

知乎

首页发现话题

搜索你感兴趣的内容...

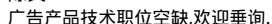
机器学习 文本挖掘 文本数据分析 估计 文本分析

关注者298 被浏览18620

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



感谢肖智博邀请回答。

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，其实也可以用经典统计学派的偏差方差的折中来解释。

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，还有两次的结果没有记录下来，不妨自己用上述步骤推算一下硬币正面向上的概率。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

相关问题

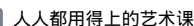
物理专业自学计算机应该学些什么？ 23
个回答

高频交易和统计/机器学习套利模型，哪个技术含量高？7 个回答

最优化问题的简洁介绍是什么？ 53 个回答

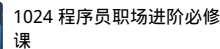
随机过程、机器学习和蒙特卡洛在金融应用中都有哪些关系？ 9 个回答

私家课·Live 推荐

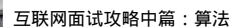


翁昕

共 13 节课



8 场 live, 6107 次参与



★★★★☆ 795 人参与



刘看山 · 知乎指南 · 知乎协议 · 应用 · 工作

侵权举报 · 网上有害信息举报专区

违法和不良信息举报：010-82716601

儿童色情信息举报专区

联系我们 © 2017 知乎

知乎用户

▲ 90

● 3 条评论

分享

★ 收藏

♥ 感谢

收起 ^

