

微博



逗比算法攻城狮

加关注

这哥们的长篇文章都是在给熟信打广告

简书

：马化腾不懂金融，结果被误导，把社交金融方向做偏了。方向偏了，下面执行的人越牛逼，局面越糟糕。对腾讯及微众银行来说，“微粒贷”是一个甜蜜的陷阱。<http://t.cn/RqWBjD>

今天 17:57 转发 | 评论

昵称：CodeMeals
园龄：3年9个月
粉丝：206
关注：19
+加关注

< 2016年4月 >

日	一	二	三	四	五	六
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
1	2	3	4	5	6	7

常用链接

我的随笔

我的评论

我的参与

最新评论

我的标签

更多链接

我的标签

数据挖掘(19)

python(13)

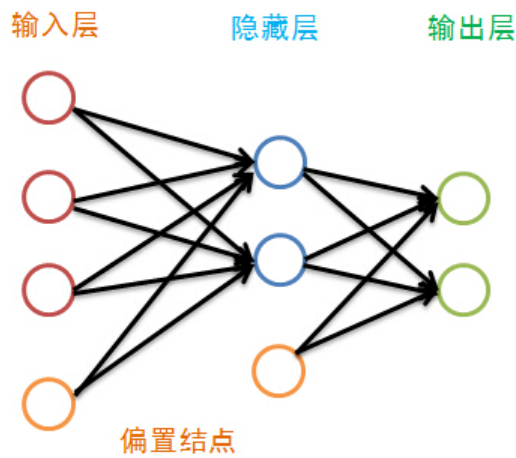
算法(10)

数据挖掘系列（9）——BP神经网络算法与实践

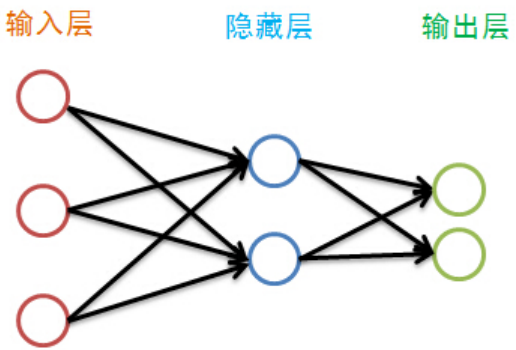
神经网络曾经很火，有过一段低迷期，现在因为深度学习的原因继续火起来了。神经网络有很多种：前向传输网络、反向传输网络、递归神经网络、卷积神经网络等。本文介绍基本的反向传输神经网络（Backpropagation 简称BP），主要讲述算法的基本流程和自己在训练BP神经网络的一些经验。

BP神经网络的结构

神经网络就是模拟人的大脑的神经单元的工作方式，但进行了很大的简化，神经网络由很多神经网络层构成，而每一层又由许多单元组成，第一层叫输入层，最后一层叫输出层，中间的各层叫隐藏层，在BP神经网络中，只有相邻的神经层的各个单元之间有联系，除了输出层外，每一层都有一个偏置结点：



虽然图中隐藏层只画了一层，但其层数并没有限制，传统的神经网络学习经验认为一层就足够好，而最近的深度学习不这么认为。偏置结点是为了描述训练数据中没有的特征，偏置结点对于下一层的每一个结点的权重的不同而生产不同的偏置，于是可以认为偏置是每一个结点（除输入层外）的属性。我们偏置结点在图中省略掉：



在描述BP神经网络的训练之前，我们先来看看神经网络各层都有哪些属性：

1. 每一个神经单元都有一定量的能量，我们定义其能量值为该结点j的输出值。 C_j ；
2. 相邻层之间结点的连接有一个权重 w_{ij} ，其值在 $[-1,1]$ 之间；
3. 除输入层外，每一层的各个结点都有一个输入值，其值为上一层所有结点按权重传递过来的能量之和加上偏置；
4. 除输入层外，每一层都有一个偏置值，其值在 $[0,1]$ 之间；
5. 除输入层外，每个结点的输出值等该结点的输入值作非线性变换；
6. 我们认为输入层没有输入值，其输出值即为训练数据的属性，比如一条记录 7 ；： $\&/\ast 0\ast 1'\ast 类别/ <$ ，那么输入层的三个结点的输出值分别为 $1,2,3$ 。因此输入层的结点数一般等于训练数据的属性个数。

java(9)
LeetCode(6)
实践中的bug(4)
数据库(3)
c++(3)
android(2)
C语言(1)
更多
随笔档案
2015年11月 (1)
2015年9月 (1)
2014年11月 (1)
2014年9月 (1)
2014年8月 (4)
2014年7月 (2)
2014年6月 (5)
2014年5月 (8)
2014年4月 (1)
2014年3月 (1)
2014年1月 (1)
2013年12月 (1)
2013年11月 (2)
2013年8月 (5)
2013年7月 (6)
2013年6月 (7)
2013年5月 (5)
2013年4月 (2)
2013年3月 (1)
2013年2月 (2)
2013年1月 (3)
2012年11月 (1)
2012年10月 (1)

训练一个BP神经网络，实际上就是调整网络的权重和偏置这两个参数，BP神经网络的训练过程分两部分：

1. 前向传输，逐层波浪式的传递输出值；
2. 逆向反馈，反向逐层调整权重和偏置；

我们先来看前向传输。

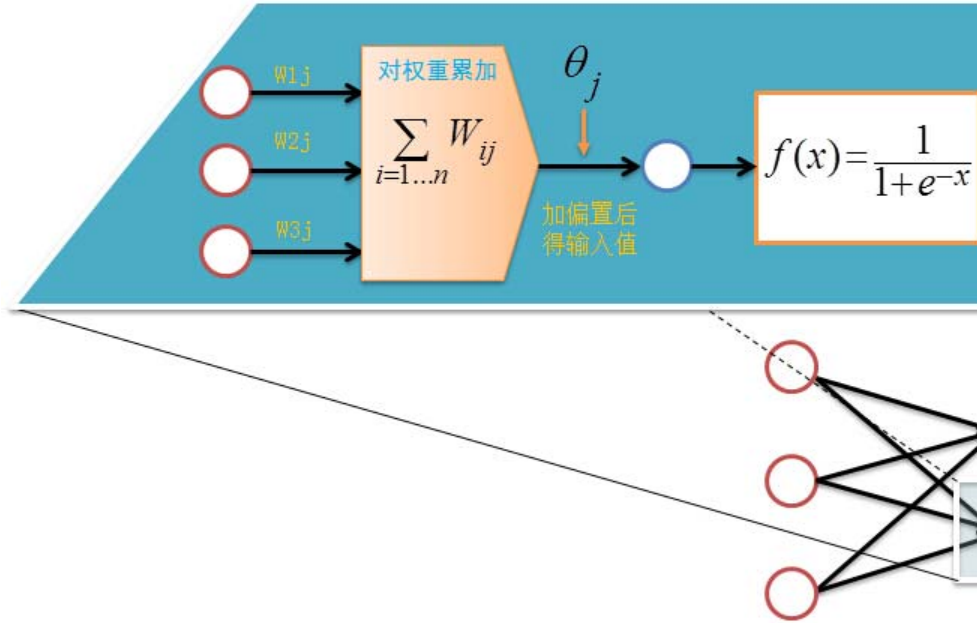
前向传输（Feed-Forward前向反馈）

在训练网络之前，我们需要随机初始化权重和偏置，对每一个权重取[-1,1]的一个随机实数，每一个偏置取[0,1]的一个随机实数，之后就开始进行前向传输。

神经网络的训练是由多趟迭代完成的，每一趟迭代都使用训练集的所有记录，而每一次训练网络只使用一条记录，抽象的描述如下：

```
while 终止条件未满足:
    for record:dataset:
        trainModel(record)
```

首先设置输入层的输出值，假设属性的个数为100，那我们就设置输入层的神经元个数为100，输入层的结点 x_i 为记录第*i*维上的属性值 Q_i 。对输入层的操作就这么简单，之后的每层就要复杂一些了，除输入层外，其他各层的输入值是上一层输入值按权重累加的结果值加上偏置，每个结点的输出值等该结点的输入值作变换



前向传输的输出层的计算过程公式如下：

$$C_j = \frac{\sum_{i=1}^n W_{ij} \cdot C_i + \theta_j}{\sum_{i=1}^n W_{ij} \cdot C_i + \theta_j}$$

对隐藏层和输出层的每一个结点都按照如上图的方式计算输出值，就完成前向传播的过程，紧接着是进行逆向反馈。

逆向反馈（Backpropagation）

逆向反馈从最后一层即输出层开始，我们训练神经网络作分类的目的往往是希望最后一层的输出能够描述数据记录的类别，比如对于一个二分类的问题，我们常常用两个神经元作为输出层，如果输出层的第一个神经元的输出值比第二个神经元大，我们认为这个数据记录属于第一类，否则属于第二类。

还记得我们第一次前向反馈时，整个网络的权重和偏置都是我们随机取，因此网络的输出肯定还不能描述记录的类别，因此需要调整网络的参数，即权重值和偏置值，而调整的依据就是网络的输出层的输出值与类别之间的差异，通过调整参数来缩小这个差异，这就是神经网络的优化目标。对于输出层：

$$C_j = \frac{\sum_{i=1}^n W_{ij} \cdot C_i + \theta_j}{\sum_{i=1}^n W_{ij} \cdot C_i + \theta_j}$$

其中 C_j 表示第*j*个结点的误差值， C_i 表示第*i*个结点的输出值， θ_j 记录输出值，比如对于2分类

2012年9月 (1)
2012年7月 (1)
友情链接
老高
新浪微博(@也爱数据挖掘)
积分与排名
积分 - 69151
排名 - 3153
最新评论
1. Re:异常检测算法--Isolation Forest
@大闹天空的程序猴同求测试数据集...
--linpeikun16
2. Re:模块度与Louvain社区发现算法
您好，我对这个方法很感兴趣，您能把这个代码分享给我吗，如果可以的话，我将会非常感激，谢谢。
--北区二排18号
3. Re:异常检测算法--Isolation Forest
"南大周志华老师在2010年提出一个异常检测算法Isolation Forest"，明明是周老师的一个学生提出的算法，这能一样么？
--yongmou-
4. Re:数据挖掘系列（4）使用weka做关联规则挖掘
博主你好，我在本地用retail.arff试了下，FPGrowth没有发现规则，输出信息如下：=== Evaluation ==Elapsed time: 0.0020sNo rules found.....
--dostar022
5. Re:如何找出知乎的所有神回复
我也在研究扒这一块数据 收藏一下
--Lumia1020

问题，我们用 $\mathbf{01}$ 表示类标 $\mathbf{1}$ ， $\mathbf{10}$ 表示类标 $\mathbf{2}$ ，如果一个记录属于类别 $\mathbf{1}$ ，那么其 $\mathbf{3} / \mathbf{30} ; /$ 。

中间的隐藏层并不直接与数据记录的类别打交道，而是通过下一层的所有结点误差按权重累加，计算公式如下：

$$\mathbf{\$}_C ; \cdot \mathbf{C} \& / \mathbf{H} \cdot \mathbf{C} ' \mathbf{3} \mathbf{D} \mathbf{\$}_D \mathbf{6} \mathbf{CD}$$

其中 $\mathbf{6}_{CD}$ 表示当前层的结点 \mathbf{j} 到下一层的结点 \mathbf{k} 的权重值， $\mathbf{\$}_D$ 下一层的结点 \mathbf{k} 的误差率。

计算完误差率后，就可以利用误差率对权重和偏置进行更新，首先看权重的更新：

$$\check{\mathbf{6}}_{EC} ; \mathbf{j} \mathbf{\$}_C \cdot \mathbf{B} \\ \mathbf{6}_{EC} ; \mathbf{6}_{EC}) \check{\mathbf{6}}_{EC}$$

其中 \mathbf{j} 表示学习速率，取值为 $\mathbf{0}$ 到 $\mathbf{1}$ ，学习速率设置得大，训练收敛更快，但容易陷入局部最优解，学习速率设置得比较小的话，收敛速度较慢，但能一步步逼近全局最优解。

更新完权重后，还有最后一项参数需要更新，即偏置：

$$\check{\mathbf{g}}_C ; \mathbf{j} \mathbf{\$}_C \\ \mathbf{g}_C ; \mathbf{g}_C) \check{\mathbf{g}}_C$$

至此，我们完成了一次神经网络的训练过程，通过不断的使用所有数据记录进行训练，从而得到一个分类模型。不断地迭代，不可能无休止的下去，总归有个终止条件

训练终止条件

每一轮训练都使用数据集的所有记录，但什么时候停止，停止条件有下面两种：

1. 设置最大迭代次数，比如使用数据集迭代 $\mathbf{100}$ 次后停止训练
2. 计算训练集在网络上的预测准确率，达到一定门限值后停止训练

使用BP神经网络分类

我自己写了一个BP神经网络，在数字手写体识别数据集MINIST上测试了一下，MINIST数据集中训练图片有 $\mathbf{12000}$ 个，测试图片 $\mathbf{20000}$ 个，每张图片是 $\mathbf{28*28}$ 的灰度图像，我对图像进行了二值化处理，神经网络的参数设置如下：

1. 输入层设置 $\mathbf{28*28=784}$ 个输入单元；
2. 输出层设置 $\mathbf{10}$ 个，对应 $\mathbf{10}$ 个数字的类别；
3. 学习速率设置为 $\mathbf{0.05}$ ；

训练经过约 $\mathbf{50}$ 次左右迭代，在训练集上已经能达到 $\mathbf{99\%}$ 的正确率，在测试集上的正确率为 $\mathbf{90.03\%}$ ，单纯的BP神经网络能够提升的空间不大了，但kaggle上已经有人有卷积神经网络在测试集达到了 $\mathbf{99.3\%}$ 的准确率。代码是去年用C++写的，浓浓的JAVA的味道，代码价值不大，但注释比较详细，可以查看[这里](#)，最近写了一个Java多线程的BP神经网络，但现在还不方便拿出来，如果项目黄了，再放上来吧。

训练BP神经网络的一些经验

讲一下自己训练神经网络的一点经验：

1. 学习速率不宜设置过大，一般小于 $\mathbf{0.1}$ ，开始我设置了 $\mathbf{0.85}$ ，准确率一直提不上去，很明显是陷入了局部最优解；
2. 输入数据应该归一化，开始使用 $\mathbf{0-255}$ 的灰度值测试，效果不好，转成 $\mathbf{01}$ 二值后，效果提升显著；
3. 尽量是数据记录随机分布，不要将数据集按记录排序，假设数据集有 $\mathbf{10}$ 个类别，我们把数据集按类别排序，一条一条记录地训练神经网络，训练到后面，模型将只记得最近训练的类别而忘记了之前训练的类别；
4. 对于多分类问题，比如汉字识别问题，常用汉字就有 $\mathbf{7000}$ 多个，也就是说有 $\mathbf{7000}$ 个类别，如果我们将输出层设置为 $\mathbf{7000}$ 个结点，那计算量将非常大，并且参数过多而不容易收敛，这时候我们应该对类别进行编码， $\mathbf{7000}$ 个汉字只需要 $\mathbf{13}$ 个二进制位即可表示，因此我们的输出层只需要设置 $\mathbf{13}$ 个结点即可。

阅读排行榜

1. 如何找出知乎的所有神回复(10925)
2. 数据挖掘系列（6）决策树分类算法(9813)
3. 数据挖掘系列（10）——卷积神经网络算法的一个实现(8269)
4. 数据挖掘系列（1）关联规则挖掘基本概念与Aprior算法(4596)
5. 数据挖掘系列（4）使用weka做关联规则挖掘(4447)

评论排行榜

1. 如何找出知乎的所有神回复(28)
2. 一个简单的多线程爬虫(26)
3. 数据挖掘系列（2）--关联规则FpGrowth算法(22)
4. 像Hacker News一样排序博客园首页文章(18)
5. 开发一个简单实用的android紧急求助软件(17)

推荐排行榜

1. 像Hacker News一样排序博客园首页文章(18)
2. PageRank实践-博客园用户PageRank排名(17)
3. 如何找出知乎的所有神回复(13)
4. 数据挖掘系列（1）关联规则挖掘基本概念与Aprior算法(11)
5. 在茫茫人海中发现相似的你——局部敏感哈希（LSH）(8)


参考文献：

Jiawei Han. 《Data Mining Concepts and Techniques》

转载请注明出处：<http://www.cnblogs.com/fengfenggirl/>

标签：[数据挖掘](#)



 **CodeMeals**
关注 - 19
粉丝 - 206
[+加关注](#)

3

0

(请您对文章做出评价)

« 上一篇：[python生成汉字图片字库](#)

» 下一篇：[数据挖掘系列（10）——卷积神经网络算法的一个实现](#)

posted @ 2014-06-20 22:59 CodeMeals 阅读(2570) 评论(6) 编辑 收藏

评论列表

#1楼 2014-06-20 23:27 jihite

高大上

支持(0) 反对(0)

#2楼 2014-06-21 10:01 Ant

高大上+1

支持(0) 反对(0)

#3楼[楼主] 2014-06-21 10:21 CodeMeals

@ Ant
哈哈，见笑了

支持(0) 反对(0)

#4楼 2014-06-21 18:13 Cloudoit

高大上+10086

支持(0) 反对(0)

#5楼 2015-11-18 10:45 songhuawang

楼主你好，你能不能把：逆向反馈（Backpropagation）再详细介绍一下。 特别这个 $E_j = O_j(1 - O_j)(T_j - O_j)$ 公式的意义

支持(0) 反对(0)

#6楼[楼主] 2015-11-28 17:21 CodeMeals

@ songhuawang
这个是求导得到的

支持(0) 反对(0)

[刷新评论](#) [刷新页面](#) [返回顶部](#)

注册用户登录后才能发表评论，请 [登录](#) 或 [注册](#)，[访问网站首页](#)。

【推荐】50万行VC++源码：大型组态工控、电力仿真CAD与GIS源码库

【推荐】融云即时通讯云-豆果美食、Faceu等亿级APP都在用



最新IT新闻：

- WebKit将不再使用前缀标记属性
 - Dropbox将允许用户摆脱本地空间限制访问所有文件
 - 新函数编程语言简化GPU应用开发
 - 游戏公司高管组团冲榜，《皇室战争》全球前5被中国人包揽
 - 过不了多久，你就能在月球上看到维基百科了
- » 更多新闻...



最新知识库文章：

- 架构漫谈（九）：理清技术、业务和架构的关系
 - 架构漫谈（八）：从架构的角度看如何写好代码
 - 架构漫谈（七）：不要空设架构师这个职位，给他实权
 - 架构漫谈（六）：软件架构到底是要解决什么问题？
 - 架构漫谈（五）：什么是软件
- » 更多知识库文章...

历史上的今天：

2013-06-20 python实现简易数据库之一——存储和索引建立

Copyright ©2016 CodeMeals

134732
Visitors