MLAPP-C6

6.频率学派的统计学

6.1 简介

我们在第5章讨论的统计推断方法被称为贝叶斯统计。或许让人惊讶的是，这种方法被有些人认为是存在争议的，尽管贝叶斯理论在非统计问题——比如医疗诊断（见2.2.3.1节），垃圾邮件过滤（见3.4.4.1节），或者飞机追踪（见18.2.1节）的应用没有争议。反对的理由在于贝叶斯学派中未区分统计模型中的参数与其他类型的未知的变量。

在统计推断中避免将参数作为未知变量的尝试已经进行，这些方法避免了使用先验分布和贝叶斯法则。这种方法被称为**频率学派统计**（frequentist statistics），**古典统计**（classical statistics）或者**正统统计**（orthodox statistics）。该方法不依赖于后验分布，取而代之的是，它基于采样分布的概念。这个分布是关于估计量的分布，每一个估计量是从一个数据集中得到的，不同的数据集都是从真实但未知的分布中采样产生的，6.2节将介绍更多细节。正是这种重复试验，导致了频率学派中对参数估计值的不确定性。

相反，在贝叶斯方法中，我们只需要将已经看到的样本数据作为条件，并不存在重复试验的概念。从而允许我们计算一次性事件的概率（**译者注**：即只可能发生一次而无法进行重复试验的事件的概率），这一点我们在2.1节已经进行过讨论。或许更加重要的是，贝叶斯方法避免了某些悖论，这些悖论折磨着频率学派的方法（见6.6节）。然而，熟悉频率学派的方法是十分重要的（尤其是6.5节的内容），因为它在机器学习中被广泛使用。

6.2 估计量的采样分布

在频率学派统计中，一个参数的估计值是通过将一个估计函数（estimator）应用到某些数据，所以。该参数被认为是固定的，数据是随机的。这与贝叶斯方法中完全是相反的（**译者注**：在贝叶斯方法中数据是观察到的确定的值，而参数是随机的）。参数估计值的不确定度可以通过计算估计量的**采样分布**（sampling distribution）来衡量。为了理解这一点，想象从一个真实的模型中采样得到许多不同的数据集，举例来说，令，其中，表示真实的参数。此处*s*=1:*S*表示采样数据集的索引号，*N*表示每个数据集的样本数量。现在将估计函数应用在每个数据集上，从而得到一系列估计值。当时，由引起的分布就是估计量的抽样分布。我们会在后面的章节中讨论使用采样分布的不同方法。但首先介绍两种计算采样分布本身的方法。

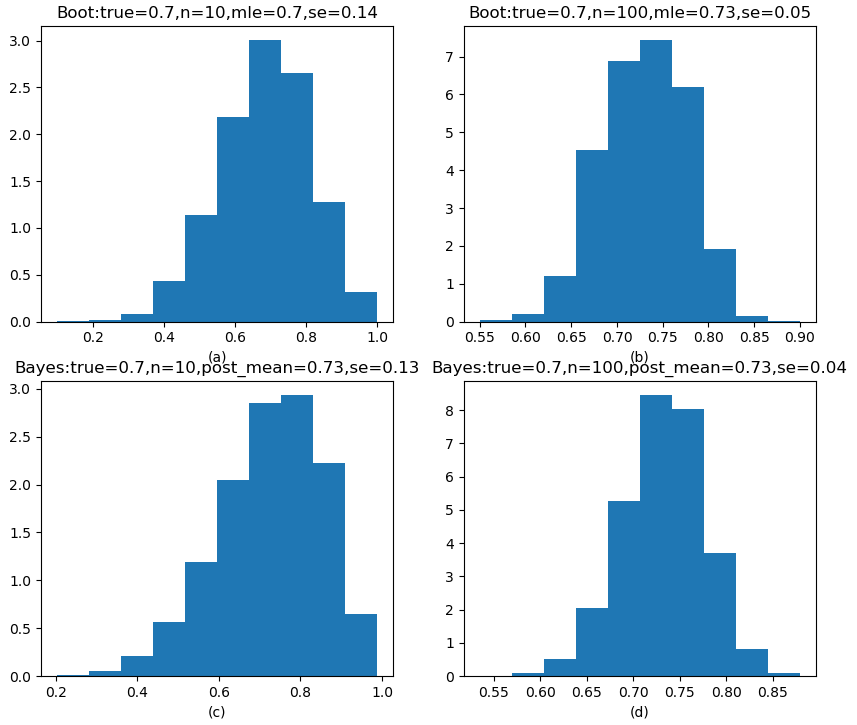


图6.1 自举法（第一行）与贝叶斯方法（第二行）的对比。N个样本产生自分布。左侧：N=10。右侧：N=100。（a-b）采用自举法对伯努利分布中参数MLE的采样分布进行近似。我们展示了B=10000次的自举结果的直方图。（c-d）从后验分布中采样得到的10000个样本的直方图。其中先验分布采样的均匀分布。图形由程序**bootstrapDemoBer**生成。

6.2.1 自举法

自举法（**bootstrap**）是近似采样分布的一种简单的蒙特卡洛技术。当估计量是真实参数的复杂函数的时候，这种方法特别有用。

自举法的思想很简单，如果我们知道真实的参数，我们可以从真实的分布中产生许多（*S*个）伪造数据集，每个数据集的大小为*N*。接着，我们可以从每个采样数据集中计算出估计量，然后使用经验分布来近似采样分布。因为是未知的，所以**含参自举**（parametric bootstrap）的思想是使用来产生样本。另一种备选方案被称为**不含参自举**（non-parametric bootstrap），它从原来的数据集中（有放回）对样本进行采样，然后像之前介绍的那样计算估计值。

图6.1（a-b）展示了一个例子，在这个例子中，我们使用含参自举法计算了伯努利分布的MLE的采样分布。（使用无参自举法得到的结果是一样的。）我们发现当*N*=10时，采样分布是非对称的，所以与高斯分布相差甚远；当*N*=100时，采样分布与高斯分布更像，这一点有理论支撑（见下文）。

一个自然的问题在于：我们通过自举法计算出来的参数估计值与从后验分布采样得到的参数估计值有什么联系呢？从概念上来说，它们是非常不同的。但当先验分布不是很强时，它们可能会十分相似。举例来说，图6.1（c-d）展示了一个例子，其中我们基于先验分布Beta（1,1）计算了参数的后验分布，然后基于后验分布进行采样。我们发现后验分布与采样分布十分相似。所以我们可以将采样分布当作一个“穷人的”后验分布（**译者注**：即拥有的先验经验很少）。

然而，令人惊讶的是，自举法可能比后验分布采样慢。原因在于自举法需要训练模型*S*次，然而在后验采样中，我们通常只需要训练模型一次（为了找到局部众数），然后在这个众数周边展开局部探索。这种局部探索通常比从头开始训练一个模型更快（**译者注**：其实就是在线训练的方式）。

6.2.2 MLE的大数定理\*

在某些情况下，一些估计量的采样分布是可以通过解析的方式计算出来的。特别地，结果表明，在某些条件下，当样本的数量趋于无穷大时，MLE的采样分布将趋向于正态分布。非正式地，上述结果能够成立的要求是模型中的每一个参数都能够“看到”足够多的数据，且模型是可被训练的。不幸的是，这个要求对于很多机器学习中的模型都无法满足。然而，让我们假设存在一个简单的设置使理论能够成立。

高斯分布的中心将会是MLE。但高斯分布的方差是多少呢？直觉上，估计量的方差与似然函数在峰值的曲率程度直接相关。（**译者注**：从概念上理解，似然函数衡量的是在某个参数下，观测数据的概率分布，但如果似然函数的定义域比较窄，那么从这个分布中采样得到数据也就会相对集中，从而，根据这些采样样本得到的参数估计值也会比较集中，即方差比较小。）如果曲率半径很大，则峰值会很“尖”，此时的方差会很小；在这种情况下，估计量可以被更容易地确定。相反，如果曲率半径比较小，则峰值部分的曲线会变得比较“平滑”，方差会比较大。

让我们对上述的观点进行形式化的表达。定义**评价函数**（score function）为对数似然函数在某个点的对数似然值的梯度：

 （6.1）

定义**观测信息矩阵**（observed information matrix）为负评价函数的梯度，或者等价地，NLL（negative log likelihood）的海森矩阵：

 （6.2）

在1维情况下，上式变成：

 （6.3）

上式仅仅只是负对数似然在点处曲率的测度。

因为我们正在研究采样分布，为一系列随机变量。**费舍尔信息矩阵**（Fisher information matrix）定义为观测信息矩阵的期望值：

 （6.4）

其中表示当函数**f**作用于采样自真实分布（由参数决定）数据时的期望值。通常情况下，我们使用表示产生数据的“真实的参数”，且假设已经知道，所以出于方便，我们仅仅书写。进一步地，我们很容易地发现，因为样本数量为*N*的对数似然只是样本数量为1时的对数似然的*N*倍。所以，我们只书写，这是被经常使用的符号。

现在令为MLE，其中。结果表明，当：

 （6.5）

我们称MLE的采样分布是**渐进正态分布**（asymptotically normal）的。

那MLE的方差又是多少呢，它可以用来衡量对MLE的信念程度？不幸的是，是不知道，所以我们不能计算出采样分布的方差。然而，我们可以将替换为以实现对采样分布的近似。因此，的标准差（standard errors）近似值为：

 （6.6）

举例来说，根据式5.60，我们知道对于伯努利采样模型的费舍尔信息为：

 （6.7）

所以，MLE的近似标准差为：

 （6.8）

其中。对比式3.27，它是在均匀先验分布下的后验标准差。

6.3 频率学派的决策论

在频率学派的决策论中，存在损失函数和似然函数，但是没有先验和后验分布，那么也就没有后验期望损失。所以没有自动的方式去推导一个最佳的估计量，这一点与贝叶斯学派不一样。取而代之的是，在频率学派的方法中，我们可以按照我们的想法自由地选择估计函数或者决策程序：。

在选择一个估计函数之后，我们定义它的期望损失或者**风险**（risk）为：

 （6.9）

其中为从“自然的分布”中采样得到的数据，这个自然的分布由参数决定。换句话说，期望是关于估计量的采样分布（**译者注**：等价于每一次采样得到的数据集）进行计算的。将上式与贝叶斯后验期望损失比较：

 （6.10）

我们发现贝叶斯方法是基于参数（未知）的平均，且以数据（已知）为条件，然而，频率学派的方法基于数据求平均（所以忽略了已经观测到的数据），以（未知）为条件。

频率学派的定义不仅不自然，而且也无法计算，因为是未知的。因此，我们并不能根据频率学派的风险去比较不同的估计量。我们在下面讨论不同的方法。

6.3.1 贝叶斯风险

那么我们该如何选择估计量呢？我们需要某种方法将转换为一个单一的量，这个量的计算并不依赖于。一种方式是为参数添加一个先验分布，然后定义一个估计量的**贝叶斯风险**（Bayes risk）或者**积分风险**（integrated risk）为：

 （6.11）

一个**贝叶斯估计量**（Bayes estimator）或者**贝叶斯决策准则**（Bayes decision rule）使期望风险最小：

 （6.12）

注意积分风险又被称为**预后验风险**（preposterior risk），因为它是在我们看到数据之前计算的。最小化上式在试验设计阶段十分有用。

我们将证明一个决策论中将贝叶斯方法与频率学派方法连接起来的重要的理论。

**定理 6.3.1** 一个贝叶斯估计量可以通过最小化每个x的后验期望损失得到。

证明。 通过调整积分的顺序，我们有：

 （6.13）

 （6.14）

 （6.15）

 （6.16）

为了最小化全局期望，我们只需要针对每一个**x**进行最小化，所以我们的决策规则是：

 （6.17）

因此，我们看到根据具体情况（如贝叶斯方法）选择最优行动是平均最优的（如在频率论方法中）。 换句话说，贝叶斯方法提供了实现频率主义目标的好方法。 事实上，人们可以进一步证明以下几点。

**定理 6.3.2** 每个可接受的决策规则都是关于某些（可能是不正确的）先验分布的贝叶斯决策规则。

该理论表明，最小化频率学派风险的最佳方式是符合贝叶斯学派观点的。

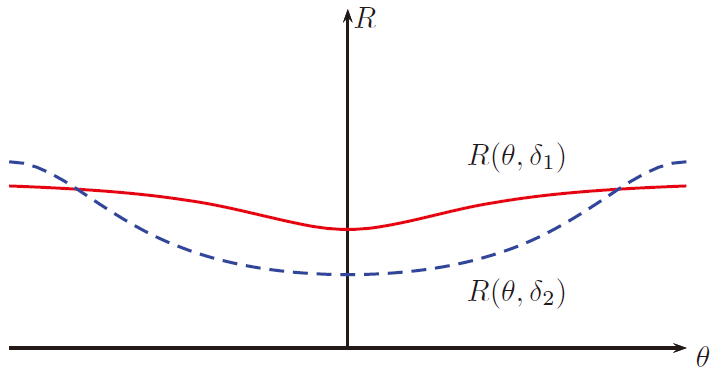


图6.2 两种决策规则和的风险函数，因为具有较低的最大风险，所以他是最小最大估计函数，尽管对于大部分，都具有较低的风险。所以最小最大估计函数过分保守。（**图形来自于原书**）

6.3.2 最小最大风险

显然，一些频率学派的学者不喜欢使用贝叶斯风险，因为它需要对先验分布作出选择（尽管它只是用来评价估计量，而不需要作为构造的一部分）。一种可选的方式如下。定义一个估计量的最大化风险（maximum risk）为：

 （6.18）

最小最大规则是指估计量应该使上式最小：

 （6.19）

举例来说，在图6.2中，我们发现，在整个的的变化范围内，具有较低的最大风险，所以它是最小最大估计量。

最小最大估计量有一定的吸引力。然而，计算它却十分困难。进一步讲，而且，它们非常悲观。事实上，结果表明，所有的最小最大估计量都等价于基于最不利先验分布（least favorable prior）的贝叶斯估计量。在大多数统计情况下（不包括博弈论），假设自然是对手，这不是一个合理的假设。

6.3.3 可接受的估计量

频率学派的决策论最基本的问题是为了计算风险，我们需要依赖于真实的分布。然而，存在一些估计量本身就比其他的差，无论的取值多少。特别地，如果对于所有的成立，我们称“控制”（dominates）了。如果对于，不等式是严格成立的，那么我们称这种控制是严格的。一个估计量被称为是**可接受的**（admissible），如果它并非严格地被其他估计量控制。

6.3.3.1 例子

下面我们介绍一个例子。考虑估计高斯分布的期望的问题。我们假设数据是从分布采样得到的，使用平方损失。相应的风险函数为MSE（均方差）。一些可能的决策规则或者估计量如下：

* ，样本均值
* ，样本中值

* ，固定值
* ，基于先验分布的后验期望：

 （6.20）

对于，我们考虑一个弱先验，*k*=1和一个强先验，*k*=5。先验分布的期望为某个固定值。我们假设已知。（所以当*k*趋向于无穷大时，与一样。）

下面我们使用解析的方式导出风险函数。（我们可以在这个示例中计算风险函数，因为我们知道真实的。）在6.4.4节，我们将展示MSE可以分解为偏差的平方与方差的和：

 （6.21）

采样均值是无偏的，所以风险为：

 （6.22）

采样的中值同样是无偏的，其方差近似等于，所以

 （6.23）

对于固定值，其方差为0，所以

 （6.24）

最后，对于后验期望，我们有

 （6.25）

 （6.26）

 （6.27）

 （6.28）

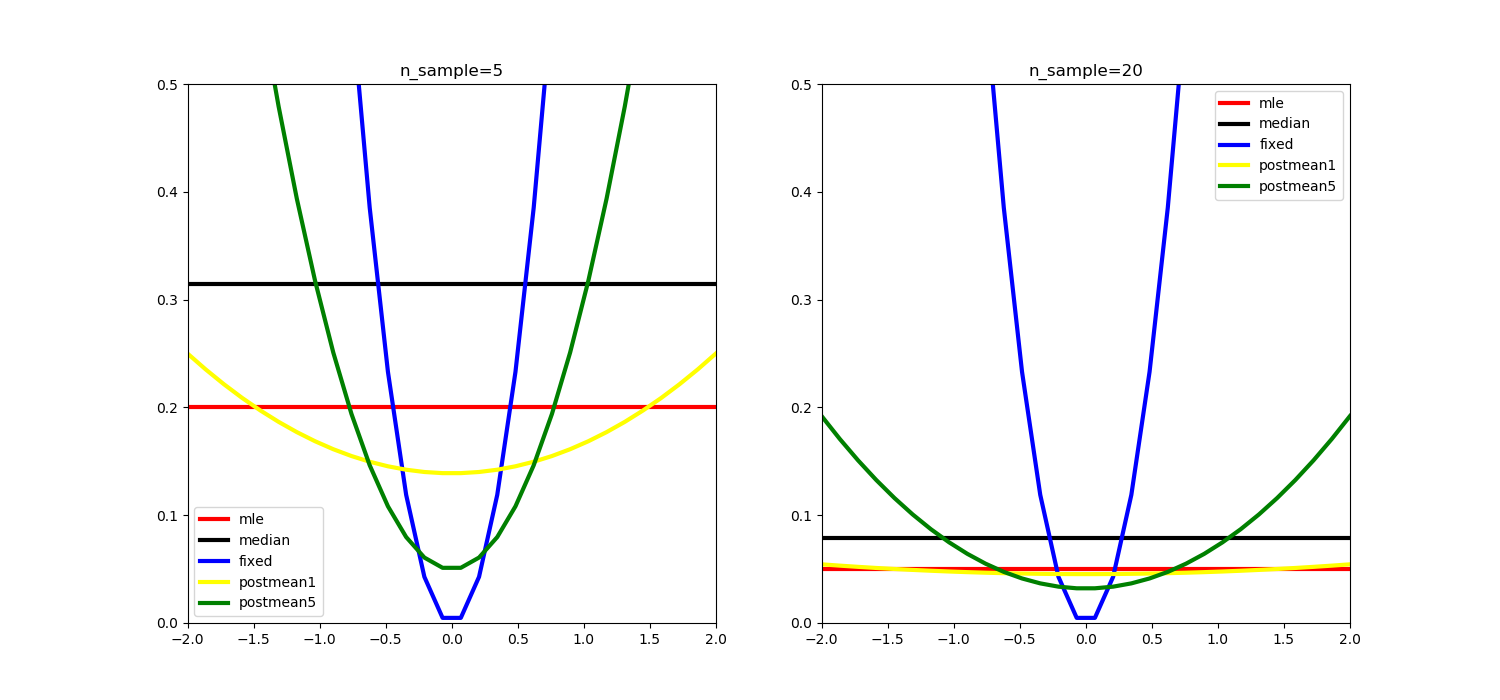


图6.3 高斯分布期望的估计值的风险函数。其中红色线为MLE。左边对应样本数量N=5。右边对应样本数量N=20。图形由程序**riskFnGauss**生成。

这些函数在图6.3中被绘制出。我们发现在一般情况下，最佳的估计量取决于未知的真实参数。如果非常靠近，则（只预测）是最好的估计量。如果在周边一个合理范围内变化，则后验期望是最好的，后验期望组合了先验的猜测和实际的数据。如果离很远，则MLE是最好的。这一切都不应该令人惊讶：假设我们先验的均值是合理的，通常需要少量的收缩（使用先验分布较弱的后验期望）。

令人惊讶的是，决策规则（采样中值）的风险对于任意的都比（采样均值）高。因此，在当前的问题下（数据是从高斯分布中采样得到的），采样中值是一个不被接受的估计量。

实际上，采样中值通常比采样均值好，因为前者对异常点更具备鲁棒性。事实上，如果数据是从拉普拉斯分布（相较于高斯分布尾部更肥）采样得到的，那么当我们使用平方差损失时，中值是贝叶斯估计量。更一般地，我们可以通过使用我们数据的灵活模型（例如混合模型或非参数密度估计器（第14.7.2节））构建稳健估计器，然后计算后验均值或中位数。

6.3.3.2 斯坦悖论

假设我们有*N*个独立同分布的变量，现在我们想估计。最明显的估计量为MLE，在这种情况下，令。结果表明当*N* ≥ 4，使用平方差损失时，这是一个不被接受的估计量。

为了展示这一点，我们完全可以构造一个更好的估计量。James-Stein估计量就是这样的一个估计量，定义为：

 （6.29）

其中，0<*B*<1为一个微调的常数。这个估计量将向全局均值进行收缩。（我们在5.6.2节使用经验贝叶斯方法推导过这个估计量。）

结果表明，当*N* ≥ 4时，相较于MLE，收缩的估计量有较低的频率学派风险（MSE）。这被称为**斯坦悖论**（Stein’s paradox）。它被称为悖论的理由由下面的例子说明。假设为学生*i*的“真实的”IQ，*X*i为测试的结果。那么为什么我对的估计要依赖于全局的均值呢，后者与其他学生的关系相关？我们可以通过使不同的维度在性质上有所不同来创造更多自相矛盾的例子，比如说为我的IQ，为温哥华的平均降雨量等等。

解决这个悖论的方式如下。如果你的目标只是估计，除了选择，你并不能作出更好的选择，但是如果你是估计整个向量，并且使用平方差作为你的损失函数，那么收缩的估计是有益的。为了说明这一点，假设我们希望从一个样本中估计。一种简单的估计是，但这样会高估结果，因为：

 （6.30）

因此，我们可以通过共享信息来实现风险的降低，哪怕是基于不相关的来源，让结果向全局的期望收缩。在章节5.6.2中，我们针对这一点给出了贝叶斯学派的解释。

6.3.3.3 被接受是不足够的

貌似，我们可以将对好的估计量的研究限制在可接受估计量的类别中。但是事实上，我们很容易地构造一个可接受估计量，我们在下面的例子中给出介绍。

**定理6.3.3** 令，考虑一个在平方损失下的估计量。令，它是一个与数据无关的常数。这是一个可接受估计量。

**证明**。假设它不是一个可接受估计量。那么存在一个风险更小的估计量，所以，其中不等式必须对任意成立。假设真实的参数为，那么，并且：

 （6.31）

因为对所有成立，同时，所以我们有，因此。所以在处，的风险避免比高的唯一方式是两者相等。所以不存在其他的估计量有更低的风险，所以是可被接受的。

6.4 估计量的理想性质

既然频率学派的决策论并没有提供选择最好的估计量的自动方式。我们需要提出其他启发式的原则。本节，我们讨论一些我们希望估计量应该有的性质。不幸的是，我们发现并不能同时满足这些性质。

6.4.1 相容性估计量

当样本的数量趋向于无穷多时，如果一个估计量收敛于真实的参数，则称这个估计量是**一致的**（consistent），即：当，（箭头是指依概率收敛）。当然，这个概念只有当数据是来自由参数指定的分布时才是有意义的，而实际上的数据通常并不满足。然而，这个理论性质却十分有用。

结果表明，MLE是一个一致性估计量。直觉上的原因是最大化似然函数等价于最小化，其中为真实的分布，为我们的估计。当且仅当，=0。

6.4.2 无偏估计

一个估计量的**偏差**（bias）定义为：

 （6.32）

其中为真实的参数值。如果偏差为0，则称估计量为**无偏的**（unbiased）。这就意味着，采样分布以真实的参数为中心。举例来说，高斯分布期望的MLE是无偏的。

 （6.33）

然而，高斯分布方差的MLE是有偏的。事实上，结果表明：

 （6.34）

然而，下面的估计量是无偏的：

 （6.35）

我们可以很容易地证明这一点：

 （6.36）

当数据量特别大时，上述两个估计量的区别可以忽略。

尽管有些时候MLE是有偏的估计量，但结果表明，MLE总是渐进式地无偏。（这对于MLE是一个一致性估计量是必要的。）

尽管无偏是一个令人满意的性质，但这一点并不总是正确的。我们会在6.4.4节讨论这一点。

6.4.3 方差最小估计量

直觉上，我们对估计量是无偏的要求是合理的（尽管接下来我们将给出一些反对这一点的讨论。）然而，仅仅无偏是不够的。举例来说，假设我们希望根据数据集去估计一个高斯分布的期望。将第一个点*x*1作为估计量，即，显然该估计量是无偏的，但相较于（同样是无偏的），前者距离更远。所以估计量的方差也是很重要的。

一个自然的问题是：估计量的方差最小能达到多少？一个著名的结论，称为Cramer-Rao lower bound，提供了任何无偏估计量方差的下限。更加精确地表达：

**定理6.4.1** （Cramer-Rao inequality）。令*X*1,…,*X*n~,为的无偏估计。在关于的不同光滑假设下，我们有

 （6.37）

其中为费舍尔信息矩阵（见6.2.2）.

结果表明，MLE能够到达Cramer Rao下确界，所以在所有的无偏估计中拥有最小的渐进方差。所以，MLE又被称为**渐进式最优**（asymptotically optimal）。

6.4.4 偏差-方差权衡

尽管使用一个无偏估计量是一个比较好的主意，但这并不总是这样的。为了说明为什么，假设我们使用平方差损失。正如我们在前面所讨论的，对应的风险函数为MSE。我们将导出关于MSE的一个有用的分解式。（所有的期望和方差都是关于真实分布进行计算的，但为了符号上的简便，我们省去了明确的条件的书写。）令表示估计量，表示估计量的期望值（当采样数据变化时）。然后，我们有：

 （6.38）

 （6.39）

 （6.40）

 （6.41）

总结如下：

MSE = variance+bias2 （6.42）

上式被称为**偏差-方差权衡**（bias-variance tradeoff）。它的含义在于，如果需要最小化平方差损失，那么使用一个有偏估计可能是明智的，只要它能降低方差。

6.4.4.1 例子：估计高斯分布的期望

下面让我们给出一个例子。假设我们希望根据数据去估计期望。我们假设这些数据是从分布采样得到的。一个明显的估计量为MLE。其偏差为0，方差为：

 （6.43）

但我们也可以使用一个MAP估计。在4.6.1节，我们展示了在一个高斯先验分布下的MAP估计为：

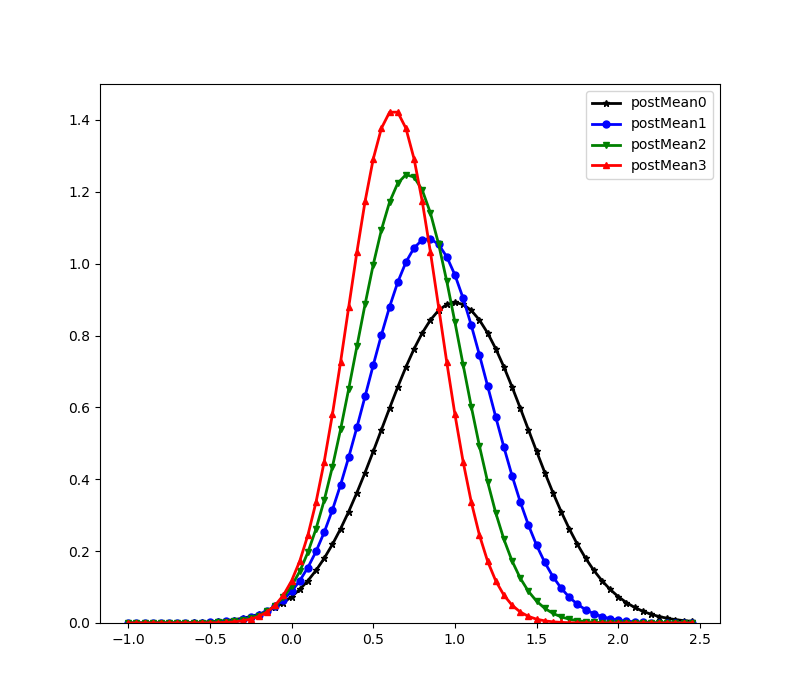
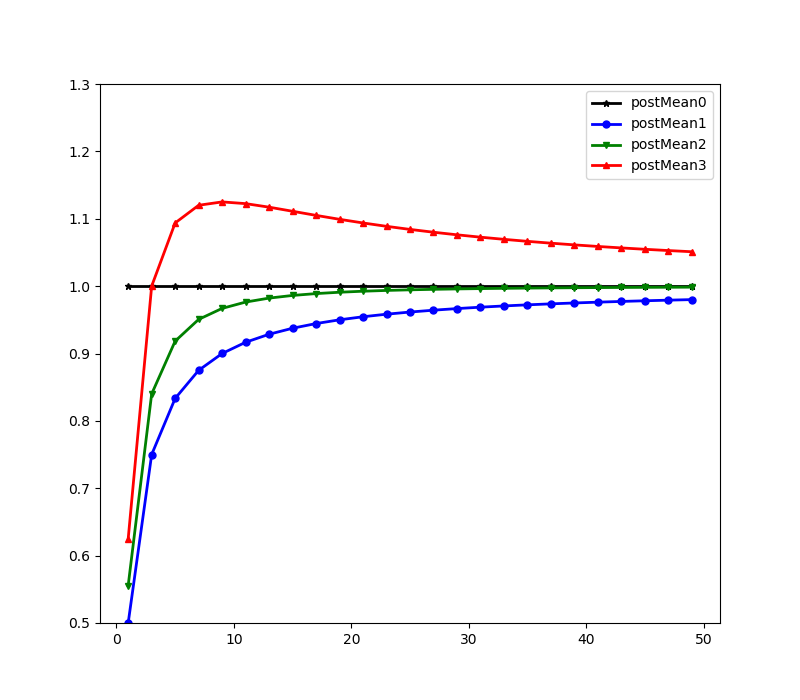
 （6.44）

其中控制着我们对MLE的信任程度。（上式又被称为后验期望，因为对于高斯分布而言，期望和众数是一样的。）MAP的偏差和方差分别为：

 （6.45）

 （6.46）

所以尽管MAP是有偏的，但其方差更小（假设*w*<1）。

（a） （b）

图6.4 左侧：含不同先验强度*k*0的MAP估计的采样分布。（*k*0=0对应MLE。）右侧：MAP与MLE的MSE比例与样本数量的关系。图形由程序**samplingDistGaussShrinkage**生成。

让我们假设先验分布有些偏差，即，然而真实值为。在图6.4（a）中，我们发现，当*k*0>0时，MAP的采样分布是有偏的，但相较于MLE，其方差也更小。

在6.4（b）中，我们绘制了与*N*的关系。我们发现对于*k*0∈{1,2}，MAP估计的MSE较MLE的要小，尤其是对于样本数量比较小的时候。*k*0=0对应MLE，*k*0=3时，对应一个强先验分布，而该先验分布的期望是错误的，所以降低了性能。调整先验分布的强度显然是很重要的，这是我们接下来讨论的主题。

6.4.4.2 例子：岭回归

另一个偏差方差权衡的例子是岭回归，我们将在7.5节进行讨论。概括地讲，在这种方法中，我们求解线性回归的MAP估计，其中基于的先验分布为。该先验分布具有的0值期望将鼓励权重变得很小，从而降低过拟合的风险；精度项控制着先验分布的强度。当时，对应MLE估计；使用将导致一个有偏估计。为了说明对方差的影响，考虑一个简单的例子。图6.5的左侧绘制了每一个单独的拟合的曲线，右侧绘制了拟合的曲线的平均值。我们发现当我们增加正则项的强度时，方差在减小，但偏差在增加。

6.4.4.3 在分类问题中的偏差-方差权衡

如果我们使用0-1损失代替平方差，上面的分析将不再适用，因为频率学派的风险将不能表达为偏差的平方与方差的和。事实上，结果表明它是偏差与方差乘积的组合。如果估计量在决策边界的正确的一侧，则偏差将会被忽略，那么降低方差将减少误分类率。但如果估计量在决策边界的错误一侧，那么偏差将会是正数，所以增加方差是有好处的。这个鲜为人知的事实说明偏差-方差权衡对于分类问题不是很有用。更好地方式是聚焦在期望损失（下文介绍），而不是直接聚焦在偏差和方差。我们可以使用交叉验证的方式近似期望损失，这一点将在6.5.3节介绍。

6.5 经验风险最小化

频率学派的决策论面临着一个最基本的问题在于它事实上无法计算风险函数，因为它依赖于真实的数据分布。（相反，贝叶斯学派中的后验期望损失可以被计算，因为它以观测到的数据而非真实的参数为条件。）然而，有一种方法可以避免这个问题，在这种方法中，我们的任务是预测可观察到的量，而非估计隐藏的变量或者参数。也就是说，我们不再关心类似于这样的损失函数，其中为真实但未知的参数，为我们的估计量。取而代之的是，我们关心形如的损失函数，其中*y*表示真实但未知的响应，表示在给定输入**x**的情况下的预测。在这种情况下，频率学派的风险变成：

 （6.47）

其中代表“自然的分布”。当然这个分布是不知道的，但是一种简单的方法是使用经验分布去近似，这个经验分布可以从一些训练数据集中得到：

 （6.48）

接着我们定义**经验风险**（empirical risk）为：

 （6.49）

在0-1损失的情况下，，上式将变成**误分类率**（misclassification rate）。在平方误差损失的情况下，，上式将变成**均方误差**（mean squared error）。我们将**经验风险最小化**（empirical risk minimization, ERM）的任务定义为寻找一个决策程序（典型的情况是一个分类规则）使得经验风险最小：

 （6.50）

在非监督问题中，我们去除所有关于y的引用，将替换为，举例来说，衡量了重构的误差。比如在向量编码（11.4.2.6节）或者PCA（12.2节）中，我们可以定义决策规则为。最后我们定义经验风险为：

 （6.51）

当然，我们总是可以通过设置最小化风险，所以编码器 - 解码器经历某种瓶颈是至关重要的。

6.5.1 正则化风险最小化

值得注意的是，如果我们关于“自然分布”的先验分布就是经验分布，那么经验风险等价于贝叶斯风险：

 （6.52）

所以，最小化经验风险通常会导致过拟合。所以通常需要在目标函数中添加一个复杂度惩罚项：

 （6.53）

其中衡量了预测函数的复杂度，控制着复杂度惩罚项的强度。这种方法被称为**正则化风险最小化**（regularized risk minimization, RRM）。值得注意的是，如果损失函数是负对数似然，正则项是负对数先验分布，那么上式将等价于MAP估计。

RRM的两个关键点在于：如何衡量模型复杂度，如何选择。对于一个线性模型，我们可以将复杂度定义为它的自由度，这一点将在7.5.3讨论。对于一般的模型，我们可以使用VC维，这一点将在6.5.4进行讨论。为了选择，我们可以使用6.5.2讨论的方法。

6.5.2 结构化风险最小化

正则化风险最小化原则说明我们应该在给定的复杂度惩罚项的情况下训练模型，其决策规则由下式给定：

 （6.54）

但是我们又该如何挑选呢？我们不能使用训练集，因为这会过分低估真实的风险，导致一个被称为**乐观训练误差**（optimism of the training error）的问题。作为一种备选方案，我么可以使用下面的规则，被称为**结构风险最小化**（structural risk minimization）原则：

 （6.55）

其中为风险的估计值。存在两种普遍使用的估计方法：交叉验证和理论风险上确界。我们在下面讨论这两种方法。

6.5.3 使用交叉验证估计风险

我们可以通过一个验证集来估计某个估计量的风险。如果我们没有一个单独的验证集，我们可以使用**交叉验证**（cross validation, CV）,关于这一点我们在1.4.8节简单讨论过。更加精确地，CV的定义如下。令*N*=||表示在训练集中样本的数量。第*k*个测试包中的数据由*k*表示，其他数据由-*k*表示。（在分层交叉验证stratified CV中，这些分包在被选择的过程中，需要尽量在每个包中不同类别的样本数量近似相等。）令为一个算法或者一个拟合函数，它以一个数据集和一个模型的索引号*m*（*m*可以是一个离散的索引号，比如说一个多项式的自由度，或者一个连续的索引号，比如一个正则项的强度）为输入，返回一个参数向量：

 （6.56）

最后，令为一个预测函数，它基于一个输入和一个参数向量，返回一个预测：

 （6.57）

所以，组合得到的训练—预测循环（fit-predict cycle）表示为：

 （6.58）

风险*f*m的*K*-flod 交叉验证定义为：

 （6.59）

值得注意的是，针对每一个分包，我们可以只训练模型一次。令为基于所有除了在分包*k*内的数据训练的函数。然后，我们可以将CV估计重新书写为：

 （6.60）

其中为将数据*i*作为测试数据的分包。换句话说，我们预测*y*i时所使用的模型，是基于除了**x**i以外的数据进行训练的。

如果*K*=*N*，这个方法被称为**留一法交叉验证**（leave one out cross validation）或者LOOCV。在这种情况下，估计风险变成

 （6.61）

其中。这要求训练模型*N*次，其中对于，我们忽略第*i*个训练样本。幸运的是，对于一些模型种类和损失函数（比如说线性模型和平方损失），我们可以只训练模型一次，然后以解析的方式“移除”第*i*个样本的影响。这被称为**广义交叉验证**（generalized cross validation, GCV）。

6.5.3.1 例子：使用CV为岭回归挑选

作为一个具体的例子，考虑在一个含惩罚项的线性回归中，选择*l*2正则的强度。我们使用如下的规则：

 （6.62）

其中为我们搜寻的一个有限区间，为使用*K*分包CV估计得到的使用的风险，由下式给定：

 （6.63）

其中为预测函数，该预测函数基于除了分包*k*中的数据进行训练，为MAP估计。图6.6（b）给出了风险的CV估计与的关系，其中损失函数为平方差。

当我们进行分类时，我们通常使用0-1损失。在这种情况下，我们优化经验风险的凸上限从而实现对的估计，但我们优化风险（的交叉验证估计）本身来实现对的估计。当我们估计时，我们可以使用非光滑版的0-1损失函数，因为我们是在（一维）空间中采用暴利搜索。（**译者注**：该段的意思是说，我们在分类问题中，对模型参数进行优化时需要采用光滑版的目标函数，而如果使用交叉验证对超参进行选择，这个选择的依据可以基于非光滑的损失函数）

当我们有超过1个或者2个调整参数时，这种方式将变得不可行。在这种情况下，我们可以使用经验贝叶斯，它允许我们使用基于梯度的方法优化大规模的超参数，而不是使用暴利搜索。通过5.6节了解更多细节。

6.5.3.2 一标准差准则

上述介绍的内容对风险进行了估计，但却没有衡量估计的不确定度。在频率学派中，一个标准的关于不确定度的测度是期望的标准差，定义为：

 （6.64）

其中为损失方差的估计：

 （6.65）

注意衡量的是*L*i在整个样本上的固有的变化，其中se衡量了我们对期望的不确定度。

假设我们对一系列模型使用CV，并且计算出它们估计风险的期望和标准差。根据含噪的风险估计值进行模型的选择时，一个启发式的方式选择那个最简单的模型，该模型的风险比那个最好的模型的风险不大于一个标准差，这被称为**一标准差准则**（one-standard error rule）。举例来说，在图6.6中，我们发现启发式方式并没有选择曲线的最低点，而是右边比最低点稍高的点，因为在经验性能一样的情况下，它对应一个正则强度更强的模型。

6.5.3.3 非概率性质下的非监督学习模型中的CV

我们正在面对一个非监督学习问题，我们必须使用一个损失函数，比如，它衡量了重构的误差。此处表示某种编码——解码组合。然而，正如我们在11.5.2节所讨论的，我们并不能使用CV来决定的复杂度，因为我们总是可以基于更加复杂的模型得到更低的损失，哪怕是在测试集上进行评估。这是因为更加复杂的模型将会较少程度地压缩数据，从而导致更少的损坏。因此，我们必须使用概率学的模型或者发明其他启发式的方法。

6.5.4 使用统计学习理论计算风险的上确界

交叉验证方法的主要问题在于慢，因为我们需要对模型训练多次。一个更快的方式是以解析的方式计算泛化误差的近似值或者边界值。这一点在**统计学习理论**（statistical learning theory, SLT）中被研究。更加精确的，SLT尝试使用经验风险来确定任意数据分布、假设、样本尺寸，假设空间大小下的风险的边界。

首先考虑假设空间的大小是有限的这一种情况，其大小为。换句话说，我们正在从一个有限的列表中选择模型/假设，而不是对实数域的参数进行优化。

**定义6.5.1** 对于任意数据分布，和任意来自分布的大小为N的数据集，我们对错误率的估计大于的概率，在最坏的情况下，其上确界为：

 （6.66）

**证明：**为了证明上式，我们需要两个有用的结论。第一，霍夫丁不等式（Hoeffding’s inequality），如果*X*1,…,*XN*~，然后对于任意，

 （6.67）

其中。第二，布尔不等式（union bound），如果*A*1,…,*A*d为一系列事件，则。

最后，为了符号上的简便，令为真实的风险， 为经验风险。

基于这些结论，我们有：

 （6.68）

 （6.69）

 （6.70）

这个边界告诉我们乐观情况下训练误差随着增加而增加，但是随着增加而下降。这与我们所预料的是一样的。

如果假设空间是无穷大的（比如说我们有实值参数），我们就不能使用。取而代之的是，我们需要使用一个叫VC（Vapnik-Chervonenkis）维的量。

我们从所有理论退出来，统计学习理论背后的关键思想其实相当简单。假设我们找到一个经验风险很低的模型。如果相对于数据规模而言，假设空间非常大，那么很大程度上我们因为十分“幸运”，而正好选择了那个十分适合数据的模型。然而，这并不意味着被选择的那个模型就具备低的泛化误差。但是如果假设空间的大小被充分限制，训练集足够的大，那么我们选择一个适应数据很好的模型的“运气”就没那么好了，所以低的经验风险往往意味着低的真实风险。

统计学习理论相较于CV的优势在于前者计算风险上限的速度要快于后者。缺点在于对于很多感兴趣的模型，很难计算出模型的VC维，并且上确界通常非常松。

如果将学习器的计算复杂度考虑在内，我们可以拓展统计学习理论。这个领域被称为**计算学习理论**（computational learning theory，COLT）。该领域大部分的工作集中在当*h*是二值分类器，损失函数是0-1损失的情况。如果我们观察到了一个低的经验风险，而且假设空间“适当地”小，然后我们称我们的估计函数是**在概率上近似正确**（probably approximately correct, PAC）的。如果一个属于PAC的函数可以使用一个多项式时间算法去确定，则称这个函数为高效PAC可学的（efficiently PAC-learnable）。

6.5.5 代理损失函数

在ERM/RRM框架内最小化损失通常不是很容易。举例来说，我们可能想要优化AUC或者F1分值。或者更加简单地，我们可能只是想最小化0-1损失，这在分类问题中常常用到。不幸的是，0-1风险是一个非常不光滑的目标函数，所以很难对它进行优化。一个备选方案是使用最大似然估计取而代之，因为对数似然是0-1损失的光滑凸上界，我们在下文进行展示。

为了说明这一点，考虑一个二值逻辑回归，令。假设我们的决策函数计算对数几率比例，

 （6.71）

输出标签对应的概率分布为：

 （6.72）

定义**对数损失**（log-loss）为：

 （6.73）

显然最小化对数损失等价于最大似然函数。

现在考虑计算最优可能的那个类别，等价于：如果，则选择，如果，则。那么0-1损失函数就变成：

 （6.74）

图6.7绘制了两个损失的图形。我们发现NLL确实是0-1损失的上确界。

对数损失为代理损失函数（surrogate loss function）的一个例子，另一个例子为铰链损失（hinge loss）：

 （6.75）

图6.7绘制了相关图形。我们发现这个函数像一个门的铰链，这也是它的名字由来。这个损失函数是一个著名的分类方法支持向量机（SVM）的基础，我们将在14.5节对这个模型进行讨论。

代理函数一般选择的是一个凸边界，因为凸函数很容易被最小化。

6.6 频率学派的病理

I believe that it would be very difficult to persuade an intelligent person that current

[frequentist] statistical practice was sensible, but that there would be much less difficulty with an approach via likelihood and Bayes’ theorem. — George Box, 1962.

频率统计显示出各种形式的怪异和不受欢迎的行为，被称为**病症**（pathologies）。为了提醒读者，我们在下文给出一些例子。

6.6.1 置信区间的反直觉行为

**置信区间**（confidence interval）是从估计量的采样分布中得到的一个区间（然而，贝叶斯中的可靠区间是从参数的后验分布中得到的，我们在5.2.2节进行过讨论）。更加精确的，一个参数的频率学派的置信区间定义为如下的（非自然的）表达式：

 （6.76）

也就是说如果我们根据采样得到未来的假想的数据，那么参数落在区间的次数占比为。

让我们退一步想想正在发生的事情。在贝叶斯学派中，我们以“我们知道的”为条件——即观测到的数据——然后基于我们不知道的求平均，即参数。在频率学派中，我们做的事情完全相反：我们将“我们不知道的”作为条件——即真实的参数值——然后对假设的未来的数据求平均。

这种关于置信区间的反直觉的定义可能导致奇怪的结果。考虑如下的例子，假设我们从下面的分布中采样得到两个整数：

 （6.77）

如果，我们得到下面每一个输出的概率为0.25：

（39,39），（39,40），（40,39），（40,40） （6.78）

令，定义下面的置信区间为：

 （6.79）

对于上面的例子，我们得到：

[39, 39],[39,39],[39,39],[40,40] （6.80）

所以公式6.79显然是一个75%的CI，因为39在上述区间中3/4中出现。然而，如果=(39,40)，那么，所以我们知道肯定是39，然而事实上，我们只有75%的“信心”。

另一个不那么不自然的例子。假设想估计伯努利分布的参数。令为样本的均值。MLE为。一个关于伯努利参数近似95%的置信区间为（这被称为Wald区间，它基于对伯努利分布的高斯近似；对于式3.27）。现在考虑一个单独的试验，其中*N*=1，*x*1=0。显然MLE=0，这是过拟合的，这一点在3.3.4.1节中介绍。但是95%的置信区间也是（0,0），貌似更加严重。可以认为，上述缺陷是由于我们用高斯逼近了真实的抽样分布，或者是由于样本量太小，或者是参数“太极端”。然而，即使对于较大的N和非极端参数，Wald区间也可能表现得很差。

6.6.2 p值被认为是有害的

假设我们想决定是否接受或者拒绝一个基准模型，也就是所谓的**原假设**（null hypothesis）。我们需要定义一个决策规则。在频率学派，标准的方式是首先计算一个值，被称为p值（p-value），它定义为（在原假设的情况下）观察到一些与实际观察到的一样大或更大的检验统计量*f*（）（如卡方统计量）的概率：

 （6.81）

这个量需要计算采样分布的**尾域概率**（tail area probability）；我们在下面给出一个例子。

在给定p值的情况下，我们定义的决策规则为：当且仅当p值小于某个阈值时，比如，我们拒绝这个原假设。如果我们这么做了，我们说观察到的检验统计量与预期检验统计量之间的差异在水平上具有统计学意义。这种方法被称为**原假设显著性检验**（null hypothesis significance, NHST）。