

## Deep Learning 读书会第七次讨论记录 ( 由@极视角小助手整理 )

下面为 2017 年 1 月 2 日 Deep Learning 读书会第六章深度前馈网络前三节。本次讨论问题由@安兴乐 整理，并组织讨论。笔记由极视角筱雅整理。如有想加入读书会讨论的，请联系小助手（微信：Extreme-Vision）。

### 讨论话题

话题一. 神经网络能够模拟“XOR”运算有什么意义？神经网络是如何模拟“XOR”运算的？（安兴乐） .....	1
话题二. 为什么需要用“激活函数”，如何确定是用什么样的“激活函数”呢？（安兴乐） .....	5
话题三. BP 算法是否能保证收敛于“最佳状态”呢？（人工智障 v1.04） .....	9
话题四.如何去衡量一个神经网络的 VC 维呢？（安兴乐） .....	12
写在最后 .....	15

### 话题一. 神经网络能够模拟“XOR”运算有什么意义？神经网络是如何模拟“XOR”运算的？（安兴乐）

**安兴乐**

之前神经网络不被看好，很大程度上是因为感知器无法模拟 XOR 运算。

**Stomachache007**

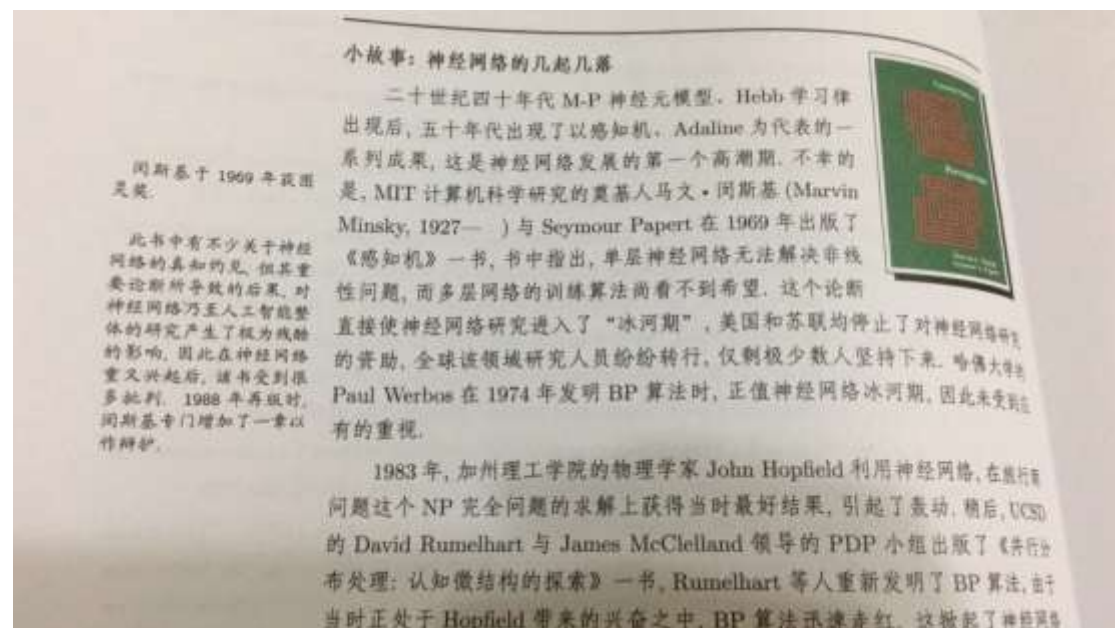
哦，这样。

**安兴乐**

实际上根据《neural network and deep learning》这本书的阐述，第一个问题也可以这样提问：为什么神经网络可以拟合所有函数？（<http://neuralnetworksanddeeplearning.com/chap4.html>）

**曲晓峰**

西瓜书上关于这段故事的介绍。



## 卷心菜+翻译+第九章

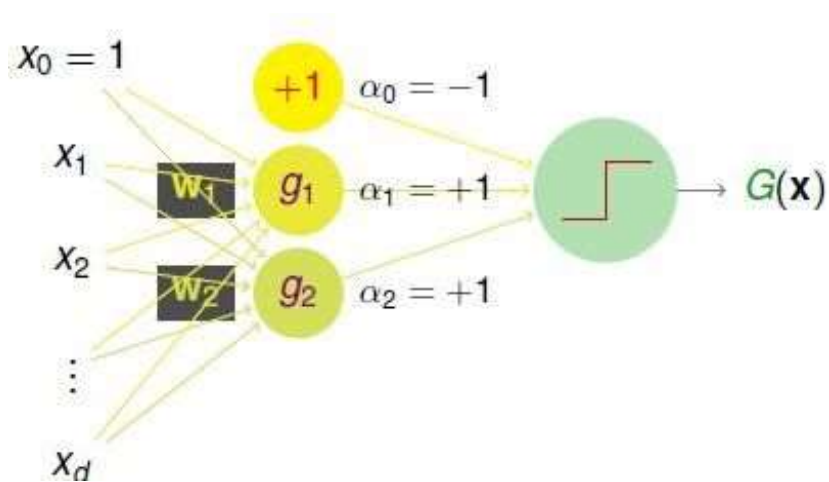
是因为神经元有多隐层吗？

## 安兴乐

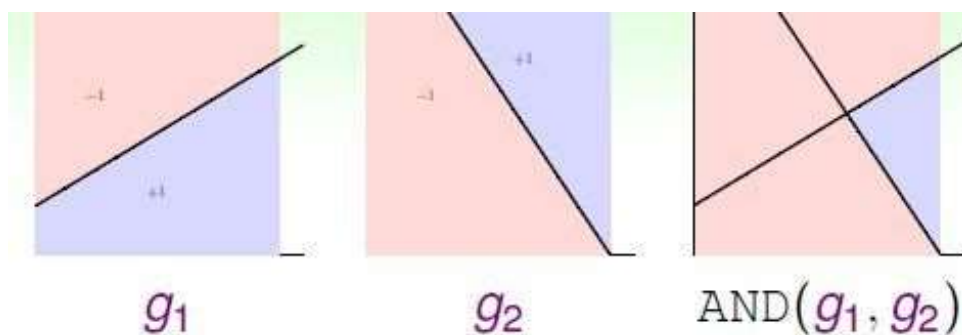
这里给出了一个可视化证明：



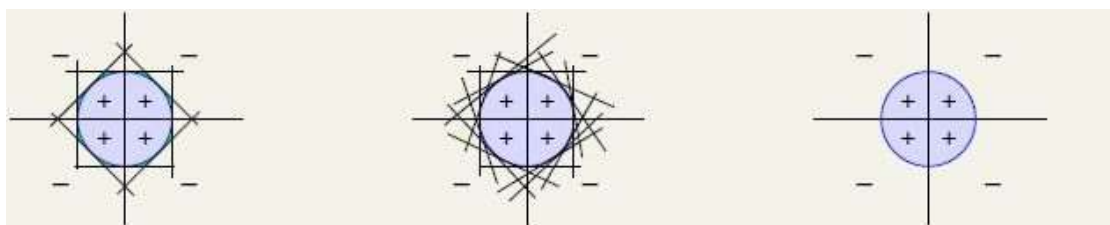
证明多层神经网络可以计算任意的函数。不过我觉得还是不够“明显”。在台大的《机器学习技法》里面给了一个相当可视化的证明过程：这是一个典型的神经网络：



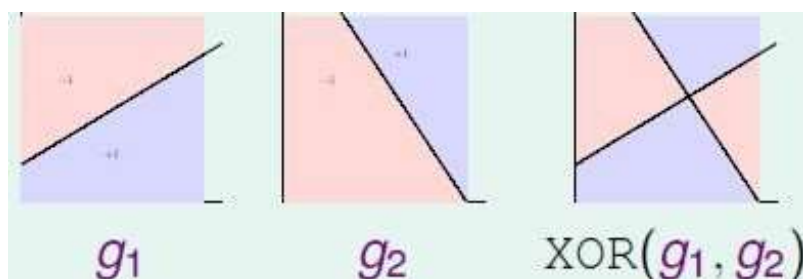
它对数据的拟合图示（同理可得 OR 运算的）：



如果在固定的一层上增加神经元的数量：



但是如何处理这样的数据呢？非线性可分的



这个时候线性不可分的数据处理就是一个问题了。这就是异或问题。很显然无法增加神经元数量来解决了。

大家想如何解决这样的问题呢？

曲晓峰

线性不可分，在 svm 里面也是靠 kernel 来做，算法有时尽，不可强求。

**安兴乐**

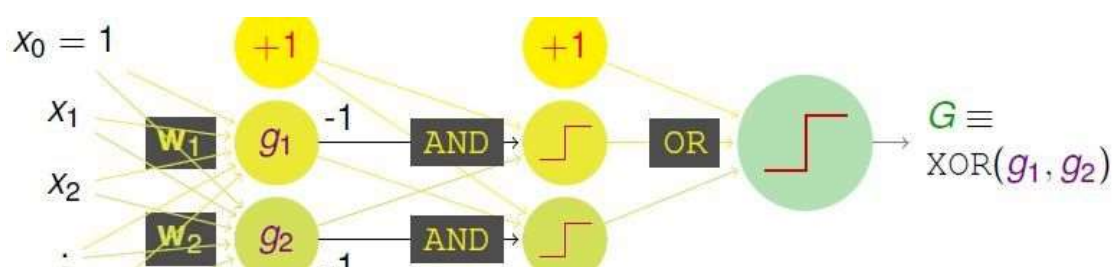
但是这里肯定不能用核函数了，都是线性单元，从数理逻辑的角度来看：我们知道  $XOR(g_1; g_2) = OR(AND(-g_1; g_2); AND(g_1; -g_2))$ ，两条线的 and 之后再 OR。

**曲晓峰**

所以还是靠多层？

**安兴乐**

也就是之后的这个 OR，需要增加一层 模拟 “OR” 的神经元来解决！



**曲晓峰**

嗯，所以一般说神经网络能逼近任意函数，至少也要三层才有这特性，还得输入数据给力。

**安兴乐**

是的。这样的话，神经网络通过增加层数就解决了长期被人诟病的 XOR 问题了。这样是为什么我们经常说的 神经网络只要够深，就能解决所有问题....

**枫**

好像必须要非线性单元才可以吧？

**安兴乐**

这就是另外一个问题了。

## AG-GROUP 元芳 第四章翻译

没有，最早的多层感知器激活函数就是简单的阈值。

### 极视角小助手

那对于第一个问题大家还有其他有疑问的地方吗？没有的话 我们继续第二个问题哈，为什么需要用“激活函数”，如何确定使用什么样的“激活函数”呢？

## 话题二. 为什么需要用“激活函数”，如何确定是用什么样的“激活函数”呢？（安兴乐）

Yisen

线性的话一直加深，还是线性啊。

安兴乐

@Yisen？从何说来

枫

书上的 171 页有讲。

安兴乐

对于单层的神经元，多个线性还是线性的。

Yisen

刚才一个童鞋说的，必须要非线性单元。

枫

倒数第二段。

## AG-GROUP 元芳 第四章翻译

链式法则决定了线性怎么线性叠加还是线性的。

枫

$y = w * x$   $z = W * y$   $z = W * w * x$

**yisen**

多层的话，也还是线性啊。

**安兴乐**

但是对于多层的神经网络，我们通过上面的证明已经可以看到，通过增加层数，是可以你和任意的函数的。大家一起来看看书吧。这个问题再重新表述：没有核函数，神经网络怎么可以比 SVM 牛逼啊？看 172 页的图示。

**Yc**

我觉得是不是没有激活函数再深也是线性可分才能表示。

**安兴乐**

我们无法通过线性函数来拟合 XOR 问题。这个时候的方法就是：通过隐层来实现特征空间的映射

**Stomachache007**

这个隐层是非线性运算？不好意思 比较小白。

**阿林**

激活函数就是映射函数？

**安兴乐**

还是个线性的。

**Stomachache007**

$\max(0, x)$  是线性运算么？

**安兴乐**

我错了，relu 不是线性的。

**Stomachache007**

灰常感谢，这个问题困扰了我好久。

**柳阳**

分段线性。

**安兴乐**

Relu 就是为了解决这个问题而来的。在第二层的线性函数之上加入了“非线性元素”来解决这个问题。

**Yisen**

就没有线性的激活函数。因为这样失去了激活的意义。

**柳阳**

有道理。

**Stomachache007**

嗯，是的。

**极视角小助手**

那为什么要使用激活函数呢？

**Stomachache007**

为了模拟非线性函数吧。

**Yisen**

引入非线性元素啊。

**阿林**

特征映射？

**安兴乐**

来解决我们的 非线性函数问题。

Clearly, we must use a nonlinear function to describe the features. Most neural

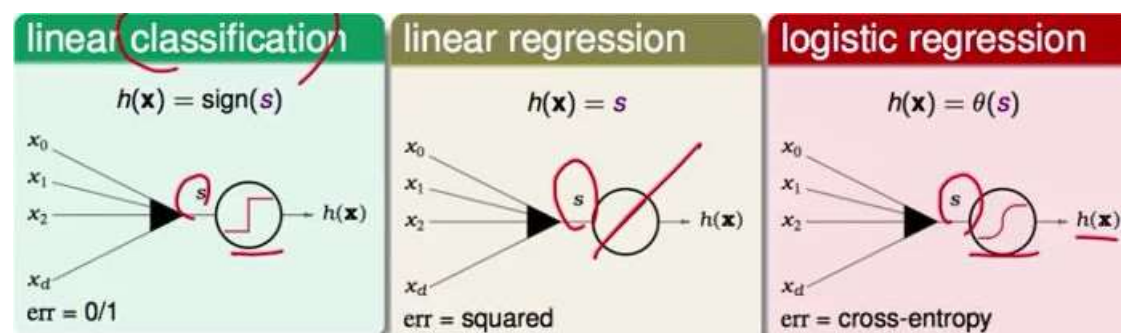
networks do so using an affine transformation controlled by learned parameters, followed by a fixed, nonlinear function called an activation function. We use that strategy here, by defining  $\mathbf{h} = g(\mathbf{W}^\top \mathbf{x} + \mathbf{c})$ , where  $\mathbf{W}$  provides the weights of a linear transformation and  $\mathbf{c}$  the biases. Previously, to describe a linear regression model, we used a vector of weights and a scalar bias parameter to describe an

极视角小助手

哦 那怎么选用激活函数呢？

安兴乐

解决问题的不同吧。



枫

现在好像基本都用 RELU 或 PRELU。

安兴乐

这是 6.3 的开篇第一问：How to choose the type of hidden unit to use in the hidden layers of the model.

极视角小助手

好的，谢谢。

曲晓峰

非线性、易求导、不阻碍导数向前传递。

安兴乐

靠猜。

枫



closer linear 函数

**曲晓峰**

后来又加上尽量避免负输出，输出尽量在规范化范围内。

**安兴乐**

用 Relu 的开始变多了 Hinton 去年发了篇关于 Relu 应用的文章 <https://arxiv.org/abs/1504.00941> 把 Relu 用在 RNN 上 ,证明不比 LSTM 差。

**枫**

RELU 的输出好像是避免了负输出，那正无穷怎么办？

**柳阳**

为什么会有正无穷？

**安兴乐**

输入有正无穷吗？

**枫**

当输入无限大的时候 RELU 也会输出无限大的 这个有问题吗？

**Yisen**

Normalization

**曲晓峰**

对呀，输入有 BN 对付

### **话题三. BP 算法是否能保证收敛于“最佳状态”呢？（人工智障 v1.04）**

**极视角小助手**

那我们继续下一个关于 BP 算法的问题哈，之前也有讨论过 BP 算法，那么 BP 算法是否能够保证收敛于“最佳状态”呢？

**Yc**

为什么要抑制负的？

**曲晓峰**

不好处理，有正负就需要叠加，单正容易做组合。收敛方面，不然没法保证最优的。

**安兴乐**

也因为 人的神经元需要一定的 “阈值” 才可以被激活吧

**极视角小助手**

那也就是不能保证啦？

**曲晓峰**

但似乎现在很多情况下是高维数据，数据维度一高，局部极小的情况就变得非常罕见了，大多数都会是鞍点，所以，越是高维状况，局部极小的情况越不严重。

**安兴乐**

最近面试的时候，有面试官让我讲讲 BP。我就吧 BP 的公式和简单伪代码写出来了，然后面试官说后面确定了再确定前面，这样子的话不断地调整一层又一层，永远也不会找到合适的权重的。我说：会收敛的。他说：不会的....然后我们开始 Rap 了。面试官：“你来证明一下会收敛”

**曲晓峰**

就是在一个甚至几个维度上的局部极小，在其它维度还有较为显著的梯度，迭代还是会很顺畅的走下去。

**Yisen**

没法证明收敛 也没法证明不收敛 从数学上。

**曲晓峰**

这个面试官其实不是问 BP，他问的其实是积分的基本理解。阿吉里斯追龟问题

**枫**

今年有篇文章好像证明了收敛 我找找

**曲晓峰**

但深度学习也确实会震荡。

**Yisen**

我猜 特定情况下的 可以 general 的 估计不行。

**安兴乐**

@枫 是吗？

**Yisen**

收敛

**阿林**

一般证明收不收敛我都不看的

**枫**

Deep Learning without Poor Local Minima

**Yc**

Bp 的公式是什么？

**枫**

<https://www.zhihu.com/question/54016305/answer/137631979>

**yc**

是指的 cost function 对 系数的导数吗？

**Yisen**

  
110

周博磊，MIT博士在读，AI方向。

110 人赞同

论文内容本身我就不多说了，证明了在线性网络上softmax上的一个conjecture，还是有些局限，想做DL偏理论的同学可以对比看下这篇和@田渊栋的那篇iclr'17。大过节的，我来多说说八卦算了：)



枫

我没看懂 能力不够

Yisen

假设都要求网络隐层中各结点的输出相互独立,才能得到结论。但是众所周知这个假设在实际情况中几乎不成立,各结点的输出都依赖于输入,因此往往强相关。

安兴乐

Got it! 工程上只要能够进入一定条件就可以了。

## 话题四.如何去衡量一个神经网络的 VC 维呢? (安兴乐)

极视角小助手

好,那我们进入最后一个问题,如何去衡量一个神经网络的 VC 维呢?

安兴乐

在简单网络中应该是 神经元个数\*权重个数。这个对于 RNN 怎么来算呢?

极视角小助手

大家有什么见解吗?

柳阳

能解释一下吗?简单网络的

安兴乐

VC 维就是“能打撒(区分)”的样本数。那么一个网络能够区分多大的数据呢?如果一个网络就是一个简单的线性函数,那么它应该可以区分 2 类。衡量一个网络的“区分”数据的能力,就是看它能否“区分”多大的数据量。这个时候简单网络的“打散数据”的能力就是 神经元个数\*权重个数。N 维空间中线性分类器和线性实函数的 VC 维是 N,错了 应该是 N+1 吧。

**柳阳**

$N+1$  是对。

**枫**

那神经网络的 VC 维就是最后的 softmax 的 VC 维吗？假设使用的是 softmax

**安兴乐**

VC 维 和 分类的类别数，之间有区别。

**枫**

什么是区分数据？

**安兴乐**

这个.....看一下 VC 维的定义吧。我好像解释的不太清楚。

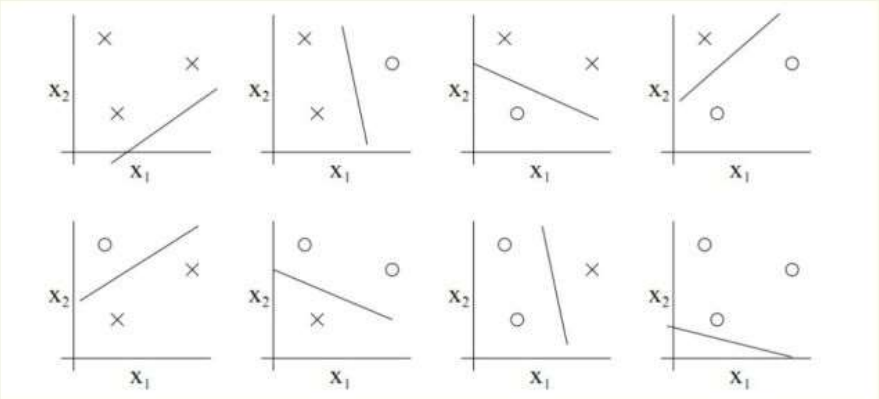
**枫**

先要介绍分散(shatter)的概念：对于一个给定集合 $S = \{x_1, \dots, x_d\}$ ，如果一个假设类 $H$ 能够实现集合 $S$ 中所有元素的任意一种标记方式，则称 $H$ 能够分散 $S$ 。

这样之后才有VC维的定义： $H$ 的VC维表示为 $VC(H)$ ，指能够被 $H$ 分散的最大集合的大小。若 $H$ 能分散任意大小的集合，那么 $VC(H)$ 为无穷大。在《神经网络原理》中有另一种记号：对于二分总体 $F$ ，其VC维写作 $VCdim(F)$ 。

通常定义之后，会用二维线性分类器举例说明为什么其VC维是3，而不能分散4个样本的集合，这里也就是容易产生困惑的地方。下面进行解释。

对于三个样本点的情况，下面的S1所有的标记方式是可以使用线性分类器进行分类的，因此其VC维至少为3（图片来自于斯坦福机器学习公开课的materials,cs229-notes4.pdf）：



是这个意思吧？

安兴乐

对！

$x_n$	$y_n = f(x_n)$
0 0 0	o
0 0 1	x
0 1 0	x
0 1 1	o
1 0 0	x

•  $\mathcal{X} = \{0, 1\}^3$ ,  $\mathcal{Y} = \{o, x\}$ , can enumerate all candidate  $f$  as  $\mathcal{H}$

以前讨论讲过吧。

James Liu

找到一篇证明加 momentum 的 GD 收敛性的文章 ( CONVERGENCE OF GRADIENT METHOD WITH MOMENTUM FOR BACK-PROPAGATION NEURAL NETWORKS\* )。反正结论是收敛到 local minima... （文章有附在 git 项目里）

## 极视角小助手

刚刚最后一个问题 如何去衡量一个神经网络的 VC 维呢？ 这里有个@纓宁 答的参考答案:被模型分散的最大样本集合的大小 。大家可以参考想想~

---

End

---

## 写在最后

非常感谢此次进行讨论交流的朋友们以及群内支持的朋友们 ,希望我们读书会能让大家学到更多，并且讨论后可以对原书有更独到的理解。

*#广告时间#*

视觉前沿资讯，将算法放至极市关注请关注极市平台公众号。

