机器学习 文本挖掘 文本数据分析 估计 文本分析

关注者 298 被浏览 18637

MLE，MAP，EM 和 point estimation 之间的关系是怎样的？

在阅读LDA和主题模型相关论文的时候，这些名词经常出现。MLE(Maximum Likelihood Estimation), MAP(Maximum…显示全部

关注问题 写回答 添加评论 分享 邀请回答 ···

查看全部 6 个回答

H.Chen
澳大利亚昆士兰大学计算机科学与技术博士在读

MLE vs MAP: the connection between Maximum Likelihood and Maximum A Posteriori Estimation

发布于 2017-09-13

▲ 0 ▼ 添加评论 分享 收藏 感谢

更多回答

陈义
广告产品技术职位空缺,欢迎垂询.

90 人赞同了该回答

感谢肖智博邀请回答。

和点估计相对应的是区间估计，这个一般入门的统计教材里都会讲。直观说，点估计一般就是要找概率密度曲线上值最大的那个点，区间估计则要寻找该曲线上满足某种条件的一个曲线段。

最大似然和最大后验是最常用的两种点估计方法。以最简单的扔硬币游戏为例，一枚硬币扔了五次，有一次是正面。用最大似然估计，就是以这五次结果为依据，判断这枚硬币每次落地时正面朝上的概率（期望值）是多少时，最有可能得到四次反面一次正面的结果。不难计算得到期望概率0.2。

用五次试验结果来估计硬币落地时正面朝上的概率显然不够可靠。这时候先验知识可以发挥一些作用。如果你的先验知识告诉你，这枚硬币是制币局制造，而制币局流出的硬币正面朝上的概率一般是0.5，这时候就需要在先验概率0.5和最大似然估计0.2之间取个折中值，这个折中值称为后验概率。这时候剩下的问题就是先验知识和最大似然估计结果各应起多大作用了。如果你对制币局的工艺非常有信心，觉得先验知识的可靠程度最起码相当于做过一千次虚拟试验，那么后验概率是(0.2 * 5 + 0.5 * 1000)/(5 + 1000) = 0.4985，如果你对制币局技术信心不足，觉得先验知识的可靠程度也就相当于做过五次试验，那么后验概率是(0.2 * 5 + 0.5 * 5)/(5 + 5) = 0.35. 这种在先验概率和最大似然结果之间做折中的方法称为后验估计方法。这是用贝耶斯观点对最大后验方法的阐述，其实也可以用用经典统计学派的偏差方差的折中来解释。

EM方法是在有缺失值时进行估计的一种方法，这是一个迭代方法，每个迭代有求期望(E)和最大化(M)两个步骤。其中M可以是MLE或者MAP。一般需要先为缺失值赋值（E步骤初始化），然后重复下面的步骤：
1）用MLE或MAP构造模型(M步骤)；
2）用所得模型估计缺失值，为缺失值重新赋值(E步骤)；
仍然以扔硬币为例，假设投了五次硬币，记录到结果中有两正一反，还有两次的数据没有记录下来，不妨自己用上述步骤演算一下硬币正面朝上的概率。需要注意，为缺失值赋值可以有两种策略，一种是按某种概率赋随机值，采用这种方法得到所谓hard EM，另一种用概率的期望值来为缺失变量赋值，这是通常所谓的EM。另外，上例中，为两个缺失记录赋随机值，以期望为0.8的0-1分布为他们赋值，还是以期望为0.2的0-1分布为他们赋值，得到的结果会不同。而赋值方法的这种差别，实际上体现了不同的先验信息。所以即便在M步骤中采用MLE，EM方法也融入了非常多的先验信息。

上面的例子中只有一个随机变量，而LDA中则有多个随机变量，考虑的是某些随机变量完全没有观测值的情况（也就是Latent变量），由于模型非常复杂，LDA最初提出时采用了变分方法得到一个

简单的模型，EM被应用在简化后的模型上。从学习角度说，以PLSA为例来理解EM会更容易一点。另外，kmeans聚类方法实际上是典型的hard EM，而soft kmeans则是通常的EM，这个在[1]中的讨论最直观易懂。

[1] Information Theory, Inference, and Learning Algorithms, http://inference.phy.cam.ac.uk/mackay/itila/

发布于 2011-11-06

知乎用户

83 人赞同了该回答

故事是这样的：
从前，有一个卖算法的小女孩有一堆数据，她首先假设这些数据的生成机制可以用某一个概率分布来描述。
这时她面对了两种选择：

1. 不做进一步假设，认为这个概率分布（如果是实数域上的数据）可以是正态分布，可以是学生t-分布，可以是拉普拉斯分布，可以是各种其他各种分布，甚至还可以是把以上分布用另一个随机变量混合起来的。在这种情况下，她要做的就是非参数统计。这个和题目中提到的四个名词关系都不大，暂且按下不表。
2. 假设形成数据的概率分布是某一族分布中的一个。每一族概率分布都可能有无数个。这时她要问自己的问题是：我们面前的这个概率分布的参数是多少。比如说她假设数据是由某种正态分布形成的，就要对这个分布的期望和方差做一个估值（estimation），也就是推论（inference）。这里就可以提到题主问题中的 point estimation 这个词了，MLE，MAP和EM都是对模型估值的方式或者方法。实际上，在对模型的参数做出估计以后，买算法的小女孩们还可以问自己另一个问题：眼前的数据确实被这个推论得知的概率分布形成的概率是多少。这时就要用到假设检验，并以此得到参数的置信区间（confidence interval），point estimation 中的“点”就是和置信区间中的“区间”相对的。

现在小女孩决定了要使用的分布族群，她还是不知道怎么估计参数的值。其实这里有三种可能的状况：

MLE，MAP，EM 和 point estimation 之间的关系是怎样的？

1. 小女孩选择的模型很简单，数据量也很大，给定一组参数以后，数据的似然函数（likelihood function）可以很明白的写出来（tractable）。这时候，一个很明显也很自然的选择就是使用最有可能生成了数据的参数值，也就是说，选择让似然函数最大化的参数值，也即 maximum likelihood estimation。如果似然函数可以直接在理论上求最大值当然好，算不出最大化的表达式，可以靠数值运算最大化也可以。因为数据量很大，模型简单（甚至足够规则），MLE 是一个不错的选择。
2. 小女孩选择的模型很复杂。似然函数虽然貌似可以写出来，但是要给指数级的项目求和，或者似然函数根本写不出来。这样的模型就不能简单最大化似然了之了。这时候，买算法的小女孩们或者老板突然发现，虽然不能直接写出模型的似然函数，要是给模型加上几个隐变量，那么给定参数下，数据与隐变量的联合分布倒是很容易算，要是知道隐变量的值，针对参数最大化似然函数也很容易。唯一的问题是，他们既不知道隐变量的值，也不知道参数的值。这时就可以用到 EM 算法了，这个两步的算法很好地解决了这个两不知的问题，也即：第一步，给定参数，对隐变量做期望，算出包括隐变量的似然函数；第二步，对这个似然函数最大化，update 参数。因为这个模型可以让似然函数递增，如果似然函数是凹函数，那就一定会收敛到最大值，如果似然函数有多个极值，则要随机化初始参数值，算很多次，选择似然最大的参数。
3. 小女孩的模型未必很复杂，但是数据非常少，与此同时，小女孩或者老板已经关注这个问题很久，对参数有一定的想法了。这时候就可以用到贝叶斯统计了：我们可以给参数定一个先验统计分布，然后用数据算出参数的后验分布（posterior probability，其实大概就是把先验的概率密度和数据的似然函数乘一乘，然后再标准化一下的问题），然后再最大化后验，这个最大化后验分布的参数值就是 maximum a posteriori 了。其实在贝叶斯统计中，最大化后验概率的参数值未必是最好的参数值，根据决策论的看法，一般会最小化某个 loss function，得到的结果

可能是后验分布的期望或者中值。不过如果参数的空间是非凸的（比如离散集合），这两个值未必在参数空间内，说不定也很不好算，所以在实际应用中，用 MAP 的也不少。

编辑于 2015-08-17

查看全部 6 个回答