Welcome

MSRA Researcher新手，现阶段从事多模态大模型预训练方向。对NLP感兴趣或对博客内容有疑惑及意见建议的，欢迎评论或添加我微信。此外如果有需要内推的同学，也欢迎来骚扰我。联系方式详见concat页面。

前往GitHub (https://github.com/Dod-o)

机器学习面试之最大熵模型

11月 22, 2018 (http://www.pkudodo.com/2018/11/)

机器学习面试之最大熵模型

作者 Dodo (http://www.pkudodo.com/author/root/) 在 (http://www.pkudodo.com/2018/11/22/1/)
机器学习 (http://www.pkudodo.com/category/%e6%9c%ba%e5%99%a8%e5%ad%a6%e4%b9%a0/) 标签
最大熵 (http://www.pkudodo.com/tag/%e6%9c%80%e5%a4%a7%e7%86%b5/),
逻辑回归 (http://www.pkudodo.com/tag/%e9%80%bb%e8%be%91%e5%9b%9e%e5%bd%92/)

在学习最大熵过程中看到的一篇非常优秀的解说，来来回回读了好几遍以后实在忍不住给转载过来了。

原文链接：https://www.jianshu.com/p/e7c13002440d (https://www.jianshu.com/p/e7c13002440d) 作者：milter

> 最大熵模型属于运用最大熵原理的多分类模型，这个模型在面试中经常会与逻辑回归一起问，比如，为什么说二者是类似的？要解答这个问题，需要对两个模型的原理都有清晰的理解，很多面试者虽然能从书上背来一两句结论，比如二者都是求的最大似然概率，但是只要深入问下去，都会面露囧色。本文试图尽可能用清晰简洁的语言说明白最大熵模型的原理，以及它与最大似然的关系。

1、分清最大熵思想与最大熵模型

我们平常说的最大熵模型，只是运用最大熵思想的多分类模型，最大熵的思想却是一种通用的思维方法。所以，理解最大熵模型只需要搞清楚两件事就可以：

- 最大熵思想是什么
- 最大熵模型是如何运用最大熵思想的

2、最大熵思想

我们知道，分类模型有判别模型和生成模型两种，判别模型是要学习一个条件概率分布 $P(y|x)$。
举例说明，x是病人身体指标，体温、血压、血糖，y是各种可能的疾病，可简化为小病、中病、大病三种。

现在，我们有一个样本x1={体温：30，血压：160，血糖：60}，那么P(y|x1)就是一个概率分布，该分布的值就是上面简化的三种，小病、中病、大病。可能的概率分布如下所示：

小病	中病	大病
1/2	1/4	1/4
1/4	1/3	5/12
1/3	1/3	1/3

当然，这样的分布有无数种，上面只是举例说明而已。那么，问题来了，在这无数种概率分布中，哪一个才是好的呢?
为了选出一个好的分布，可以做如下两步：

- 1、看看以往的病例中，指标x1={体温：30，血压：160，血糖：60}和三种病之间的关系，如果没有这样的病例，也就是说我们没有过往的经验可以参考，那么，就直接选一个熵最大的分布就是，也就是上面表格中的第三个分布，因为均匀分布总是同类分布中熵最大的分布。
- 2、如果查看以往病例后，我们得到一个经验，指标x1={体温：30，血压：160，血糖：60}有1/2的概率是小病，于是我们有了一定的经验知识，此时，最好的分布就是**符合这个经验知识的前提下，熵最大**的分布，显然，第一个分布就是最好的分布。

以上，我们就是运用了最大熵的思想。总结来说，最大熵的思想是，当你要猜一个概率分布时，如果你对这个分布一无所知，那就猜熵最大的均匀分布，如果你对这个分布知道一些情况，那么，就猜满足这些情况的熵最大的分布。

3、运用最大熵思想来做多分类问题

现在，我们来看最大熵模型是如何运用最大熵思想的。
还是上面的例子，假设我们不只有一个x1样本，而是有x1、x2、...、xN个样本。并且知道每一个样本所得的病y1、y2、...、yN，yi是小病、中病、大病三者之一。这个时候，我们要怎么运用最大熵思想呢?

首先，我们要认真考虑一下这个例子和第2部分中的例子的不同之处，在第2部分的例子中，我们只有一个样本x1，并且假设我们有关于x1的先验知识，那就是1/2的概率是小病，要求的概率分布只有一个，那就是P(y|x1)。
现在，我们有N个样本和它们的标签。这些**标签就是我们现在的先验知识**，即，对于xi，我们知道它的标签是yi，这个先验知识与第2部分例子中的已知的1/2概率不再是同一种形式了。

其次，此时我们要求的模型P(y|x)已经不是一个概率分布，而是**无数个概率分布**，因为，每一个x都会对应一个P(y|x)。但是，这无数个分布可以用一个关于x的函数来表示，即 P(y|x) ~x。这样，我们只要求出这个函数的形式和它的参数值，就算求出了模型P(y|x)。
在后面的叙述中，P(y|x)有时代表某个x下y的条件概率分布，有时也指无数个分布的集合，即关于x的函数。请注意辨别。
请思考一下，在这种情况下，如何贯彻最大熵思想来求解条件概率 P(y|x)?

首先，我们回顾一下最大熵思想：

> 当你要猜一个概率分布时，如果你对这个分布一无所知，那就猜熵最大的均匀分布，如果你对这个分布知道一些情况，那么，就猜满足这些情况的熵最大的分布。

其实，我们只要两步就可以贯彻最大熵思想：

1、找出满足现有情况的分布P(y|x)。

虽然我们现在对P(y|x)的形式和参数还是一无所知，但这并不妨碍我们从概率分布的层面上去考察它的一些特点。也就是P(y|x)要满足的一些约束，这些约束，就是对我们已知的先验知识的拟合。我们的先验知识就是N个训练样本（xi，yi）。假设我们通过观察这N个样本，发现了一个事实：

当体温小于38，血压小于100，血糖小于30时，总是得小病。这就是一个综合后的先验知识。我们可以据此定义一个特征函数：

f(x,y) = 1 当且仅当 x ={体温小于38，血压小于100，血糖小于30}，y=小病

将f(x,y)运用到任一个样本（xi，yi）上，我们就可以知道该样本是不是满足上述事实。你可以认为，f(x,y)是对样本是否符合某个事实的判定函数。

也许你还是会对这个特征函数感到迷惑，请暂时放下迷惑，只要相信，这一切都是为了让我们能更加形式化地定义：什么样的P(y|x)是满足现有情况的。
根据已有的N个样本，我们可以算出P(x,y)的经验分布P~(x,y)和P(x)的经验分布P~(x)。
然后，我们就可以统计下，在这个经验分布中，f(x,y)的期望是多少，如下所示


这个期望表示什么意思呢？它表示的是，就我们的经验分布来看，满足 x ={体温小于38，血压小于100，血糖小于30}，y=小病这一事实的样本占总体样本的比率。比如说是1/3，表示从我们的经验分布看，一个样本有1/3的概率是符合这个事实的。
那么，我们求出的P(y|x)也要符合这个期望值才能算是满足现有情况。至此，我们终于找到一个衡量P(y|x)是否满足现有情况的指标。但是，还有最后一个问题，我们的P(y|x)是条件分布，衡量分布是否满足现有情况时，需要联合分布。

这个问题，很好解决，我们有了x的经验分布P~(x)，将这个经验分布乘以P(y|x)就可以近似表示我们的P(y|x)背后的联合分布，据此，我们可以写出P(y|x)要满足的一个约束：


我们求出的P(y|x)满足这个约束条件，就相当于满足了现有的 x ={体温小于38，血压小于100，血糖小于30}，y=小病这一事实。这个事实来自于我们对N个样本的观察总结。

当然，观察N个样本，我们还可以得出其他事实，每一个事实都可以按照上述步骤，为P(y|x)施加一个“紧箍咒”。这个事实总结的越准确，我们就越能窥见要求的P(y|x)的模样。

2、使得P(y|x)的熵最大化

假设我们现在已经找出了所有满足上面约束条件的P(y|x)，现在，我们要运用最大熵思想来从中找出熵最大的P(y|x)。

这里运用最大熵思想时，我们要将P(y|x)看做是无数个概率分布的集合，即**每一个x，都对应一个特定的概率分布P(y|x)，每一个概率分布都会有一个熵，此时，所谓的最大熵，就是最大化这些所有的概率分布的熵的和，由于每个x都有一个经验概率P~(x)，我们还需要对所有这些熵进行加权求和，以此表示哪一个概率分布的熵的最大化更加重要。**如下所示：

4、求解最大熵模型

对前面的内容进行整理，我们得出如下的优化数学问题：

注意：这里的 i=1,...,n表示我们对样本观察得出的n个事实，可不是N个样本哦。同时，这里通过加负号的办法，将最大化问题转化成了最小化问题。

为啥我们喜欢最小化而不喜欢最大化呢，因为最小化相当于就坡往下滚，省劲，爽，最大化相当于上坡，费劲，累。哈哈，如果你当真我就服了你！

到这一步，我们稍微回顾一下。

- 我们到现在都不知道P(y|x)的函数形式是什么，参数有多少，我们仅仅是从概率分布的抽象层面上进行讨论，确定它要满足的一些约束
- 每一个约束都来自于我们凭借已有知识，对N个样本进行观察总结得出的一个事实。
- 按照最大熵思想求P(y|x)时，我们是对所有可能的概率分布的熵进行了加权求和。然后最大化这个和，而不是某个单一的概率分布的最大熵。

那么，剩下的工作就应该由数学家帮我们搞定了，因为我们已经将问题完全形式化为一个约束最优化问题了！

由于下面涉及很多具体的数学，在叙述时，我尽可能不涉及具体的数学，而只从整体的思路上说，以防止具体的数学干扰我们对模型的理解。
具体的求解方法和svm的求解是一致的，利用拉格朗日函数转为求解对偶问题。对偶问题如下所示：

求解分为两步：
第一步是求对偶问题里面的最小化问题，该问题求解完成后，我们可以看到P(y|x)的形式，如下所示：

形式虽然有了，但是里面的参数w还没有具体确定，第二步的最大化就是来确定参数w的。

第二步，最大化。将P_w(y|x)带入L(P,w)，最大化该函数的值，也就是求对偶问题外层的最大化问题，从而求出具体的w。

至此，最大熵模型解答完毕！

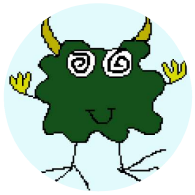
5、 说好的与逻辑回归的相似处呢?

Note：由于下面的讨论比较笼统，建议结合李航书观看，风味更佳。

到这里，我们都没有涉及到最大熵与逻辑回归的相似的讨论。因为这个问题也会涉及到很多的数学问题。让我们从整体上来看一下。

在第4部分中，我们首先求解对偶问题的最小化，得出了P(y|x)的形式P_w(y|x)，然后我们要求最大化，当我们将P_w(y|x)代入上面的L(P,w)后，经过一系列的变形推导，我们惊奇的发现我们求最大化时，实际上与我们直接求解P_w(y|x)关于样本数据的对数似然最大化是一样一样的。

至此，我们发现，**我们从最大熵的思想出发得出的最大熵模型，最后的最大化求解就是在求P(y|x)的对数似然最大化。逻辑回归也是在求条件概率分布关于样本数据的对数似然最大化。二者唯一的不同就是条件概率分布的表示形式不同。**



Dodo

3条评论

匿名 **发布于7:04 上午 - 1月 28, 2023**

To the pkudodo.com admin, Your posts are always a great read.

()

回复

匿名 **发布于10:10 上午 - 11月 23, 2021**

e46xti

()

回复

匿名 **发布于7:28 下午 - 12月 20, 2019**

大佬，请问下你的这个博客使用什么搭建的？感觉好清爽的

()

回复

发表评论

内容(链接请去除http协议头，否则会被误判为垃圾评论)

设置显示昵称(选填)

发送信息

近期文章

- 论文|记忆网络之Memory Networks (http://www.pkudodo.com/2019/06/14/1-13/)
- 梳理|对话系统中的DST (http://www.pkudodo.com/2019/06/09/1-12/)
- 论文阅读|How Does Batch Normalizetion Help Optimization (http://www.pkudodo.com/2019/05/23/1-11/)

- 学习规划|机器学习和NLP入门规划 (http://www.pkudodo.com/2019/03/20/1-10/)
- 面试体会|微软、头条、滴滴、爱奇艺NLP面试感想 (http://www.pkudodo.com/2019/03/10/1-9/)

文章归档

- 2019年6月 (http://www.pkudodo.com/2019/06/)
- 2019年5月 (http://www.pkudodo.com/2019/05/)
- 2019年3月 (http://www.pkudodo.com/2019/03/)
- 2018年12月 (http://www.pkudodo.com/2018/12/)
- 2018年11月 (http://www.pkudodo.com/2018/11/)
- 2018年10月 (http://www.pkudodo.com/2018/10/)
- 2016年9月 (http://www.pkudodo.com/2016/09/)

标签

DST (http://www.pkudodo.com/tag/dst/)

K近邻 (http://www.pkudodo.com/tag/k%e8%bf%91%e9%82%bb/)

LSI (http://www.pkudodo.com/tag/lsi/)

memory network (http://www.pkudodo.com/tag/memory-network/)

MQTT (http://www.pkudodo.com/tag/mqtt/) NLP (http://www.pkudodo.com/tag/nlp/)

s3c2440 (http://www.pkudodo.com/tag/s3c2440/) SLU (http://www.pkudodo.com/tag/slu/)

SVM (http://www.pkudodo.com/tag/svm/) VSM (http://www.pkudodo.com/tag/vsm/)

体会 (http://www.pkudodo.com/tag/%e4%bd%93%e4%bc%9a/)

决策树 (http://www.pkudodo.com/tag/%e5%86%b3%e7%ad%96%e6%a0%91/)

分类器 (http://www.pkudodo.com/tag/%e5%88%86%e7%b1%bb%e5%99%a8/)

实现 (http://www.pkudodo.com/tag/%e5%ae%9e%e7%8e%b0/)

对话系统 (http://www.pkudodo.com/tag/%e5%af%b9%e8%af%9d%e7%b3%bb%e7%bb%9f/)

感想 (http://www.pkudodo.com/tag/%e6%84%9f%e6%83%b3/)

感知机 (http://www.pkudodo.com/tag/%e6%84%9f%e7%9f%a5%e6%9c%ba/)

支持向量机 (http://www.pkudodo.com/tag/%e6%94%af%e6%8c%81%e5%90%91%e9%87%8f%e6%9c%ba/)

文章相似度 (http://www.pkudodo.com/tag/%e6%96%87%e7%ab%a0%e7%9b%b8%e4%bc%bc%e5%ba%a6/)

最大熵 (http://www.pkudodo.com/tag/%e6%9c%80%e5%a4%a7%e7%86%b5/)

朴素贝叶斯 (http://www.pkudodo.com/tag/%e6%9c%b4%e7%b4%a0%e8%b4%9d%e5%8f%b6%e6%96%af/)

机器学习 (http://www.pkudodo.com/tag/%e6%9c%ba%e5%99%a8%e5%ad%a6%e4%b9%a0/)

爬虫 (http://www.pkudodo.com/tag/%e7%88%ac%e8%99%ab/)

统计学习方法 (http://www.pkudodo.com/tag/%e7%bb%9f%e8%ae%a1%e5%ad%a6%e4%b9%a0%e6%96%b9%e

统计方法学习 (http://www.pkudodo.com/tag/%e7%bb%9f%e8%ae%a1%e6%96%b9%e6%b3%95%e5%ad%a6%e

综述 (http://www.pkudodo.com/tag/%e7%bb%bc%e8%bf%b0/)

网易云歌单 (http://www.pkudodo.com/tag/%e7%bd%91%e6%98%93%e4%ba%91%e6%ad%8c%e5%8d%95/)

规划 (http://www.pkudodo.com/tag/%e8%a7%84%e5%88%92/)

详解 (http://www.pkudodo.com/tag/%e8%af%a6%e8%a7%a3/)

逻辑回归 (http://www.pkudodo.com/tag/%e9%80%bb%e8%be%91%e5%9b%9e%e5%bd%92/)

逻辑斯蒂回归 (http://www.pkudodo.com/tag/%e9%80%bb%e8%be%91%e6%96%af%e8%92%82%e5%9b%9e%e

面试 (http://www.pkudodo.com/tag/%e9%9d%a2%e8%af%95/)