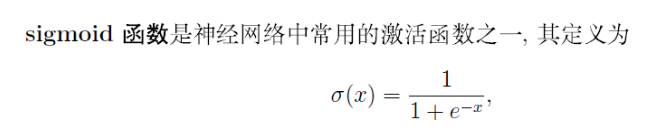
# 关于word2vec的一些理解？

**主要功能和作用：**word2vec主要是通过把常见的词语向量化，使得词语和词语之间可以进行一定的相似度计算，并且可以计算p(w|content(wi))以及p(content(wi)|w)的概率。通过Word2vec的一些算法和原理推测，word2vec应该具有如下功能：

1. 计算词语之间的相似度，即输出一个词语的相似词。
2. 计算词语之间的关联度。
3. 由周围的词语推测中心词，由中心词语推测周边词语。

# 二、word2vec中的数学知识

## sigmoid函数



### 1.什么是激活函数，为什么要引入激活函数？

如果不用激活函数（其实相当于激活函数是f(x) = x），在这种情况下你每一层输出都是上层输入的线性函数，很容易验证，无论你神经网络有多少层，输出都是输入的线性组合，与只有一个隐藏层效果相当，这种情况就是多层感知机（MLP）了。

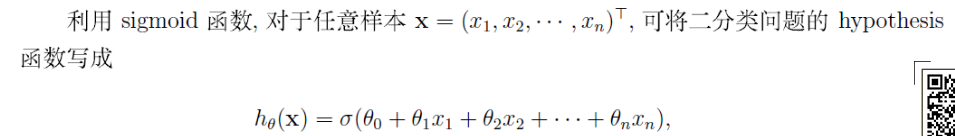
#### 什么是MLP？

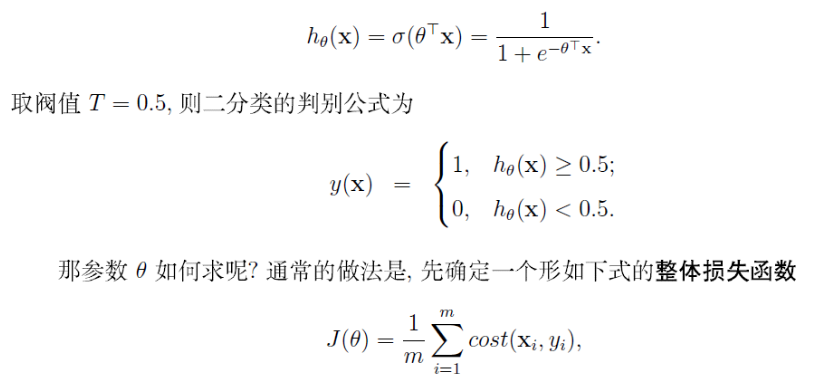
MLP是一个多层的神经网络，类似于多项图，在神经网络中，有从前向后的推导，也有从后向前根据损失函数调整各个结点权重的算法，在MLP中如果每个神经元的激活函数都是线性函数，那么，任意层数的MLP都可被约简化成一个等价的只有一个隐藏层的神经网络。所以尽量要选择比较有可微性的函数来当做激活函数。

### 2. sigmoid函数的导函数

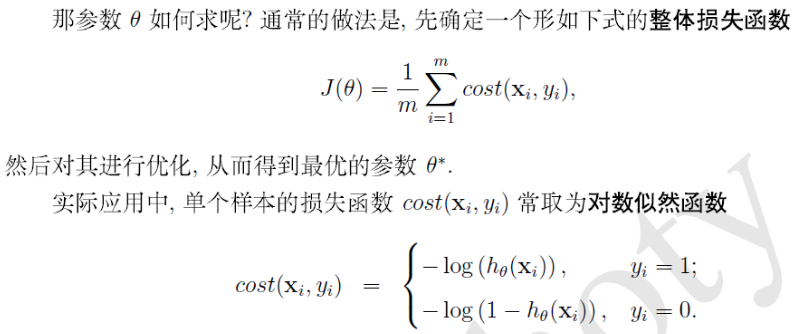
## （二）逻辑回归

生活中的分类问题相当多，例如判别是否是垃圾邮件，是否得了某种疾病，当然也有许多的多分类问题，二分类问题就是多分类问题中的简单版本。主要流程如下：

1.最后是要求出来一个概率值，并且通过一个自己设置的阈值来判断，如果最后的计算值大于阈值就是正例要么就是负例。

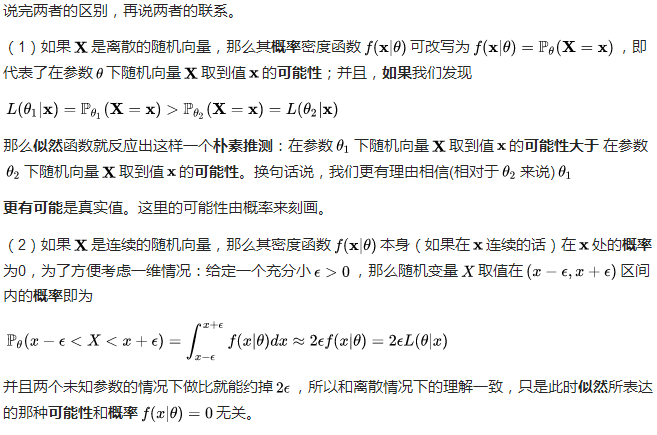
2. 

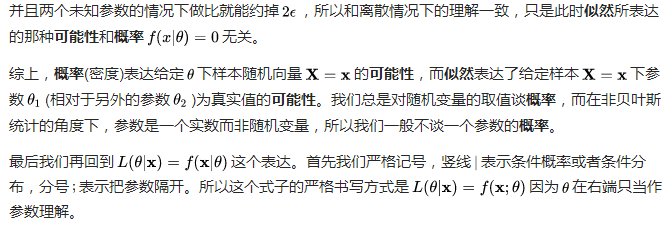
3.如何求参数数值呢？



1. 什么是似然函数？为什么要用对数似然函数？

**划重点起始：知乎解释如下：链接为：https://www.zhihu.com/question/54082000**





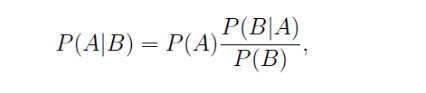
我的理解猜测是这样子的。

既然我们是要求参数的数值，那么一开始可以先给定一些随机数值，因为这么似然函数可以比较不同参数下，x出现的概率，那么，就可以逐渐逼近一个实际的参数值，好吧，好像理解错了。之所以叫做似然函数的概率，应该是因为概率一个是用来形容未知量X的吧，这里说的概率是关于参数的，那么也就是说，称作概率不合适？

涉及到似然函数的许多应用中，更方便的是使用似然函数的自然对数形式，即“对数似然函数”。求解一个函数的极大化往往需要求解该函数的关于未知参数的[偏导数](https://baike.baidu.com/item/%E5%81%8F%E5%AF%BC%E6%95%B0)。由于[对数函数](https://baike.baidu.com/item/%E5%AF%B9%E6%95%B0%E5%87%BD%E6%95%B0)是单调递增的，而且对数似然函数在极大化求解时较为方便，所以对数似然函数常用在最大似然估计及相关领域中。

## （三）bayes公式

我觉得贝叶斯推导的关键之处就是p(a∩b)=p(b∩a)

所以有

## （四）huffman树

理解的稍微有点问题，但是也弄懂点，知识点如下：

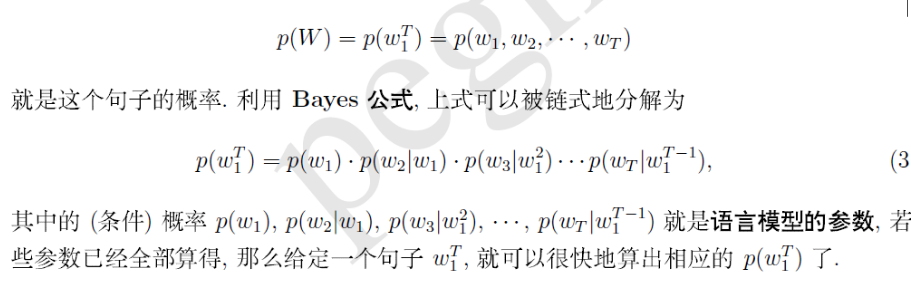
1. 通过构建huffman树，可以实现所有的词语都出现在叶子结点并且保证他有一个固定的01编码，并且能够保证出现频率比较高的词语出现在距离根节点更近的地方。
2. 我的理解是，从根节点到目标节点一直在经过一个二分的分类器，所以可以有效的减少计算次数，同时，还没理解了。说不下去了

# 三、统计语言模型

## （一）什么是统计语言模型？

我的理解是，统计语言模型就是通过将一句话进行分词之后，计算这个句子是一个自然语言的概率，就是说这句话能被人或者机器程序所理解，如果这句话不是一个可理解的话，没什么计算的必要。

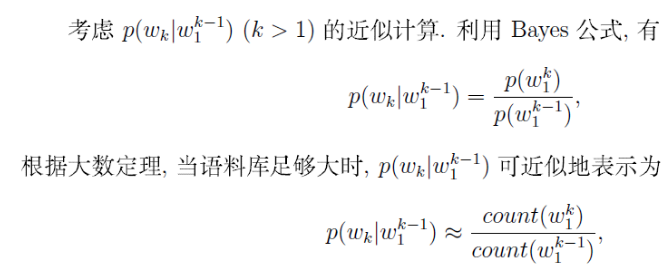
## 如何进行概率计算呢？



的意思是，第三个单词w3在前两个单词之后出现的概率。

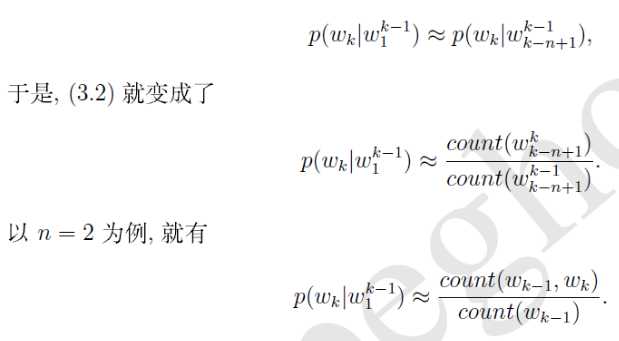
## （三）n-gram模型

### 1.为什么要出现n-gram模型？



当语料库非常大，并且k很大的时候，计算量过大，在这种情况下假设的是，第n个词语出现的概率和前面n-1个单词都有关系，计算量过于庞大，但是n-gram模型假设第N个单词出现的概率只和前面的x个单词有关系。所以简化了数据量。

### 2.n-gram模型的公式：



### 语料和词典的区别？

词典是从语料中提取出来的不包含重复词语的词组，但是语料是包含全部的内容，其中包含重复的词组。

# 四、[基于 Hierarchical Softmax 的模型](http://blog.csdn.net/itplus/article/details/37969979)

## （一）skip-gram模型

## （二）CBOW模型