MLAPP-C5

5.贝叶斯统计

5.1 引言

我们已经看到了各种不同的概率模型，并且讨论了如何基于数据去训练这些模型，比如说：基于不同的先验分布，我们讨论了如何计算参数的最大后验（MAP）估计。我们还讨论了如何计算全后验概率分布，以及在某些特定情况下的后验预测分布（在后面的章节中，我们将讨论在一般情况下的计算方法）。

使用后验分布来概况我们对一系列未知参数的所有了解情况是**贝叶斯统计**（Bayesian statistics）中的核心思想。本章，我们将讨论这种统计方法的更多细节。第6章，我们将讨论另一种统计方法，它属于频率学派，又被称为古典学派。

5.2 关于后验分布的相关总结

后验分布总结了我们关于未知参数的所有情况。本节，我们将讨论一些简单的统计量，这些统计量可以从一个概率分布（比如一个后验概率分布）中推导出来，且相较于完整的联合分布，这些具备总结性质的统计量通常更加容易理解和可视化。

5.2.1 最大后验估计

通过计算后验分布的期望、中值和众数，我们可以很容易得到一个未知变量（**译者注：在贝叶斯理论中，参数被作为变量看待**）的**点估计**（point estimate）。在5.7节，我们将讨论如何通过决策论的方法来决定选择哪种方法。典型情况下，对于一个实数变量而言，后验分布的期望或中值是最合适的选择，对于离散变量而言，后验分布的边缘分布所构成的向量是最好的选择。然而，后验分布的众数，又被称为MAP估计，是最普遍的选择，因为对它的计算将退化为对一个优化问题的求解，而对这个优化问题的求解通常存在有效的算法。更进一步地讲，如果我们将先验分布只是作为一种正则项（6.5将介绍更多细节），那么MAP估计也可以不属于贝叶斯统计学派。

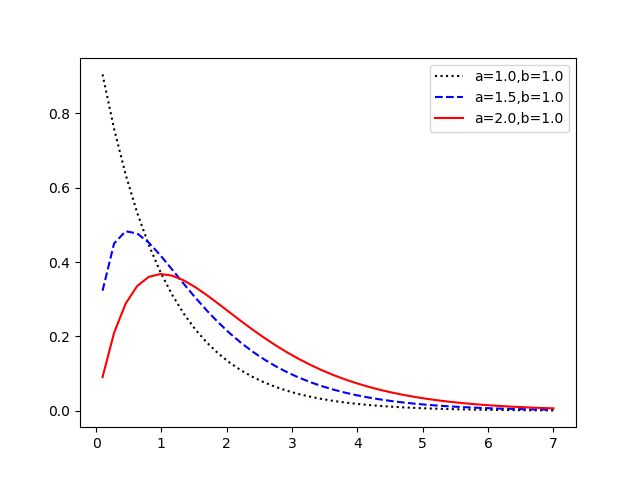
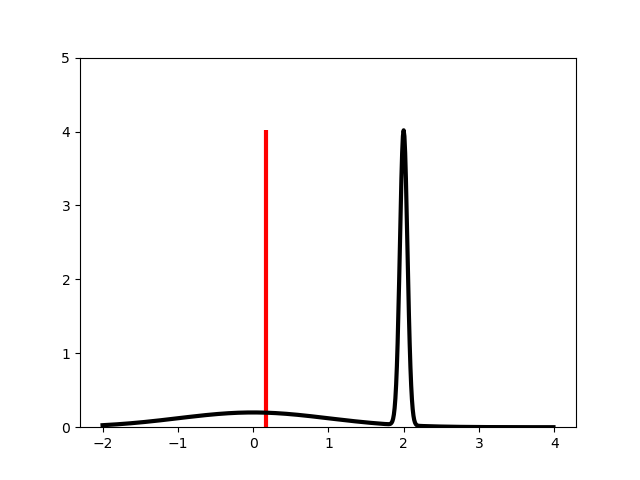
尽管MAP估计的计算很具有吸引力，但有一点很重要的是MAP估计存在各种缺点，我们将在后面作简明讨论。这将导致我们在本章后面的内容中讨论更加健壮的贝叶斯方法（包括本书的其他地方）。

5.2.1.1 无法对不确定度进行衡量

MAP估计以及其他**点估计**（point estimate）方法（比如后验期望或者中值）的最明显的缺点是无法提供对估计值不确定度的衡量。在很多应用中，了解对一个估计值的信任程度是很重要的一件事。正如在5.2.2节中所讨论的那样，我们可以从后验分布中推导出不确定度。

5.2.1.2 基于MAP估计的预测将导致过拟合

在机器学习中，相较于估计模型的参数，我们更关心利用模型进行预测时的精度。然而，如果我们不知道参数估计值的不确定度，那么我们将对最终的预测分布过分相信。我们在第3章中看到过几个例子，在后面的章节中我们将看到更多。在那些需要规避风险的领域，对预测结果过分相信是一个特别严重的问题，我们在5.7节将看到更多例子。



（a） （b）

图5.1 （a）一个双峰分布，其众数并非分布的典型代表。红色垂直线段为期望，是对整个分布的一个较好的概括，因为它更接近整个分布的大部分概率质量。图形由程序**bimodalDemo**生成。（b）一个倾斜的分布，其众数与期望存在较大差异。图形由程序**gammaPlotDemo**生成。

5.2.1.3 众数并非一个具备代表性的点

将后验分布的众数作为对变量后验分布的总结通常是一种非常不合适的选择，因为与期望或者中数相比，众数通常是一个分布的非典型代表。关于这一点，图5.1（a）给出了在1维连续空间中的说明，图中的问题在于众数只是一个单独的点，而期望和中值考虑了整个的分布空间。另一个例子在图5.1（b）中作出说明：其中众数为0，而期望并非为0。这种倾斜的分布在我们对方差进行推理时经常出现，尤其是在分层模型中。在这种情况下，MAP估计（包括MLE）显然是一个非常差的估计。

如果众数不是对后验分布的一个好的总结，那么我们又该如何对后验分布进行概括呢？方法就是使用决策论，关于这一点我们将在5.7节进行讨论。其基本思想是指定一个损失函数，它表示当真实值是，而估计值是时所需要承担的损失。如果我们使用0-1损失函数，此时最优的估计值是后验分布的众数。0-1损失意味着，只有在你不犯错的情况下，你才能得分，否则你将一无所获，在这种损失函数的情况下，不存在“中间状态”。对于连续变量，我们通常更倾向于使用平方误差损失，其对应的最优估计值为后验分布的期望，关于这一点我们将在5.7节讨论。或者我们使用一个更加健壮的损失函数，，其对应的最优估计为后验分布的中值。

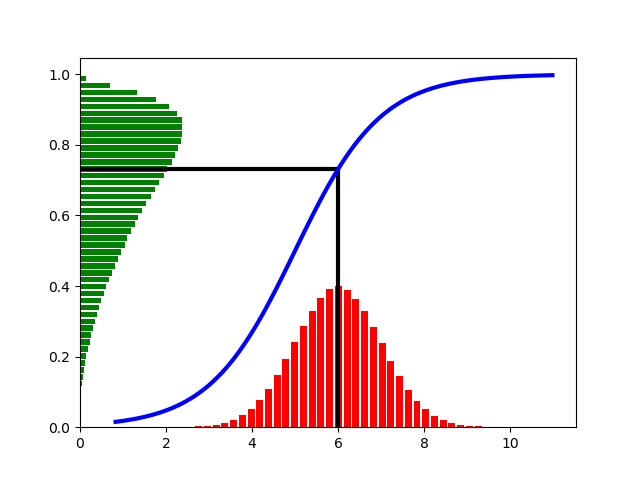


图5.2 在一个非线性变换的情况下的概率密度的转换。注意变换后的分布的众数并非原先众数的变换。图形由程序**bayesChangeOfVar**生成。

5.2.1.4 MAP估计在再参数化过程中具备可变性\*

一个更加微妙的问题在于，我们得到的MAP估计与我们对一个概率分布进行参数化的方式有关。从一种表达方式转变为另一种表达方式，将导致估计值的改变，而这一点是我们所不希望发生的，因为我们对事物的测量基准是任意的（比如，当我们考察测量距离时，我们可能使用厘米或者英尺作为单位）。

为了理解这个问题，假设我们希望计算*x*的后验分布。如果我们定义*y*=*f*(*x*)，根据式2.87，我们可以得到关于*y*的分布，为了方便，我们将式2.87重写：

 （5.1）

其中被称为雅各比，它衡量了在函数*f*的作用下，单位体积的尺寸变化。令为*x*的MAP估计。通常情况下，关于*y*的最大后验估计并不能通过得到。比如说，令，*y*=*f*(*x*)。其中：

 （5.2）

我们可以使用蒙特卡洛近似法得到关于*y*的分布（见2.7.1节）。结果如图5.2所示。我们发现原来的高斯分布因为sigmoid函数的非线性而被压缩。特别需要注意的是，我们发现变换后分布的众数并非等于原先众数的变换。

为了看清楚在MAP估计中这个问题是如何出现的，考虑Michael Jordan提供的如下例子。伯努利分布由参数即它的期望进行确定，所以，其中。假设我们有一个关于参数的，定义在单位区间内的均匀先验分布。如果没有采样数据（即没有似然值），则MAP估计就是先验分布的众数，即0和1之间的任意数。我们将会展示不同的参数化将会导致众数的变化。

首先令，所以。则先验分布为：

 （5.3）

上式中，所以新的众数为：

 （5.4）

现在我们令，新的先验分布为：

 （5.5）

上式中，所以新的众数为：

 （5.6）

所以最大后验概率估计取决于参数化的方式。MLE不会遇到这样的问题，因为似然函数是一个函数，而非一个概率密度。贝叶斯推断也不会遇到这个问题，因为测量基准的变化已经被考虑到了对参数空间的积分中去了。

一种解决这个问题的方法是优化下面的目标函数：

 （5.7）

其中为与相关的费舍尔信息矩阵（见6.2.2节）。这个估计值与参数化的方式无关，理由在相关文献中记载。不幸的是，优化式5.7是一件非常困难的事情，从而降低了整个方法的吸引力。

5.2.2 可靠区间

除了点估计，我们经常需要对我们的可信度进行一个衡量。我们可以将该变量（标量）后验分布的“宽度”作为一种可信度的标准测度。这个宽度可以使用一个的**可靠区间**（credible interval）去衡量，该区间是一个（连续）区域（表示下界和上界），包含整个概率质量的。比如说：



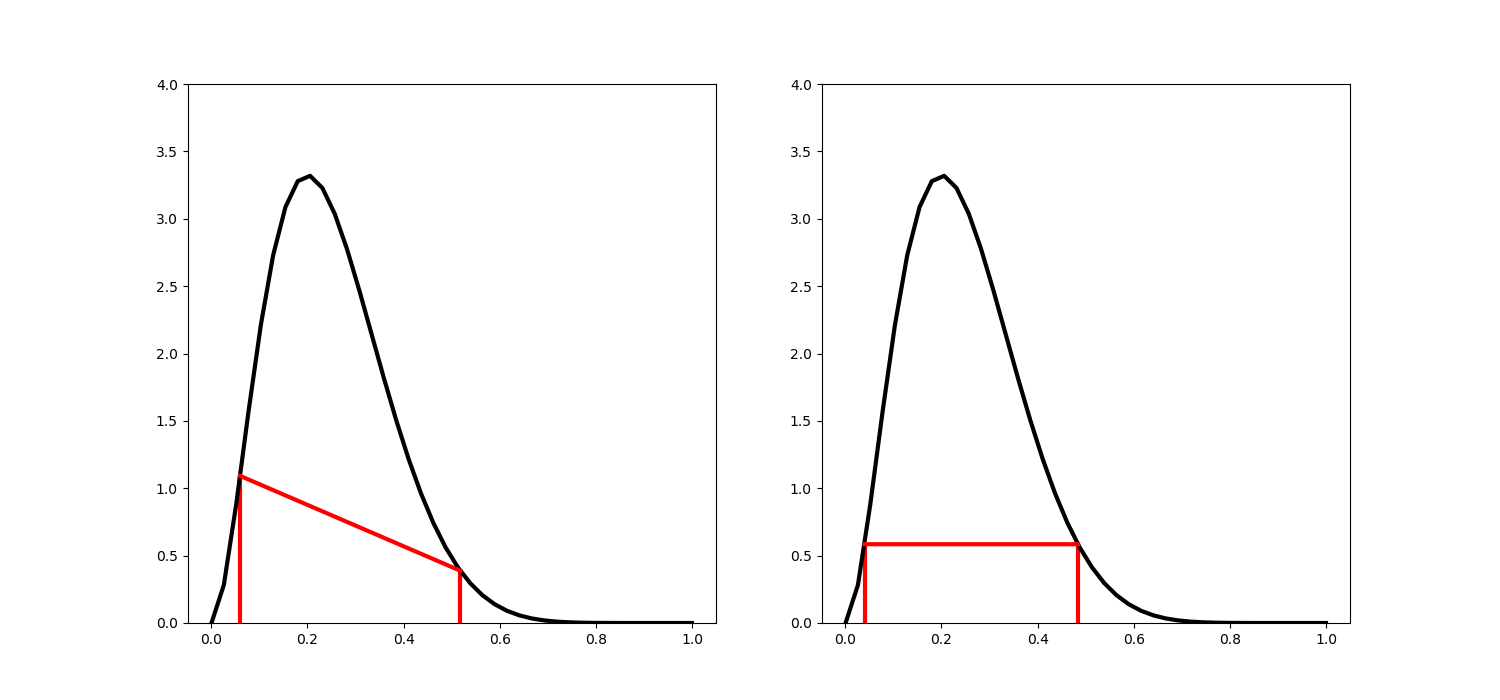
这样的区间会有很多，所以我们选择一个区间使得在分布的两个尾端其概率质量占比分别为，这个区间被称为**中心区间**（central interval，CI）。

如果后验分布的形式是一个已知函数，那么我们可以使用和求解中心区间，其中*F*为后验分布的累积密度函数。比如说，如果后验分布是一个高斯分布，当时，我们有，，其中表示高斯分布的累积概率函数。这一点在图2.3（c）中给出说明。这验证了一个在标注可靠区间时的一个常规操作，即，其中表示后验期望，表示后验标准差，2是1.96的一个比较好的近似。

当然，后验分布不一定总是高斯分布。举例来说，在我们的硬币试验中，如果我们使用一个均匀先验分布，并且观察到在*N*=100次试验中观察到*N*1=47次正面朝上。则对应的后验分布属于一个Beta分布，。我们发现95%的后验可靠区间为（0.3749,0.5673）（程序**betaCredibleInt**给出案例）。

如果我们不知道函数的形式，但是我们可以从后验分布中采集样本，然后使用蒙特卡洛近似法：我们只是对*S*个样本进行排序，然后找到在有序列表中处在位置（**译者注**：个人觉得是排在）的那个样本。当，这个值将收敛于正确的分位数。程序**mcQuantileDemo**给出了一个样例。

人们经常将贝叶斯可靠区间与统计学派的置信区间混淆。然而，它们并非一样，关于这一点我们将在6.6.1节进行讨论。通常情况下，可靠区间是人们想要计算的，但置信区间是人们实际计算的，因为大部分人都只被教授关于频率学派而非贝叶斯学派的相关知识。幸运的是，计算可靠区间的方法与计算置信区间的方法一样简单（程序**betaCredibleInt**给出一个实现案例）。



（a） （b）

图5.3 （a）Beta(3,9)分布的中心区域；（b）Beta(3,9)分布的HPD。CI为[0.06，0.52]，HPD为[0.04，0.48]。图形由程序**betaHPD**生成。

5.2.2.1 最高后验密度区域\*

中心区间的问题在于在该区间之外可能存在更高的概率密度。图5.3（a）说明了这一点，我们发现在中心区间的左边界的左侧（中心区间外部）的概率密度依然比右边界的左侧（中心区间内部）的概率密度高。。

这个现象导出了另一个可选参数，被称为**最高后验密度**（highest posterior density）或者**HPD**区域。这个区域定义为在概率质量占比为的区域内最有可能的点。更加正式的表达为，我们找到概率密度函数的阈值，使其满足：

 （5.9）

定义HPD为：

 （5.10）

在1维情况下，HPD区域又被称为**最高密度区间**（highest density interval）或者**HDI**。举例来说，图5.3（b）展示了分布Beta(3,9)的95%的HDI，为（0.04,0.48）。我们发现这个区间比CI要窄，尽管它也包含95%的概率质量；更进一步，每一个处在HDI内的点的概率密度都比区间外的概率密度高。

（a） （b）

图5.4 （a）多峰后验分布的中心区域（b）HPD区域。图形来自于原书。

对于一个单峰分布，HDI将会是围绕众数的包含95%概率质量的最狭窄的区间。为了说明这一点，我们假设整个概率分布构成一个容器，容器内装满了水，我们从峰值开始逐渐降低我们的水位，直到95%的水显露出来，只有5%的质量的水被遮挡。这个方法导出了计算1维情况下HDIs的一个简单算法：只需要找到那些包含95%概率质量的区间并且确保该区间宽度最小。如果我们知道概率分布的逆累积密度函数，那么这个问题就可以通过1维数值优化去求解，或者如果我们有一些样本，我们可以通过搜寻排序好的数据点，从而解决这个问题。（程序**betaHPD**给出一个实现案例）。

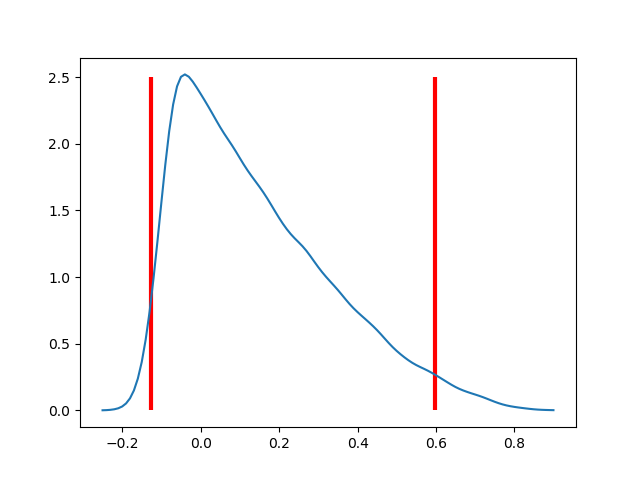
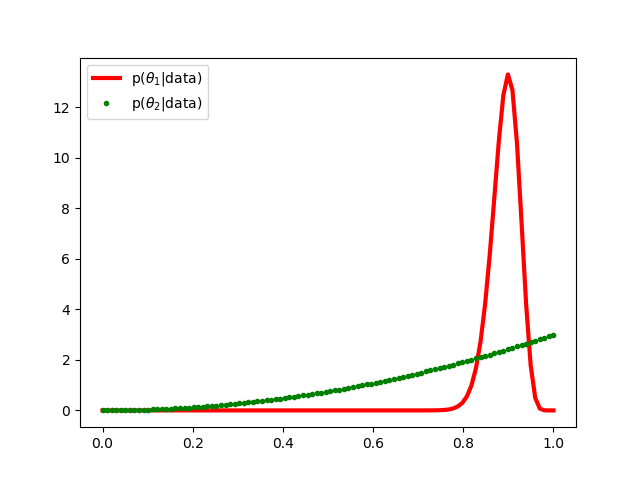
如果后验分布是多峰分布，HDI甚至可能不是一个连续的区间：见图5.4（b）中展示的案例。然而，对多峰后验分布进行总结往往是很困难的。

5.2.3 对比例差异的推断

有些时候，我们可能有多个参数，我们对计算关于这些参数的函数的后验分布比较感兴趣。举例来说，假设你想从亚马逊上买某样东西，有两个卖家为之出了同样的价格。卖家1有90个正面评价和10个负面评价。卖家2有2个正面评价和0个负面评价。那你应该去买哪家商品呢？

表面看来，你应该选择卖家2，但我们不能确信卖家2更好，因为它的评论数太少了。本节，我们将描绘一种关于这个问题的贝叶斯方法。相似的方法可以用在其他不同的情景之下。

令和表示两个卖家未知的可信赖程度。既然我们对它们的情况了解并不多，我们可以赋予这两个变量均匀先验分布。后验分布服从，。



（a） （b）

图5.5 （a）的精确后验分布。（b）关于的MC近似。我们使用了核密度估计方法得到了一个光滑的密度分布曲线。垂线是95%的CI。图形由程序**amazonSellerDemo**生成。

我们希望计算。为了简单起见，定义作为两种变量之间的差。（另一种可选的方案是我们可能希望基于对数几率比例进行计算。）我们可以通过数值积分的方式计算目标值：

 （5.11）

我们发现（**译者注**：本人算出来是0.713），意味着你最好选择卖家1！代码**amazonSellerDemo**给出实现细节。

一种更简单的方式是通过蒙特卡洛方法对后验分布进行估计。这个过程十分容易，因为和在后验分布中是独立的，且两者都服从Beta分布，进而我们可以使用标准的方式进行采样。分布如图5.5（a）所示，图5.5（b）展示了的MC估计，以及95%HPD（译者注：此处应该是CI）。关于的MC估计可以通过统计样本中的占比；这个结果为0.718（**译者注**：本人算出来是0.713），这与精确值十分相近。（见**amazonSellerDemo**程序）。

5.3 贝叶斯模型选择

在图1.18中我们发现，当使用阶数过高的多项式进行拟合时容易出现过拟合，反之出现欠拟合。同样，在图7.8（a）中，我们将会发现，使用过小的正则化参数会导致过拟合，反之则出现欠拟合。一般情况下，当我们面对一系列具有不同复杂度的模型时（比如含参模型的分布族），我们该如何选择最好的那一个呢？这被称为**模型选择**（model selection）问题。

一种方法是使用交叉验证的方法去估计候选模型的泛化误差，然后挑选出似乎是最好的那一个模型。然而，这种方法要求训练每个模型*K*次，其中*K*表示交叉验证过程中分包的数量。一个更加有效的方式是计算模型的后验分布，

 （5.12）

通过上式，我们可以很容易地计算出MAP模型，。这被称为**贝叶斯模型选择**（Bayesian model selection）。

如果我们使用关于模型的均匀先验分布，这将等价于挑选使下式最大的模型

 （5.13）

这个变量被称为**边缘似然**（marginal likelihood）或者**积分似然**（integrated likehood）又或者叫模型*m*的**证据**（evidence）。关于如何计算这个积分我们将在5.3.2节介绍，但首先我们将给出这个变量含义的直观解释。

5.3.1 贝叶斯奥卡姆剃刀

或许你会觉得如果使用来作为模型选择的依据，那么我们总会选择那些具有最多参数的模型。事实上，如果我们使用来对模型进行选择，其中表示模型*m*的参数的MLE或者MAP估计，因为具备更多参数的模型将更好地适应数据，因此会得到更高的似然值。然而，如果我们对整个参数空间进行积分，而不是只是选择最大化这些参数，我们将自动地避免过拟合的发生：具备更多参数的模型不一定具备更高的边缘似然。这被称为**贝叶斯奥卡姆剃刀**（Bayesian Occam’s razor）效应，该术语在**奥卡姆剃刀**（Occam’s razor）原则之后命名，该原则指出在足够解释数据的前提下，我们应该选择最简单的模型。

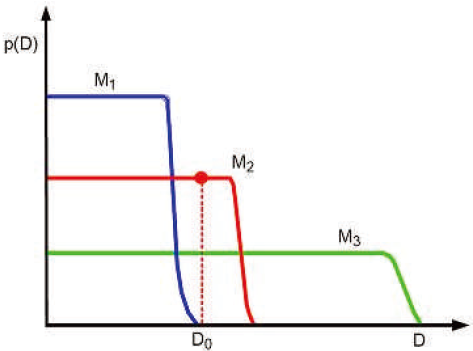


图5.6 贝叶斯奥卡姆剃刀原理的原理示意图。更广的（绿色）曲线代表复杂的模型，较窄的（蓝色）曲线代表简单的模型，中间（红色）曲线则刚刚好。**图形来自于原书。**

一种理解贝叶斯奥卡姆剃刀的方式是注意到边缘似然可以按照如下的方式重写，基于概率论中的链式法则，我们有：

 （5.14）

式中为了书写上的简洁，我们省去了条件变量**x**。这与似然函数的留一法交叉验证估计（章节1.4.8）十分相似，因为我们在给定所有历史数据的基础上预测每一个未来的点。（当然，数据的顺序在上述表达式中没有关系。）如果一个模型太过复杂，它会在“很早”的案例中产生过拟合，并且在剩下的案例中产生不好的预测。

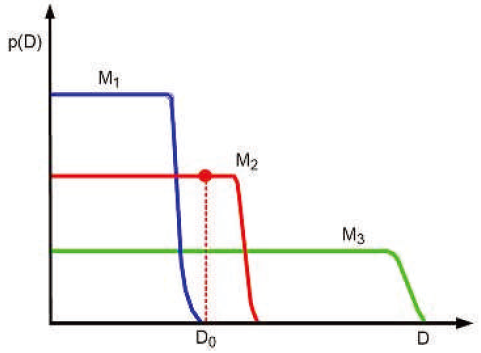


图5.6 贝叶斯奥卡姆剃刀的失意性说明。宽（绿色）的线对应一个复杂的模型，窄（蓝色）的线对应一个简单的模型，中间（红色）的线刚好正确。图形来自于原书。

另一种理解贝叶斯奥卡姆剃刀效应的方法是注意到概率值的和必须为1。所以，其中求和是基于所有可能的数据集。而复杂的模型务必会预测很多东西，从而使得整个概率质量分布变得稀疏，这样我们无法像使用一个更加简单的模型那样，在任意的给定数据集上得到更大的概率值。这通常被称为**概率质量守恒**（conservation of probability mass）原则，在图5.6中给出说明。在水平轴，我们按照模型的复杂度顺序绘制出了所有可能的数据集。在垂直轴中，我们绘制了3个可能模型的预测结果：一个简答的模型*M*1；一个适度复杂的模型*M*2；一个复杂的模型*M*3。我们同时通过一条垂直直线表示实际观察到的数据集0。模型1因为过于简单，所以赋予观察数据集0较低的概率值。模型3同样赋予数据集0较低的概率值，因为它可以预测更多的数据集，因此它将它的概率质量分散的更加宽广和稀疏。模型2则刚刚好：它以一个合理的可信度对数据集0进行预测，但并没有预测更多的数据集。所以模型2是最有可能的模型。

给出一个贝叶斯奥卡姆剃刀的例子，考虑图5.7中的数据。我们绘制了阶数为1,2和3的多项式拟合曲线，来适应*N*=5个数据点。针对模型本身我们使用一个高斯先验分布（见7.6节），图中向我们展示了关于模型的后验分布。因为没有足够的数据去适应一个复杂的模型，所以MAP模型为*d* = 1。图5.8展示了当*N*=30时会发生的情况。此时，*d* = 2是正确的模型（实际上数据就是从2次函数生成的）。

另一个例子在图7.8（c）中展示，该图绘制了在多项式岭回归模型中，与之间的关系，其中的取值区间与在交叉验证实验中的取值区间一致。我们发现最大证据发生的地方与最小均方差发生的地方几乎是一样的，而这个点就是我们通过CV方法选择的点。

当使用贝叶斯方法时，我们并未被限制在一个有限网格点上进行证据的评估。取而代之的是，我们可以使用数值优化的方式去搜寻。这种方法被称为**经验贝叶斯**（empirical Bayes）或者**第II类极大似然**（type II maximum likehood）（5.6节给出更多细节）。图7.8（b）给出了一个例子：我们发现曲线与CV估计有一个相似的形状，但它可以以一个更加高效地方式进行计算。

5.3.2 计算边缘似然（证据）

当我们讨论对一个固定模型进行参数推断时，我们通常写成：

 （5.15）

上式省略了归一化常数，这种方式是有效的，因为关于参数是常数。然而，当我们比较不同模型的性能时，我们需要知道如何计算边缘似然。通常情况下，这会是非常困难的事情，因为我们需要基于整个参数空间进行积分，但当我们有一个共轭先验时，事情将会变得简单。

令为我们的先验分布，其中为一个非归一化分布，为先验分布的归一化常数。令表示似然函数，其中包含在似然函数中任意的常数项。最后令为我们的后验分布，为未归一化后验分布，为后验分布的归一化常数。我们有：

 （5.16）

 （5.17）

 （5.18）

假设相关的归一化常数是容易解决的，我们有一个简单的方式去计算边缘似然。我们将在下文给出例子。

5.3.2.1 Beta-binomial 模型

让我们将上述结果应用在Beta-binomial模型。既然我们知道，其中，我们知道其归一化常数为。所以：

 （5.19）

 （5.20）

 （5.21）

所以：

 （5.22）

 （5.23）

Beta-Bernoulli模型的边缘似然与上述与一样的，除了缺少项之外。

5.3.2.2 Dirichlet-multinoulli 模型

按照Beta-Bernoulli的推理方法，我们可以得到关于Dirichlet-multinoulli模型的边缘似然函数：

 （5.24）

其中：

 （5.25）

所以我们可以对上式采取不一样的形式书写，该形式更多地出现在相关文献中：

 （5.26）

我们将会在后面的内容中看到更多关于上式的应用。

5.3.2.3 高斯-威舍特-高斯模型

考虑多变量似然函数，同时对应一个服从逆威舍特分布的共轭先验分布（这被称为高斯-威舍特-高斯模型，因为我们可以将高斯-威舍特先验分布与高斯似然函数组合）。令Z0为先验分布的归一化常数，Z*N*为后验分布的归一化常数，令为似然函数的归一化常数。我们很容易地发现：

 （5.27）

 （5.28）

 （5.29）

这个公式在后面的章节中将被证明是有用的，比如我们在考虑混合高斯模型时。

5.3.2.4 对数边缘似然函数的BIC近似

通常情况下，计算5.13的积分是非常困难的事情。一种简单但是非常著名的近似方法被称为**贝叶斯信息准则**（Bayesian information criterion, BIC），其形式如下：

 （5.30）

其中表示模型的**自由度**（degrees of freedom），为模型的MLE。不难发现上式是一个**含惩罚项的对数似然**（penalized log likelihood），其中的惩罚项取决于模型的复杂度。章节8.4.2将给出BIC的推导。

以线性回归为例，该模型将会在7.3节介绍，其MLE等于，，其中。其相应的对数似然函数为：

 （5.31）

所以相应的BIC值为（丢弃常数项）：

 （5.32）

其中*D*为模型中变量的数目。在统计学的相关文献中，通常使用BIC的另一种替代方案，被称为BIC成本（cost）（因为我们希望最小化这个参数）：

 （5.33）

在线性回归模型中，上式为：

 （5.34）

BIC方法与**最小描述长度**（minimum description length，MDL）原则具有紧密的联系，该原则根据“模型对数据的适应程度减去模型的复杂程度”为模型赋分。

与BIC或者MDL相似的一个表达式被称为**赤池信息量准则**（Akaike information criterion，AIC），定义为：

 （5.35）

上式是在频率学派的框架内推导出来的，不能解释为边缘似然函数的近似。然而，该式的形式与BIC十分相似。我们发现AIC的惩罚项比BIC小。从而导致AIC可以选择更加复杂的模型。然而，这可以导致更加好的预测精度。

5.3.2.5 先验分布的影响

有时我们并不知道如何设置先验分布。当我们在基于后验分布进行推理时，先验分布的细节可能并没有影响，因为似然函数会“压倒”先验分布。但当我们计算边缘似然时，先验分布将扮演一个非常重要的角色，因为我们需要基于所有可能的参数设置值对似然函数求加权平均，其权重就是先验分布。

在图5.7和5.8中，我们说明了对于线性回归所进行的模型选择，我们使用了一个先验分布。其中参数为一个可调节参数，它控制着先验分布的强度。这个参数具有很大的影响，关于这一点我们会在7.5节讨论。更加直观的感受，如果很大，那么权重将被约束到很小的值，所以我们需要使用一个复杂的模型去适应数据，这个复杂的模型具有很多取值较小的参数。相反，如果取值较小，我们将更倾向于简单的模型，因为每一个参数都被允许在更大的幅度范围内变化。

如果先验分布未知，正确的贝叶斯方法是为先验分布再赋予一个先验分布。也就是说，我们需要同时为参数和权重赋予先验分布。为了计算边缘似然，我们需要对所有未知参数进行积分，也就是说，我们需要计算：

 （5.36）

当然，这需要指定**超先验**（hyper-prior）。幸运的是，在贝叶斯方法中先验分布的层次越高，最终结果对先验分布设置的敏感性越弱。所以我们通常可以设置超先验分布为无信息先验分布。

一个计算的捷径是对进行优化而非对它进行积分。也就是说，我们使用

 （5.37）

其中

 （5.38）

这个方法被称为**经验贝叶斯**（empirical Bayes, EB），在5.6节将讨论更多的细节。它是图5.7和5.8中使用的方法。

5.3.3 贝叶斯因子

假设我们基于模型的先验分布是均匀的，即。那么模型选择等价于选择具有更高边缘似然的模型。现在假设我们只有两个模型，分别称它们为**虚假设**（null hypothesis）*M*0和**备择假设**（alternative hypothesis）*M*1。定义**贝叶斯因子**（Bayes factor）为边缘似然函数的比例：

 （5.39）

（上式与似然函数比十分相似，除了我们需要对模型的参数进行积分，这将允许我们对比具有不同复杂度的模型。）如果*BF*1,0>1，那么我们会选择模型1，否则选择模型0。

|  |  |
| --- | --- |
| 贝叶斯因子BF(1,0) | 解释 |
| BF < 1/100 | Decisive evidence for M0 |
| BF < 1/10 | Strong evidence for M0 |
| 1/10 < BF < 1/3 | Moderate evidence for M0 |
| 1/3 < BF < 1 | Weak evidence for M0 |
| 1 < BF < 3 | Weak evidence for M1 |
| 3 < BF < 10 | Moderate evidence for M1 |
| BF > 10 | Strong evidence for M1 |
| BF > 100 | Decisive evidence for M1 |

表5.1 解释贝叶斯因子的Jeffreys尺度

当然，也有可能*BF*1,0只是稍微大于1。在那种情况下，我们并没有足够的信心确保模型1就是最好的。Jeffreys提出了一个解释贝叶斯尺度因子大小的证据尺度，如表5.1所示。这个概念与频率学派中的*p*值（*p*-value）相对应。另一种可选的方案是，我们可以只是将贝叶斯因子转换为模型的后验分布。如果*p*(*M*1)=*p*(*M*0)=0.5，我们有

 （5.40）

5.3.3.1 例子：测试一个硬币是否是均匀的

假设我们观察了一些硬币投掷试验，并且想知道这些实验结果是由一个均匀的硬币（即），还是一个可能不均匀的硬币（其中可能是[0,1]间的任何一个值）产生的。我们将第一个模型设置为*M*0，第二个模型设置为*M*1。模型*M*1的边缘似然函数为：

 （5.41）

其中*N*为硬币投掷的次数。在*M*1模型的情况下，使用Beta先验分布，边缘似然为：

 （5.42）

我们在图5.9（a）中绘制了与正面朝上的次数*N*1的关系。假设*N*=5，。（只要，曲线的形状对两者的取值并不敏感。）如果我们观察到2次或者3次正面朝上，则硬币是均匀的这一假设*M*0比*M*1更有可能，因为*M*0是一个更加简单的模型（它没有自由参数）——如果硬币是不均匀的，那么最终试验产生的结果几乎刚好是一半正面/一半反面会是一个非常可疑的巧合。然而，当正面与反面的结果发生极端的倾斜，我们将支持硬币是不均匀的这一假设。需要注意的是，如果我们绘制对数贝叶斯因子，它们的形状将完全一样，因为是一个常数。

图5.9（b）展示了对边缘似然的近似值BIC。我们发现曲线与精确的对数边缘似然曲线形状一样，这对于模型选择的任务已经完全足够了，因为绝对的数值已经无关紧要了。特别地，除非试验数据是压倒性的，否则我们将更倾向于更简单的模型。

5.3.4 Jeffreys-Lindley 悖论\*

对于模型选择/假设检验的任务，当我们使用一个不合适的先验分布（比如先验分布的积分不为1）时，问题可能会出现，尽管出于其他的目的，这些先验分布可能是可以被接受的。举例来说，考虑检验两个假设。为了定义的边缘概率密度，我们使用下面的混合模型：

 （5.43）

只有在和是合适的（归一化后的）密度函数时，上式才有意义。在这种情况下，后验分布为：

 （5.44）

（5.45）

现在假设我们使用一个不合适先验分布，和。然后我们有：

 （5.46）

 （5.47）

其中为模型*i*的积分或边缘似然。令。所以：

 （5.48）

所以通过选择和，我们可以任意地改变后验分布。注意使用合适的但模糊的先验分布会导致同样的问题。特别地，贝叶斯因子总是会支持那些更加简单的模型，因为一个复杂的模型如果其先验分布是非常分散的，那么基于该模型观察到的数据的概率将会很低。这被称为Jeffreys-Lindley悖论。

所以当进行模型选择的时候，选择合适的先验分布是十分重要的事情。然而，值得注意的是，如果模型*M*0和*M*1在某个参数子集上具有同样的先验分布，那么这部分先验分布将会是不合适的，因为相应的归一化常数将会被消掉。

5.4 先验分布

贝叶斯统计最饱受争议的地方就是它对先验分布的依赖。贝叶斯学派认为这是不可避免的，因为没有人是一块白板（tabula rasa或者blank slate）：所有的推理都必须基于对这个世界的某些特定假设。然而，或许我们想将一个人的先验假设的影响最小化。我们在下面将讨论几种方法。

5.4.1 无信息先验分布

如果我们对参数的相关信息没有足够的信心，我们通常会使用一个**无信息**（uninformative）或者**非信息**（non-informative）先验，并将最终的预测全部“寄托”在数据上。

事实上，设计无信息先验是有一些麻烦的。举一个比较困难的例子，考虑一个伯努利分布的参数。有些人可能会觉得，最有可能的无信息先验分布是均匀分布，即Beta(1,1)。但在这种情况下的后验期望，然而MLE为。所以有些人会觉得所谓的无信息先验分布并非完全意义上的无信息。

显然，通过降低伪计数的大小，我们可以减轻先验分布的影响。基于上面的讨论，最符合无信息先验的分布为：

 （5.49）

上式为在0和1位置的两个等质量点的混合。这又被称为Haldane先验。注意Haldane并非是一个合适的先验分布，意味着它的积分并非等于1。然而，只要我们观察到至少1次正面朝上和1次反面朝上的试验结果，那么后验分布就是合适的。

在5.4.2.1节，我们将发现“正确的”无信息先验分布事实上是Beta(1/2,1/2)。显然在实践中这三个先验分布的区别非常有可能被忽略。一般情况下，建议进行某种**敏感性分析**（sensitivity analysis），在这个过程中，我们检测随着模型假设（其中包括先验分布的选择、似然函数的选择和数据预处理方式）的改变，所导致的结论或者预测结果的改变。如果结果对模型假设的敏感性相对较小，那么我们对结果应该有信心。

5.4.2 Jeffreys priors\*

重新回到上节关于硬币朝上的概率的均匀先验分布的例子。这个先验分布意味着一次正面朝上的概率可以是0和1之间的任何数。那么连续两次正面朝上的概率的多少呢，是？它具备的先验分布为，这与均匀分布相距甚远。

Harold Jeffreys设计了一个创建非信息先验分布的通用技术。结果被称为Jeffreys prior。关键的地方在于如果是无信息先验分布，那么关于的任意形式的再参数化，比如，其分布应该还是无信息的。现在，通过变量的形式变换，我们有：

 （5.50）

所以先验分布一般情况下都会改变。然而，让我们选择

 （5.51）

其中为**费舍尔信息**（Fisher information）：

 （5.52）

这是一个对负对数似然期望曲率的测度，所以它衡量了MLE的稳定性（见6.2.2节）。现在

 （5.53）

对*x*进行平方和求期望操作，得到：

 （5.54）

 （5.55）

然后，我们发现变换后的先验分布为：

 （5.56）

所以与是一样的。

直觉上，我们可以按照如下的方式理解Jeffreys先验：它赋予那些似然函数比较“平坦”的参数值较小的先验权重（因为费舍尔信息衡量的是对数似然的曲率），所以阻止了先验有过度的影响。

5.4.2.1 例子：伯努利和multinoulli分布的Jeffreys先验

假设。对于单个样本的对数似然为：

 （5.57）

**赋分函数**（score function）只是对数似然的梯度：

 （5.58）

**观测信息**（observed information）为对数似然的二阶导：

 （5.59）

费舍尔信息为信息的期望值：

 （5.60）

所以Jeffreys先验分布为：

 （5.61）

现在我们考虑multinoulli随机变量，该变量具有*K*个状态。我们可以发现其对应的Jeffreys先验分布为：

 （5.62）

注意上式与一般的选择，比如不同。

5.4.2.2 例子：位置和尺度参数的Jeffreys先验

位置参数（比如高斯分布的期望）的Jeffreys先验分布为。所以，这是一个**平移不变先验**（translation invariant prior）的案例，它满足在任意区间[*A*, *B*]的概率质量等于具备相同宽度的区间[*A*-*c*, *B*-*c*]内的概率质量。即：

 （5.63）

上式在的情况下成立，该先验分布可以由一个方差无限大的高斯分布近似。注意这不是一个合适的先验分布，因为它的积分不等于1。只要后验分布是合适的，那么使用不合适的先验分布也是可以的，这种情况当我们看到*N* ≥ 1个数据点时会发生，因为只要我们看到一个数据点，我们就可以“钉住”分布的位置。

类似的，尺度参数（比如高斯分布的方差）的Jeffreys先验分布为：。这是一个**尺度不变先验**（scale invariant prior）的例子，它满足如下的性质：在区间[*A*, *B*]内的概率质量等于任意其他的区间[*A*/*c*, *B*/*c*]内的概率质量，其中常数因子*c* > 0。（比如，如果我们将单位从米改成英尺，我们不希望影响最终的推理。）这一点可以通过使用下式达到：

 （5.64）

为了说明这一点，注意到：

 （5.65）

 （5.66）

我们可以使用一个退化的Gamma分布（2.4.5节）对该分布进行近似，。该先验分布也不合适，但只要我们看到*N* ≥ 2个数据，那么后验分布就会是合适的（因为我们至少需要两个数据去估计方差）。

5.4.3 鲁棒先验

在很多情况下，我们对先验分布不是很有信心，所以我们希望确保它对我们的结果不会有过分的影响。这一点可以通过使用**鲁棒先验**（robust priors）实现，这种先验分布的尾部较肥，从而避免强迫结果接近先验的期望值。

让我们考虑一篇文献中的例子。假设。我们观察到了*x* = 5并且想基于此对参数进行估计。其MLE显然是，这貌似是合理的。如果使用均匀先验分布，那后验分布期望也会是。但是现在假设我们已经知道先验分布的中值为0，先验分布的分位数分别是-1和1，所以。同时让我们假设先验分布是光滑且单峰的。

我们可以很容易地得到满足上述先验约束的分布为。但在这种情况下，后验分布的期望为3.43，貌似不是很满意。

现在假设我们使用一个柯西先验。这同样满足上述的先验约束。但是我们发现（使用数值积分的方式：程序**robustPriorDemo**给出实现）后验期望近似为4.6，这个结果貌似更加合理。

5.4.4 混合共轭先验

鲁棒先验虽然有用，但计算成本较高。共轭先验计算简单，但缺乏鲁棒性，并且在对我们的先验知识进行编码的过程中并不灵活。然而，结果表明共轭先验的混合同样是共轭的，并且可以近似任意形式的先验分布。因此这类的先验分布在计算便捷性和灵活性方面提供了一种权衡。

举例来说，我们需要对投掷硬币的试验进行建模，我们认为硬币要么是均匀的，要么是偏向于某一面。这个模型不能通过一个beta分布进行建模。然而，我们可以使用两个beta分布的混合对其进行建模。举例来说，我们可能使用

 （5.67）

如果来自于第一个分布，那么硬币是均匀的，但如果它来自于第二个分布，则它偏向于正面朝上。

我们可以通过引入一个潜在指示变量*z*来表示模型的混合，其中*z* = *k*意味着来自于混合模型中的分量*k*。先验分布具备如下的形式

 （5.68）

其中每一个是共轭先验，被称为（先验）混合权重。同时我们发现后验分布也可以写成共轭分布的混合：

 （5.69）

其中为后验混合权重，由下式给定：

 （5.70）

其中为混合分量*k*的边缘似然（见章节5.3.2.1）。

5.4.4.1 例子

假设我们使用混合先验分布：

 （5.71）

其中和。同时我们观察到*N*1次正面和*N*0次反面。后验分布为：

 （5.72）

如果*N*1=20，*N*0=10，然后使用式5.23，后验分布为：

 （5.73）

图5.10给出了说明。

5.4.4.2 应用：在DNA和蛋白质序列中找到守恒区域

我们在前面的内容中提到，Dirichlet-multinomial模型在基因序列分析中被广泛使用。现在让我们展示一个简单的例子来说明一些比较成熟的机制。特别的，考虑2.3.2.1中的序列logo。现在假设我们希望找到基因组编码区域的位置。迫于进化的压力，这些位置通常情况下在整个序列中具有相同的字母。所以我们需要找到哪些列是“纯”的或者近似“纯”的，直观上理解就是这一列基本上都是As，Ts，Cs或者Gs。一种方式是找到熵比较小的列，它们往往具备确定性的分布（比较纯）。

但是如果我们希望对关于纯度的估计值进行一个可信度的度量。如果我们相信相邻位置的基因倾向于一起保守（译者注：即都不发生突变）。在这种情况下，我们令*Z*t=1，如果在位置*t*的基因是保守的，否则令*Z*t=0。然后，我们可以使用马尔科夫链为相邻两个位置添加一个依赖关系（见17章查看更多细节）。

在任何情况下，我们需要定义一个似然模型，，其中为第t列向量，表示(A,C,G,T)的计数。我们很自然地选择multinomial分布进行建模，其参数为。因为每一列都有一个不同的分布，我们希望对参数进行积分，计算出边缘似然：

 （5.74）

但是关于的先验分布又该如何选择呢？当*Z*t=0，我们可以使用一个均匀先验分布，即，但如果*Z*t=1，我们又该如何选择呢？毕竟，如果一个列是保守的，它可能是一个（近似）纯的列，可以全是As，Cs，Gs或者Ts。一种很自然的选择是使用混合狄利克雷先验分布，每一个分量都倾向于四种状态中的一种：

 （5.75）

因为这是共轭先验，所以我们可以很容易地计算。

5.5 分层贝叶斯

为了计算后验分布，一个关键的要求是指定先验分布，其中为超参数。如果我们不知道如何设置，那又该怎么办呢？在某些情况下，我们可以使用非信息先验分布，这一点我们在前文已经讨论过。一个更加符合贝叶斯学派的做法是为先验分布再添加一个先验分布！如果用图模型（第10章）来表示的话，我们可以用下图来解释这种情况：

 （5.76）

这是一个**分层贝叶斯模型**（hierarchical Bayesian model）的例子，又被称为**多层模型**（multi-level model），因为我们存在多层的未知量。我们在下面给出一个简单的例子，在本书的其他地方还会看到更多的例子。

5.5.1 例子：癌症发病率建模

考虑一个问题，我们需要预测在不同城市中，癌症的发病率。特别地，假设我们统计了不同城市的人口，以及在这些城市中因癌症而死亡的人数*xi*。我们假设，我们希望估计癌症的发病率。一种方式是单独对每个城市的癌症发病率进行估计，但这会带来数据稀疏的问题（由于较小而导致估计值较小）。另一种方式是假设是一样的，这被称为**参数绑定**（parameter tying）。最终的混合MLE为。但假设每个城市的癌症发病率都一样显得过于极端。一种折中的方式假设每个城市的癌症发病率相似但又有可能与每个城市自身相关。这个假设可以通过如下方式建模：从一些相似的分布中采样得到，即。则联合概率分布为：

 （5.77）

其中。

需要注意的是，我们需要从数据中推理而不是将它绑定为常数，这一点很重要。因为如果将其作为常数，那么将会是条件独立的，那么参数之间将不再存在信息流动。相反，将其作为未知（隐）变量，那些统计数据匮乏的城市可以从数据富余的城市**借统计强度**（borrow statistical strength）。

假设我们计算联合后验分布。通过它我们可以知道后验边缘分布。在图5.11（a）中，我们绘制了后验分布的期望（蓝色条状），以及人口水平的平均数（红色线，这代表着的平均数）。我们发现，对于那些样本尺寸较小的城市，其后验期望值向混合估计值靠拢的程度更大。举例来说，城市1和20都只有0个因癌症而死亡的观察案例，但城市20的人口数量更少，所以它的发病率更加接近基于人口水平的估计（离水平红色线更近）。

图5.11（b）展示了后验分布95%的可靠区间。我们发现城市15的人口数很大（53637）,其后验分布的不确定度较小。因此这座城市对参数的后验估计的影响最大，进而影响到其他城市对发病率的估计。城市10和城市19，具有最高的MLE，同样具有最高的后验不确定度，这就反映了一个事实：如此高的估计值与先验分布是冲突的（即它是从其他城市估计出来的）。

在上面的例子中，我们为每个城市都赋予了一个参数，来对目标变量的概率进行建模。通过假设伯努利速率参数为协变量的函数，我们可以对多相关逻辑回归任务建立模型。这被称为多任务学习（multi-task learning），我们将在9.5节讨论更多细节。

5.6 经验贝叶斯

在分层贝叶斯模型中，我们需要计算多层潜变量的后验分布。举例来说，在两层模型中，我们需要计算

 （5.78）

在某些情况下，我们用解析的方式对进行积分，最终遗留一个更加简单的问题，即只需要计算。

作为一种计算上的捷径，我们可以用超参数的点估计来近似它的后验分布，其中。因为在维度上相较于要小得多，发生过拟合的风险更小，所以关于参数我们可以安全地使用一个均匀先验分布。然后估计值将变成：

 （5.79）

其中括号中的量被称为边缘或者积分似然，有时也被称为证据。整个方式被称为**经验贝叶斯**（empirical Bayes，EB）或者**第二类极大似然**（type-II maximum likelihood）。在机器学习中，有时也被称为**证据过程**（evidence procedure）。

经验贝叶斯违背了如下原则：先验分布的选择应该独立于数据。然而，我们可以仅仅将它看作是分层贝叶斯模型的一种近似，只不过这种近似方法在计算量上更少，正如我们在单层模型的推理中，将MAP作为推理结果的近似。事实上，我们可以构造一个分层结构，在这个结构中，我们进行的积分越多，那么整个结构就越显得“贝叶斯”。

值得注意的是，EB具备一个很好的频率学派性质，所以它也被广泛应用于非贝叶斯学派。

5.6.1 例子：beta-binomial模型

让我们回到癌症发病率的例子。我们可以用解析的方式对进行积分，然后直接得到边缘似然函数，为：

 （5.80）

 （5.81）

有不同方式求关于*a*, *b*的最大值。

在计算出*a*, *b*之后，我们可以基于超参数的点估计来计算后验概率分布，使用的方式为共轭分析。最终的结果为：每个参数的后验期望值为它的局部MLE和先验期望的的加权平均，后者依赖于。但是是根据所有的数据估计出来的，所以每个都受全部数据影响。

5.6.2 例子：高斯-高斯模型

我们现在学习另一个与癌症发病率相似的例子，除了数据是实数外。我们将使用一个高斯似然和高斯先验。这将允许我们用解析的方式得到结果。

特别地，假设我们的数据来自多个相关的组。比如，*xij*可能是学生*i*在学校*j*的测试成绩，其中*j*=1 : *D*，*i*=1 : *Nj*。我们希望每个学校的平均分。然而，因为采样样本*Nj*在某些学习可能很小，我们使用分层贝叶斯模型来解决这个问题，其中我们假设采样自一个平常的先验分布。

则联合分布具备如下的形式：

 （5.82）

为了简单起见，我们假设是已知的。我们在后文解释如何估计。一旦我们估计出了，我们就可以计算出的后验分布。为了实现这一点，我们对上式进行一些简化，*Nj*个测量值*xij*和方差为的高斯分布等价于一个测量值为方差为的一个高斯分布。我们得到：

 （5.83）

根据上式以及4.4.1节得到的结论，其后验分布为：

 （5.84）

 （5.85）

参数控制着向全局期望收缩的程度。如果数据依赖于组别j（比如：因为采样样本Nj很大），然后相较于就会很小，所以就会较小，那么我们在估计时就会将更多的权重赋予。然而，那些样本数量比较小的组将会更多地被正则化（向全局期望收缩）。我们会在下面看到例子。

如果对于所有的组别j，我们有，则后验期望将变成

 （5.86）

这与6.3.3.2节介绍的James Stein 估计器具备一样的形式。

5.6.2.1 例子：预测棒球成绩

我们现在给出一个适用于棒球击球平均数的收缩的例子。我们观察到在前*T*=45场比赛中*D*=18个选手的击中数目。我们称击中数目为*bj*。我们假设，其中为球员*j*“真正的”击球成功率。我们的目标是估计。其MLE显然等于，其中表示经验击打平均值。然而，采用EB方法可以使我们做的更好。

为了使用上文介绍的高斯收缩方法，我们要求似然值服从高斯分布，其中已知。（我们将中的下标*j*省略，是因为我们假设*Nj*=1，*xj*已经代表球员*j*的平均值）。然而，在这个例子中，我们的似然值服从二项式分布。尽管*xj*的期望满足，但其方差却不是常数：

 （5.87）

现在让我们针对使用**方差稳定性变换**（variance stabilizing transform），从而更好地与高斯分布假设接近：

 （5.88）

现在我们可以近似地有。基于式5.86我们使用高斯收缩来估计，其中方差，然后再使用逆变换：

 （5.89）

结果如图5.12（a-b）所示。在图（a）中，我们绘制了MLE和后验期望。我们发现所有的估计值都向全局期望0.265收缩。在图（b）中，我们绘制了真实值，MLE和后验期望。（其中真实值是从大量的独立实验中估计出来的。）我们发现，整体而言，收缩估计较MLE更加接近真实的参数值。特别地，当我们使用收缩估计值时，其均方误差比MLE时的均方误差小三倍多。

5.6.2.2 估计超参数

本章，我们将给出一个估计超参数的例子。假设初始值对所有的分组都一样。在这种情况下，我们可以得到关于EB的一个封闭解。根据式4.126，我们有：

 （5.90）

所以其边缘似然为：

 （5.91）

所以，我们使用常规的MLE去估计超参数。对于，我们有：

 （5.92）

上式为全局期望。

对于方差，我们可以使用矩匹配（与高斯分布的MLE等价）的方式：模型的方差与经验方差相等：

 （5.93）

所以。因为我们知道肯定是正数，所以通常使用下面的修正后的估计值：

 （5.94）

所以收缩因子为：

 （5.95）

当不一样时，我们将不再可能导出一个闭合解。

5.7 贝叶斯决策论

我们已经了解了概率论是如何表达和更新我们关于这个世界的信念程度的。然而，我们最终的目的是将这种信念程度转化为具体的行为。本节我们将讨论进行决策的最优的方式。

我们可以将任意统计决策问题形式化为一个与自然界做的游戏。在这个游戏中，自然界挑选了一个状态、参数或者标签，但这个状态对我们是隐藏的。然后，它又产生一个观测值，这个值我们是可以看到的。然后我们需要作出一个决策，也就是说我们需要选从**行为空间**（action space）中选择一个行为*a*。最后我们需要降低某一类**损失**（loss），*L*(*y*,*a*)，该损失衡量了我们的行为*a*与自然的隐藏状态*y*之间的相容性。比如说，我们可能使用误分类损失或者平方损失。我们将在下面看到一些例子。

我们的目标是设计出一套决策程序，它为每一个可能的输入指定了一个最优的行为。通过最优化，我们得到的行为应该最小化期望损失：

 （5.96）

在经济学领域，更常用的术语是**效能函数**（utility function）；这仅仅是损失值的负数，*U*(*y*, *a*) = -*L*(*y*, *a*)。所以上述结果变成：



上式被称为**最大期望效能原则**（maximum expected utility principle），也就是我们所谓的**理性行为**（rational behavior）的本质。

值得注意的是，关于上式中的“期望”存在两种不同的解释。在我们接下来讨论的贝叶斯学派中，“期望”是指在给定已知的数据的情况下*y*的期望值。在频率学派中，“期望”是指我们希望在未来看到的关于*y*和**x**的期望值。

在贝叶斯学派下的决策论中，已知观测值x，其最优的行为*a*定义为使下列**后验期望损失**（posterior expected loss）最小的值：

 （5.98）

（如果*y*是连续值，比如说，我们希望估计一个参数向量，我们应该将求和转换为积分。）因此，**贝叶斯估计量**（Bayes estimator），又被称为**贝叶斯决策准则**（Bayes decision rule）由下式给定：

 （5.99）

5.7.1 常规损失函数下的贝叶斯估计量

本节，我们将讨论那些在机器学习中常见的损失函数下，如何构建贝叶斯估计量。

5.7.1.1 最大后验估计使0-1损失最小

0-1损失定义为：

 （5.100）

上式损失函数通常在分类问题中被使用，其中*y*为真实的类别标签，为估计值。

举例来说，在二分类问题中，我们可以按照如下方式书写损失矩阵：

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  | 0 | 1 |
|  | 1 | 0 |

（在5.7.2节，我们将介绍这种损失函数的泛化版本，从而实现对非对角线的两种损失进行不同的惩罚。）

后验期望损失定义为：

 （5.101）

所以为了最小化期望损失，所选择的决策为后验分布的众数或者MAP估计

 （5.102）

5.7.1.2 拒绝行为

在有些分类问题中，决策可能非常不确定，在这种情况下，我们选择**拒绝行动**（reject action），在这种情况下，我们拒绝将样本归为任意一类，取而代之的是说“我不知道”。这种模糊的情况可以通过类似于人类专家去解决。图5.13给出了说明。这在像医疗和金融等需要**风险规避**（risk averse）的领域十分有用。

我们可以按照如下的方式形式化拒绝行为的定义。令*a* = *C*+1代表着选择了“拒绝行为”，选择对应着选择其中某个类别。假设我们定义损失函数为：

 （5.103）

其中为拒绝选择所造成的损失，表示因替换错误而导致的损失。结果表明，最优的行为是：如果最有可能的类别的概率小于，你应该拒绝选择；否则选择最有可能的那个类别。

5.7.1.3 后验期望最小化*l*2（平方）损失

对于连续空间的参数，一个更加合适的损失函数为**平方差**（squared error）,*l*2损失，或者叫**二次损失**（quadratic loss），定义为：

 （5.104）

后验期望损失定义为：

 （5.105）

所以最佳的决策为后验期望：

 （5.106）

上式通常被称为**最小均方误差**（minimum mean squared error）估计或者MMSE估计。

在线性回归模型中，我们有

 （5.107）

在这种情况下，在给定一些训练集的情况下，最佳的估计值为：

 （5.108）

也就是说，我们仅仅基于参数的后验分布的期望值进行点估计。需要注意的是，上述求解最优行为的过程与选择什么样的**w**的先验分布无关。

5.7.1.4 后验中值最小化*l*1（绝对值）损失

*l*2损失函数惩罚的是估计值与真实值偏差的平方，所以其对异常点比较敏感。一种更加健壮的方案是使用绝对值或者*l*1损失，（见图5.14）。其最优的估计值为后验中值，也就是说决策值*a*应该满足。

5.7.1.5 监督学习

考虑一个预测函数，假设我们有一个损失函数，它表示当真实值是*y*而预测值是时的损失。假设未知的真实状态为（即原本数据生成的过程），我们预测的行为是（比如预测函数），我们定义损失函数为：

 （5.109）

上式被称为**泛化误差**（generalization error）。我们的目标是最小化后验期望损失，由下式给定：

 （5.110）

这与在6.47节介绍的频率学派中的风险是一致的。

5.7.2 假正例与假负例之间的权衡

本节，我们将集中讨论二值决策问题，比如说假设检验，二分类，目标/事件监测等等。在这类问题中，存在两种我们可能犯的错误：假正例（false positive）又被称为误报警（false alarm），这类错误会在真实值是*y*=0，而预测值是时发生的；或者是假负例（false negative）（漏检 ，missed detection），这类错误在真实值是*y*=1，而预测值是时发生的。0-1损失对待这两种错误是平等的。然而，我们可以考虑一个更加一般化的损失矩阵：

|  |  |  |
| --- | --- | --- |
|  |  |  |
| *y*=1 | 0 | *L*FN |
| *y*=0 | *L*FP | 0 |

表

其中*L*FN为一个假负例所造成的成本，*L*FP为一个假正例所造成的成本。则两种行为所造成的后验损失期望为：

 （5.111）

 （5.112）

所以我们会选择，当且仅当：

 （5.113）

 （5.114）

如果*L*FN=*cL*FP，我们很容易地得到，当且仅当，其中，我们选择。举例来说，如果一个假负例所带来的成本是一个假正例所带来的成本，即*c*=2，那么我们将使用阈值2/3来决定是否选择正例。

接下来，我们将讨论ROC曲线，它提供了一种研究FP-FN权衡的方式，在这种方式中我们不需要指定决策阈值。

5.7.2.1 ROC曲线及相关内容

假设我们正在解决一个二值决策问题，比如分类，假设检验，目标检测等等。同时，我们假设存在一个含标签的数据集。令为我们的决策规则，其中*f*(**x**)为我们对于*y*=1这个结论的信念程度（该值应该与*p*(*y*=1|**x**)呈单调关系，但不需要是一个概率值），表示某个阈值参数。对于每一个给定的阈值，我们可以应用决策规则计算出真正例，假正例，真负例和假负例的个数，如表5.2所示。这个表征错误的表格被称为**混淆矩阵**（confusion matrix）。

从这个表格中，我们可以计算出**真正例率**（true positive rate,TPR），又被称为**敏感度**（sensitivity），**召回率**（recall rate）或者**命中率**（hit rate），计算方式为。我们同样可以计算出**假正例率**（false positive rate，FPR），又被称为**误报警率**（false alarm rate），或者**一类误差率**（type I error rate），计算方式为。这些以及其他类型的定义见表5.3和5.4。在计算我们的损失函数时，我们可以以任意的方式去组合这些错误。

然而，相较于基于一个固定的阈值去计算TPR和FPR，我们可以基于不同的阈值去计算这些值，然后绘制出TPR和FPR的关系曲线。这被称为受试者工作特征曲线（receiver operating characteristic，ROC）。图5.15（a）展示了一个例子。通过设置阈值=1，任何系统都可以到达左下角的点（FPR=0,TPR=0），也就是说我们将所有的样本都分类为负例；类似地，如果我们设置=0，那么所有的样本将会被判定为正例，此时我们将到达曲线的右上角（FPR=1,TPR=1）。如果一个系统的性能与一个随机系统相当，那么通过选择一个合适的阈值，我们将得到对角线TPR=FPR上的任意一个点。一个能够完美地将正例与负例分开的系统是存在一个阈值使系统能够到达坐标系的左上角（FPR=0,TPR=1）；通过改变阈值，这样的系统将会靠近左坐标轴，其次是顶部坐标轴，如图5.15（a）所示。

ROC曲线的质量通常使用一个单独的数值去衡量，即曲线下方面积（area under curve）或者AUC，AUC的取值越高，曲线的质量越好；最大值为1。另一种衡量ROC曲线质量的统计量为同等误差率（equal error rate，EER），又被称为交换率（cross over rate），定义为满足FPR=FNR的值。因为FNR=1-TPR，我们可以在坐标轴左上角和右下角之间绘制一条直线，它与ROC曲线的交点即为EER（见图5.15（a）中的点A和B）。越小的EER取值意味着系统性能越好，显然其最小值为0。

5.7.2.2 精度召回率曲线

当我们尝试去探测一个小概率事件时（比如检索一个相关文本或者在一张图片中找到一张脸），在这些问题中，负例的数量将会是巨大的。所以比较与所提供的信息十分有限，因为FPR的取值将会很小。所以，在ROC曲线中，所有的“决策”将主要集中在左边。在这种情况下，通常我们绘制TPR与假正例数目的关系曲线，而不是去假正例率的关系曲线。

然而，在某些情况下，我们关于“负例”的选择是不确定的。举例来说，当我们在一张图片中检测目标时，如果检测器工作的方式是对图中的斑点进行分类，那么需要检测的斑点的数目——及真负例的数量是算法的一个参数，而不是问题定义的部分。所以我们希望使用一个测度，该测度只讨论正例。

精度定义为，召回率定义为。精度衡量了在我们判断为正例的样本中有多少是真的正例，召回率衡量了在所有正例中，我们实际检测出了多少。如果为预测的标签值，为真实的标签值，我们可以通过下式去估计精度和召回率。

 （5.115）

精度召回率曲线的绘制方法是在不同的阈值下，绘制精度与召回率的关系。图5.15（b）给出了曲线的例子。靠近右上角的曲线性能更好的系统。

评价该曲线性能的方式是使用平均精度（基于召回率的加权平均），它近似地曲线下的面积。一种可选的方案是，我们可以标记在固定召回率的情况下的精度，比如在前K=10个项目被召回时的精度。这被称为在K分值时的平均精度（average precision at K）。这种测量方式在信息检索系统中被广泛使用。

5.7.2.3 F-scores

在给定阈值的情况下，我们可以计算一个精度和召回率。这两个值通常被组合成一个统计值F score或者F1 score，它是精度和召回率的调和平均值：

 （5.116）

使用5.115，我们可以将上式写成：

 （5.117）

这在信息检索系统中常被使用。

为了理解为什么我们使用调和平均和不是算术平均值（P+R）/2，考虑如下的场景。假设我们召回了所有样本（译者注：将所有样本都预测为正例），所以R=1。此时精度由盛行率（prevalence）p(y=1)给定。假设盛行率十分低，不妨设置为p(y=1)=10-4。精度与召回率的算术平均为（P+R）/2=（10-4+1）/2≈50%。相反，调和平均值仅仅是。

在多分类问题中（比如，文本分类问题），有两种方式形成F1分值的泛化版。第一个被称为宏平均F1（macro-averaged F1），定义为，其中为将类别c区别于其他所有类别时得到的F1。另一个被称为微平均F1（micro-averaged F1），其中F1定义为将不同类别的统计计数进行混合后的计算结果。

表5.5给出了上述两个定义的区别。我们发现类别1的精度为0.5，类别2的精度为0.9.其宏平均精度为0.7，然而其微平均精度为0.83.后者更加接近类别2的精度，因为类别2是类别1的5倍。为了赋予每个类别相同的权重，使用宏平均。

5.7.2.4 误发现率\*

假设我们现在正在使用一个高流通率的设备，比如一个基因表达微阵列或者一个射电望远镜去探索一个罕见现象。我们将会作出很多二值决策，其中，N可能是非常大的数。这被称为**多重假设检验**（multiple hypothesis testing）。注意此处与标准的二值分类问题的区别，前者在对*y*i进行分类时基于的是整个数据集，而不仅仅是**x**i。所以这是一个联立的分类问题，我们希望比一系列单独分类问题所做出的结果好。

我们又该如何设置阈值呢？一种很自然的方法是尝试最小化假正例的期望数。在贝叶斯方法中，该值可以通过下式进行计算：

 （5.118）

其中为你关于这个目标展示了相关问题现象的信念程度。我们接着定义**误发现率**（false discovery rate）的后验期望值：

 （5.119）

其中为发现的项目的数量。给定一个期望的FDR容忍度，我们可以通过调整阈值来实现这个目标；这被称为直接后验概率方法（direct probability approach），用于控制FDR。

为了控制FDR，同时估计所有pi的值，而不是单独地估计，是十分有帮助的（比如使用一个分层贝叶斯模型）。这有利于混合统计强度，所以能够降低FDR。

5.7.3 其他的主题\*

待定