最大熵模型

# 前言

没有前言

# 涉及概念

样本空间，随机变量，概率质量函数，随机变量的期望，联合概率，条件概率，最大熵模型，信息量，熵，条件熵，拉格朗日乘子法，拉格朗日对偶性，极大似然估计法等

# 概念简介

在正式介绍最大熵模型前，先要介绍几个重要的概念和工具，以便后续阅读时方便理解。

## 3.1熵和信息量

### 3.1.1信息量

设随机变量X为投掷一枚骰子的点数，则其样本空间为{1，2，3，4，5，6}，概率质量函数为P{X=i}=1/6，i属于[1，6]（请默视这个随机变量10秒，本章都会以这个随机变量X为例子）。这个随机变量X的一个事件发生（例如X=2，骰子的点数为2）时所携带的信息大小称作该随机事件的“信息量”，公式如下：



即随机变量的某个事件发生的“信息量”为该事件的概率的倒数再取对数。其中对数可以以2，e，10为底。以2为底时单位为比特（bit）；以e为底时单位为奈特(nat)；以10为底时单位为哈特(hat)。由这个公式得知投掷一枚骰子得到的各个结果的信息量都是log6（bit）。

### 3.1.2熵的定义

前文中提到“信息量”是指随机变量的一个事件e发生时产生的信息大小，而一个随机变量的取值有很多种，那么一个随机变量平均能够产生多少“信息量”呢？这就引入了“熵”的概念。一个随机变量的“熵”就是该随机变量的平均“信息量”：



“熵”的取值范围[0，log|X|]，一个随机变量的“熵”越大，则表示该随机变量的结果越不可预期。如果上述随机变量X在点数为2时的概率是100%。那么这个随机变量的结果就是完全可预期的结果2，经过“熵”公式计算得到这种情况下H(X)=0为最小值。那么什么情况下结果最不可预期呢？那就是我们生活中最普通的骰子了，每种情况的概率都是相等的平均的，这种情况下的熵值为最大值log6。

### 3.1.3条件熵

在已知条件变量Y的情况下，随机变量X的条件熵为：



## 3.2最大熵原理

最大熵模型的思想为：在学习概率模型时，所有可能的概率模型中，“熵”最大的模型就是最好的模型，从信息论的角度这种模型保存了最大的不确定性。这样在选择最好的概率模型时包含了两个标准：

1. 首先必须是可能的模型，即满足所有已知的限制条件；
2. 其次在满足已知的限制条件的所有概率模型中选取“熵”最大的一个。

再以上述随机变量X为例：在不加任何限制的情况下很显然随机变量X最好的概率模型就是所有的概率都是均匀的，概率模型为：



而这正是“熵”最大的情况；现在给随机变量X加上一些条件限制，假设有一个骰子P(2)=0.5，即点数为2的概率为50%；如果还要使用这个骰子，我们希望这个骰子的其他点数的概率为多少呢？我们肯定希望其他几个点数的概率是均匀的（都为10%）。这样至少取其他几种情况时概率是相等的，而这个骰子除去点数2之外其他的点数出现的可能也是不可预期的。在限制条件P(2)=0.5的概率下，“熵”最大的模型为：



以上就是最大熵模型的思想，总结出来就是一句话“在满足已知条件的情况下（点数为2的概率为50%），其它概率分布是均匀的（其它5个点数的概率都为10%）的概率模型为最好的模型。

## 3.3拉格朗日乘子法

拉格朗日乘子法的功能：求目标函数在约束函数的条件下的极值的问题。通过引入拉格朗日乘子，构造拉格朗日函数，将有约束条件的极值问题转化为无约束条件的极值问题。具体原理请查阅相关书籍。

拉格朗日乘子法的用途：用于求解在有限制条件下的“熵”最大的概率模型，前文描述的最大熵原理中可以得出这是一个有约束条件的最优值求解的问题：约束条件为当前已知条件，要求解的最优值为“熵”最大的概率模型。

后续本文会在最大熵模型求最优概率模型时用到拉格朗日乘子法。过程中根据概率模型的熵定义一个目标函数，为每个特征构造一个约束条件，然后引入拉格朗日乘子将目标函数和约束条件构造成拉格朗日函数。通过求偏导的方法解出概率模型关于约束条件和拉格朗日乘子的表达式。

## 3.4极大似然估计法

有限制条件下的最大“熵”问题，通过拉格朗日乘子法转变成了求无约束条件下的极值问题，转变为求对偶函数最大化问题，又通过证明得到对偶函数和对数似然函数是等价的（中间有很多公式推导，本文不做赘述，有兴趣可参阅）。所以问题可以转化为以似然函数为目标函数的最优化问题，进而通过迭代算法求解。

# 最大熵模型

本节将从4个方面介绍最大熵模型，分别是：“模型”“特征和约束”“概率公式”“迭代算法”。在正式介绍之前，我们先总结前面简单介绍了什么：

1. 最大熵原理：将最大熵模型总结为有约束条件下的“熵”的极值问题。
2. 拉格朗日乘子法：解上面的问题，给出概率模型关于拉格朗日乘子和约束条件的表达式。
3. 极大似然估计法：通过证明最大熵模型的对偶函数和对数似然函数等价，将问题转变为以似然函数为目标函数的最优化问题，迭代求出引入的拉格朗日橙子。

## 4.1模型

给定一个训练样例集合S={<x1,y1>,<x2,y2>,…,<xn,yn>}，每个训练样例是一对x，y。可以根据该样例集合S计算出联合分布P(XY)的经验分布和边缘分布P(X)的经验分布，用P~(XY)和P~(X)表示：





最大熵模型是一个条件概率模型P(Y|X)。这个条件概率模型p(Y|X)满足样例集合中所有的约束条件以及条件熵最大。条件熵公式如公式（3）。这里可以给出等价的有约束条件的最优解问题为：



公式中以边缘分布的经验分布P~(x)，近似等于边缘分布P(x)。

## 4.2特征函数和约束函数

上一小节给出了有约束条件的最优解问题，但并没有给出约束条件是什么。本小节将给出特征函数和约束函数的定义。

### 4.2.1特征函数

最大熵模型引入特征函数来描述输入x和输出y的某一个事实，其定义如下：



### 4.2.2约束函数

约束函数由特征函数引入，特征函数f(x,y)关于联合分布关于经验分布P~(x,y)的期望，等于特征函数关于模型P(Y|X)和P~(X)的期望：



其中：





上一小节的最优解问题转化为：



可以看出约束条件有两个，第一个是所有的特征函数关于先验概率的期望和关于概率模型的期望相等；第二个是条件概率的归一化，对于随机变量X的每个情况x发生的情况下，条件变量Y的所有可能情况y发生的条件概率之和为1。

## 4.3拉格朗日乘子法

求解概率模型的计算公式，这里将用到前文讲到的拉格朗日乘子法，先将上一小节中总结的最优解求最大值问题，转化为等价的求最小值问题：



注意第一个式子中将log1/P(y|x)变成了-logP(y|x)，并将负号提到了公式最前面，这样求最大值问题就变成了等价的求最小值问题。

### 4.3.1拉格朗日函数

引入拉格朗日乘子w0,w1,…,wn，并定义拉格朗日函数L(P,w)如下：



式（14）的极小值问题等价于式子（15）的极小值问题，因为在满足约束条件的情况下，式（15）等价于式（14）的第一个公式。因此上述有约束条件下的最优解问题就转化为无约束条件下的最优解问题：



### 4.3.2极小极大问题

不幸的是L(P,w)可能没有最小值，因为当不满足前述约束条件时，在Ep(fi)-Ep~(fi)小于0的情况下，只要另wi去无穷大，那么L(P,w)的值就趋向于负无穷大。因此应该限制wi的取值，在wi取值使L(P,w)最大化时，求P是这种情况下L(P,w)最小值的情况。由于在满足式（14）条件约束的情况下，wi取任何值L(P,w)都等价于条件“熵”公式，所以最优解问题又等价于：



这样就将原始的最优解问题等价于拉格朗日函数的极小极大问题。

### 4.3.3极大极小问题

由于拉格朗日函数是凸函数，所以其极小极大问题的解于其对偶问题得解是等价的，拉格朗日极大极小问题为：



这样就将拉格朗日函数的极小极大问题转化为其对偶问题极大极小问题，求解这个极大极小问题得到的结果，就是原始问题的结果。该问题的最优解要先求内部最小问题minL(P,w)，再求解外部最大问题。

#### 4.3.3.1内部极小化

将w看作参数，则L(P,w)是关于P的函数，P取何值时L(P,w)的值最小呢？



求解方法很简单：求L(P,w)关于P的偏导数，并令该偏导数为0，可求出P关于w的表达式，该表达式就是关于P的函数L(P,w)取极小值时的解（偏导数的求法这里不做赘述，如有感兴趣的可以自己推导）。P关于w的表达式为：



其中Z(x)为归一化因子，它确保了Y的所有可能情况在X=x时的条件变量之和为1。

#### 4.3.3.2外部极大化

将上面得到的公式（20）带入到公式（15）中得：



得到的公式将拉格朗日函数转变为关于w的函数。外部极大化然后只需要求出相应的w，使得拉格朗日函数最大。再将这个求得的w代入回式（20），得到的条件模型P(Y|X)就是最终要求的最大熵模型。

## 4.3 极大似然估计

已知训练数据的经验分布P~(XY)，条件概率公式P(Y|X)的对数似然函数为：



当条件概率满足最大熵原理时，将公式（20）带入到公式（22）得：



我们可以看出公式23和公式21是一样的，说明根据经验分布P~(XY)的对数似然函数的极大化等价于上述拉格朗日函数的外部极大化，因此上述拉格朗日函数外部极大化问题可以通过极大似然估计法求得。

迭代算法：

输入：特征函数f1,…,fn；初始参数w1,…,wn；经验分布P~(XY)；

输出：最优条件概率分布P(Y|X)

算法步骤：

1. 根据当前参数值和特征函数计算当前条件概率模型P(Y|X)
2. 迭代每一个特征fi，分别计算其在经验概率和当前条件概率模型下的期望Ep~和Ep
3. 根据特征fi的期望Ep~和Ep，更新参数wi



其中M为特征fi在所有x，y中出现的次数

1. 如果不是所有的wi都收敛，跳转到步骤1

最终可求得最优的wi，并将wi带入到公式（20），可以得到条件概率的最大熵模型P(Y|X)

# 总结

本文讲述了最大熵原理和最大熵模型，以及介绍了最大熵模型解的推导。

1. 由最大熵原理得出最大熵模型是一个有限制条件的最优值求解问题；
2. 然后通过引入拉格朗日乘子，将原始问题转化为拉格朗日函数的极小极大问题；
3. 通过其对偶性转变为求拉格朗日函数的极大极小问题；
4. 通过解内部极小问题求出概率模型关于拉格朗日乘子的表达式，代入后转变成求关于拉格朗日乘子的函数极大化问题；
5. 通过证明外部极大化问题等价于对数似然函数极大化问题，进而通过极大似然估计算法来求使拉格朗日函数极大化的乘子。
6. 最终迭代算法得到乘子带入到概率模型公式可得到最终的概率模型。