# 人工神经网络

<http://blog.csdn.net/fengbingchun/article/details/50274471>

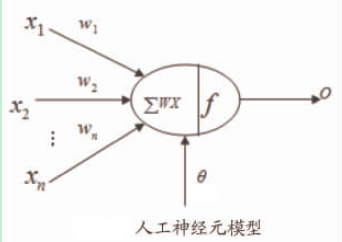
http://blog.csdn.net/zhongkejingwang/article/details/44514073

## 简介

人工神经网络（Artificial Neural Network，ANN）简称神经网络(NN)，是基于生物学中神经网络的基本原理，在理解和抽象了人脑结构和外界刺激响应机制后，以网络拓扑知识为理论基础，模拟人脑的神经系统对复杂信息的处理机制的一种数学模型。该模型以并行分布的处理能力、高容错性、智能化和自学习等能力为特征，将信息的加工和存储结合在一起，以其独特的知识表示方式和智能化的自适应学习能力，引起各学科领域的关注。它实际上是一个有大量简单元件相互连接而成的复杂网络，具有高度的非线性，能够进行复杂的逻辑操作和非线性关系实现的系统。

## 人工神经元结构

人工神经元的研究源于脑神经元学说，19世纪末，在生物、生理学领域，Waldeger等人创建了神经元学说。人工神经网络是由大量处理单元经广泛互连而组成的人工网络，用来模拟脑神经系统的结构和功能。而这些处理单元我们把它称作人工神经元。人工神经网络可看成是以人工神经元为节点，用有向加权弧连接起来的有向图。在此有向图中，人工神经元就是对生物神经元的模拟，而有向弧则是轴突----突触----树突对的模拟。有向弧的权值表示相互连接的两个人工神经元间相互作用的强弱。人工神经元结构如下图所示：

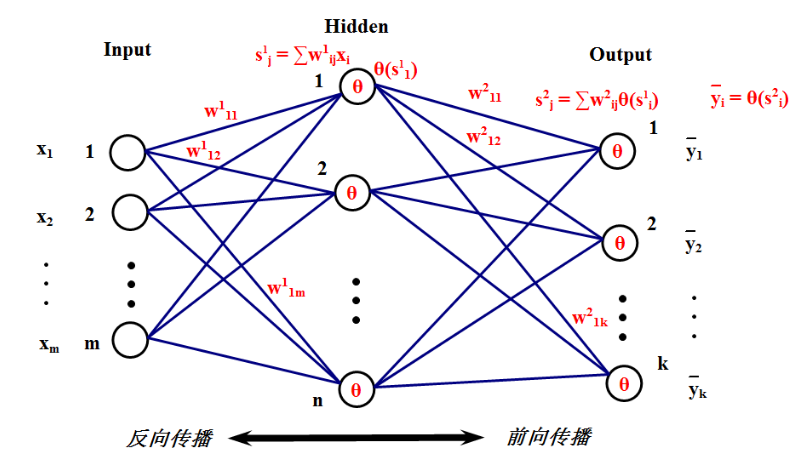


## 人工神经网络算法（BP算法）

1.3.1 BP算法简介

BP神经网络，BP即Back Propagation的缩写，也就是反向传播的意思。它可以对组成前向多层网络的各人工神经元之间的连接权值进行不断的修改，从而使该前向多层网络能够将输入它的信息变换成所期望的输出信息。同时也可以在修改各人工神经元的连接权值时，所依据的是该网络的实际输出与其期望的输出之差，将这一差值反向一层一层的向回传播，来决定连接权值的修改。

1.3.2 BP算法的数学原理



为了简单起见，这里只介绍只有一个隐层的BP网络，多个隐层的也是一样的原理。这个网络的工作原理应该很清楚了，首先，一组输入来到输入层，然后通过与隐层的连接权重产生一组数据作为隐层的输入，然后通过隐层节点的激活函数后变为其中表示隐层的第j个节点产生的输出，这些输出将通过隐层与输出层的连接权重产生输出层的输入，这里输出层的处理过程和隐层是一样的，最后会在输出层产生输出，这里j是指输出层第j个节点的输出。这只是前向传播的过程。在这里，先解释一下隐层的含义，可以看到，隐层连接着输入和输出层，它到底是什么？它就是特征空间，隐层节点的个数就是特征空间的维数，或者说这组数据有多少个特征。而输入层到隐层的连接权重则将输入的原始数据投影到特征空间，比如就表示这组数据在特征空间中第j个特征方向的投影大小，或者说这组数据有多少份量的j特征。而隐层到输出层的连接权重表示这些特征是如何影响输出结果的，比如某一特征对某个输出影响比较大，那么连接它们的权重就会比较大。关于隐层的含义就解释这么多，至于多个隐层的，可以理解为特征的特征。

既然在输出层产生输出了，现在就来分析一下什么东西在影响输出。显然，输入的数据是已知的，变量只有那些个连接权重了，那这些连接权重如何影响输出呢？现在假设输入层第i个节点到隐层第j个节点的连接权重发生了一个很小的变化Δ，那么这个Δ将会对产生影响，导致也出现一个变化Δ，然后产生Δ，然后传到各个输出层，最后在所有输出层都产生一个误差Δe。所以说，权重的调整将会使得输出结果产生变化，那么如何使这些输出结果往正确方向变化呢？这就是接下来的任务：如何调整权重。对于给定的训练样本，其正确的结果已经知道，那么由输入经过网络的输出和正确的结果比较将会有一个误差，如果能把这个误差降到最小，那么就是输出结果靠近了正确结果，就可以说网络可以对样本进行正确分类了。怎样使得误差最小呢？首先，把误差表达式写出来，为了使函数连续可导，这里最小化均方根差，定义损失函数如下：



在这里我们用随机梯度下降来最小化L。随机梯度下降也就是对每个训练样本都使权重往其负梯度方向变化。现在的任务就是求L对连接权重w的梯度。

用表示输入层第i个节点到隐层第j个节点的连接权重，表示隐层第i个节点到输出层第j个节点的连接权重，表示隐层第j个节点的输入，表示输出层第j个几点的输入，区别在右上角标，1表示第一层连接权重，2表示第二层连接权重。那么有：



由于



所以



代入前面式子可得



接下来只需要求出即可。

由于对所有输出层都有影响，所以



由于



所以



代入前面的式子可得：



现在记



则隐层为



输出层为









到这一步，可以看到是什么反向传播了吧？没错，就是误差e！

反向传播过程是这样的：输出层每个节点都会得到一个误差e，把e作为输出层反向输入，这时候就像是输出层当输入层一样把误差往回传播，先得到输出层δ，然后将输出层δ根据连接权重往隐层传输，即前面的式子：



现在再来看第一层权重的梯度：



第二层权重梯度：



可以看到一个规律：每个权重的梯度都等于与其相连的前一层节点的输出（即和）乘以与其相连的后一层的反向传播的输出（即和）。这样反向传播得到所有的以后，就可以更新权重了。

# 决策树

<https://www.cnblogs.com/pinard/p/6053344.html>

《机器学习》清华大学出版社 周志华著

## 简介

决策树是一类常见的机器学习方法，它是基于数结构来进行决策的，这恰是人类在面临决策问题时一种很自然的处理机制。一般的，一颗决策树包括一个根结点、若干个内部结点和若干个叶结点；叶结点对应于决策结果，其他每个结点则对应于一个属性测试；每个结点包含的样本集合根据属性测试的结果被划分到子结点中；根结点包含样本全集。从根结点到每个叶结点的路径对应了一个判定测试序列。决策树学习的目的是为了产生一颗泛化能力强，即处理未见示例能力强的决策树。

## 决策树算法

2.2.1信息增益

“信息熵”是度量样本集合纯度最常用的一种指标。假定当前样本集合D中第k类样本所占的比例为（k=1,2,...,），则D的信息熵定义为



的值越小，则D的纯度越高。

假定离散属性a有V个可能的取值，若使用a 来对样本集D进行划分，则会产生V个分支结点，其中第v个分支结点包含了D中所有在属性a上取值为的样本，记为。可根据定义计算出的信息熵，再考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重，可计算出用属性a对样本集D进行划分所获得的“信息增益”



ID3算法就是用信息增益大小来判断当前节点应该用什么特征来构建决策树，用计算出的信息增益最大的特征来建立决策树的当前节点。

著名的C4.5决策树算法不直接使用信息增益，而是使用“增益率”来选择最优划分属性，其定义为：



其中



2.2.2 CART算法的原理

分类回归树(CART,Classification And Regression Tree)算法采用一种二分递归分割的技术，将当前的样本集分为两个子样本集，使得生成的的每个非叶子节点都有两个分支。CART分类树算法使用基尼系数来代替信息增益比，基尼系数代表了模型的不纯度，基尼系数越小，则不纯度越低，特征越好。这和信息增益(比)是相反的。

具体的，在分类问题中，假设有K个类别，第k个类别的概率为, 则数据集D基尼系数的表达式为:



属性a的基尼指数定义为：



CART具体的算法流程：

算法输入是训练集D，基尼系数的阈值，样本个数阈值，输出是决策树T。我们的算法从根节点开始，用训练集递归的建立CART树。

（1）对于当前节点的数据集为D，如果样本个数小于阈值或者没有特征，则返回决策子树，当前节点停止递归。

（2）计算样本集D的基尼系数，如果基尼系数小于阈值，则返回决策树子树，当前节点停止递归。

（3）计算当前节点现有的各个特征的各个特征值对数据集D的基尼系数，。

（4）在计算出来的各个特征的各个特征值对数据集D的基尼系数中，选择基尼系数最小的特征A和对应的特征值a。根据这个最优特征和最优特征值，把数据集划分成两部分D1和D2，同时建立当前节点的左右节点，做节点的数据集D为D1，右节点的数据集D为D2.

（5）对左右的子节点递归的调用1-4步，生成决策树。

对于生成的决策树做预测的时候，假如测试集里的样本A落到了某个叶子节点，而节点里有多个训练样本。则对于A的类别预测采用的是这个叶子节点里概率最大的类别。

由于决策时算法很容易对训练集过拟合，而导致泛化能力差，为了解决这个问题，我们需要对CART树进行剪枝，即类似于线性回归的正则化，来增加决策树的返回能力。CART采用的办法是后剪枝法，即先生成决策树，然后产生所有可能的剪枝后的CART树，然后使用交叉验证来检验各种剪枝的效果，选择泛化能力最好的剪枝策略。

# 随机森林

http://blog.csdn.net/zrjdds/article/details/50133843

## 简介

随机森林是一种比较新的机器学习模型。上世纪八十年代Breiman等人发明分类树的算法（Breiman et al. 1984），通过反复二分数据进行分类或回归，计算量大大降低。2001年Breiman把分类树组合成随机森林（Breiman 2001a），即在变量（列）的使用和数据（行）的使用上进行随机化，生成很多分类树，再汇总分类树的结果。随机森林在运算量没有显著提高的前提下提高了预测精度。随机森林对多元公线性不敏感，结果对缺失数据和非平衡的数据比较稳健，可以很好地预测多达几千个解释变量的作用（Breiman 2001b），被誉为当前最好的算法之一（Iverson et al. 2008）。

随机森林顾名思义，是用随机的方式建立一个森林，森林里面有很多的决策树组成，随机森林的每一棵决策树之间是没有关联的。在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类。

## 随机森林基本原理

随机森林由LeoBreiman（2001）提出，它通过自助法（bootstrap）重采样技术，从原始训练样本集N中有放回地重复随机抽取k个样本生成新的训练样本集合，然后根据自助样本集生成k个分类树组成随机森林，新数据的分类结果按分类树投票多少形成的分数而定。其实质是对决策树算法的一种改进，将多个决策树合并在一起，每棵树的建立依赖于一个独立抽取的样品，森林中的每棵树具有相同的分布，分类误差取决于每一棵树的分类能力和它们之间的相关性。特征选择采用随机的方法去分裂每一个节点，然后比较不同情况下产生的误差。能够检测到的内在估计误差、分类能力和相关性决定选择特征的数目。单棵树的分类能力可能很小，但在随机产生大量的决策树后，一个测试样品可以通过每一棵树的分类结果经统计后选择最可能的分类。

## 随机森林实现过程

随机森林中的每一棵分类树为二叉树，其生成遵循自顶向下的递归分裂原则，即从根节点开始依次对训练集进行划分；在二叉树中，根节点包含全部训练数据， 按照节点纯度最小原则，分裂为左节点和右节点，它们分别包含训练数据的一个子集，按照同样的规则节点继续分裂，直到满足分支停止规则而停止生长。若节点n上的分类数据全部来自于同一类别，则此节点的纯度I(n)=0，纯度度量方法是Gini准则，即假设P(X)是节点n上属于类样本个数占训练。

具体实现过程如下：

（1）原始训练集为N，应用bootstrap法有放回地随机抽取k个新的自助样本集，并由此构建k棵分类树，每次未被抽到的样本组成了k个袋外数据；

（2）设有个变量，则在每一棵树的每个节点处随机抽取个变量( n )，然后在中选择一个最具有分类能力的变量，变量分类的阈值通过检查每一个分类点确定；

（3）每棵树最大限度地生长, 不做任何修剪；

（4）将生成的多棵分类树组成随机森林，用随机森林分类器对新的数据进行判别与分类，分类结果按树分类器的投票多少而定。



# 非线性回归

<http://blog.csdn.net/loveliuzz/article/details/78024355>

<http://blog.csdn.net/AriesSurfer/article/details/41310525>

高正明，赵娟，贺升平 多项式回归分析及回归方程的显著性检验[J] 中国核科学技术进展报告（第二卷）2011.10

## Logistic Regression

1、概率

（1）定义：概率（Probability）：对一件事情发生的可能性的衡量。

（2）取值范围：

（3）计算方法：根据个人置信、根据历史数据、根据模拟数据

（4）条件概率：在事件B已经发生的情况下，事件A发生的概率等于事件A、B同时发生的概率除以B事件发生的概率。



1. Logistic Regression (逻辑回归)

Logistic回归为概率型非线性回归模型，是研究二分类观察结果与一些影响因素之间关系的一种多变量分析方法。通常的问题是，研究某些因素条件下某个结果是否发生，比如医学中根据病人的一些症状来判断它是否患有某种病。

Regression问题的常规步骤为：

寻找h函数（即hypothesis）；

构造J函数（损失函数）；

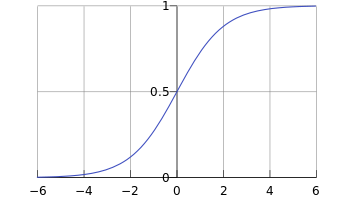
想办法使得J函数最小并求得回归参数（θ）。

（1）构造预测函数h

Logistic回归虽然名字里带“回归”，但是它实际上是一种分类方法，主要用于两分类问题（即输出只有两种，分别代表两个类别），所以利用了Logistic函数（或称为Sigmoid函数），函数形式为：



Sigmoid 函数在有个很漂亮的“S”形，如下图所示（引自维基百科）：



➀对于线性边界的情况，边界形式如下：



构造预测函数为：



函数的值有特殊的含义，它表示结果取1的概率，因此对于输入x分类结果为类别1和类别0的概率分别为：

 （1）

构造损失函数J

Cost函数和J函数如下，它们是基于最大似然估计推导得到的：

****



下面详细说明推导的过程：

1. 式综合起来可以写成：



取似然函数为：



对数似然函数为：



最大似然估计就是求使取最大值时的θ，其实这里可以使用梯度上升法求解，求得的θ就是要求的最佳参数。但是，由于，此式乘了一个负的系数-1/m，所以取最小值时的θ为要求的最佳参数，所以梯度下降法求的最小值。

更新过程：







更新过程可以写成：



下面介绍向量化的过程：

约定训练数据的矩阵形式如下，x的每一行为一条训练样本，而每一列为不同特征取值：

，，





g(A)的参数A为一列向量，所以实现g函数时要支持列向量作为参数，并返回列向量。

更新过程可以改为：



综上所述，向量化后更新的步骤如下：

1. 求；
2. ；
3. 。

## 多项式回归

4.2.1多项式拟合的普适性

根据高等数学理论，若连续函数在的某一领域内具有直到（n+1）阶的导数，则函数在处可展开为n阶Taylor级数形式：

 其中是与之间的某个值，这说明任意连续的在领域内具有阶导数的函数均可以简化为多项式函数表达式：



因此，对于任意母体的n组统计独立的样本数据（）（），在母体规律存在的前提下，必然存在一组系数，使式表征母体分布规律，因此多项式拟合对统计数据的普适的拟合方法，适用于任意母体规律存在的样本数据回归分析过程。

4.2.2 多项式回归参数的最小二乘估计

对于任意实际问题的n组观测数据（）（），有



设





代入上式可得到其矩阵表达形式：



根据最小二乘估计原则，回归参数的估计值应使估计值与观测值之间的残差达到最小，即：



其中即有：



即有当存在时：



此外，零均值正态分布的随机误差的方差估计量为：



其中样本残差：



