机器学习系列-强填EM算法在理论与工程之间的鸿沟（下）

原创 夕小瑶 夕小瑶的卖萌屋 2017-03-16

前言

在上一篇文章《机器学习系列-强填EM算法在理论与工程之间的鸿沟（上）》中，小夕用优（恐）雅（怖）的数学理论来向读者解释了EM算法的工作原理。那么从工程角度出发的EM算法又是怎样的呢？

EM算法在工程上有很多应用场景，例如：
1、 **半监督学习**：即利用包含缺失类别标签的数据的混合数据集训练分类器。
2、 **数据预处理**：给缺失某一维特征的值的数据补上缺失值。
3、 **聚类**：对，聚类。
4、 **隐马尔科夫模型**：训练隐马尔科夫模型中的参数。
5、 …

场景辣么多，理论却只有一个。因此讨厌数学的攻城狮可能会记住很多场景下的EM算法，而喜欢数学（最起码不要跟数学打起来）的攻城狮则以不变应万变，早已看透一切，2333。

小夕搬出大栗子：

比如，我们要做文档分类。我们手头有10000篇文章，其中只有600篇标好了类别，其余9400篇均没有类别标签。那么如何训练出一个尽可能高精度的分类器呢？

> 诶？有人可能想，既然9400篇文档都没有标签，难道这些没有标签的数据都会有助于提高分类器的精度？怎么可能呢？
>
> 其实很好理解呀。虽然有些文档没有类别标签，但是这些文档的内容就包含分类信息啊。这里的信息指的是“词共现”，或者广义上说“特征共现”。比如我们利用有标签的文档发现“么么哒”是非常有助于文档分类的强特征，然而我们又在没有标签的文档中发现“么么哒”经常与“抱抱”一起出现！也就是共现！那么就可以从很大程度上说明“抱抱”也是有助于文档分类的强特征。
>
> 举个生动的事实，在UseNet语料库中做新闻类别分类，若要达到70%的精度，则需要2000篇有类别标记的文档。但是，如果我们有600篇有类别标记的文档，还有10000篇无类别标记的文档，那么同样可以达到70%的精度。

攻城狮眼中的EM算法

在攻城狮眼中，上面那个栗子显然是一个半监督学习问题（即数据集中既有有类别标记的样本，也有无类别标记的样本），因此显然可以搬出来EM算法呀。

在攻城狮眼中，EM算法非常简单：
1、 **仅利用有标签**的数据，训练一个朴素贝叶斯分类器。
2、 利用训练好的分类器给无类别标签的数据打上标签，顺便记下分类器对该标签的把握。然后将所有标签的把

握求和，得到值sum。
3、 利用**全部**数据重新训练朴素贝叶斯分类器。
4、 重复2、3步，直到sum不再变化（或者说近似于不再变化）。

诶？明明思路很简单啊，怎么会跟上一篇中那么多恐怖的公式扯上关系呐！

然而，机智的你有没有想过，算法为什么要这样写呢？这就是关键啦。

好桥梁，小夕造

首先，我们在理论EM中的目标是最大化似然函数！而你还记不记得小夕前面讲过，其实最大化后验概率的本质工作就是最大化似然函数呢？

诶？发现了没有~在工程上，我们在**第2步**中收集分类器对每个标签的把握并求和，那不就是收集的整个数据集的后验概率嘛！不就是在近似计算似然函数嘛！

因此，显然，在工程上的**第4步**，也就是不停的重复2、3步，肯定会让分类器的精度越来越大呀，因此分类器会对每个标签的把握越来越大！因此这不就是相当于理论上的最大化似然函数嘛！

再想，在工程上，**第3步**的训练朴素贝叶斯分类器的本质是什么？不就是训练朴素贝叶斯分类器的参数嘛！而朴素贝叶斯分类器的参数是什么？不就是先验概率跟每个类别下的每个特征的每个值的后验概率嘛！而先验概率不用管了，那每个类别下的每个特征的每个值的后验概率合在一起是什么？不就是理论EM算法中的每个随机变量的概率分布模型的参数嘛！恍然大悟啊有没有？！

路人某：┐(ﾟ_ﾟ)┌并没有。

小夕：(ﾉ°Д°)ﾉ ┻ /(.□ . \)

好吧，给你几分钟时间接受一下**训练分类器的理论意义竟然是计算随机变量所服从的概率分布模型的参数**这个事实。

工程EM的第2、3、4步竟然完完全全的卡到了理论EM算法的相应位置。那么理论EM算法还有哪一步没有对应上呢？当然是参数θ的初始化啦~相信机智的你已经想到了，那就是工程EM中的**第1步**所做的事情啦。

细心的你又有没有留意到什么不同之处呢？

藏的再深也要挖出来！

如果能留意到，那就非常厉害了。还记得理论EM中，我们计算似然函数的过程中，是要计算无标签样本的每种标签取值的概率之和的！对，就是下面这货：

$$\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)$$

（我叫图片，不叫公式）

然而，我们在工程上计算似然函数则是**先用分类器预测一个类别**，然后叠加**该类别的**后验概率！

这意味着什么呢？显然意味着忽略了样本为其他类别的概率呀！这样做，肯定导致导致计算出的后验概率没有那么准，但是，却极大的提高了计算效率！

因此，本质上讲，工程上，半监督学习中的EM算法不过是简化了计算、优化了初始化的理论EM模型罢了╮(╯▽╰)╭

建造桥梁好辛苦，坐等小红包买瓶水\(//▽//)\