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。

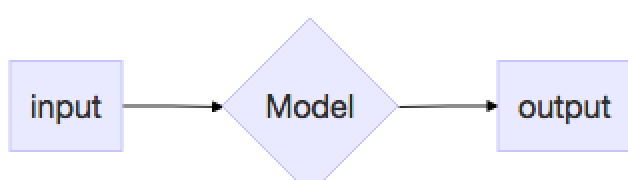


图 1：预测模型

输入是一个向量 $X = [x_0, x_1, x_2, x_3]$. 有 3 个可能的输出：a, b, c. 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：

$$P(a|\mathbf{X}), P(b|\mathbf{X}), P(c|\mathbf{X})$$

当然，概率只是一个实数，位于封闭的区间 [0,1].

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

	$\boldsymbol{a}$	$\boldsymbol{b}$	$\boldsymbol{c}$
$\boldsymbol{x_0}$	.1	.4	.3
$\boldsymbol{x_1}$	.2	.3	.4
$\boldsymbol{x_2}$	.3	.2	.1
$\boldsymbol{x_3}$	.4	.1	.2

3.2 产出一个输出

给定一个输入向量 $x = [x_0, x_1, x_2, x_3]$. 我们的模型将使用上述权重为每个输出 a, b, c 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。

$$\tilde{a} = \sum_{i=0}^{3} \omega_{i,a} x_i$$

$$\tilde{b} = \sum_{i=0}^{3} \omega_{i,b} x_i$$

$$\tilde{c} = \sum_{i=0}^{3} \omega_{i,c} x_i$$

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：

$$\{\tilde{a}: 5, \tilde{b}: 7, \tilde{c}: 9\}$$

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：

$$P(a|\mathbf{X}), P(b|\mathbf{X}), P(c|\mathbf{X})$$

X 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 x，我们可以更准确地说明我们的目标：

$$P(a|x), P(b|x), P(c|x)$$

到目前为止，我们只有 $\{\tilde{a}: 5, \tilde{b}: 7, \tilde{c}: 9\}$. 要将每个值转换成概率，即, 在 [0,1] 上的一个非特殊的数字，仅仅除以总和即可。

$$P(a|x) = \frac{5}{5+7+9} = \frac{5}{21}$$

$$P(b|x) = \frac{7}{5+7+9} = \frac{7}{21}$$

$$P(c|x) = \frac{9}{5+7+9} = \frac{9}{21}$$

最后，作为有效的概率分布，所有数字必须总和为 1。

$$\frac{5}{21} + \frac{7}{21} + \frac{9}{21} = 1 \checkmark$$

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 $\{\tilde{a}: -5, \tilde{b}: 7, \tilde{c}: 9\}$. 所有这些都不成立了。

$$P(a|x) = \frac{-5}{-5+7+9} = \frac{-5}{11}$$

$$P(b|x) = \frac{7}{-5+7+9} = \frac{7}{11}$$

$$P(c|x) = \frac{9}{-5+7+9} = \frac{9}{11}$$

$\frac{-5}{11}$ 不是一个有效的概率，因为它不会落入区间 [0,1].

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（e）。 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）。

$$\tilde{a} = -5 \rightarrow e^{-5}$$
$$\tilde{b} = 7 \rightarrow e^{7}$$
$$\tilde{c} = 9 \rightarrow e^{9}$$

我们的归一化概率，即有效概率现在看起来如下：

$$P(a|x) = \frac{e^{-5}}{e^{-5}+e^{7}+e^{9}}$$

$$P(b|x) = \frac{e^{7}}{e^{-5}+e^{7}+e^{9}}$$

$$P(c|x) = \frac{e^{9}}{e^{-5}+e^{7}+e^{9}}$$

统一表示为：

$$P(y|x) = \frac{e^{\bar{y}}}{\sum_y e^{\bar{y}}} \ \ for\ y = a, b, c$$

这便是 softmax 函数。

3.5 与 sigmoid 函数的关系

softmax 在 $n > 2$ 个不同的输出上输出有效的概率分布，sigmoid 只是在 $n = 2$ 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 p 和 q，在给定输入 x ，我们可以写出 sigmoid 如下所示：

$$P(y|\mathbf{x}) = \frac{e^{\bar{y}}}{\sum_y e^{\bar{y}}} \ \ for\ y = p, q$$

但是，请注意，我们只需要计算 p 的概率，至于 $P(y = q|\mathbf{x}) = 1 - P(y = p|\mathbf{x})$ 在这个注释里，我们来重新展开表达式 $P(y = p|\mathbf{x})$：

$$P(y = p|\mathbf{x}) = \frac{e^{\bar{p}}}{e^{\bar{p}} + e^{\bar{q}}}$$

然后，将分子和分母除以 $e^{\bar{p}}$：

$$\begin{aligned} P(y = p|\mathbf{x}) &= \frac{e^{\bar{p}}}{e^{\bar{p}} + e^{\bar{q}}} \\ &= \frac{\frac{e^{\bar{p}}}{e^{\bar{p}}}}{\frac{e^{\bar{p}}}{e^{\bar{p}}} + \frac{e^{\bar{q}}}{e^{\bar{p}}}} \\ &= \frac{1}{1 + e^{\bar{q} - \bar{p}}} \end{aligned}$$

最后，我们可以将其代回原来的补充：

$$\frac{1}{1 + e^{\bar{q} - \bar{p}}} = 1 - \frac{1}{e^{\bar{p} - \bar{q}}}$$

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 $(\tilde{p}, \tilde{q})$。因此，我们可以直接固定这些值之一，设定 $\tilde{q} = 0.$

$$P(y = p|\mathbf{x}) = \frac{1}{1 + e^{-\tilde{p}}}$$

这便是 sigmoid 函数。最后：

$$P(y = q|\mathbf{x}) = 1 - P(y = p|\mathbf{x})$$

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 $\sum_i w_i x_i$ 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：在输入条件下预测产生每个输出的概率,即 $P(Y = y|\mathbf{x})$ ，接下来，我们将重新讨论条件概率本身的定义：

$$P(B|A) = \frac{P(A, B)}{P(A)}$$

就个人而言，我觉得这有点难以解释。 让我们重新变形，以获得更直观的东西。

$$P(A, B) = P(A)P(B|A)$$

这可以解释为：

同时观察（给定值）A 和 B 的概率，即，A 和 B 的联合概率等于观察 A 发生的概率乘以在 A 发生的情况下观察 B 发生的概率。

例如，假设女孩出生的概率是 0.55，并且一个女孩欢数学的可能性是 0.88。因此，

$$P(sex = girl, likes = math) = 0.55 * 0.88 = 0.484$$

现在，我们按照上面的定义重写我们的原始模型输出。

$$P(y|\mathbf{x}) = \frac{P(y,\mathbf{x})}{P(\mathbf{x})} = \frac{e^{\tilde{y}}}{\sum_{y} e^{\tilde{y}}} = \frac{e^{\left(\sum_i w_i x_i\right)_{\tilde{y}}}}{\sum_{y} e^{\left(\sum_i w_i x_i\right)_{\tilde{y}}}} \quad (1)$$

记住，我们指数化每一个非规范化概率 $\tilde{y}$，以将其转换为严格正数。从技术上讲，这个数字应该被称为 $\tilde{P}(y,\mathbf{x})$, 由于其值可能 > 1 ，因此不是一个有效的概率，因此，我们需要再为方程式链引入另外一个术语：

$$\frac{P(y,\mathbf{x})}{P(\mathbf{x})} = \frac{\tilde{P}(y,\mathbf{x})}{normalizer}$$

这是一个类似 $\frac{0.2}{1} = \frac{3}{15}$ 的算术等式。

在等式左边：

- 分子是有效的联合概率分布。
- 分母为“观察 $\mathbf{x}$ 的任何值的概率”是 1。

在等式右边：

- 分子是严格的正的非规范化概率分布。
- 分母是一个常数，可以确保

$$\frac{\tilde{P}(a,\mathbf{x})}{normalizer} + \frac{\tilde{P}(b,\mathbf{x})}{normalizer} + \frac{\tilde{P}(c,\mathbf{x})}{normalizer}$$

其和为 1. 实际上，这个“规范化器”被称为分区函数；我们将在下面回到这里。

考虑到这一点，让我们进一步分解我们的 softmax 方程的分子。

$$\begin{aligned} e^{\tilde{y}} &= e^{\left(\sum_i w_i x_i\right)} \\ &= e^{(w_0 x_0 + w_1 x_1 + w_2 x_2 + w_3 x_3)} \\ &= e^{(w_0 x_0)} e^{(w_1 x_1)} e^{(w_2 x_2)} e^{(w_3 x_3)} \\ &= \tilde{P}(a,\mathbf{x}) \quad (2) \end{aligned}$$

引理：鉴于我们的输出函数[1] 执行取幂，以便在可能的模型输出上获得有效的条件概率分布，因此我们对该函数[2]的输入应该是加权模型输入元素[3]的总和。

1. softmax 函数
2. $\tilde{a},\tilde{b},\tilde{c}$ 其中之一
3. 模型输入元素为 $[x_0,x_1,x_2,x_3]$。加权模型输入元素是 $w_0x_0,w_1x_1,w_2x_2,w_3x_3$。

不幸的是，如果我们首先能够把握住我们所遇到的 $\tilde{P}(a,\mathbf{x})=\prod_i e^{(w_i x_i)}$ 这样事实就好了。下面介绍 Gibbs 分布。

3.7 Gibbs 分布

吉布斯分布给出了一组结果的非规范联合概率分布，类似于在上面计算的 $e^{\bar{a}},e^{\bar{b}},e^{\bar{c}}$,表示为：

$$\tilde{P}_{\Phi}(\mathrm{X}_1,\ldots,\mathrm{X}_n)=\prod_{i=1}^{k}\Phi_i(\mathrm{D}_i)$$

$$\Phi=\{\Phi_1(\mathrm{D}_1),\ldots,\Phi_k(k)\}$$

其中 Φ 定义了一组因子。

3.7.1 因子

一个因子是一个函数：

- 将一组随机变量列表作为输入。 该列表被称为该因子的范围。
- 返回每个 随机变量可以采用的值的 唯一组合 的值，即对于其范围内，交叉积空间中的每个条目。

例如，一个范围为 $\{A,\mathrm{B}\}$ 的因子可能看起来像：

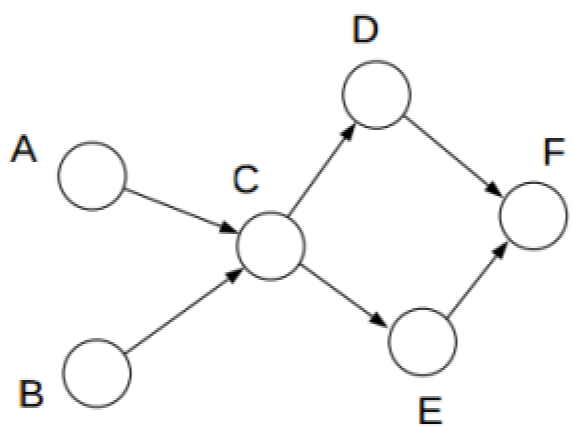
A	B	**Φ**
$\boldsymbol{a^0}$	$\boldsymbol{b^0}$	20
$\boldsymbol{a^0}$	$\boldsymbol{b^1}$	25
$\boldsymbol{a^1}$	$\boldsymbol{b^0}$	15
$\boldsymbol{a^1}$	$\boldsymbol{b^1}$	4

3.7.2 概率图模型

推测复杂系统的行为（通常）会计算其可能结果的联合概率分布。 例如，假设我们有一个商业问题，其中：

- 星期几（A）和营销渠道（B）影响客户注册的概率（C）。
- 顾客注册会影响我们的年度经常性收入（D）和年终雇佣预测（E）。
- 我们的 ARR 和雇佣预测会影响我们将为节日派对（F）订购多少蛋糕。

我们可以这样绘制我们的系统：



我们的目标是计算 $P(A,B,C,D,E,F)$ 。在大多数情况下，我们只会拥有我们系统小的子集数据；例如，我们曾经运行过一个受控实验来调查 A,B,C 其间的关系，或者问卷调查，他们喜欢在圣诞节吃多少蛋糕。如果完全不合理，获得一个相当复杂系统的联合概率分布，是非常少见的。

为了计算这个分布，我们把它分成几部分。每个部分都是描述系统某些子集细节行为的一个因子。(例如，一个因子可能会给出你在给定的一天观察到的次数（$A > 3pm, B = Facebook, C > 50signup$)）。为此，我们说：

如果存在一组因子 Φ，则在图 G 上所需的非归一化概率分布 $\tilde{P}$ 因子分解为：

$$\tilde{P} = \tilde{P}_{\Phi} = \prod_{i=1}^{k} \phi_i(D_i)$$

其中 $\Phi = \{\phi_i(D_1), \dots, \phi_k(D_k)\}$ G 是 Φ 的推理图。

这个引理的前半部分只不过是重新定义非规范化的吉布斯分布。 下半部分进行扩展，我们注意到：

由一系列因子得到的推理图是一个相当完美的图，其中我们在因子域超集中的每个变量周围绘制一个圆圈，并在给定因子域中同时出现的变量之间绘制线条。

$\phi(\mathrm{A, B})$ ，$\phi(\mathrm{B, C})$ 这两个因子的“因子域超集”为 $\{\mathrm{A, B, C}\}$ 。推理图将有用线连接的三个圆圈， A 连接 B ，B 连接 C 。

最后，它接着是：

1. 给定带有变量的商业问题 A, B, C, D, E, F --我们可以画该问题的一幅图。

2. 我们可以建立描述这个问题子集行为的因子。 实际上，这些只是小的子集。

3. 如果由我们的因子推理的图形看起来像我们所绘制的图形，我们可以将我们的系统表示为一个因子积。

不幸的是，在我们最初的模型中所产生的因子积结果 P 仍然是非规范化的，像 $e^{\bar{a}}, e^{\bar{b}}, e^{\bar{c}}$ 一样。

3.7.3 配分函数

配分函数是分母，即 “归一化”，是我们的 softmax 函数。它用于将非规范化概率分布转换为归一化（即有效）概率分布。 真正的吉布斯分布如下：

$$\tilde{P}_{\Phi}(\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n}) = \prod_{i=1}^{k} \phi_i(D_i)$$

$$P_{\Phi}(\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n}) = \frac{1}{\mathrm{Z}_{\Phi}} \tilde{P}_{\Phi}(\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n})$$

其中 Z_{Φ} 是配分函数。

要计算此函数，我们只需将非标准化表中的所有值相加。 给定 $\tilde{P}_{\Phi}(\mathrm{X}_1, \dots, \mathrm{X}_\mathrm{n})$ ：

A	B	**Φ**
$\boldsymbol{a^0}$	$\boldsymbol{b^0}$	20
$\boldsymbol{a^0}$	$\boldsymbol{b^1}$	25
$\boldsymbol{a^1}$	$\boldsymbol{b^0}$	15
$\boldsymbol{a^1}$	$\boldsymbol{b^1}$	4

$$\begin{aligned} \mathrm{Z}_{\Phi} &= 20 + 25 + 15 + 4 \\ &= 64 \end{aligned}$$

我们的有效概率分布变为：

A	B	**Φ**
$\boldsymbol{a^0}$	$\boldsymbol{b^0}$	$\frac{20}{64}$
$\boldsymbol{a^0}$	$\boldsymbol{b^1}$	$\frac{25}{64}$
$\boldsymbol{a^1}$	$\boldsymbol{b^0}$	$\frac{15}{64}$
$\boldsymbol{a^1}$	$\boldsymbol{b^1}$	$\frac{4}{64}$

这是 softmax 函数的分母。

3.8 Softmax 回归

再次申明，我们的模型的目标是在输入条件下预测产生每个输出的概率，即：

$$P(a|\mathbf{X}), P(b|\mathbf{X}), P(c|\mathbf{X})$$

在机器学习训练数据中，我们给出了联合概率分布的构建模块。例如，观察共同投入和产出的分类帐。我们推测每个输入元素以不同的程度会影响每个可能的输出，即我们将它乘以一个权重。接下来，我们对每个结果 $w_i x_i$ 进行取幂，即因子, 然后乘以结果（或者，我们可以指数化因子的线性组合，即机器学习中的特征）：这给我们在所有（输入和输出）变量上的非规范化联合概率分布。

我们想要的是在输入条件下，在可能输出上一个有效的概率分布，即 $P\{y|\mathbf{X}\}$ 。而且，我们的输出是一个单一的，在 $\{a, b, c\}$ 中 “1-of-k”变量（与一系列变量相反）。这是 softmax 回归的几乎逐字定义。

Softmax 回归也称为多项回归，或多类逻辑回归。二进制逻辑回归是 softmax 回归的一个特殊情况，与 Sigmoid 是 softmax 的特殊情况相同。

为了计算我们的条件概率分布，我们得到等式（1）：

$$P(y|\mathbf{X}) = \frac{P(y, \mathbf{X})}{P(\mathbf{X})} = \frac{e^{\tilde{y}}}{\sum_y e^{\tilde{y}}} = \frac{e^{\left(\sum_i w_i x_i\right)_{\tilde{y}}}}{\sum_y e^{\left(\sum_i w_i x_i\right)_{\tilde{y}}}} = \frac{\tilde{P}(y, \mathbf{X})}{normalizer}$$

换句话说，在输入条件下，产生每个输出的概率相当于：

1. softmax 函数

2. 由配分函数归一化的输入元素的乘积的一个指数化因子。

3.8.1 我们的配分函数取决于 $\mathbf{X}$

为了计算在条件 $\mathbf{X}$ 下，在 y 上的分布，我们的配分函数变成依赖 $\mathbf{X}$ 。换句话说，对于给定的输入 $x = [x_0, x_1, x_2, x_3]$，我们的模型计算条件概率 $P(y|x)$. 如果对 softmax 函数进行迂腐的重述，虽然这可能看起来是一个微不足道的事情，但是要注意的是，我们的模型是有效地计算一系列条件分布——每个对应独特的输入 x 。

3.9 条件随机场

以这种方式构建我们的模型，使我们自然地扩展到其他类别的问题。 想象一下，我们正在尝试为给定会话中的每个单词分配一个标签，标签可能包括：“neutral”，“offering an olive branch”和“them is fighting words”。现在我们的问题与原来的模型在一个关键的方式有所不同，和另一个可能的关键方式：

1. 我们的结果现在是一系列的标签。不再是“1-of-k”。在会话“hey there jerk shit roar”中可能的标签序列为：“neutral”，“neutral”，“them is fighting words”，“them is fighting words”，“them is fighting words”。

2. 单词之间可能存在关系，这可能会影响最终输出标签序列。例如，对于每一个单词来说，当叙述这个词时，这个人举起他的拳头，是一个在前还是另一个在前？换句话说，我们构建表明我们输入元素之间的空间关系的因子（即特征）。我们这样做是因为我们认为这些关系可能影响最终输出（当我们说我们的模型“假设特征之间的依赖关系”，这就是我们所要表达的）。

条件随机场输出函数是一个像 softmax 一样的函数。换句话说，如果我们为我们的会话分类任务构建一个 softmax 回归，其中：

1. 我们的输出是一系列标签。

2. 我们的特征是一堆（空间启发）的交互的特征，一个

“sklearn.preprocessing.PolynomialFeatures”

我们基本上只是建立一个条件随机场。

当然，在输入条件下建立输出的完整分布的模型，其中我们的输出再次是一系列标签，导致组合爆炸非常快（例如，5 个字的语句已经有 $3^5 = 243$ 个可能的输出）。为此，我们使用一些动态编程技巧来确保我们在合理的时间内计算 $P\{y|x\}$.

3.10 隐式马尔科夫模型及其以外

最后，隐式马尔可夫模型是朴素贝叶斯条件随机场的 softmax 回归：每一对中的前者通过对一系列标签进行建模而建立在后者上。这个图 1 可以更深入地了解这些关系：

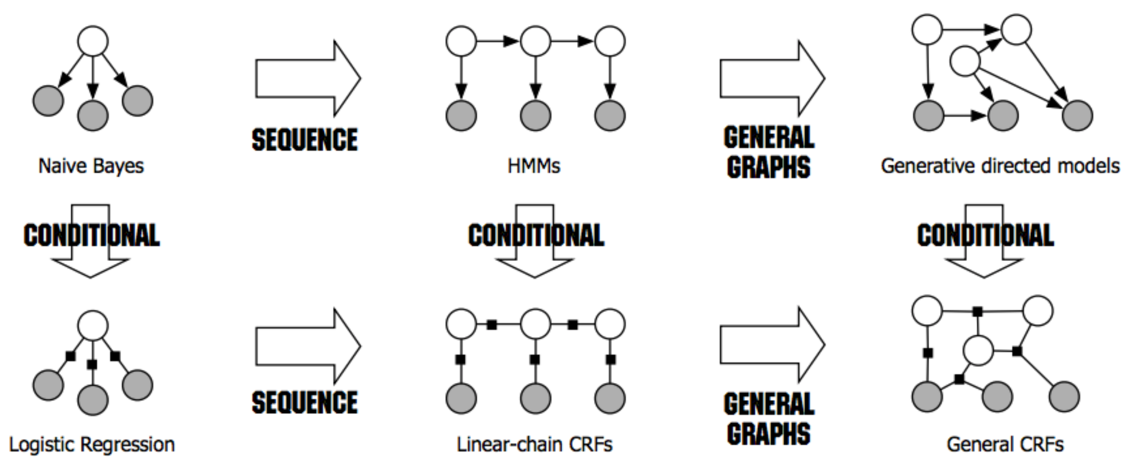


图 1

3.11 e 从哪里来？

方程式（2）表明，softmax 的分子，即输入元素的指数化线性组合，等于吉布斯分布因子乘积给出的输入和输出的非归一化联合概率。

但是，仅适用于以下两个条件之一成立：

1. 我们的因子是 e^z 形式。

2. 我们的因子采取任何形式，我们“预期”这种形式将在 softmax 函数内被指数化。

记住，这个取幂的点是将我们的加权输入元素“在算术上成为有效的概率”，即，使它们严格为正。这就是说，据我所知，没有任何一个因子产生一个严格正数。那么先是 - 鸡还是鸡蛋（指数还是 softmax）？

实际上，我并不真的确定，但我确实相信我们可以安全地将 softmax 分子和非规范化的 Gibbs 分布作为等价物，并将其简单地归结为：将它称为你将要的东西，我们需要一个指数在区间 $[0,1]$ 来放置这个东西。

第四章 总结

这次学习使典范机器学习模型，激活函数和条件概率的基本公理之间的关系变得更加清晰。 有关更多信息，请参考以下资源，特别是 Daphne Koller 的关于概率图形模型的材料。 非常感谢您阅读这篇文章。