MLAPP 读书笔记 - 05 贝叶斯统计 (Bayesian statistics)

> A Chinese Notes of MLAPP，MLAPP 中文笔记项目
> https://zhuanlan.zhihu.com/python-kivy

记笔记的人：cycleuser

2018年05月29日12:22:15

5.1 概论

之前咱们已经看到过很多不同类型的概率模型了,然后也讲了如何用这些模型拟合数据,比如讲到了如何去进行最大后验估计(MAP)参数$\hat{\theta} = \arg\max p(\theta|D)$,使用各种不同先验等等.还讲了全后验(full posterior)$p(\theta|D)$,以及一些特定情境下的后验预测密度(posterior predictive density) $p(x|D)$(后面的章节中会介绍更多通用算法).

对未知变量,使用后验分布来总结我们知道的信息,这就是贝叶斯统计学的核心思想.在本章要对这个统计学上的方法进行更详细讲解,在本书第六章,也就是下一章,会讲解另一种统计方法,也就是频率论统计,或者也叫做经典统计(frequentist or classical statistics).

5.2 总结后验分布

后验分布$p(\theta|D)$总结了我们关于未知量θ的全部信息.在本节我们会讨论一些从一个概率分布,比如后验等,能推出来的简单量.这些总结性质的统计量通常都比全面的联合分布(full joint)更好理解也更容易可视化.

5.2.1 最大后验估计(MAP estimation)

通过计算后验均值(mean)/中位数(median)或者众数(mode)我们可以很容易地对一个未知量进行点估计(point
estimate).在本书5.7,会讲到如何使用决策理论(decision theory)在这些不同方法之间进行选择.通常后验均值或者中位数最适合用于对实数值变量的估计,而后验边缘分布(posterior marginals)的向量最适合对离散向量进行估计.不过后验模,也就是最大后验估计(MAP estimate)是最流行的,因为这将问题降低到了一个优化问题(optimization problem),这样就经常能有很多高效率的算法.另外,最大后验分布还可以用非贝叶斯方式(non-Bayesian terms)来阐述,将对数先验(log prior)看作是

正则化工具(regularizer)(更多内容参考本书6.5).

虽然最大后验估计的方法在计算上很有吸引力,但还是有很多缺陷的,下面就要详细讲一讲.这也为对贝叶斯方法的详尽理解提供了动力,本章后文以及本书其他地方会讲到.

此处参考原书图5.1

5.2.1.1 无法测量不确定度

最大后验估计(MAP estimation)最明显的一个缺陷就在于没有对不确定性提供任何量度,其他的各种点估计比如后验均值或者中位数也都有这个问题.在很多应用中,都需要知道对一个估计值到底能信任多少.对于后验置信度的度量可以推导出来,在本书5.2.2中有.

5.2.1.2 可能导致过拟合

最大后验分布估计中进行插值可能导致过拟合.在机器学习里面,对于预测准确性往往比模型参数的可解释性更看重.不过如果对参数的不确定度不能建模的话,就可能会导致预测分布过分有信心(overconfident).在本书第三章有若干这类例子,后面还会看到更多.在风险敏感的情况下,对预测过分置信就可能很是个问题了,具体参考本书5.7.

5.2.1.3 众数(mode)是非典型点(untypical point)

选择众数(mode)来作为一个后验分布的总结通常是非常差的选择,因为众数在一个分布通常不典型的,而不像均值或者中位数那样有代表意义.这一点如图5.1(a)所示,图中是一个一维连续空间.最基本的问题是众数是在测量值为0的位置的一个点,而均值和中位数都把空间的体积(volume)考虑了进去.另外一个例子如图5.1(b)所示:这里面的众数是0,但均值则是非零的.对方差之类的参数进行推测的时候就很容易遇到这种特别完犊子的分布形态,尤其是在分层模型(hierarchical models)中.这种情况下最大后验估计(MAP)和最大似然估计(MLE)很明显都是很差的估计.

如果众数不是一个好的选择,那该怎么去对一个后验分布进行总结概括呢?答案就是使用决策理论(decision theory),在本书5.7会讲到.基本思想就是设置一个损失函数(loss function),其中的$L(\theta, \hat{\theta})$表示的意思是在真实值是θ而你估计值是$\hat{\theta}$的时候造成的损失(loss).如果使用0-1二值化量化损失,即$L(\theta, \hat{\theta}) = I(\theta \neq \hat{\theta})$,那么最优估计就是后验众数了.0-1损失函数就意味着你只使用那些没有误差的点,其他有误差的就不要了:这样这种损失函数的情况下就没有部分置信(partial credit)了.对于连续值变量,通常用平方误差损失函数$L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$;对应的最优估计就是后验均值,这也会在本书5.7讲到.或者也可以使用一个更健壮的损失函数,也就是绝对值损失函数:
$L(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$,这时候最佳估计就是后验中位数了.

此处参考原书图5.2

5.2.1.4 最大后验估计(MAP estimation)对重参数化(reparameterization)是

可变的(not invariant)*

最大后验估计(MAP estimation)的一个更严重问题就是得到的结果依赖于如何对概率分布进行参数化.从一种参数形式转换到另一种等价的参数形式,就会改变结果,这很显然略坑,因为测量单位可以是任意的(比如说测量距离的时候用的单位可以使厘米或者英寸).

为了便于理解,假设要计算x的后验分布.如果定义$y = f(x)$,可以根据等式2.87得到y的分布,这里重复一下:
$p_y(y) = p_x(x)|\frac{dx}{dy}|$(5.1)

上面的$|\frac{dx}{dy}|$叫做雅可比项(Jacobian),衡量的是经过函数f之后单位体积的变化规模.设$\hat{x} = \arg\max_x p_x(x)$是对x的最大后验估计(MAP estimation).通常来说这并不意味着通过函数$f(\hat{x})$就能得到$\hat{y} = \arg\max_y p_y(y)$.例如,设$x \sim N(6,1), y = f(x)$,其中有:
$f(x) = \frac{1}{1+\exp(-x+5)}$(5.2)

然后使用蒙特卡罗模拟方法(Monte Carlo simulation)(参考本书2.7.1)来推导出y的分布.结果如图5.2所示.可见原来的正态分布被S型非线性函数给"压缩"了.具体来说就是变换后分布的众数和变换之前的众数是不相等的.

在最大似然估计的过程中遇到这种问题会怎么样呢?设想下面这个例子,参考资料 Michael Jordan.伯努利分布(Bernoulli distribution)通常是使用均值(mean)μ来参数化的,也就是$p(y = 1|\mu) = \mu, y \in \{0,1\}$.假设在单位间隔(unit interval)上用一个均匀先验(uniform prior):$p_\mu(\mu) = 1I(0 \leq \mu \leq 1)$.如果没有数据,最大后验估计正好就是先验的众数(mode)可以在0和1之间的任意位置.不同的参数化就会在这个区间中选择任意的不同的点.

首先设$\theta = \sqrt{\mu}, \mu = \theta^2$.新的先验就是:

$p_\theta(\theta) = p_\mu(\mu)|\frac{d\mu}{d\theta}|$(5.3)

对于$\theta \in [0,1]$,新的众数就是:

$\hat{\theta}_{MAP} = \arg\max_{\theta \in [0,1]} 2\theta = 1$(5.4)

然后设$\phi = 1 - \sqrt{1-\mu}$,新的先验就是:

$p_\phi(\phi) = p_\mu(\mu)|\frac{d\mu}{d\theta}| = 2(1-\phi)$(5.5)

对于$\phi \in [0,1]$,新的众数就是:

$\hat{\theta}_{MAP} = \arg\max_{\theta \in [0,1]} 2 - 2\phi = 0$(5.6)

所以最大后验估计明显是依赖参数化设置的.最大似然估计(MLE)就不会有这个问题,因为似然率函

数是一个函数而不是概率密度.贝叶斯方法也不会有这个问题,因为测量过程的变化在对参数空间上进行积分的时候都考虑进去了.

这个问题的一个解决方案就是优化下面的目标函数(objective function):

$\hat{\theta} = \arg\max_\theta p(D|\theta)p(\theta)|I\theta|^{-\frac{1}{2}}$ (5.7)

这里的$I(\theta)$就是与$p(x|\theta)$相关的费舍信息矩阵(Fisher information matrix)(具体参考本书6.2.2).这个估计与参数化相独立,具体原因如(Jermyn 2005; Druilhet and Marin 2007)所述.然而很悲剧,对等式5.7进行优化通常都很难,这就极大降低了这个方法的吸引力了.

5.2.2 置信区间

除点估计外,我们经常还要对置信度进行衡量.标准的置信度衡量值是某个(标量 scalar)值θ,对应的是后验分布的"宽度(width)".这可以使用一个$100(1-\alpha)$置信区间(credible interval),是一个(连续(contiguous))区域$C = (l, u)$(这里的l和u的意思是lower和upper的缩写,表示的就是下界和上界),这个区域包含了$1-\alpha$的后验概率质量,即:

$C_\alpha(D) = (l, u) : P(l \leq \theta \leq u|D) = 1 - alpha$(5.8)

可能会有很多个这样的区间,我们要选一个能满足在每个尾部(tail)有$(1-\alpha)/2$概率质量的一个区间,这个区间就叫做中央区间(central interval).

此处参考原书图5.3

如果一个后验分布有已知的函数形式,就可以使用$l = F^{-1}(\alpha/2), u = F^{-1}(1-\alpha/2)$来计算得到后验中央区间,其中的F就是后验的累积密度函数(cdf).例如,如果后验是正态分布,
$p(\theta|D) = N(0,1), \alpha = 0.05$,这样就能得到了
$l = \Phi(\alpha/2) = -1.96, u = \Phi(1-\alpha/2) = 1.96$,其中的$\Phi$表示的是正态分布的累积密度函数(cdf).如图2.3(c)所示.这也证明了实际应用中使用$\mu \pm 2\sigma$做置信区间的可行性,其中的μ表示的是后验均值,σ表示的是后验标准差,2是对1.96的一个近似.

当然了,后验分布也不可能总是正态分布的.比如对于抛硬币的那个例子来说,如果使用一个均匀先验,然后观察出N=100次试验中有$N_1 = 47$次人头朝上,那么后验分布就是一个β分布,即
$p(\theta|D) = Beta(48, 54)$.然后会发现其中95%的后验置信区间位于(0.3749,0.5673)(可以参考本书配套的PMTK3当中的betaCredibleInt,里面是单行MATLAB代码来进行这个计算).

如果我们不知道其函数形式,但可以从后验中进行取样,那么也可以使用蒙特卡罗方法近似来对后验的量进行估计:简单排序S个样品,然后找到其中一个出现在排序列表中的α/S位置上的样本.由于$S \to \infty$,这就会收敛到真实值了.具体可以参考本书配套PMTK3当中的mcQuantileDemo.

人们总容易对贝叶斯方法置信区间(Bayesian credible intervals)和频率论的信心区间(frequentist

confidence intervals)产生混淆.这两个可不是一回事,本书6.6.1会详细讲到.一般来说,贝叶斯的置信区间往往是人们要去计算的,而实际上他们通常计算的却都是频率论的信心区间,这是因为大多数人都是学习的频率论的统计学,而不是贝叶斯方法统计学.好在计算贝叶斯置信区间和计算频率论信心区间的方法都不太难,这部分可以参考本书配套PMTK3当中的betaCredibleInt来查看如何在MATLAB中进行计算.

5.2.2.1 最高后验密度区(Highest posterior density regions)*

此处参考原书图5.4

中央区间的一个问题在于可能有不在这个区间内的点却有更高的概率密度,这如图5.3(a)所示,其中左侧置信区间边界之外的点就比右侧区间外的点有更高的概率.

这也导致了另外一个有用的变量,即最高后验密度(highest posterior density,缩写为HPD)区域.这个可以定义为组成总体概率质量的$100(1-\alpha)\%$最可能的点(的集合).更正式来说,可以定义概率密度函数(pdf)上的阈值(threshold)p^*,满足:

$$1-\alpha = \int_{\theta:p(\theta|D)>p^*} p(\theta|D)d\theta \text{(5.9)}$$

然后就可以定义最高后验密度区(HPD)为:

$$C_\alpha(D) = \{\theta : p(\theta|D) > p^*\} \text{(5.10)}$$

在一维情况下,最高后验密度区(HPD)也叫做最高密度区间(highest density interval,缩写为HDI).例如图5.3(b)所示就是一个$Beta(3,9)$分布的95%的最高密度区间(HDI),也就是(0.04,0.48).可以发现这要比置信区间(CI)要更窄一些,虽然也是包含了95%的概率质量;另外,这个区间内的每个点都比区间外的点有更高的概率.

对一个单峰分布(unimodal distribution)来说,最高密度区间(HDI)是围绕众数包含95%概率质量的最窄区间了.可以想象一些灌水的逆过程,将水平面逐渐降低,直到95%的质量暴露出来,而只有5%的部分被淹没.这就使得在一维情况下对最高密度区间(HDI)的计算有了一个很简单的算法:搜索每个点,使区间包含95%的概率质量,然后又有最小的宽度.这可以通过一维数值优化来实现,只要知道概率分布累积密度函数(cdf)的逆函数就可以了,或者如果有一部分样本的话,可以在排序过的数据点上进行搜索(具体可以参考PMTK3当中的betaHPD).

如果后验分布是多峰(multimodal)分布,那么最高密度区间(HDI)可能就不是单个连续区间了,如图5.4(b)所示.不过对多峰后验分布进行总结通常都挺难的.

5.2.3 比例差别的推导(Inference for a difference in proportions)

有时候有很多个参数,然后要计算这些参数的某些函数的后验分布.例如,假设你要从亚马逊买个东

西,然后有两个不同的卖家,提供同样的价格.第一个卖家有90个好评,10个差评,第二个卖家有2个好评没有差评,那你从哪个卖家那里买呢?

此处查看原书图5.5

刚一看,好像你应该选第二个卖家,但也不见得很有信心,因为第二个卖家的评价太少了.本章咱们就用贝叶斯分析(Bayesian analysis)来处理一下这个问题.类似的方法还可以用于其他背景下的不同群体之间的比率或比值的对比.

设θ_1, θ_2分别是两个卖家的可信度,都是未知的.没有啥其他信息,就都用均匀先验$\theta_i \sim Beta(1,1)$.这样后验分布就是$p(\theta_1|D_1) = Beta(91,11), p(\theta_2|D_2) = Beta(3,1)$.咱们要计算的是$p(\theta_1 > \theta_2|D$.为了方便起见,这里定义一个比率的差值$\delta = \theta_1 - \theta_2$.(或者也可以用对数比(log-odds ratio)).使用下面的数值积分就可以计算目标变量:

$$p(\delta > 0|D) = \int_0^1 \int_0^1 I(\theta_1 > \theta_2) Beta(\theta_1|y_1+1, N_1-y_1+1) Beta(\theta_2|y_2+1, N_2-y_2+ \tag{5.11}$$

经过计算会发现$p(\delta > 0|D) = 0.710$,也就意味着最好还是从第一个卖家那里买!具体可以参考本书配套的PMTK3中的amazonSellerDemo查看代码.(这个积分也可以以解析形式解出来(Cook 2005).)

解决这个问题有一个更简单的方法,就是使用蒙特卡洛方法抽样来估计后验分布$p(\delta|D)$.这就容易多了,因为在后验中θ_1, θ_2两者是相互独立的,而且都遵循β分布,可以使用标准方法进行取样.分布$p(\theta_i|D_i)$如图5.5(a)所示,而对$p(\delta|D)$的蒙特卡罗方法(MC)估计,总共95%的最高后验密度(HPD)区域如图5.5(b)所示.对$p(\delta > 0|D)$的蒙特卡罗方法估计是通过计算样本中$\theta_1 > \theta_2$的部分,这样得到的值是0.718,和真实值已经非常接近了.(具体可以参考本书配套的PMTK3中的amazonSellerDemo查看代码.)

5.3 贝叶斯模型选择

在图1.18中,我们看到了使用过高次多项式拟合导致了过拟合,而用过低次数多项式又导致了欠拟合.类似的,在图7.8(a)中可以看到使用的归一化参数(regularization parameter)太小会导致过拟合,而太大又导致欠拟合.一般来说,面对着一系列不同复杂性的模型(比如不同的参数分布族)可选的时候,咱们该怎么选择呢?这就是所谓的模选择问题(model selection problem).

一种方法是使用交叉验证(cross-validation),估算所有备选模型的泛化误差(generalization error),然后选择最优模型.不过这需要对每个模型拟合K次,K是交叉验证的折数(CV folds).更有效率的方法是计算不同模型的后验:

$$p(m|D) = \frac{p(D|m)p(m)}{\sum_{m \in M}} p(m, D) \tag{5.12}$$

然后就很容易计算出最大后验估计模型(MAP model),$\hat{m} = \arg\max p(m|D)$,这就叫做贝叶斯模型选择(Bayesian model selection).

如果对模型使用均匀先验,即$p(m) \propto 1$,这相当于挑选能够让下面的项目最大化的模型:

$$p(D|m) = \int p(D|\theta)p(\theta|m)d\theta \text{(5.13)}$$

这个量也叫做模型m的边缘似然率(marginal likelihood),积分似然率(integrated likelihood)或者证据(evidence).具体如何进行积分在本书5.3.2有讲解.不过首先来对这个量的含义给出一下直观解释.

5.3.1 贝叶斯估计的奥卡姆剃刀

有人可能会觉得使用$p(D|m)$来选择模型可能总会偏向于选择有最多参数的模型.如果用$p(D|\hat{\theta}_m)$来选择模型,那么确实如此,其中的$\hat{\theta}_m$是模型m的参数的最大似然估计(MLE)或者最大后验估计(MAP),因为有更多参数的模型会对数据有更好的拟合,因此就能得到更高的似然率(higher likelihood).不过如果对参数进行积分,而不是最大化,就能自动避免过拟合了:有更多参数并不必然就有更高的边缘似然率(marginal likelihood).这就叫做贝叶斯奥卡姆剃刀效应(Bayesian Occam's razor effect, MacKay 1995b; Murray and Ghahramani 2005),这是根据奥卡姆剃刀原则得名的,其要旨是说应该选择能解释数据的最简单模型.

理解贝叶斯奥卡姆剃刀的一种方式是将边缘似然率改写成下面的形式,基于概率论的链式规则(等式2.5):

$$p(D) = p(y_1)p(y_2|y_1)p(y_3|y_{1:2})\dots p(y_N|y_{1:N-1}) \text{(5.14)}$$

上式中为了简单起见去掉了关于x的条件.这个和留一交叉验证法(leave-one-out cross-validation)估计似然率(likelihood)(本书1.4.8)看着很相似,因为也是给出了之前的全部点之后预测未来的每个点的位置.(当然了,上面表达式中,数据的顺序就没有影响了.)如果一个模型太复杂,在早期样本的时候就可能会过拟合,对后面余下的样本的预测也可能很差.

此处查看原书图5.6

另外一种理解贝叶斯奥卡姆剃刀效应的思路是参考概率总和积累起来必然是1.这样就有$\sum_{D'} p(D'|m) = 1$,其中的求和是在全部可能的数据点集合上进行的.复杂的模型可能进行很多预测,就必须把概率质量分散得特别细(thinkly),然后对任意给定数据就不能得到简单模型一样大的概率.这也叫做概率质量守恒原则(conservation of probability mass principle),如图5.6所示.水平方向的是所有可能的数据集,按照复杂性递增排序(以某种抽象概念来衡量).在纵轴上投下的是三个可能的概率模型:M_1, M_2, M_3复杂性递增.实际观测到底数据为竖直线条所示的D_0.图示可知,第一个模型太简单了,给D_0的概率太低.第三个模型给D_0的概率也很低,因为分布的更宽更窄.第二个模型就看上去正好,给已经观测到底数据给出了合理的置信程度,但又没有预测更多.因此第二个模型是最可选的模型.

图5.7中的数据也是贝叶斯奥卡姆剃刀的一个样例.其中的多项式次数分别为1,2,3,对于N=5个数据点.另外还展示了各个模型的后验,其中使用了正态分布先验(更多细节参考本书7.6).由于没有足够的数据,不能判断复杂模型,所以最大后验估计模型(MAP model)就是d=1.图5.8展示的是当样本规模扩大到N=30的时候的情况.这时候就很明显了d=2是最佳模型(实际上这时候的数据就是用一个二次函数生成的).

另一个例子是图7.8(c),其中是对数投图的$\log p(D|\lambda)$和$\log(\lambda)$,对应的是多项式岭回归模型(polynomial ridge regression model),其中的λ和交叉验证试验中用的值有相同的范围.可见最大证据(maximum evidence)大概就出现在测试的均方误差(MSE)最小的时候,也就是对应着交叉验证所选择的点.

当使用贝叶斯方法的时候,我们不仅可以仅对有限数值网格来计算证据(evidence).可以用数值优化来找到$\lambda^* = \arg\max_\lambda p(D|\lambda)$.这个方法也叫做经验贝叶斯(empirical Bayes)或者二类最大似然估计(type II maximum likelihood)(参考本书5.6).样例可以参考本书图7.8(b):可以看到曲线和交叉验证估计的形状相似,但计算效率要更高.

此处查看原书图5.7

5.3.2 计算边缘似然率(证据)

当我们讨论推导一个混合模型的参数的时候,通常会写:
$p(\theta|D,m) \propto p(\theta|m)p(D|\theta,m)$(5.15)

然后忽略掉归一化常数$p(D|m)$.因为$p(D|m)$相对于θ来说是恒定的,所以这样也有效.不过如果对比模型的话,需要知道如何去计算边缘似然率(marginal likelihood)$p(D|m)$.一般来说这就挺麻烦,因为必须要对所有可能的参数值来进行积分,但是如果有了一个共轭先验,就很容易计算了.

设$p(\theta) = q(\theta)/Z_0$是先验,然后$q(\theta)$是一个未归一化的分布(unnormalized distribution),而Z_0是针对这个先验的归一化常数.设$p(D|\theta) = q(D|\theta)/Z_l$是似然率(likelihood),其中的$Z_l$包含了似然函数中的任意常数项.最终设$p(\theta|D) = q(\theta|D)/Z_N$是后验,其中的$q(\theta|D) = q(D|\theta)q(\theta)$是未归一化的后验,而$Z_N$是这个后验的归一化常数.则有:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \qquad (5.16)$$

$$\frac{q(\theta|D)}{Z_N} = \frac{q(D|\theta)q(\theta)}{Z_l Z_0 p(D)} \qquad (5.17)$$

$$p(D) = \frac{Z_N}{Z_0 Z_l} \qquad (5.18)$$

所以只要归一化常数能算出来,就可以很简单地计算出边缘似然率了.接下来会给出一些例子.

此处查看原书图5.8

5.3.2.1 β-二项模型(Beta-binomial model)

先把上面的结论用到β-二项模型上面.已知了
$p(\theta|D) = Beta(\theta|a', b'), a' = a + N_1, b' = b + N_0$.这个后验的归一化常数是$B(a', b')$.
因此有:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{5.19}$$

$$= \frac{1}{p(D)}[\frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}][\binom{N}{N-1}\theta^{N_1}(1-\theta)^{N_0}] \tag{5.20}$$

$$= \binom{N}{N-1}\frac{1}{p(D)}\frac{1}{B(a,b)}[\theta^{a+N_1-1}(1-\theta)^{b+N_0-1}] \tag{5.21}$$

因此有:

$$\frac{1}{B(a+N_1, b+N_0)} = \binom{N}{N-1}\frac{1}{p(D)}\frac{1}{B(a,b)} \tag{5.22}$$

$$p(D) = \binom{N}{N-1}\frac{B(a+N_1, b+N_0)}{B(a,b)} \tag{5.23}$$

β-伯努利分布模型(Beta-Bernoulli model)的边缘似然函数和上面的基本一样,唯一区别就是去掉了$\binom{N}{N-1}$这一项.

5.3.2.2 狄利克雷-多重伯努利模型(Dirichlet-multinoulli model)

和上面β-伯努利模型类似,狄利克雷-多重伯努利模型(Dirichlet-multinoulli model)的边缘似然函数如下所示:
$p(D) = \frac{B(N+\alpha)}{B(\alpha)}$(5.24)
其中的
$B(\alpha) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\Sigma_k\alpha_k))}$(5.25)
把上面两个结合起来写成如下所示形式:
$p(D) = \frac{\Gamma(\Sigma_k\alpha_k)}{\Gamma(N+\Sigma_k\alpha_k)}\prod_k\frac{\Gamma(N_k+\alpha_k)}{\Gamma(\alpha_k)}$(5.26)
这个等式在后文中会有很多用处.

5.3.2.3 高斯-高斯-威沙特分布(Gaussian-Gaussian-Wishart model)

设想使用了一个共轭正态逆威沙特分布(NIW)的多元正态分布(MVN).设Z_0是先验的归一化项(normalizer),Z_N是后验的归一化项,$Z_t = (2\pi)^{ND/2}$是似然函数的归一化项.然后很明显就能发

现:

$$p(D) = \frac{Z_N}{Z_0 Z_1} \tag{5.27}$$

$$= \frac{1}{\pi^{ND/2}} \frac{1}{2^{ND/2}} \frac{(\frac{2\pi}{k_N})^{D/2} |S_N|^{-v_N/2} 2^{(v_0+N)D/2} \Gamma_D(v_N/2)}{(\frac{2\pi}{k_0})^{D/2} |S_0|^{-v_0/2} 2^{v_0 D/2} \Gamma_D(v_0/2)} \tag{5.28}$$

$$= \frac{1}{\pi^{ND/2}} (\frac{k_0}{k_N})^{D/2} \frac{|S_0|^{-v_0/2} \Gamma_D(v_N/2)}{|S_N|^{-v_N/2} \Gamma_D(v_0/2)} \tag{5.29}$$

这个等式后面也会用得上.

5.3.2.4 对数边缘似然函数的贝叶斯信息标准估计(BIC approximation to log marginal likelihood)

一般来说,直接计算等式5.13里面的积分还挺难的.一种简单又流行的近似方法是使用贝叶斯信息量(Bayesian information criterio,缩写为BIC),形式如下所示(Schwarz 1978):

$$BIC \triangleq \log p(D|\hat{\theta}) - \frac{dof(\hat{\theta})}{2} \log N \approx \log p(D) \tag{5.30}$$

上式中的$dof(\hat{\theta})$是模型中的自由度个数(number of degrees of freedom),而$\hat{\theta}$是模型的最大似然估计(MLE).这有一种类似惩罚对数似然函数(penalized log likelihood)的形式,其中的惩罚项(penalty term)依赖于模型的复杂度.具体信息可以从本书8.4.2查看贝叶斯信息量评分的推导过程.

以一个线性回归为例.如本书7.3所示,最大似然估计为$\hat{w} = (X^T X)^{-1} X^T y$, $\hat{\sigma}^2 = RSS/N$, $RSS = \sum_{i=1}^{N} (y_i - \hat{w}_{mle}^T x_i)^2$.对应的对数似然函数为:

$$\log p(D|\hat{\theta}) = -\frac{N}{2} \log(2\pi\hat{\sigma}^2) - \frac{N}{2} \tag{5.31}$$

因此对应的贝叶斯信息量(BIC)评分为(去掉了常数项):

$$BIC = -\frac{N}{2} \log(\hat{\sigma}^2) - \frac{D}{2} \log(N) \tag{5.32}$$

其中的D是模型中的变两个数．在统计学中，通常对BIC有另外的一种定义,称之为贝叶斯信息量损失(BIC cost,因为目的是将其最小化):

$$BIC - cost \triangleq -2 \log p(D|\hat{\theta}) + dof(\hat{\theta}) \log(N) \approx -2 \log p(D) \tag{5.33}$$

在线性回归的情况下,这就变成了:

$$BIC - cost = N \log(\hat{\sigma}^2) + D \log(N) \tag{5.34}$$

贝叶斯信息量(BIC)方法非常类似于最小描述长度原则(minimum description length,缩写为MDL),这个原则是根据模型拟合数据的程度以及定义复杂度来对模型进行评分.更多细节参考(Hansen and Yu 2001).

还有一个和BIC/MDL非常相似的概念叫做赤池信息量(Akaike information criterion,缩写为AIC),定义如下所示:

$$AIC(m,D) \triangleq \log p(D|\hat{\theta}_{MLE}) - dof(m) \text{(5.35)}$$

这个概念是从频率论统计学的框架下推导出来的,不能被解释为对边缘似然函数的近似.虽然它的形式和BIC很相似.可以看出AIC当中的惩罚项(penalty)要比BIC里面小.这就导致了AIC会挑选比BIC更复杂的模型.不过这也会导致更好的预测精度.具体参考(Clarke et al. 2009, sec 10.2)来了解更多讨论以及这类信息量.

5.3.2.5 先验的效果

有时候咱也不知道该怎么设置先验.在进行后验推导的时候,先验的细节可能也不太重要,因为经常是似然率(likelihood)总会覆盖了先验.不过当计算边缘似然函数的时候,先验扮演的角色就重要多了,因为要对所有可能的参数设定上的似然函数进行平均,然后用先验来做权重.

图5.7和5.8所示的是线性回归的模型选择问题,使用了先验$p(w0 = N(0, \alpha^{-1}I)$.其中的$\alpha$是一个用来控制先验强度的调节参数.这个参数效果很大,具体参考本书7.5所属.只管来说,如果α很大,意味着权重被"强迫"降低,所以对于一个复杂模型需要使用很多小参数来拟合数据(比如高维度多项式拟合).反过来如果α很小,就更倾向于使用简单模型,因为每个参数都可以放大很多倍.

如果先验未知,那么正确的贝叶斯过程就是先对先验给出一个先验.也就是对超参数(hyper-parameter)α和参数w给出一个先验.要计算边缘似然函数,就要对所有未知量进行积分,也就是要计算:

$$p(D|m) = \int\int p(D|w)p(w|\alpha, m)p(\alpha|m)dwd\alpha \text{(5.36)}$$

当然,这就需要实现制定一个超先验(hyper-prior,即对先验的先验).好在在贝叶斯层次(Bayesian hierarchy)中越高的层次就对先验设置越不敏感.所以通常就是用无信息的超先验.

计算的一个捷径就是对α进行优化,而不是去积分.也就是用下面的近似:

$$p(D|m) \approx \int p(D|w)p(w|\hat{\alpha}, m)dw \text{(5.37)}$$

其中的

$$\hat{\alpha} = \arg\max_{\alpha} p(D|\alpha, m) = \arg\max_{\alpha} \int p(D|w)p(w|\alpha, m)dw \text{(5.38)}$$

这个方法就叫做经验贝叶斯(empirical Bayes,缩写为EB),更多相关细节在本书5.6.这正是图5.7和5.8中所用的方法.

贝叶斯因数(Bayes factor,BF(1,0))	解析

BF<1/100	M_0决定性证据
BF<1/10	M_0强证据
1/10<BF<1/3	M_0中等证据
1/3<BF<1	M_0弱证据
1<BF<3	M_1弱证据
3<BF<10	M_1中等证据
BF>10	M_1强证据
BF>100	M_1决定性证据

表格5.1 Jeffrey 对贝叶斯因数的证据规模解析

5.3.3 贝叶斯因数(Bayes factors)

设模型先验是均匀分布的,即$p(m) \propto 1$.那么这时候模型选择就等价于选择具有最大边缘似然率的模型了.接下来假设有两个模型可选,分别是空假设(null hypothesis)M_0和替换假设(alternative hypothesis)M_1.然后就可以定义边缘似然函数之比就叫做贝叶斯因数(Bayes factor):

$$BF_{1,0} \triangleq \frac{p(D|M_1)}{p(D|M_0)} = \frac{p(M_1|D)}{p(M_0|D)} / \frac{p(M_1)}{p(M_0)} \text{(5.39)}$$

(这个跟似然率比值(likelihood ratio)很相似,区别就是整合进来了参数,这就可以对不同复杂度的模型进行对比了.)如果$BF_{1,0} > 1$,我们就优先选择模型1,繁殖就选择模型0.

当然了,也可能这个$BF_{1,0}$就只比1大一点点.这时候也不能确信模型1就一定更好.Jeffreys(1961)提出了一系列的证据范围来解析贝叶斯因数的不同值,如表格5.1所示.这个大概就相当于频率论统计学中的p值(p-value)在贝叶斯统计学中的对应概念.或者也可以把这个贝叶斯因数转换成对模型的后验.如果$p(M_1) = p(M_0) = 0.5$,则有:

$$p(M_0|D) = \frac{BF_{0,1}}{1+BF_{0,1}} = \frac{1}{BF_{1,0}+1} \text{(5.40)}$$

5.3.3.1 样例:测试硬币是否可靠

加入我们观察了一些抛硬币试验,然后想要判断所用硬币是否可靠,可靠的意思就是两面朝上的概率都是一半,即有$\theta = 0.5$,不可靠就是θ可以在[0,1]中间的任意位置取值了.咱们把两种模型表示为M_0, M_1.M_0下的边缘似然函数为:

$$p(D|M_) = (\frac{1}{2})^N \text{(5.41)}$$

其中的N就是抛硬币次数.然后M_1下的边缘似然函数使用一个β分布先验:

$$p(D|M_1) = \int p(D|\theta)p(\theta)d\theta = \frac{B(\alpha_1+N_1,\alpha_0+N_0)}{B(\alpha_1,\alpha_0)} \quad (5.42)$$

此处参考原书图5.9

对$\log p(D|M_1)$和人头朝上次数N_1的关系进行投图得到了图5.9(a),设$N=5,\alpha_1=\alpha_0=1$.(这个曲线的形状对α_0,α_1并不敏感,因为这两个相等了.)如果观察到了两次或者三次人头朝上,那么就更倾向认为硬币没问题,即M_0的概率大于M_1,因为M_0是个更简单的模型(没有自由参数).如果硬币有问题而出现人头背面各自一半情况只是碰巧就是很可疑的巧合了.不过随着次数越来越极端,就可能更倾向硬币有问题的假设了.如果把贝叶斯因数的对数函数$\log BF_{1,0}$投图,会发现有一样的形状,因为$\log p(D|M_0)$是一个常量(constant).参考练习3.18.

图5.9(b)是对$\log p(D|M_1)$的贝叶斯信息量(BIC)估计,这个不可靠的硬币样例参考本书5.3.3.1.很明显这个曲线几乎和对数边缘似然函数的形状一模一样,这对于模型选择的目的来说就足够用了,因为绝对范围都没有影响.具体来说就是倾向于用更简单的模型,除非数据本身特别支持更复杂的模型.

5.3.4 杰弗里斯 - 林德利悖论(Jeffreys-Lindley paradox)*

当使用不适当先验(improper priors,比如积分不为1的先验)来进行模型选择和假设测试的时候,会遇到各种问题,即便这些先验对于其他目的来说可能还是可用的.例如,假如测试假设$M_0:\theta\in\Theta_0$和$M_1:\theta\in\Theta_1$.要定义在θ上的边缘分布密度(marginal density),使用下面的混合模型:

$$p(\theta) = p(\theta|M_0)p(M_0) + p(\theta|M_1)p(M_1) \quad (5.43)$$

只有当$p(\theta|M_0),p(\theta|M_1)$都是适当(归一化)的密度函数的时候,上式才有意义.这种情况下,后验为:

$$p(M_0|D) = \frac{p(M_0)p(D|M_0)}{p(M_0)p(D|M_0)+p(M_1)p(D|M_1)} \quad (5.44)$$

$$= \frac{p(M_0)\int_{\Theta_0}p(D|\theta)p(\theta|M_0)d\theta}{p(M_0)\int_{\Theta_0}p(D|\theta)p(\theta|M_0)d\theta + p(M_1)\int_{\Theta_1}p(D|\theta)p(\theta|M_1)d\theta} \quad (5.45)$$

然后假设使用了不适当先验,即有$p(\theta|M_0)\propto c_0, p(\theta|M_1)\propto c_1$.则:

$$p(M_0|D) = \frac{p(M_0)c_0\int_{\Theta_0}p(D|\theta)d\theta}{p(M_0)c_0\int_{\Theta_0}p(D|\theta)d\theta + p(M_1)c_1\int_{\Theta_1}p(D|\theta)d\theta} \quad (5.46)$$

$$= \frac{p(M_0)c_0l_0}{p(M_0)c_0l_0+p(M_1)c_1l_1} \quad (5.47)$$

上式中$l_i = \int_{\Theta_i} p(D|\theta)d\theta$是模型i的整合/边缘似然函数(integrated or marginal likelihood).然后设$p(M_0) + p(M_1) = \frac{1}{2}$.则有:

$$p(M_0|D) = \frac{c_0 l_0}{c_0 l_0 + c_1 l_1} = \frac{l_0}{l_0 + (c_1/c_0) l_1} \text{(5.48)}$$

然后就可以任意选择c_1, c_0来改变后验.要注意,如果使用了适当(proper)但很模糊(vague)先验也能导致类似问题.具体来说,贝叶斯因数总会倾向于选择更简单的模型,因为使用非常分散的先验的复杂模型观察到数据的概率是非常小的.这就叫做杰弗里斯 - 林德利悖论(Jeffreys-Lindley paradox).

所以在进行模型选择的时候选择适当先验是很重要的.不过还是要注意,如果M_0, M_1在参数的一个子集上分享了同样的先验,那么这部分先验就可能是不适当的,因为对应的归一化常数会被约掉无效.

5.4 先验

贝叶斯统计最受争议的就是对先验的依赖.贝叶斯主义者认为这是不可避免的,因为没有人是白板一片(tabula rasa/blank slate):所有的推测都必须是以客观世界的某些假设为条件的.不过人们还是希望能够尽量缩小事先假设的影响.接下来就简要讲一下实现这个目的的几种方法.

5.4.1 无信息先验

如果关于θ应该是啥样没有比较强的事先认识,通常都会使用无信息先验(uninformative or non-informative prior),然后就去"让数据自己说话".

不过设计一个无信息先验还是需要点技巧的.例如,伯努利参数$\theta \in [0, 1]$.有人可能会觉得最无信息的先验应该是均匀分布$Beta(1, 1)$.不过这时候后验均值就是$E[\theta|D] = \frac{N_1 + 1}{N_1 + N_0 + 2}$,其中的最大似然估计(MLE)为$\frac{N_1}{N_1 + N_0}$.因此就可以发现这个先验也并不是完全无信息的.

很显然,通过降低伪计数(pseudo counts)的程度(magnitu),就可以降低先验的影响.综上所述,最无信息的先验应该是:

$$\lim_{c \to 0} Beta(c, c) = Beta(0, 0) \text{(5.49)}$$

上面这个是在0和1两个位置有质量的等价点的混合(a mixture of two equal point masses at 0 and 1),参考(Zhu and Lu 2004).这也叫做Haldane先验(Haldane prior).要注意这个Haldane先验是一个不适当先验,也就是积分不为1.不过只要能看到至少有一次人头朝上以及至少有一次背面朝上,这个后验就是适当的.

在本书5.4.2.1,会论证"正确"的无信息先验是$Beta(\frac{1}{2}, \frac{1}{2})$.很明显这三个先验在实际使用的时候差异很可能是可以忽略不计的.通常来说,都可以进行敏感度分析,也就是检测建模假设的变化对模型

结论和预测变化的影响,这种敏感度分析包含了对先验的选择,也包含了似然率的选择以及对数据的处理过程.如果结论是对于模型假设来说并不敏感,那么就可以对结果更有信心了.

5.4.2 杰弗里斯先验论(Jeffreys priors)*

哈罗德 杰弗里斯(Harold Jeffreys)设计了一个通用技巧来创建无信息先验.得到的就是杰弗里斯先验论(Jeffreys prior).关键之处是观察到如果$p(\phi)$是无信息的,那么对这个先验重新参数化,比如使用某函数h得到的$\theta = h(\phi)$,应该也是无信息的.然后就可以更改方程变量:

$$p_\theta(\theta) = p_\phi(\phi)|\frac{d\phi}{d\theta}| \quad (5.50)$$

所以先验会总体上有变化.然后选择

$$p_\phi(\phi) \propto (I(\phi))^{\frac{1}{2}} \quad (5.51)$$

其中的$I(\phi)$是费舍信息矩阵(Fisher information matrix):

$$I(\phi) \triangleq -E[(\frac{d\log p(X|\phi)}{d\phi})^2] \quad (5.52)$$

这是对预期的负对数似然函数的度量,也是对最大似然估计的稳定性的度量,参考本书6.2.2.于是就有:

$$\frac{d\log p(x|\theta)}{d\theta} = \frac{d\log p(x|\phi)}{d\phi}\frac{d\phi}{d\theta} \quad (5.53)$$

开平方,然后去对x的期望,就得到了:

$$I(\theta) = -E[(\frac{d\log p(X|\theta)}{d\theta})^2] = I(\phi)(\frac{d\phi}{d\theta})^2 \quad (5.54)$$

$$I(\theta)^{\frac{1}{2}} = I(\phi)^{\frac{1}{2}}|\frac{d\phi}{d\theta}| \quad (5.55)$$

所以能发现变换后的先验是:

$$p_\theta(\theta) = p_\phi(\phi)|\frac{d\phi}{d\theta}| \propto (I(\phi))^{\frac{1}{2}}|\frac{d\phi}{d\theta}| = I(\theta)^{\frac{1}{2}} \quad (5.56)$$

所以就说明$p_\theta(\theta)$和$p_\phi(\phi)$是一样的.
接下来举几个例子.

5.4.2.1 样例:伯努利模型和多重伯努利模型的杰弗里斯先验论(Jeffreys priors)

设X服从伯努利分布,即$X \sim Ber(\theta)$.一个单独样本(single sample)的对数似然函数为:

$$\log p(X|\theta) = X|log\theta(1-X)\log(1-\theta) \tag{5.57}$$

得分函数(score function)正好就是对数似然函数的梯度:

$$s(\theta) \triangleq \frac{d}{d\theta}\log p(X|\theta) = \frac{X}{\theta} - \frac{1-X}{(1-\theta)^2} \tag{5.58}$$

观测信息(observed information)就是对数似然函数的二阶导数(second derivative):

$$J(\theta) = -\frac{d^2}{d\theta^2}\log p(X|\theta) = -s'(\theta|X) = \frac{X}{\theta^2} + \frac{1-X}{(1-\theta)^2} \tag{5.59}$$

费舍信息矩阵(Fisher information)正好就是期望信息矩阵(expected information):

$$I(\theta) = E[J(\theta|X)|X \sim \theta] = \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta(1-\theta)} \tag{5.60}$$

因此杰弗里斯先验(Jeffreys' prior)为:

$$p(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} = \frac{1}{\sqrt{\theta(1-\theta)}} \propto Beta(\frac{1}{2}, \frac{1}{2}) \tag{5.61}$$

然后考虑有K个状态的多重伯努利随机变量.很明显对应的杰弗里斯先验(Jeffreys' prior)为:

$$p(\theta) \propto Dir(\frac{1}{2}, \ldots, \frac{1}{2}) \tag{5.62}$$

要注意这个和更明显的选择 $Dir(\frac{1}{K}, \ldots, \frac{1}{K}), Dir(1, \ldots, 1)$是不一样的.

5.4.2.2 样例:局部和缩放参数的杰弗里斯先验论(Jeffreys priors)

可以对一个局部参数(location parameter)取杰弗里斯先验(Jeffreys prior),比如以正态分布均值为例,就是$p(\mu) \propto 1$.这是一个转换不变先验(translation invariant prior),满足下述性质,即分配到任意区间[A,B]的概率质量等于分配另一个等宽区间[A-c,B-c]的概率质量.也就是:

$$\int_{A-c}^{B-c} p(\mu)d\mu = (A-c) - (B-c) = (A-B) = \int_A^B p(\mu)d\mu \tag{5.63}$$

这可以通过使用$p(\mu) \propto 1$来实现,可以使用一个有限变量的正态分布来近似,即$p(\mu) = N(\mu|0, \infty)$.要注意这也是一个不适当先验(improper prior),因为积分不为1.只要后验适当,使用不适当先验也没问题,也就是有$N \geq 1$个数据点的情况,因为只要有一个单独数据点就可以"确定"区域位置了.

与之类似,也可以对缩放参数(scale parameter)取杰弗里斯先验(Jeffreys prior),比如正态分布的方差,$p(\sigma^2 \propto 1/\sigma^2)$.这是一个缩放独立先验(scale invariant prior)满足下面的性质:分配到任意区间[A,B]的概率质量等于分配另一个缩放区间[A/c,B/c]的概率质量,其中的c是某个常数$c > 0$.(例如,如果将单位从米换成影驰,还不希望影响参数推导过程等等).这可以使用:

$p(s) \propto 1/s$(5.64)

要注意:

$$\int_{A/c}^{B/c} = [\log s]_{A/c}^{B/c} = \log(B/c) - \log(A/c) \qquad (5.65)$$

$$= \log(B) - \log(A) = \int_A^B p(s)ds \qquad (5.66)$$

可以使用退化γ分布(degenerate Gamma distribution)来近似,(参考本书2.4.4),即$p(s) = Ga(S|0,0)$.这个先验$p(s) \propto 1/s$也是不适当先验,但只要有观测到超过$N \geq 2$个数据点,就能保证后验是适当的,也就可以了.(因为只要最少有两个数据点就可以估计方差了.)

5.4.3 健壮先验(Robust priors)

在很多情况下,我们对先验并不一定很有信心,所以就需要确保先验不会对结果有过多的影响.这可以通过使用健壮先验(robust priors,Insua and Ruggeri 2000)来实现,这种先验通常都有重尾(heavy tailes),这就避免了过分靠近先验均值.

考虑一个来自(Berger 1985,p7)的例子.设x服从正态分布,即$x \sim N(\theta, 1)$.观察到了 x=5 然后要去估计θ.最大似然估计(MLE)是$\hat{\theta} = 5$,看上去也挺合理.在均匀先验之下的后验均值也是这个值,即$\bar{\theta} = 5$.不过如果假设我们知道了先验中位数(median)是0,而先验的分位数(quantiles)分别是-1和1,所以就有$p(\theta/le - 1) = p(-1 < \theta/le 0) = p(0 < \theta/le 1) = p(1 < \theta) = 0.25$.另外假设这个先验是光滑单峰的.

很明显正态分布的先验$N(\theta|0, 2.19^2)$满足这些约束条件.但这时候后验均值就是3.43了,看着就不太让人满意了.

然后再考虑使用柯西分布作先验(Cauchy prior)$T(\theta|0,1,1)$.这也满足上面的先验约束条件.这次发现用这个先验的话,后验均值就是4.6,看上去就更合理了(计算过程可以用数值积分,参考本书配套的PMTK3当中的 robustPriorDemo中的代码).

5.4.4 共轭先验的混合

健壮先验很有用,不过计算开销太大了.共轭先验可以降低计算难度,但对我们把已知信息编码成先验来说,往往又不够健壮,也不够灵活.不过,共轭先验(conjugate priors)的混合(mixtures)还是共轭的(参考联系5.1),然后还可以对任意一种先验进行近似((Dallal and Hall 1983; Diaconis and Ylvisaker 1985).然后这样的先验就能在计算开销和灵活性之间得到一个很不错的折中.

还以之前抛硬币模型为例,考虑要检查硬币是否作弊,是两面概率都一样,还是有更大概率人头朝上.这就不能用一个β分布来表示了.不过可以把它用两个β分布的混合来表示.例如,就可以使用:

$$p(\theta) = 0.5Beta(\theta|20,20) + 0.5Beta(\theta|30,10) \text{(5.67)}$$

如果θ来自第一个分布,就说明没作弊,如果来自第二个分布,就说明有更大概率人头朝上.

可以引入一个潜在指示器变量(latent indicator variable)z来表示这个混合,其中的z=k的意思就是θ来自混合成分(mixture component)k.这样先验的形式就是:

$$p(\theta) = \sum_k p(z=k)p(\theta|z=k) \text{(5.68)}$$

其中的每个$p(\theta|z=k)$都是共轭的,而$p(z=k)$就叫做先验的混合权重(prior mixing weights).(练习5.1)可知后验也可以写成一系列共轭分布的混合形式:

$$p(\theta|D) = \sum_k p(z=k)p(\theta|D,z=k) \text{(5.69)}$$

其中的$p(Z=k|D)$是后验混合权重(posterior mixing weights),如下所示:

$$p(Z=k|D) = \frac{p(Z=k)p(D|Z=k)}{\sum_{k'} p(Z=k')p(D|Z=k')} \text{(5.70)}$$

这里的量$p(D|Z=k)$是混合成分k的边缘似然函数,可以参考本书5.3.2.1.

5.4.4.1 样例

假如使用下面的混合先验:
$$p(\theta) = 0.5Beta(\theta|a_1,b_1) + 0.5Beta(\theta|a_2,b_2) \text{(5.71)}$$

其中$a_1 = b_1 = 20, a_2 = b_2 = 10$.然后观察了$N_1$次的人头,$N_0$次的背面.后验就成了:

$$p(\theta|D) = p(Z=1|D)Beta(\theta|a_1+N_1,b_1+N_0) + p(Z=2|D)Beta(\theta|a_2+N_1,b_2+N \text{(5.72)}$$

如果$N_1 = 20, N_0 = 10$,那么使用等式5.23,就得到了后验如下所示:

$$p(\theta|D) = 0.346Beta(\theta|40,30) + 0.654Beta(\theta|50,20) \text{(5.73)}$$

如图5.10是对此的图示.

此处参考原书图5.10

5.4.4.2 应用:在DNA和蛋白质序列中找到保留区域(conserved regions)

之前提到过狄利克雷-多项式模型(Dirichlet-multinomial models)在生物序列分析领域用的很广泛.现在就举个例子来看一下.还要用到本书2.3.2.1当中提到的序列标识(sequence logo).现在假设我们想找到基因中代表了编码区域的位置.这类位置在不同的序列中往往都有同样的字母,这是因为进化原因导致的.所以需要找纯净(pure)或者近似纯净的列(columns),比如都是碱基A/T/C/G当中的

一种.一种方法是查找低信息熵的列(low-entropy columns)这些列的分布几乎都是确定的.

然后假设我们相对纯净度都估计的置信程度进行衡量.如果我们认为邻近区域是一同保留的,这就很有用了.这时候设置如果区域t是保留的则$Z_t = 1$,反之则$Z_t = 0$.可以在临近的Z_t变量之间加入一个依赖项(dependence),要用一个马尔科夫链(Markov chain),这部分参考本书第17章有详细内容.

无论任何情况,都要定义一个似然率模型(likelihood model)$p(N_t|Z_t)$,其中的N_t是第t列的(ACGT)碱基计数组成的向量.通常设置这是一个参数为θ_t的多项式分布.因为每一列(column)都有不同的分布,所以要对θ_t积分,然后计算边缘似然函数:

$$p(N_t|Z_t) = \int p(N_t|\theta_t)p(\theta_t|Z_t)d\theta_t \text{(5.74)}$$

但是对θ_t用什么先验呢?当$Z_t = 0$的时候可以使用一个均匀显眼,即$p(\theta|Z_t = 0) = Dir(1,1,1,1)$,可是如果$Z_t = 1$呢?如果一列被保留了,就应该是纯粹(或者近似纯粹)由ACGT四种碱基中的一种组成的.很自然的方法就是使用狄利克雷先验的混合,每一个都朝向四维单形(simplex)中的一角倾斜(tilted),即:

$$p(\theta|Z_t = 1) = \frac{1}{4}Dir(\theta|(10,1,1,1)) + \ldots + \frac{1}{4}Dor(\theta|(1,1,1,10)) \text{(5.75)}$$

由于这也是共轭的,所以$p(N_t|Z_t)$计算起来很容易.参考(Brown et al. 1993)来查看一个在现实生物序列问题中的具体应用.

5.5 分层贝叶斯(Hierarchical Bayes)

计算后验$p(\theta|D)$的一个关键要求就是特定的先验$p(\theta|\eta)$,其中的η是超参数(hyper-parameters).如果不知道怎么去设置η咋办呢?有的时候可以使用无信息先验,之前已经说过了.一个更加贝叶斯风格的方法就是对先验设一个先验!用本书第十章的图模型的术语来说,可以用下面的方式来表达:$\eta \rightarrow \theta \rightarrow D$(5.76)

这就是一个分层贝叶斯模型(hierarchical Bayesian model),也叫作多层模型(multi-level model),因为有多层的未知量.下面给一个简单样例,后文还会有更多其他的例子.

5.5.1 样例:与癌症患病率相关的模型

考虑在不同城市预测癌症患病率的问题(这个样例来自Johnson and Albert 1999, p24).具体来说,加入我们要测量不同城市的人口,N_i,然后对应城市死于癌症的人口数x_i.假设$x_i \sim Bin(N_i, \theta_i)$,然后要估计癌症发病率$\theta_i$.一种方法是分别进行估计,不过这就要面对稀疏数据问题(低估了人口少即N_i小的城市的癌症发病率).另外一种方法是假设所有的θ_i都一样;这叫做参数绑定(parameter tying.结果得到的最大似然估计(MLE)正好就是$\hat{\theta} = \frac{\Sigma_i x_i}{\Sigma_i N_i}$.可是很明显假设所有城市癌症发病率都

一样有点太牵强了.有一种折中的办法,就是估计θ_i是相似的,但可能随着每个城市的不同而又发生变化.这可以通过假设θ_i服从某个常见分布来实现,比如β分布,即$\theta_i \sim Beta(a,b)$.这样就可以把完整的联合分布写成下面的形式:

$$p(D,\theta,\eta|N) = p(\eta)\prod_{i=1}^{N} Bin(x_i|N_i,\theta_i)Beta(\theta_i|\eta) \tag{5.77}$$

上式中的$\eta = (a,b)$.要注意这里很重要的一点是要从数据中推测$\eta = (a,b)$;如果只是随便设置成一个常数,那么θ_i就会是有件独立的(conditionally independent),在彼此之间就没有什么信息联系了.与之相反的,若将η完全看做一个未知量(隐藏变量),就可以让数据规模小的城市从数据规模大的城市借用统计强度(borrow statistical strength).

要计算联合后验$p(\eta,\theta|D)$.从这里面可以得到后验边缘分布$p(\theta_i|D)$.如图5.11(a)所示,图中的蓝色柱状是后验均值$E[\theta_i|D]$,红色线条是城市人口均值$E[a/(a+b_|D]$(这代表了θ_i的均值).很明显可以看到后验均值朝向有小样本N_i的城市的汇总估计方向收缩.例如,城市1和城市20都观察到有0的癌症发病率,但城市20的人口数较少,所以其癌症发病率比城市1更朝向人口估计方向收缩(也就是距离水平的红色线更近).

图5.11(b)展示的是θ_i的95%后验置信区间.可以看到城市15有特别多的人口(53637),后验不确定性很低.所以这个城市对η的后验估计的影响最大,也会影响其他城市的癌症发病率的估计.城市10和19有最高的最大似然估计(MLE),也有最高的后验不确定性,反映了这样高的估计可能和先验相违背(先验视从所有其他城市估计得到的).

上面这个例子中,每个城市都有一个参数,然后对相应的概率进行建模.通过设置伯努利分布的频率参数为一个协变量的函数,即$\theta_i = sigm(w_i^T x)$,就可以对多个相关的逻辑回归任务进行建模了.这也叫做多任务学习(multi-task learning),在本书9.5会有详细讲解.

5.6 经验贝叶斯(Empirical Bayes)

在分层贝叶斯模型中,我们需要计算多层的潜在变量的后验.例如,在一个两层模型中,需要计算:
$$p(\eta,\theta|D) \propto p(D|\theta)p(\theta|\eta)p(\eta) \tag{5.78}$$

有的时候可以通过分析将θ边缘化;这就将问题简化成只去计算$p(\eta|D)$了.

作为计算上的简化,可以对超参数后验进行点估计来近似,即$p(\eta|D) \approx \delta_{\hat{\eta}}(\eta)$,其中的$\hat{\eta} = \arg\max p(\eta|D)$.因为$\eta$通常在维数上都比$\theta$小很多,这个模型不太容易过拟合,所以我们可以安全地来对η使用均匀显眼.这样估计就成了:

$$\hat{\eta} = \arg\max p(D|\eta) = \arg\max[\int p(D|\theta)p(\theta|\eta)d\theta] \tag{5.79}$$

其中括号里面的量就是边缘似然函数或整合似然函数(marginal or integrated likelihood),也叫证据(evidence).这个方法总体上叫做经验贝叶斯(Empirical Bayes,缩写为EB)或这也叫座第二类最大似

然估计(Type II Maximum Likelihood).在机器学习里面,也叫作证据程序(evidence procedure).

经验贝叶斯违反了先验应该独立于数据来选择的原则.不过可以将其视作是对分层贝叶斯模型中的推导的一种近似,计算开销更低.就好比是讲最大后验估计(MAP estimation)看作是对单层模型$\theta \rightarrow D$的推导的近似一样.实际上,可以建立一个分层结构,其中进行的积分越多,就越"贝叶斯化":

方法(Method)	定义(Definition)
最大似然估计(Maximum Likelihood)	$\hat\theta=\arg\max_\theta p(D
最大后验估计(MAP estimation)	$\hat\theta=\arg\max_\theta p(D
经验贝叶斯的最大似然估计(ML-II Empirical Bayes)	$\hat\theta=\arg\max_\eta \int p(D
经验贝叶斯的最大后验估计(MAP-II)	$\hat\theta=\arg\max_\eta \int p(D
全贝叶斯(Full Bayes)	$p(\theta,\eta

要注意,经验贝叶斯(EB)也有很好的频率论解释(参考Carlin and Louis 1996; Efron 2010),所以在非贝叶斯模型中也被广泛使用.例如很流行的詹姆斯-斯坦因估计器(James-Stein estimator)就是用经验贝叶斯推导的,更多细节参考本书6.3.3.2.

5.6.1 样例:β-二项模型

还回到癌症发病率的模型上.可以积分掉θ_i,然后直接写出边缘似然函数,如下所示:

$$p(D|a,b) = \prod_i \int Bin(x_i|N_i,\theta_i)Beta(\theta_i|a,b)d\theta_i \qquad (5.80)$$

$$= \prod_i \frac{B(a+x_i, b+N_i-x_i)}{B(a,b)} \qquad ((5.81))$$

关于a和b来最大化有很多方法,可以参考(Minka 2000e).

估计完了a和b之后,就可以代入到超参数里面来计算后验分布$p(\theta_i|\hat{a},\hat{b},D)$,还按照之前的方法,使用共轭分析.得到的每个$\theta_i$的后验均值就是局部最大似然估计(local MLE)和先验均值的加权平均值,依赖于$\eta = (a,b)$;但由于η是根据所有数据来估计出来的,所以每个θ_i也都受到全部数据的影响.

5.6.2 样例:高斯-高斯模型(Gaussian-Gaussian model)

接下来这个例子和癌症发病率的例子相似,不同之处是这个例子中的数据是实数值的(real-valued).使用一个高斯(正态)似然函数(Gaussian likelihood)和一个高斯(正态)先验(Gaussian prior).这样就

能写出来解析形式的解.

设我们有来自多个相关群体的数据.比如x_{ij}表示的就是学生i在学校j得到的测试分数,j的取值范围从1到D,而i是从1到N_j,即$j = 1:D, i = 1:N_j$.然后想要估计每个学校的平均分θ_j.可是样本规模N_j对于一些学校来说可能很小,所以可以用分层贝叶斯模型(hierarchical Bayesian model)来规范化这个问题,也就是假设θ_j来自一个常规的先验(common prior)$N(\mu, \tau^2)$.

这个联合分布的形式如下所示:
$$p(\theta, D|\eta, \sigma^2) = \prod_{j=1}^{D} N(\theta_j|\mu, \tau^2) \prod_{i=1}^{N_j} N(x_{ij}|\theta_j, \sigma^2) \quad (5.82)$$

上式中为了简化,假设了σ^2是已知的.(这个假设在练习24.4中.)接下来将估计η.一旦估计了$\eta = (\mu, \tau)$,就可以计算θ_j的后验了.要进行这个计算,只需要将联合分布改写成下面的形式,这个过程利用值x_{ij}和方差为σ^2的N_j次高斯观测等价于值为$\bar{x}_j \triangleq \frac{1}{N_j}\sum_{i=1}^{N_j} x_{ij}$方差为$\sigma_j^2 \triangleq \sigma^2/N_j$的一次观测这个定理.这就得到了:

$$p(\theta, D|\hat{\eta}, \sigma^2) = \prod_{j=1}^{D} N(\theta_j|\hat{\mu}, \hat{\tau}^2) N(\bar{x}_j|\theta_j, \sigma_j^2) \quad (5.83)$$

利用上面的式子,再利用本书4.4.1当中的结论,就能得到后验为:

$$p(\theta_j|D, \hat{\mu}, \hat{\tau}^2) = N(\theta_j|\hat{B}_j\hat{\mu} + (1-\hat{B}_j)\bar{x}_j, (1-\hat{B}_j)\sigma_j^2) \quad (5.84)$$

$$\hat{B}_j \triangleq \frac{\sigma_j^2}{\sigma_j^2 + \hat{\tau}^2} \quad (5.85)$$

其中的$\hat{\mu} = \bar{x}, \hat{\tau}^2$下面会给出定义.

$0 \leq \hat{B}_j \leq 1$这个量控制了朝向全局均值(overall mean)μ的收缩程度(degree of shrinkage).如果对于第j组来说数据可靠(比如可能是样本规模N_j特别大),那么σ_j^2就会比τ^2小很多;因此这样$\hat{B}_j$也会很小,然后就会在估计θ_j的时候给$\bar{x}_j$更多权重.而样本规模小的群组就会被规范化(regularized),也就是朝向全局均值μ的方向收缩更严重.接下来会看到一个例子.

如果对于所有的组j来说都有$\sigma_j = \sigma$,那么后验均值就成了:"
$$\hat{\theta}_j = \hat{B}\bar{x} + (1-\hat{B})\bar{x}_j = \bar{x} + (1-\hat{B})(\bar{x}_j - \bar{x}) \quad (5.86)$$

这和本书在6.3.3.2中讨论到的吉姆斯-斯坦因估计器(James Stein estimator)的形式一模一样.

5.6.2.1 样例:预测棒球得分

接下来这个例子是把上面的收缩(shrinkage)方法用到棒球击球平均数(baseball batting averages,引自 Efron and Morris 1975).观察D=18个球员在前T=45场比赛中的的击球次数.把这个击球次数设为b_i.假设服从二项分布,即$b_j \sim Bin(T, \theta_j)$,其中的$\theta_j$是选手j的"真实"击球平均值.目标是要顾及出来这个θ_j.最大似然估计(MLE)自然是$\hat{\theta}_j = x_j$,其中的$x_j = b_j/T$是经验击球平均值.不过可

以用经验贝叶斯方法来进行更好的估计.

要使用上文中讲的高斯收缩方法(Gaussian shrinkage approach),需要似然函数是高斯分布的,即对于一直的σ^2有$x_j \sim N(\theta_j, \sigma^2)$.(这里去掉了下标i因为假设了$N_j = 1$而$x_j$已经代表了选手j的平均值了.)不过这个例子里面用的是二项似然函数.均值正好是$E[x_j] = \theta_j$,方差则是不固定的:

$$var[x_j] = \frac{1}{T^2} var[b_j] = \frac{T\theta_j(1-\theta_j)}{T^2} \text{(5.87)}$$

所以咱们对x_j应用一个方差稳定变换(variance stabilizing transform 5)来更好地符合高斯假设:
$$y_i = f(y_i) = \sqrt{T}\arcsin(2y_i - 1) \text{(5.88)}$$

然后应用一个近似$y_i \sim N(f(\theta_j), 1) = N(\mu_j, 1)$.以$\sigma^2 = 1$代入等式5.86来使用高斯收缩对$\mu_j$进行估计,然后变换回去,就得到了:

$$\hat{\theta}_j = 0.5(\sin(\hat{\mu}_j/\sqrt{T}) + 1) \text{(5.89)}$$

此处参考原书图5.12

这个结果如图5.12(a-b)所示.在图(a)中,投图的是最大似然估计(MLE)$\hat{\theta}_j$和后验均值$\bar{\theta}_j$.可以看到所有的估计都朝向全局均值0.265收缩.在图(b)中,投图的是θ_j的真实值,最大似然估计(MLE)$\hat{\theta}_j$和后验均值$\bar{\theta}_j$.(这里的θ_j的真实值是指从更大规模的独立赛事之间得到的估计值.)可以看到平均来看,收缩的估计比最大似然估计(MLE)更加靠近真实值.尤其是均方误差,定义为
$MSE = \frac{1}{N}\sum_{j=1}^{D}(\theta_j - \bar{\theta}_j)^2$,使用收缩估计的$\bar{\theta}_j$比最大似然估计的$\hat{\theta}_j$的均方误差小了三倍.

5.6.2.2 估计超参数

在本节会对估计η给出一个算法.加入最开始对于所有组来说都有$\sigma_j^2 = \sigma^2$.这种情况下就可以以闭合形式(closed form)来推导经验贝叶斯估计(EB estimate).从等式4.126可以得到:
$$p(\bar{x}_j|\mu, \tau^2, \sigma^2) = \int N(\bar{x}_j|\theta_j, \sigma^2)N(\theta_j|\mu, \tau^2)d\theta_j = N(\bar{x}_j|\mu, \tau^2 + \sigma^2) \text{(5.90)}$$

然后边缘似然函数(marginal likelihood)为:

$$p(D|\mu, \tau^2, \sigma^2) = \prod_{j=1}^{D} N(\bar{x}_j|\mu, \tau^2 + \sigma^2) \text{(5.91)}$$

接下来就可以使用对正态分布(高斯分布)的最大似然估计(MLE)来估计超参数了.例如对μ就有:

$$\hat{\mu} = \frac{1}{D}\sum_{j=1}^{D}\bar{x}_j = \bar{x} \text{(5.92)}$$

上面这个也就是全局均值.

对于方差,可以使用矩量匹配(moment matching,相当于高斯分布的最大似然估计):简单地把模型方差(model varianc)等同于经验方差(empirical variance):

$\hat{\tau}^2 + \sigma^2 = \frac{1}{D}\sum_{j]1}^{D}(\bar{x}_j - \bar{x})^2 \triangleq s^2$(5.93)

所以有$\hat{\tau}^2 = s^2 - \sigma^2$.因为已知了$\tau^2$必然是正的,所以通常都使用下面这个修订过的估计:

$\hat{\tau}^2 = \max(0, s^2 - \sigma^2) = (s^2 - \sigma^2)_+$(5.94)

这样就得到了收缩因子(shrinkage factor):

$\hat{B} = \frac{\sigma^2}{\sigma^2+\tau^2} = \frac{\sigma^2}{\sigma^2+(s^2-\sigma^2)_+}$(5.95)

如果σ_j^2各自不同,就没办法以闭合形式来推导出解了.练习11.13讨论的是如何使用期望最大化算法(EM algorithm)来推导一个经验贝叶斯估计(EB estimate),练习24.4讨论了如何在这个分层模型中使用全贝叶斯方法.

5.7 贝叶斯决策规则(Bayesian decision rule)

之前的内容中,我们已经看到了概率理论可以用来表征和更新我们对客观世界状态的信念.不过我们最终目标是把信念转化成行动.在本节,讲的就是用最有方法来实现这个目的.

我们可以把任何给定的统计决策问题规范表达成一个与自然客观世界作为对手的游戏(而不是和其他的玩家相对抗,和玩家对抗就是博弈论范畴了,可以参考Shoham and Leyton-Brown 2009).在这个游戏中,自然客观世界会选择一个状态/参数/标签,$y \in Y$,对我们来说是未知的,然后生成了一次观察$x \in X$,这是我们看到的.接下来我们就要做出一次决策(decision),也就是要从某个行为空间(action space)中选择一个行动a.最终会得到某种损失(loss),$L(y, a)$,这个损失函数测量了咱们选择的行为a和自然客观世界隐藏的状态y之间的兼容程度.例如,可以使用误分类损失(misclassitication loss)$L(y, a) = I(y \neq a)$,或者用平方误差损失(squared loss) $L(y, a) = (y - a)^2$.接下来是一些其他例子.

我们的目标就是设计一个决策程序或者决策策略(decision procedure or policy)$\delta : X \to A$,对每个可能的输入指定了最优行为.这里的优化(optimal)的意思就是让行为能够使损失函数期望最小:

$\delta(x) = \arg\min_{a \in A} E[L(y, a)]$(5.96)

在经济学领域,更常见的属于是效用函数(utility function),其实也就是损失函数取负值,即$U(y, a) = -L(y, a)$.这样上面的规则就成了:

$\delta(x) = \arg\max_{a \in A} E[U(y, a)]$(5.97)

这就叫期望效用最大化规则(maximum expected utility principle),是所谓理性行为(rational behavior)的本质.

这里要注意"期望(expected)"这个词,是可以有两种理解的.在贝叶斯统计学语境中的意思是给定了

已经看到的数据之后,对y的期望值(expected value)后面也会具体讲.在频率论统计学语境中,意思是我们期待在未来看到y和x的期望值,具体会在本书6.3当中讲解.

在贝叶斯决策理论的方法中,观察了x之后的最优行为定义是能后让后验期望损失(posterior expected loss)最小的行为.

$$\rho(a|x) \triangleq E_{p(y|x)}[L(y,a)] = \sum_y L(y,a)p(y|x) \text{(5.98)}$$

(如果y是连续的,比如想要估计一个参数向量的时候,就应该把上面的求和替换成为积分.)这样就有了贝叶斯估计器(Bayes estimator),也叫做贝叶斯决策规则(Bayes decision rule):

$$\delta(x) = \arg\min_{a \in A} \rho(a|x) \text{(5.99)}$$

5.7.1 常见损失函数的贝叶斯估计器

这一节我们展示了对一些机器学习中常遇到的损失函数如何构建贝叶斯估计器.

5.7.1.1 最大后验估计(MAP estimate)最小化0-1损失

0-1损失(0-1 loss)的定义是:

$$L(y,a) = I(y \neq a) = \begin{cases} 0 & \text{if} \quad a = y \\ 1 & \text{if} \quad a \neq y \end{cases} \text{(5.100)}$$

这通常用于分类问题中,其中的y是真实类标签(true class label),而$a = \hat{y}$是估计得到的类标签.

例如,在二分类情况下,可以写成下面的损失矩阵(loss matrix):

	\hat y=1	\hat y=0
y=1	0	1
y=0	1	0

此处查看原书图master/Figure/5.13

(在本书5.7.2,会把上面这个损失函数进行泛化,就可以应用到对偏离对角位置的两种错误的惩罚上了.)

后验期望损失为:

$$\rho(a|x) = p(a \neq y|x) = 1 - p(y|x) \text{(5.101)}$$

所以能够最小化期望损失的行为就是后验众数(posterior mode)或者最大后验估计(MAP estimate):

$$y^{\triangle}(x) = \arg\max_{y \in Y} p(y|x) \tag{5.102}$$

5.7.1.2 拒绝选项(Reject option)

在分类问题中,$p(y|x)$是非常不确定的,所以我们可能更倾向去选择一个拒绝行为(reject action),也就是拒绝将这个样本分类到任何已有的指定分类中,而是告知"不知道".这种模糊情况可以被人类专家等来处理,比如图5.13所示.对于风险敏感的领域(risk averse domains)比如医疗和金融等,这是很有用处的.

接下来讲这个拒绝选项用正规化语言表达一下.设选择一个$a = C + 1$对应的就是选择了拒绝行为,然后选择$a \in \{1, \ldots, C\}$对应的就是分类到类标签中去.然后就可以定义下面的损失函数:

$$L(y = j, a = i) = \begin{cases} 0 & \text{if} \quad i = j \quad i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if} \quad i = C + 1 \\ \lambda_s & \text{otherwise} \end{cases} \tag{5.103}$$

此处查看原书图master/Figure/5.14

5.7.1.3 后验均值(Posterior mean)最小化l_2(二次)损失函数

对于连续参数,更适合使用的损失函数是平方误差函数(squared error),也成为l_2损失函数,或者叫二次损失函数(quadratic loss),定义如下:

$$L(y, a) = (y - a)^2 \tag{5.104}$$

后验期望损失为:

$$\rho(a|x) = E[(y - a)^2 | x|] = E[y^2|x] - 2aE[y|x] + a^2 \tag{5.105}$$

这样最优估计就是后验均值:

$$\frac{\partial}{\partial a}\rho(a|x) = -2E[y|x] + 2a = 0 \Longrightarrow \hat{y} = E[y|x] = \int yp(y|x)dy \tag{5.106}$$

这也叫做最小均值方差估计(minimum mean squared error,缩写为MMSE).

在线性回归问题中有:

$$p(y|x, \theta) = N(y|x^T w, \sigma^2) \tag{5.107}$$

这时候给定某个训练集D之后的最优估计就是:

$$E[y|x, D] = x^T E[w|D] \tag{5.108}$$

也就是将后验均值参数估计代入.注意不论对w使用什么样的先验,这都是最优选择.

5.7.1.4 后验中位数(Posterior median)最小化l_1(绝对)损失函数

l_2(二次)损失函数以二次形式惩罚与真实值的偏离,因此对异常值(outliers)特别敏感.所以有一个更健壮的替换选择,就是绝对损失函数,或者也叫做l_1损失函数$L(y,a)=|y-a|$(如图5.14所示).这里的最优估计就是后验中位数,也就是使得$P(y<a|x)=P(y\geq a|x)=0.5$的a值,具体证明参考本书练习5.9.

5.7.1.5 监督学习(Supervised learning)

设想有一个预测函数$\delta:X\rightarrow Y$,然后设有某个损失函数$l(y,y')$,这个损失函数给出了预测出y'而真实值是y的时候的损失.这样就可以定义采取行为δ(比如使用这个预测器)而未知自然状态为θ(数据生成机制的参数)的时候的损失:

$$L(\theta,\delta)\triangleq E_{(x,y)\sim p(x,y|\theta)}[l(y,\delta(x)]=\sum_x\sum_y L(y,\delta(x))p(x,y|\theta) \text{(5.109)}$$

这就是泛化误差(generalization error).咱们的目标是最小化后验期望损失,即:

$$\rho(\delta|D)=\int p(\theta|D)L(\theta,\delta)d\theta \text{(5.110)}$$

这和公式6.47当中定义的频率论中的风险(risk)相对应.

5.7.2 假阳性和假阴性的权衡

本章关注的是二分类决策问题(binary decision problems),比如假设检验(hypothesis testing),二分类,对象事件监测等等.这种情况下就有两种错误类型:假阳性(false positive,也叫假警报false alarm),就是我们估计的$\hat{y}=1$而实际上真实的是$y=0$;或者就是假阴性(false negative,也叫做漏检测missed detection),就是我们估计的是$\hat{y}=0$而实际上真实的是$y=1$.0-1损失函数同等对待这两种错误.可以用下面这个更通用的损失矩阵来表征这种情况:

	\hat y=1	**\hat y=0**
y=1	0	L_{FN}
y=0	L_{FP}	0

上面的L_{FN}就是假阴性的损失,而L_{FP}是假阳性的损失.两种可能性微的后验期望损失为:

$$\rho(\hat{y}=0|x)=L_{FN}p(y=1|x) \qquad (5.111)$$
$$\rho(\hat{y}=1|x)=L_{FP}p(y=0|x) \qquad (5.112)$$

因此应选 $\hat{y}=1$ 当且仅当:

$$\rho(\hat{y}=0|x) > \rho(\hat{y}=1|x) \tag{5.113}$$
$$\frac{p(y=1|x)}{p(y=0|x)} > \frac{L_{FP}}{L_{FN}} \tag{5.114}$$

如果$L_{FN} = cL_{FP}$,很明显(如练习5.10所示)应该选$\hat{y}=1$,当且仅当$p(y=1|x)/p(y=0|x) > \tau$,其中的$\tau = c/(1+c)$(更多细节参考 Muller et al. 2004).例如,如果一个假阴性的损失是假阳性的两倍,就设$c=2$,然后在宣称预测结果为阳性之前要先使用一个2/3的决策阈值(decision threshold).

接下来要讨论的是ROC曲线,这种方式提供了学习FP-FN权衡的一种方式,而不用必须去选择特定的阈值设置.

5.7.2.1 ROC 曲线以及相关内容

	真实1	真实0	Σ
估计1	TP真阳性	FP假阳性	$\hat{N}_+ = TP + FP$
估计0	FN假阴性	TN真阴性	$\hat{N}_- = FN + TN$
Σ	$N_+ = TP + FN$	$N_- = FP + TN$	$N = TP + FP + FN + TN$

表5.2 从混淆矩阵(confusion matrix)可推导的量.N_+是真阳性个数,$\hat{N}_+$是预测阳性个数,N_-是真阴性个数,$\hat{N}_-$是预测阴性个数.

	$y=1$	$y=0$
$\hat{y}=1$	$TP/N_+ = TPR = sensitivity = recall$	$FP/N_- = FPR = typeI$
$\hat{y}=0$	$FN/N_+ = FNR = missrate = typeII$	$TN/N_- = TNR = specifity$

表5.3 从一个混淆矩阵中估计$p(\hat{y}|y)$.缩写解释:FNR = false negative rate 假阴性率, FPR = false positive rate 假阳性率, TNR = true negative rate 真阴性率, TPR = true positive rate 真阳性率.

如果$f(x) > \tau$是决策规则,其中的$f(x)$是对$y=1$(应该与$p(y=1|x)$单调相关,但不必要是概率函数)的信心的衡量,τ是某个阈值参数.对于每个给定的τ值,可以应用局Ce规则,然后统计真阳性/假阳性/真阴性/假阴性的各自出现次数,如表5.2所示./这个误差表格也叫作混淆矩阵(confusion matrix).

从这个表中可以计算出真阳性率(TPR),也叫作敏感度(sensitivity)/识别率(recall)/击中率(hit rate),使用$TPR = TP/N_+ \approx p(\hat{y}=1|y=1)$就可以计算得到.还可以计算假阴性率(FPR),也叫作误报率(false alarm rate)/第一类错误率(type I error rate),利用

$FPR = FP/N_- \approx p(\hat{y} = 1|y = 0)$.这些定义以及相关概念如表格5.3和5.4所示.在计算损失函数的时候可以任意组合这些误差.

不过,与其使用某个固定阈值τ来计算真阳性率TPR和假阳性率FPR,还不如使用一系列的阈值来运行监测器,然后投影出TPR关于FPR的曲线,作为τ的隐含函数.这也叫受试者工作特征曲线(receiver operating characteristic curve, 简称ROC曲线),又称为感受性曲线(sensitivity curve),如图5.15(a)就是一例.任何系统都可以设置阈值为1即$\tau = 1$来实现左下角的点(FPR = 0, TPR = 0),这样也就是所有的都分类成为阴性了;类似的也可以设置阈值为0即$\tau = 0$,都跑到右上角去,即(FPR = 1, TPR = 1),也就是都分类成阳性.如果一个系统在概率层面(chance level)上运行,就可以通过选择适当阈值来实现对角线上的$TPR = FPR$的任一点.一个能够完美区分开阴性阳性的系统的阈值可以使得整个出在图的左上方,即(FPR = 0, TPR = 1);通过变换阈值,这样的系统就可以从左边的轴移动到顶部轴,如图5.15(a)所示.

一个ROC曲线的质量通常用一个单一数值来表示,也就是曲线所覆盖的面积(area under the curve,缩写为AUC).AUC分数越高就越好,最大的显然就是1了.另外一个统计量是相等错误率(equal error rate,缩写为EER),也叫做交错率(cross over rate),定义是满足FPR = FNR的值.由于FNR=1-TPR,所以可以画一条线从左上角到右下角然后看在哪里切穿ROC曲线就是了(参考图5.15(a)中的A和B两个点).EER分数越低越好,最小显然就是0.

此处查看原书图5.15

	$y = 1$	$y = 0$
$\hat{y} = 1$	$TP/\hat{N}_+ = precision = PPV$	$FP/\hat{N}_+ = FDP$
$\hat{y} = 0$	$FN/\hat{N}_-$	$TN/\hat{N}_- = NPV$

表5.4 从一个混淆矩阵中估计的量.缩写解释:FDP = false discovery probability 错误发现概率, NPV = negative predictive value 阴性预测值, PPV = positive predictive value阳性预测值.

5.7.2.2 精确率-识别率曲线(Precision recall curves)

探测小概率的罕见事件(比如检索相关文档或在图中查找面孔)时候,阴性结果的数量会非常大.这样再去对比真阳性率$TPR = TP/N_+$和假阳性率$FPR = FP/N_-$就没啥用处了,因为假阳性率FPR肯定会很小的.因此在ROC曲线上的所有行为都会出现在最左边.这种情况下,通常就把真阳性率TPR和假阳性个数投一个曲线,而不用假阳性率FPR.

不过有时候"阴性(negative)"还不太好定义.例如在图像中探测一个对象(参考本书1.2.1.3),如果探测器通过批量分块来进行探测,那么分块检验过的数目,也就是真阴性的数目,是算法的一个参数,而并不是问题本身定义的一部分.所以就要用一个只涉及阳性的概念来衡量.(注:这样就是问题定义控制的,而不是算法决定了.)

精确率(precision)的定义是$TP/\hat{N}_+ = p(y=1|\hat{y}=1)$,识别率(recall)的定义是$TP/N_+ = p(\hat{y}=1|y=1)$.精确率衡量的是检测到的阳性中有多大比例是真正的阳性,而识别率衡量的是在阳性中有多少被我们检测到了.如果$\hat{y}_i \in \{0,1\}$是预测的分类标签,而$y_i \in \{0,1\}$是真实分类标签,就可以估计精确率和识别率,如下所示:

$$P = \frac{\sum_i y_i \hat{y}_i}{\sum_i \hat{y}_i}, R = \frac{\sum_i y_i \hat{y}_i}{\sum_i y_i} \text{(5.115)}$$

精确率-识别率曲线(precision recall curves)就是随着阈值参数τ的变化对精确率与识别率直接的关系投图得到的曲线.如图5.15(b)所示.在图中曲线尽量往右上角去就好了.

这个曲线可以用一个单一数值来概括,也就是均值精确率(mean precision,在识别到的值上求平均值),近似等于曲线下的面积.或者也可以用固定识别率下的精确率来衡量,比如在前K=10个识别到的项目中的精确率.这就叫做在K分数(K score)的平均精确率.这个指标在评估信息检索系统的时候用得很广.

	Class 1			**Class 2**			**Pooled**	
	$y=1$	$y=0$		$y=1$	$y=0$		$y=1$	$y=0$
$\hat{y}=1$	10	10	$\hat{y}=1$	90	10	$\hat{y}=1$	100	20
$\hat{y}=0$	10	970	$\hat{y}=0$	10	890	$\hat{y}=0$	20	1860

表5.5 展示了宏观和微观平均的区别.y是真实类别标签,$\hat{y}$是名义标签(called label).在这个样例里面,宏观平均精确率是$[10/(10+10)+90/(10+90)]/2 = (0.5+0.9)/2 = 0.7$.微观平均精确率是$100/(100+20) \approx 0.83$.此表参考了(Manning et al. 2008)的表格13.7.

5.7.2.3 F分数(F-scores)*

对于固定阈值,可以计算单个的精确率和识别率的值.然后可以用这些值来计算出一个单个的统计量,就是F分数(F score),也叫做F_1分数(F_1 score),是精确率和识别率的调和均值(harmonic mean):

$$F_1 \triangleq \frac{2}{1/P+1/R} = \frac{2PR}{R+P} \text{(5.116)}$$

使用等式5.115,就可以把上面的式子写成下面的形式:

$$F_1 = \frac{2\sum_{i=1}^N y_i \hat{y}_i}{\sum_{i=1}^N y_i + \sum_{i=1}^N \hat{y}_i} \text{(5.117)}$$

这个量在信息检索测量里面用得很广泛.

要理解为啥用调和均值(harmonic mean)而不用算数均值(arithmetic mean)$(P+R)/2$,可以考虑下面的情况.设识别了全部的项目,也就是识别率$R=1$.精确率可以通过有效率(prevalence)

$p(y=1)$来得到.加入有效率很低,比如$p(y=1)=10^{-4}$.这时候P和R的算数均值为$(P+R)/2=(10^{-4}+1)/2\approx 50\%$.与之相对,调和均值则为$\frac{2\times 10^{-4}\times 1}{1+10^{-4}}\approx 0.2\%$.

在多类情况下(比如文档分类问题),有两种办法来泛化F_1分数.第一种就叫宏观平均F1分数(macro-averaged F1),定义是$\sum_{c=1}^{C}F_1(c)/C$,其中的$F_1(c)$是将类别c与其他分类区分开这个过程的F1分数.另一重定义叫微观平均F1分数(micro-averaged F1),定义是将每个类的情形分析表(contingency table)集中所有计数的F1分数.

表5.5给出了一个样例,可以比对这两种平均F1分数的区别.可见类别1的精确率是0.5,类别2是0.9.宏观平均精确率就是0.7,而微观平均精确率是0.83.后面这个更接近类别2的精确率而远离类别1的精确率,这是因为类别2是类别1的五倍.为了对每个类给予同样的权重,就要用宏观平均.

5.7.2.4 错误发现率(False discovery rates)*

假设要用某种高通量(high throughput)测试设备去发现某种罕见现象,比如基因在微观上的表达,或者射电望远镜等等.就需要制造很多二进制决策.形式为$p(y_i=1|D)>\tau$,其中的$D=\{x_i\}_{i=1}^{N}$,N 可能特别大.这种情况也叫做多重假设检验(multiple hypothesis testing).要注意这和标准的二分类问题的不同之处在于是要基于全部数据而不仅仅是x_i来对y_i进行分类.所以这是一个同时分类问题(simultaneous classification problem),这种问题下我们就希望能比一系列独立分类问题有更好的效果.

该怎么设置阈值τ呢?很自然的方法是尽量降低假阳性的期望个数.在贝叶斯方法中,可以用下面的方式计算:

$$FD(\tau,D)\triangleq\sum_i(1-p_i)I(p_i>\tau) \quad (5.118)$$

$$(1-p_i):pr.error$$

$$I(p_i>\tau):discovery$$

其中的$p_i\triangleq p(y_i=1|D)$是你对目标物体会表现出问题中情形的信心.用如下方式定义后验期望错误发现率(posterior expected false discovery rate):

$$FDR(\tau,D)\triangleq FD(\tau,D)/N(\tau,D) \quad (5.119)$$

上式中的$N(\tau,D)=\sum_i I(p_i>\tau)$,是发现项目数.给定一个理想的错误发现率(FDR)的容忍度(tolerance),比如$\alpha=0.05$,就可以调整τ来实现这个要求I这也叫做控制错误发现率(FDR)的直接后验概率手段(direct posterior probability approach),参考(Newton et al. 2004; Muller et al. 2004).

为了控制错误发现率FDR,更有帮助对方法是联合起来估计各个p_i(比如可以使用本书5.5当中提到的分层贝叶斯模型),而不是单独估计.这样可以汇集统计强度,从而降低错误发现率(FDR).更多内容参考(Berry and Hochberg 1999).

5.7.3 其他话题*

这一部分讲一点和贝叶斯决策规则相关的其他主题.这就没地方去详细讲了,不过也都给出了参考文献啥的,读者可以自己去进一步学习.

5.7.3.1 情境强盗(Contextual bandits)

单臂强盗(one-armed bandit)是对老虎机(slot machine)的俗称,这东西在赌场中很常见.游戏是这样的:你投进去一些钱,然后拉动摇臂,等到机器停止运转;如果你很幸运,就会赢到钱.现在加入有K个这样的机器可选.那你该选哪个呢?这就叫做一个多臂强盗(multi-armed bandit),就可以用贝叶斯决策理论来建模了:有K个可能的行为,然后每个都有位置的奖励(支付函数,payoff function)r_k.建立并维护一个置信状态(belief state)$p(r_{1:K}|D) = \prod_k p(r_k|D)$,就可以推出一个最优策略了;这可以通过编译成一系列的吉廷斯指数(Gittins Indices ,参考 Gittins 1989).这个优化解决了探索-利用之间的权衡(exploration-exploitation tradeoff),这一均衡决定了在决定随着胜者离开之前要将每个行为尝试多少次.

然后考虑扩展情况,每个摇臂以及每个玩家,都有了一个对应的特征向量;就叫x吧.这个情况就叫做情境强盗(contextual bandit 参考Sarkar 1991; Scott 2010; Li et al. 2011).比如这里的手臂可以指代要展示给用户的广告或者新闻文章,而特征向量表示的是这些广告或者文章的性质,比如一个词汇袋,也可以表示用户的性质,比如人口统计信息.如果假设奖励函数有一个线性模型,$r_k = \theta_k^T x$,就可以构建一个每个摇臂的参数的分布$p(\theta_k|D)$,其中的D是一系列的元组(tuples),形式为(a, x, r),制定对应的要比是否被拉动,以及其他特征是什么,还有就是输出的结果是什么(如果用户点击广告了就令$r = 1$,否则令$r = 0$).后面的章节中,我们会讲从线性和逻辑回归模型来计算$p(\theta_k|D)$的各种方法.

给定了一个后验,我们必须决定对应采取的行动.常见的一种期发放时,也叫做置信上界(upper confidence bound,缩写为UCB)的思路是要采取能够将下面这个项目最大化的行为:

$$K \triangleq \arg\max_{k=1}^{K} \mu_k + \lambda\sigma_k \tag{5.120}$$

上式中$\mu_k = E[r_k|D], \sigma_k^2 = var[r_k|D]$,而$\lambda$是一个调节参数,在探索(exploration)和利用(exploitation)之间进行权衡.指关节度就是应该选择我们觉得会有好结果的行为(μ_k大),以及/或者选择我们不太确定的行为(σ_k大).

还有个更简单的方法,叫做汤姆森取样(Thompson sampling),如下所述.每一步都选择一个概率等于成为最优行为选择概率的行为k:

$$p_k = \int I(E[r|a, x, \theta] = \max_{a'} E[r|a', x, \theta]) p(\theta|D) d\theta \tag{5.121}$$

可以简单地从后验$\theta_t \sim p(\theta|D)$中取出单一样本来对此进行估计,然后选择$k^* = \arg\max_k E[r|x, k, \theta^t]$.这个方法不仅简单,用起来效果还不错(Chapelle and Li 2011).

5.7.3.2 效用理论(Utility theory)

假设你是一个医生,要去决定是不是对一个病人做手术.设想这个病人有三种状态:没有癌症/罹患肺癌/罹患乳腺癌.由于行为和状态空间都是连续的,就可以按照下面这个损失矩阵(loss matrix)来表达损失函数$L(\theta, a)$:

	做手术	不做手术
没有癌症	20	0
肺癌	10	50
乳腺癌	10	60

这些数字表明,当病人有癌症的时候不去做手术是很不好的(取决于不同类型的癌症,损失在50-60),因为病人可能因此而去世;当病人没有患上癌症的时候不进行手术就没有损失(0);没有癌症还手术就造成了浪费(损失为20);而如果患上了癌症进行手术虽然痛苦但必要(损失10).

很自然咱们要去考虑一下这些数字是从哪里来的.本质上这些数字代表了一个冒险医生的个人倾向或者价值观,甚至可能有点任意性:有的人可能喜欢巧克力冰淇淋,而有人喜欢香草口味,这类情况下并没有正确/损失/效用函数等等.可是也有研究(DeGroot 1970)表明,任意一组的持续倾向都可以转换成一个标量的损失/效用函数.这里要注意,这个效用可以用任意的尺度来衡量,比如美元啊等等,反正只和对应情况下有影响的值相关.

5.7.3.3 序列决策理论(Sequential decision theory)

之前我们讲的都是单次决策问题(one-shot decision problems),就是每次都是做出一个决策然后就游戏结束了.在本书10.6,我们会把这个繁华到多阶段或者序列化的决策问题上.这些问题在很多商业和工程背景下都会出现.这些内容和强化学习的内容紧密相关.不过这方面的进一步讨论超出了本书的范围了.

练习略.