088 | 基础文本分析模型之三：EM算法

2018-04-25 洪亮劼

AI技术内参

进入课程 >

讲述：初明明

时长 06:34　大小 3.01M

周一我们分享的模型是“概率隐语义分析”（Probabilistic Latent Semantic Indexing），或者简称为 PLSA，这类模型有效地弥补了隐语义分析的不足，在 LDA 兴起之前，成为了有力的文本分析工具。

不管是 PLSA，还是 LDA，其模型的训练过程都直接或者间接地依赖一个算法，这个算法叫作“**期望最大化**”（Expectation Maximization），或简称为 **EM 算法**。实际上，EM 算法是针对隐参数模型（Latent Variable Model）最直接有效的训练方法之一。既然这些模型都需要 EM 算法，我们今天就来谈一谈这个算法的一些核心思想。

EM 和 MLE 的关系

EM 算法深深根植于一种更加传统的统计参数方法：**最大似然估计**（Maximum Likelihood Estimation），有时候简称为 **MLE。绝大多数的机器学习都可以表达成为某种概率模型的 MLE 求解过程。**

具体来说，MLE 是这样构造的。首先，我们通过概率模型写出当前数据的“似然表达”。所谓的“似然”表达，其实也就是在当前模型的参数值的情况下，看整个数据出现的可能性有多少。可能性越低，表明参数越无法解释当前的数据。反之，如果可能性非常高，则表明参数可以比较准确地解释当前的数据。因此，**MLE 的思想其实就是找到一组参数的取值，使其可以最好地解释现在的数据。**

针对某一个模型写出这个 MLE 以后，就是一个具体的式子，然后看我们能否找到这个式子最大值下的参数取值。这个时候，整个问题往往就已经变成了一个优化问题。从优化的角度来说，那就是针对参数求导，然后尝试把整个式子置零，从而求出在这个时候的参数值。

对绝大多数相对比较简单的模型来说，我们都可以根据这个流程求出参数的取值。比如，我们熟悉的利用高斯分布来对数据进行建模，其实就可以通过 MLE 的形式，写出用高斯建模的似然表达式，然后通过求解最优函数解的方式得到最佳的参数表达。而正好，这个最优的参数就是样本的均值和样本的方差。

然而，并不是所有的 MLE 表达都能够得到一个“**解析解**”（Closed Form Solution），有不少的模型甚至无法优化 MLE 的表达式，那么这个时候，我们就需要一个新的工具来求解 MLE。

EM 算法的提出就是为了简化那些求解相对比较困难模型的 MLE 解。

有一点需要说明的是，EM 算法并不能直接求到 MLE，而只能提供一种近似。多数无法直接求解的 MLE 问题都属于**非凸（Non-Convex）问题**。因此，**EM 能够提供的仅仅是一个局部的最优解，而不是全局的最优解。**

EM 算法的核心思想

理解了 EM 和 MLE 的关系后，我们来看一看 EM 的一些核心思想。因为 EM 算法是技术性比较强的算法，我建议你一定要亲自去推演公式，从而能够真正理解算法的精髓。我们在这里主要提供一种大体的思路。

EM 算法的一种解释是这样的。首先，我们可以通过代数变形，为每一个数据点的似然公式找到一个新的概率分布，而这个概率分布是通过一个隐含变量来达到的。很明显，在理论上，我们可以通过把这个隐含变量积分掉来达到恢复原始的 MLE 公式的目的。

然而，这里遇到的一个大的阻碍就是，在 MLE 公式里面，有一个求对数函数（log）在这个积分符号外面。这就导致整个式子无法进行操作。通俗地讲，EM 就是要针对这样的情况，试图把这个在积分符号之外的求对数函数拿到积分符号里面。能够这么做，是因为有一个不等式，叫“**杨森不等式**”。你不需要去理解杨森不等式的细节，大体上这个不等式是说，函数的期望值要大于或等于先对函数的变量求期望然后再对其作用函数。

于是，在这样的一个不等式的引领下，我们刚才所说的积分，其实就可以被看作是对某一个函数求期望值。而这个函数，恰好就是模型的似然表达。通过杨森不等式，我们可以把对数函数拿到积分符号里面，这样当然就无法保持等号了，也就是说，这一步的操作不是一个等值操作。利用杨森不等式之后的式子其实是原来的式子，也就是含有隐含变量的 MLE 式的一个“**下限**”（Lower Bound）。

利用杨森不等式，从而写出一个原始的 MLE 的下限，是标准的 EM 算法以及一系列基于变分 EM（Variational EM）算法的核心思想。这么做的目的其实就是把对数函数从积分的外面给拿到里面。

当我们有了这个下限之后，我们就可以套用 MLE 的一切流程了。注意，这时候，我们有两组未知数。一组未知数是我们**模型的参数**，另外一组未知数就是**模型的隐含变量**。于是，当得到下限之后，我们就需要对这两组未知数分别求导，并且得到他们的最优表达。

当我们按照当前的模型参数，对模型的隐含变量所对应的概率分布求解后，最优的隐含变量的概率分布就等于隐含变量基于数据的后验概率。什么意思呢？意思就是说，如果我们把隐含变量的取值直接等于其后验概率分布，就得到了当前的最优解。这个步骤常常被叫作“**E 步**”。

在进行了 E 步之后，我们再按照当前的隐含变量，求解这个时候最佳的模型参数。这常常被认为是“**M 步**”。**一次 E 步，一次 M 步则被认为是 EM 算法的一个迭代轮回。**

EM 算法貌似很神秘，但如果我们理解了整个流程的精髓，就可以把这个算法总结为：EM 算法是**利用杨森不等式得到 MLE 的一个下限，并且优化求解模型参数和模型的隐含变量的一个过程。**

掌握了这个精髓，我们就可以看到，为什么 LDA 和 PLSA 等隐变量模型需要利用 EM 或者类似 EM 的步骤进行求解。第一，这些模型的 MLE 都有一个对数函数在积分符号外面，使得这个过程无法直接求解。第二，这些模型本身就有隐含变量，因此不需要额外制造新的隐含变量。

总结

今天我为你介绍了一个经常用于求解概率图模型的 EM 算法。

一起来回顾下要点：第一，我们回顾了 EM 算法和 MLE 算法的关系；第二，我们讨论了 EM 算法的核心思想。

最后，给你留一个思考题，EM 算法在实际应用中有哪些问题呢？

欢迎你给我留言，和我一起讨论。

上一篇　087 | 基础文本分析模型之二：概率隐语义分析

下一篇　089 | 为什么需要Word2Vec算法?

精选留言 (3)

写留言

马勇(Dani...

2018-10-05

1

最好有公式

展开

林彦

2018-04-27

EM算法是不是有收敛速度慢，每一步的计算比较复杂的问题？

展开

罗马工匠

2018-04-25

还是有公式好理解一点。另外问题的答案能否放评论区呢？em算法除了局部最优，还有其他问题么？