$

求得参数后，给定观测样本数据，一个新的样本值$\tilde{x}$发生的概率是

$$p(\tilde{x}|X) = \int_{\theta\in\Theta} p(\tilde{x}|\hat{\theta}_{MAP})p(\theta|X)d\theta = p(\tilde{x}|\hat{\theta}_{MAP})$$

4. 区别与联系

当先验分布是均匀分布时，MAP和MLE是等价的，运用MAP实际上就是使用先验概率对似然函数进行修正，MLE可以被看作是MAP的特殊情形。

先验服从均匀分布时，被称为无信息先验（non-informative prior），通俗的说就是“让数据自己说话”，此时贝叶斯方法等同于频率方法。

随着数据的增加，先验的作用越来越弱，数据的作用越来越强，参数的分布会向着最大似然估计靠拢。而且可以证明，最大后验估计的结果是先验和最大似然估计的凸组合。