# 从第一原则推导出Softmax

**——Softmax与sigmoid函数：Softmax原理推导**

## 第一章 引言

这篇文章的最初目标是探讨softmax和sigmoid函数之间的关系。事实上，这种关系似乎一直不能达到：“一个在分子中有一个指数！一个有一个总和！一个在分母中有一个1！”当然，这两者有着不同的名字。

一旦衍生出来，受到条件概率公理本身启发，我很快就意识到，如何将这种关系推回到更一般的模型框架中。因此，这篇文章首先探讨了sigmoid是什么,它仅仅是 softmax 的一个特殊情况，以及在 Gibbs 分布，因数乘积和概率图模型中的每一个基础。接下来，我们继续展示这个框架如何进行自然扩展，去定义典型的模型类，如softmax回归，条件随机场，朴素贝叶斯和隐式马尔可夫模型。

## 第二章 目标

这是一个预测模型。它是一个接收输入并产生输出的菱形。

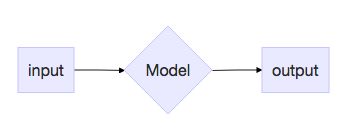


图1：预测模型

输入是一个向量 . 有3个可能的输出：.我们模型的目标是在输入条件下预测产生每个输出的概率，即：

当然，概率只是一个实数，位于封闭的区间 .

## 第三章 综述

### 3.1 输入如何影响输出？

我们的输入是4个数字的列表；每一个在不同程度上影响每一个可能的输出。我们称这种影响为“权重”。4个输入乘以3个输入等于12个不同的权重。他们可能看起来像这样：

表1：权重表

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | .1 | .4 | .3 |
|  | .2 | .3 | .4 |
|  | .3 | .2 | .1 |
|  | .4 | .1 | .2 |

### 3.2 产出一个输出

给定一个输入向量 . 我们的模型将使用上述权重为每个输出 产生一个数字。每个输入元素的效果将会被相加。之后再说明其原因。

这些总和将决定我们的模型产生什么样的结果。最大的数字获得胜利。例如，给定：

我们的模型将有最好的机会生产c。

### 3.3 转化为概率

我们之前说过，我们的目标是获得以下结果：

用粗体表示来任何输入值，鉴于我们现在具有特定的输入值，即 ，我们可以更准确地说明我们的目标：

到目前为止，我们只有 . 要将每个值转换成概率，即,在 上的一个非特殊的数字，仅仅除以总和即可。

最后，作为有效的概率分布，所有数字必须总和为1。

### 3.4 如果我们得到的值为负呢？

如果我们的初始非归一化概率之一是负的，即 .所有这些都不成立了。

不是一个有效的概率，因为它不会落入区间 .

为了确保所有非规范化概率都是正数，我们必须首先将它们传递给一个函数作为输入，这个函数将实数作为输入，并产生严格的正实数作为输出。这只是一个指数;现在我们选择欧拉数（）。 这个选择的理由将在后面解释（尽管注意到任何正指数将会满足我们所说的目的）。

我们的归一化概率，即有效概率现在看起来如下：

统一表示为：

这便是 softmax 函数。

### 3.5 与 sigmoid 函数的关系

softmax 在 个不同的输出上输出有效的概率分布，sigmoid 只是在 在情况下。因此，sigmoid 仅仅是 softmax 的一个特殊情况。通过这个定义，假设我们的模型只产生两个可能的输出 和 ，在给定输入 ，我们可以写出 sigmoid 如下所示：

但是，请注意，我们只需要计算 的概率，至于 在这个注释里，我们来重新展开表达式 ：

然后，将分子和分母除以 ：

最后，我们可以将其代回原来的补充：

我们的方程是不确定的，因为有比方程式（1个）更多的未知数（2个）。因此，我们的方程将有无数的解 。因此，我们可以直接固定这些值之一，设定

这便是 sigmoid 函数。最后：

### 3.6为什么是非标准化的概率的总和？

我们都认为规范线性组合 的语义是理所当然的，但是我们为什么首先进行求和呢？

为了回答这个问题，我们首先重申我们的目标：