<http://blog.csdn.net/zdy0_2004/article/details/45798223>

<http://blog.csdn.net/zb1165048017/article/details/50863934>

受限玻尔兹曼机详解 <http://www.tuicool.com/articles/jyqIfe6>

<http://blog.csdn.net/itplus/article/details/19408143>

Hinton教授的PPT教程注释 <http://www.cnblogs.com/shouhuxianjian/p/4529189.html>

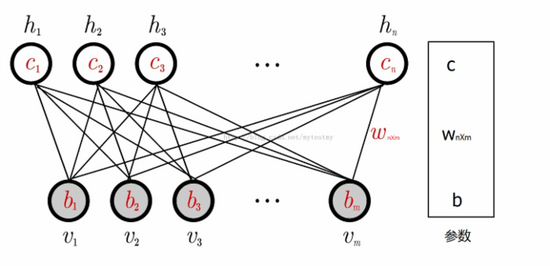
1.能量模型定义

在说RBM之前，先来说点其他的，就是能量模型。

能量模型是个什么样的东西呢？直观上的理解就是，把一个表面粗糙又不太圆的小球，放到一个表面也比较粗糙的碗里，就随便往里面一扔，看看小球停在碗的哪个地方。一般来说停在碗底的可能性比较大，停在靠近碗底的其他地方也可能，甚至运气好还会停在碗口附近（这个碗是比较浅的一个碗）；能量模型把小球停在哪个地方定义为一种状态，每种状态都对应着一个能量，这个能量由能量函数来定义，小球处在某种状态的概率（如停在碗底的概率跟停在碗口的概率当然不一样）可以通过这种状态下小球具有的能量来定义（换个说法，如小球停在了碗口附近，这是一种状态，这个状态对应着一个能量E，而发生“小球停在碗口附近”这种状态的概率p，可以用E来表示，表示成p=f(E)，其中f是能量函数），这就是我认为的能量模型。

2.RBM理解

很多文献上说受限波尔兹曼是一个无向图，从贝叶斯网的观点看，受限波尔兹曼网络也可以看作一个双向的有向图，即从输入层节点可以计算隐层节点取某一种状态值的概率，反之亦然。



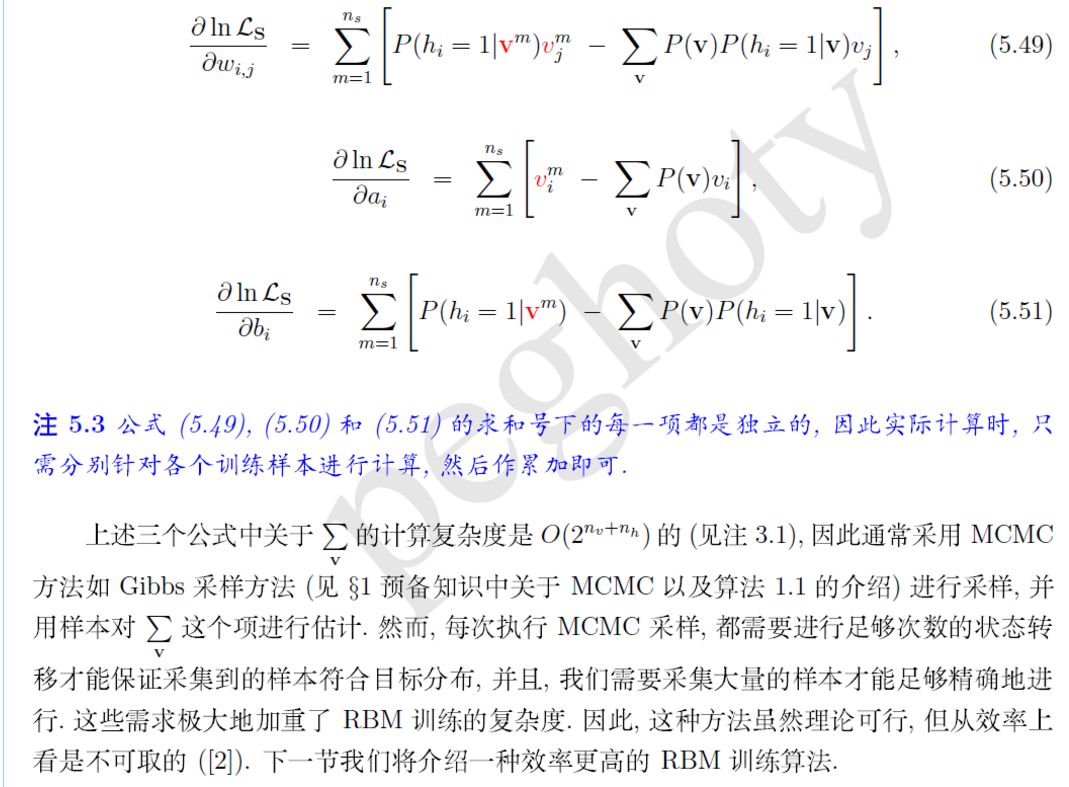
可视节点的意思就是训练的数据集，隐藏节点的一组取值就是这个训练样本编码后的值。

层内无连接，层间全连接。只有两层的神经网络，这是一个二分图。

可见层用来描述一个观察到的特征，但是隐藏层的意义不明确，可以理解为对表象的特征提取层。

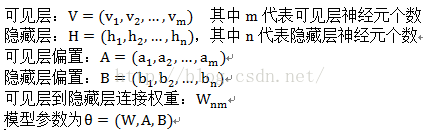
“限制”的意思是层内没有连接。

RBM模型是一个基于能量的模型，因此我们首先要引入一个能量函数，由这个能量函数能够得到一系列概率分布函数。



3.RBM证明

我们首先要知道玻尔兹曼机是一种基于统计力学提出来的能量模型，是一个随机神经元模型，首先介绍一下各个参数



已知条件：利用能量函数可以得到可见层和隐藏层的联合概率分布

http://img.blog.csdn.net/20160312111645463?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

然后我们利用联合概率计算边缘分布函数可以得到可见层V的分布函数

http://img.blog.csdn.net/20160930163058088?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

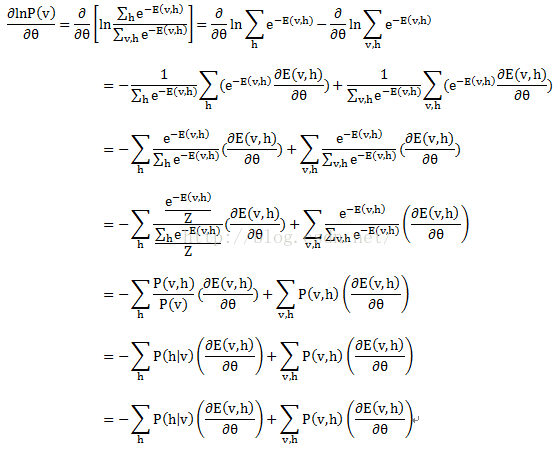
直接把和拿到分子的原因是，分母为归一化常量，为常量。

④接下来我们就是要可见层数据最大性的拟合总体样本空间的分布，我们假设总体分布为q(x)，总空间为Ω，然后我们用KL距离来计算样本分布和可见层分布的相似度

关于KL距离请看我前面写过的博客，它是一种不对称距离。既然我们要求P(x)拟合Q(x)损失的能量，那么就用下式计算



然后我们会发现，只有当P(x)最大的时候，才能让KL距离最小，这时候损失能量最小，因为总样本空间是固定的，所以这个Q(x)也是固定，只需管P(x)即可。接下来我们来求这个P(v)的最大值，求关于概率函数的最大值，我们很快就能想到极大似然函数，当然求解的时候用取对数方法求解

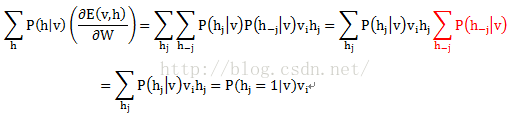


☆ ☆ ☆ ☆ ☆ ☆一定要注意，这个式子就是我们需要解决的问题，最好把它列在稿纸一边） ☆ ☆ ☆ ☆ ☆

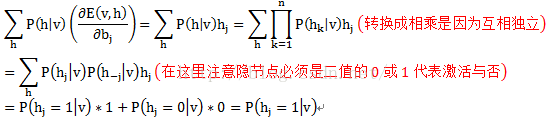
⑥接下来我们对模型参数中的权重W和偏置A，B分别求其偏导数，记住上式中，能量函数 E 在上面说过了，主要就是求前面的条件概率和联合概率

这里提前定义一个东西http://img.blog.csdn.net/20160312192917973?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center

先来逐步求解上式中第一项对模型参数中权重W，偏置A和B的偏导：

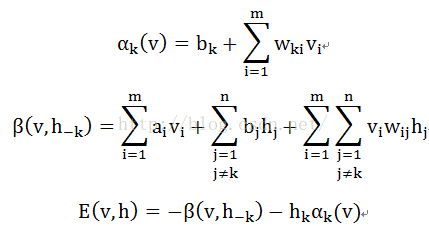


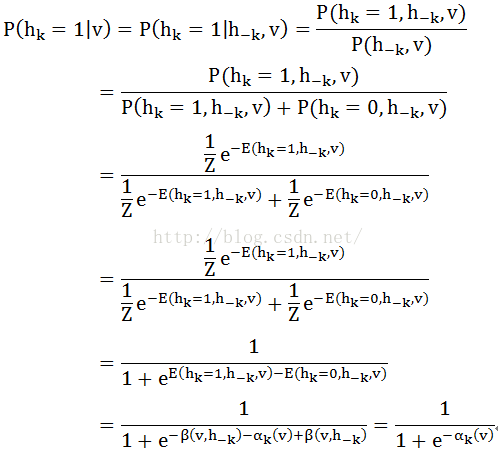
http://img.blog.csdn.net/20160312200854287?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center



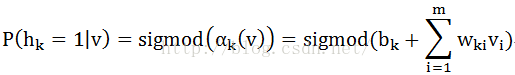
细心的话会发现这个P(hj=1|v)没有求出来，接下来我们求这一项

这里我们将能量函数拆开成两个式子





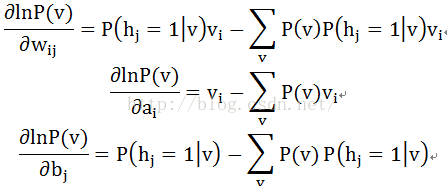
随后也就得到了最终的激活概率



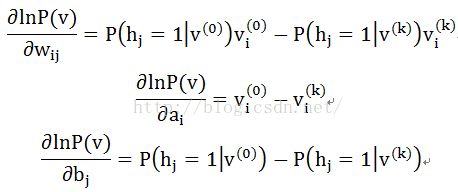
⑦第⑤步的第一项我们计算完毕，再来计算一下第二项

http://img.blog.csdn.net/20160312215345186?watermark/2/text/aHR0cDovL2Jsb2cuY3Nkbi5uZXQv/font/5a6L5L2T/fontsize/400/fill/I0JBQkFCMA==/dissolve/70/gravity/Center（这里面第二个∑ 已经在上面计算出来了）

⑧至此，我们基本求出了第⑤步中需要计算的所有参数，但是有一个没求出来，写出来以后就会发现



⑨我们可以发现上面有一项∑P(v)是没有计算出来的，这个时候，我们就必须使用某些采样方法对此项进行估计，大牛Hinton提出了对比散度算法，也就是把可见层输入数据当做起点，经过K次吉布斯采样的样本当做终点，近似计算上面三个式子，然后就变成了



4.工作原理

玻尔兹曼机学习算法是一个无监督学习算法，不像大家典型使用的BP算法，它会涉及到一个输入向量和一个理想的输出（作为目标或者标签），而在玻尔兹曼机学习中我们只需要输入向量即可。该算法试图要做的就是建立一个有关输入向量集合的模型，并通过该集合来考虑输出向量。

我们要做的就是最大化这些概率的积，这些概率也就是在训练集合中玻尔兹曼机指派的二值向量的集合。这等于最大化由玻尔兹曼机指派到训练向量的log概率的和；如果我们按照下面的方式来运行玻尔兹曼机，那么它同样的等于最大化我们获得的N个训练样本的概率：首先，让网络在N次不同的时间上在没有外部输入的情况下稳定到它的平稳分布；然后对可视向量进行采样，然后让网络再次稳定下来，然后接着对可视向量进行采用，以此类推。