数学 计算机 统计学 回归分析 计量经济学

最大似然估计和最小二乘法怎么理解？

关注问题 写回答 1 条评论 分享 邀请回答 ···

41 个回答

默认排序

司马懿
经济学、博弈论 话题的优秀回答者

375 人赞同了该回答

谢邀，这个问题下的答案很多是直接从机器学习领域过来回答的，很有启发性，让我了解了在别的领域是如何理解这两种方法的。论及本质，其实两者只是用不同的度量空间来进行的投影，如同@颢卿 的答案所提到的那样，OLS的度量是L2 norm distance，而极大似然的度量是Kullback-Leibler divergence.

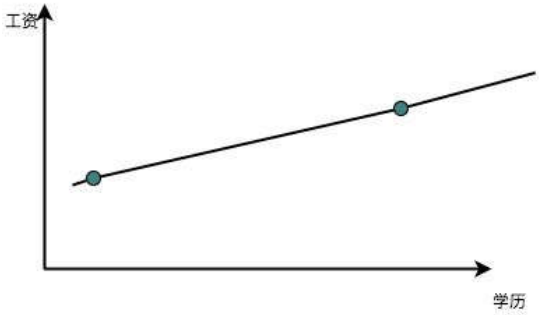
不过这种高度抽象的统一框架，主要功能就是让人听起来很优雅很爽，满足了人对形式美的追求，缺点也很明显：

1. 不在概率论方面下一些功夫不太能真正理解
2. 无法直接拿过来应用

所以在大多数情况下，我们介绍这两种方法的时候，可能并不需要讲解这么抽象的东西。好，下面我们开始说人话^_^

设想一个例子，教育程度和工资之间的关系。我们观察到的数据无非就是一个教育程度，对应着一个工资。我们希望的自然是找到两者之间的规律：如果把教育程度的初中、高中、大学、研究生及博士定义为1234的话，我们希望找到类似于工资=1000 +2000x教育程度 的这种规律，其中1000和2000是我们需要从数据里面发现的，前者称之为底薪，后者称之为教育增量薪水。

如果我们就观察到两个数据，那解起来很简单，直接把两个数据带进去，二元一次方程组，就得到底薪和教育程度增量薪水之间的关系。这个在图上就体现为两点决定一条直线：

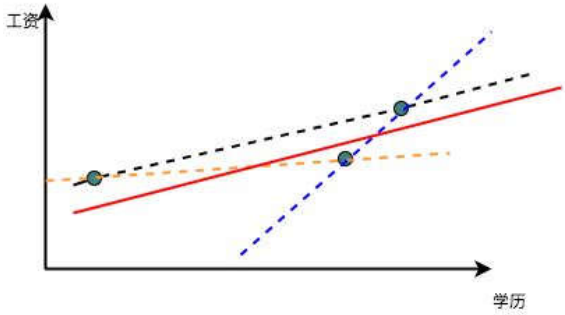


但是如果现在有三个数据，怎么办呢？如果这三个点不在一条线上，我们就需要作出取舍了，如果我们取任意两个点，那么就没有好好的利用第三个点带来的新信息，并且因为这三个点在数据中的地位相同，我们如何来断定应该选用哪两个点来作为我们的基准呢？这就都是问题了。这个时候我们最直观的想法就是『折衷』一下，在这三个数据，三条线中间取得某种平衡作为我们的最终结果，类似于图中的红线这样：

▲ 375 42 条评论 收藏 感谢 收起

相关推荐

相关问题



那怎么取平衡呢？那我们现在必须引入误差的存在，也就是我们要承认观测到的数据中有一些因素是不可知的，不能完全的被学历所解释。而这个不能解释的程度，自然就是每个点到红线在Y轴的距离。

但是我们尽管痛苦的承认了有不能解释的因素，但是我们依然想尽可能的让这种『不被解释』的程度最小，于是我们就想最小化这种不被解释的程度。因为点可能在线的上面或者下面，故而距离有正有负，取绝对值又太麻烦，于是我们就直接把每个距离都取一个平方变成正的，然后试图找出一个距离所有点的距离的平方最小的这条线，这就是最小二乘法了，简单粗暴而有效。

而极大似然则更加的有哲理一些。还用上面的例子，我们观察到了三个点，于是我们开始反思，为什么我们观察到的是这三个点而不是另外三个？大千世界，芸芸众生，这么多人都有不同的工资，不同的学历，但是偏偏这三个点让我给观察到了。这肯定说明了某种世界的真相。

最大似然估计和最小二乘法怎么理解？

什么世界的真相呢？因为我们观察到了这三个点，反过来说，冥冥之中注定了这三个点被我们观察到的概率可能是最大的。所以我们希望找到一个特定的底薪和教育增量薪水的组合，让我们观察到这三个点的概率最大，这个找的过程就是极大似然估计。

具体的做法很简单，因为底薪和教育增量薪水虽然我们不知道，但是它一定存在，所以是个固定的值，能够随机变动的就是我们观察不到的神秘误差，那么给定一组底薪和教育增量薪水，必然存在一个唯一的误差与之对应，共同组合成了我们看到的数据。比如说，我们观察到一个人是：

高中毕业（学历变量=2）工资 4500，如果我们假定工资=1000 +2000x教育程度的话，那么理论上工资应该是5000，而我们观察到了4500，所以这个时候误差为500。而误差=500，根据我们假设的误差的概率函数，总是存在一个概率与之相对应的（这个概率的分布我们可以假设）。而极大似然估计，就是把我们观察到每个样本所对应的误差的概率乘到一起，然后试图调整参数以最大化这个概率的乘积。

其背后的直觉是：假想有一个神秘的超自然力量，他全知全能，自然也知道真实的数据背后的规律。他在你抽样之前先做了一次复杂的计算，把无数个可能的抽样中，最可能出现的那个抽样展示给你。于是你根据这个抽样，逆流而上，倒推出来了数据背后的真实规律。

总结一句话，最小二乘法的核心是权衡，因为你要在很多条线中间选择，选择出距离所有的点之和最短的；而极大似然的核心是自恋，要相信自己是天选之子，自己看到的，就是冥冥之中最接近真相的。^_^

编辑于 2017-11-07

知乎用户

475 人赞同了该回答

最大似然估计：现在已经拿到了很多个样本（你的数据集中所有因变量），这些样本值已经实现，最大似然估计就是去找到那个（组）参数估计值，使得前面已经实现的样本值发生概率最大。因为你手头上的样本

▲ 375 42 条评论 收藏 感谢 收起

化，是个连乘积，只要取对数，就变成了线性加总。此时通过对参数求导数，并令一阶导数为零，就可以通过解方程（组），得到最大似然估计值。
最小二乘：找到一个（组）估计值，使得实际值与估计值的距离最小。本来用两者差的绝对值汇总并使之最小是最理想的，但绝对值在数学上求最小值比较麻烦，因而替代做法是，找一个（组）估计值，使得实际值与估计值之差的平方加总之后的值最小，称为最小二乘。“二乘”的英文为least square，其实英文的字面意思是“平方最小”。这时，将这个差的平方的和式对参数求导数，并取一阶导数为零，就是OLSE。

发布于 2014-03-27

渣君
债市小渣毛

523 人赞同了该回答

说的通俗一点啊，最大似然估计，就是**利用已知的样本结果**，**反推最有可能（最大概率）导致这样结果的参数值。**
例如：一个麻袋里有白球与黑球，但是我不知道它们之间的比例，那我就有放回的抽取10次，结果我发现我抽到了8次黑球2次白球，我要求最有可能的黑白球之间的比例时，就采取最大似然估计法：
我假设我抽到黑球的概率为p,那得出8次黑球2次白球这个结果的概率为：
P(黑=8)=p^8*（1-p）^2,现在我想要得出p是多少啊，很简单，使得P(黑=8)最大的p就是我要求的结果，接下来求导的的过程就是求极值的过程啦。
可能你会有疑问，为什么要ln一下呢，这是因为ln把乘法变成加法了，且不会改变极值的位置（单调性保持一致嘛）这样求导会方便很多~

同样，这样一道题：设总体X的概率密度为
已知 X1,X2..Xn是样本观测值，求θ的极大似然估计

这也一样啊，要得到 X1,X2..Xn这样一组样本观测值的概率是
P{x1=X1,x2=X2,...xn=Xn}=

f(X1,θ)f(X2,θ)...f(Xn,θ)
然后我们就求使得P最大的θ就好啦，一样是求极值的过程，不再赘述。

发布于 2014-03-29

亲爱的龙哥
Quant at Bluewood Capital

98 人赞同了该回答

最小二乘法可以从Cost/Loss function角度去想，这是统计（机器）学习里面一个重要概念，一般建立模型就是让loss function最小，而最小二乘法可以认为是 loss function = （y_hat -y）^2的一个特例，类似的想各位说的还可以用各种距离度量来作为loss function而不仅仅是欧氏距离。所以loss function可以说是一种更一般化的说法。

最大似然估计是从概率角度来想这个问题，直观理解，似然函数在给定参数的条件下就是观测到一组数据realization的概率（或者概率密度）。最大似然函数的思想就是什么样的参数才能使我们观测到目前这组数据的概率是最大的。

类似的从概率角度想的估计量还有矩估计（moment estimation）。就是通过一阶矩 二阶矩等列方程，来反解出参数。

各位有人提到了正态分布。最大似然估计和最小二乘法还有一大区别就是，最大似然估计是需要有**分布假设**的，属于参数统计，如果连分布函数都不知道，又怎么能列出似然函数呢？ 而最小二乘法则没有这个假设。 二者的相同之处是都把估计问题变成了最优化问题。但是最小二乘法是一个凸优化问题，最大似然估计不一定是。

发布于 2015-01-08

王相及
经济学

205 人赞同了该回答

我尽量不写数学，用通俗语言说一说OLS（最小二乘）和MLE（最大似然）的本质。

1. OLS其实就是 linear projection（线性投影），是Hilbert 空间中的被解释变量在一组解释变量上的线性投影。（这句话你可能看不懂，没事，先看下边的。）

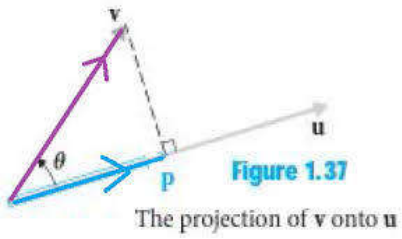


Figure 1.37
The projection of v onto u

如上图，我们有两个向量，v, u, 那么p 就是v 在u 上的线性投影，记作： p= L(v|u) = b*u

b 我们称作“系数”。因为是在u 上投影，所以 p = b* u

当然，为了做出来这个投影，我们必须定义“内积”（点乘，下面用x.y 表示x 和y的内积）的概念。这里，如果p 是 v 在u 上的投影，那么必须满足下面两个条件：

1）u . (v-p) = 0（垂直条件，也就是说，u 和 v-p “垂直”）

2）p = b*u（p必须在u 张成的子空间中)

好了，那么我们现在到Hilbert 空间，**这个Hilbert空间其实就是很多很多随机变量的集合，并且定义了内积的概念。**

怎么给随机变量定义内积？如果x, y 是两个(多维）随机变量（列向量），那么定义x.y = E(xy') (y' 表示y的转置）。此时，如果有一个多维随机变量x，和一个随机变量y，我们把y 投影到 x上，就会有L(y|x) = x b，这个b 就是我们在OLS中想要求得的系数。

怎么求这个系数？由垂直条件可知： x.(y - L(y|x)) = x.(y - x b) =0 所以：b = (x.x)^-1 x.y = (Exx')^-1 E(xy)。

最后，投影是如何跟“最小二乘”扯上关系？最小二乘，其实就是最小方差。在最上面的图中，投影变量p（拟合值）是在u（解释变量）张成的子空间中，距离v（被解释变量）最“近”的那个向量。这个“近”（距离的概念），是需要用内积来定义。而我说的 x.y = Exy' 这种定义内积的方法，正好能推导出来用“方差”来定义距离的方法。所以投影得到了，最小二乘也实现了。

2. MLE可以看作一种特殊情况下的Bayesian 估计，具体来说，就是在prior 是 diffuse（无知的）情况下，让posterior 分布取得极大值的系数值。

我们有一些理论模型，记作 "model"，这个model 是什么，在很多实践中，就是一个模型中关键系数的值是什么这样的问题（不同的系数的值，我们称作不同的model）。我们现在又观测到一组数据，记作"observation"。那么问题来了，**给定一个model（一组关键系数的值）**，必然会有关于observation 的分布密度函数，所以我们知道P(observation|model)（给定一个model，observation的条件分布）的函数形式。

我们真正关心的，却是 P(model|observation) 的函数形式，也就是给定了当前的observation（observation是实际观测到的，是确定下来的），到底不同的model的概率是什么。当然，一个很贪心的做法，就是找到那个能把P(model|observation) 取到最大值的model（给定某个观测，最有可能的model）。

现在根据贝耶斯原理，

P(model|observation) = [P(observation|model) * P(model)]/ P(observation)

其中P(observation) 不太重要，因为我们想知道不同model 是如何影响 P(model|observation) 的，或者是贪心的求P(model|observation)的最大值。而P(observation)已经固定下来了，不随model改变，所以我们无视他。

我们如果知道 P(model)（所谓的Prior) 的函数形式，那么就没有什么问题了。此时的P(model|observation)是一个关于model 的函数。报告这个P(model|observation)作为model的函数的函数形式，就叫贝耶斯估计。可是，这需要我们知道P(model)。实际中我们不知道这个玩意，所以一般我们猜一个。

我们如果承认不
此时求P(mode

MLE。

3。二者区别。从上面可见，OLS 是把所有变量扔到线性空间中，求线性投影的系数：它并不需要什么信息。而MLE 是需要我们知道一个完整的理论模型（否则P(observation|model) 根本就不知道是什么）。由于一般大家接触的都是线性模型，所以二者区别不大。当模型无法变成线性状态时（比如censored data, logit/probit 之类的），此时OLS此时报告的仍然是线性投影，我们却没有用到这些“非线性”的信息，因此MLE的选项就好很多。

不论任何时候，OLS报告的都是线性投影（准确的说，是对线性投影的“估计”值），都是 "best linear predictor"。当你加上了一些假设，（比如 在 y = x b + u 这样的理论模型中，你假设了 E(xu) = 0 这样的经典计量经济学假设），此时OLS报告的还是线性投影，只不过，这个线性投影正好等于模型中的"b"。

如果在模型 y = x b + u 中，E(xu) != 0，不满足经典计量假设。那么此时你用上了OLS，得到的是 y = x a + e 这样的模型，你是知道了a，而且很容易知道E(xe) = E(x(y-x a)) = x. (y-x a) = 0（线性投影的垂直条件）。但是这个a 却不是你一开始设定模型时想要知道的b。

编辑于 2017-05-22