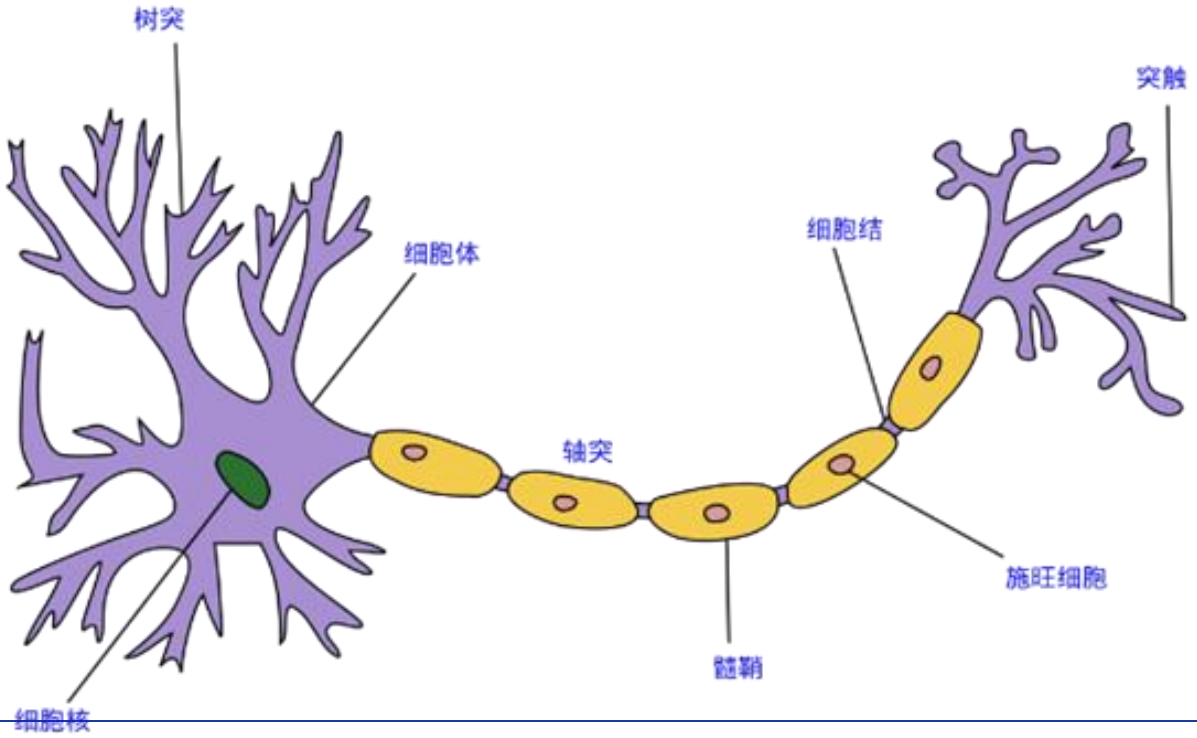
南開大學

第七讲 构建人工神经网络预测模型

提　纲

1. 什么是人工神经网络

- 神经元的结构



1. 什么是人工神经网络

◆ 生物神经网络

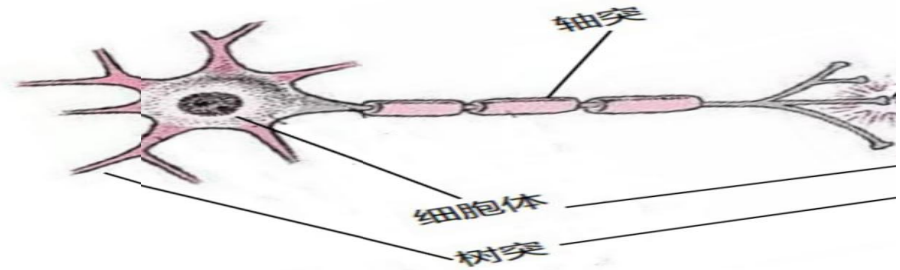
人类的大脑大约有1.4×10^{11}个神经元，每个神经元通过约10^3~10^5个突触与其他多个神经元连接形成庞大而复杂的神经网络，即生物神经网络。生物神经网络中各神经元之间连接的强弱会随着外部刺激信号发生变化，每个神经元会综合按照接收到的多个刺激信号呈现出兴奋或抑制状态。

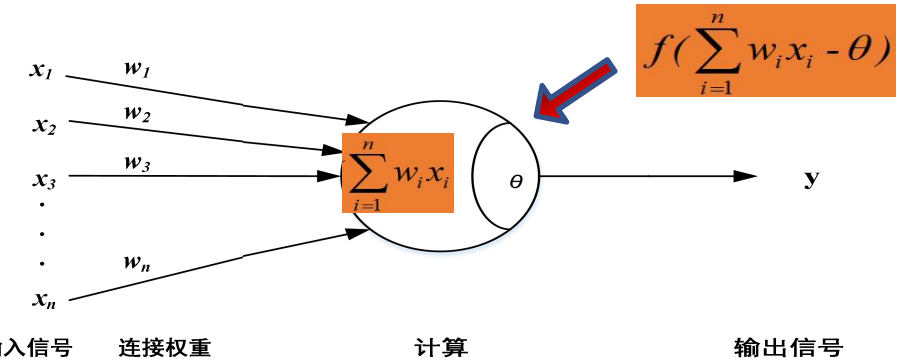
神经元是人类大脑处理信息的最小单位，大脑的学习过程就是神经元之间连接强度接收外部刺激相应地做出自适应变化的过程，各神经元所处状态的整体情况决定了大脑处理信息的结果。

1. 什么是人工神经网络

◆ 人工神经元

人工神经元正是模拟了人脑的神经元，它的基本模型包含输入，输出与计算三部分。输入可以类比为神经元的树突，输出可以类比为神经元的轴突，计算则可以类比为细胞核。





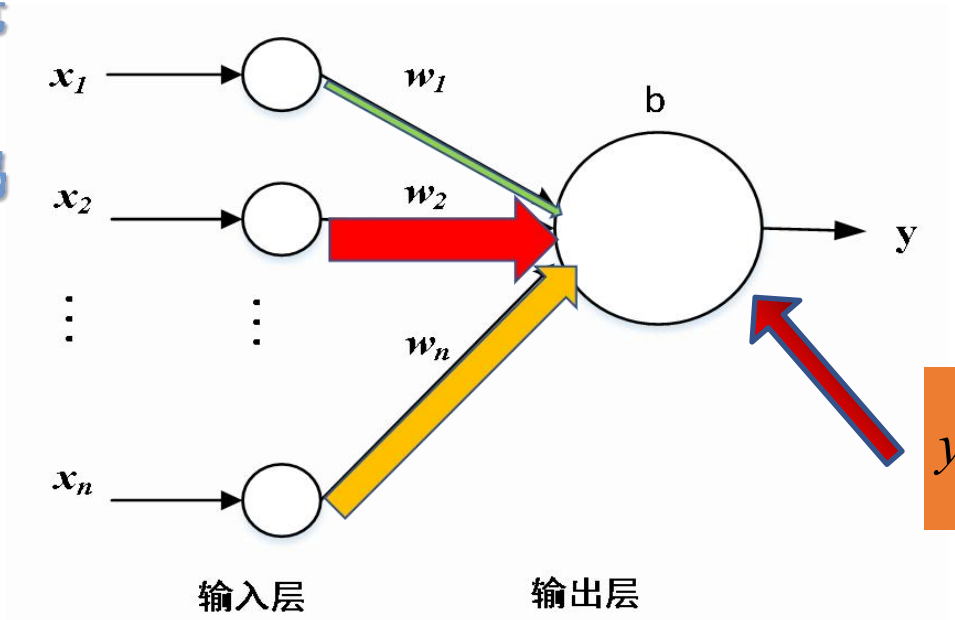
1. 什么是人工神经网络

◆ 感知机模型

爸爸

妈妈

你



$$sgn(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

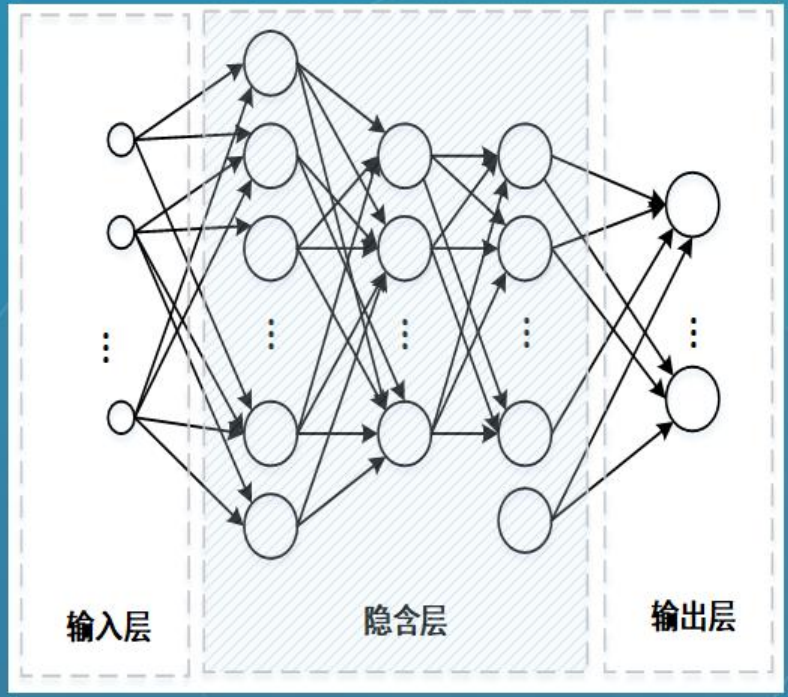
$$y = sgn\left(\sum_{i=1}^{n} w_i x_i + b\right)$$

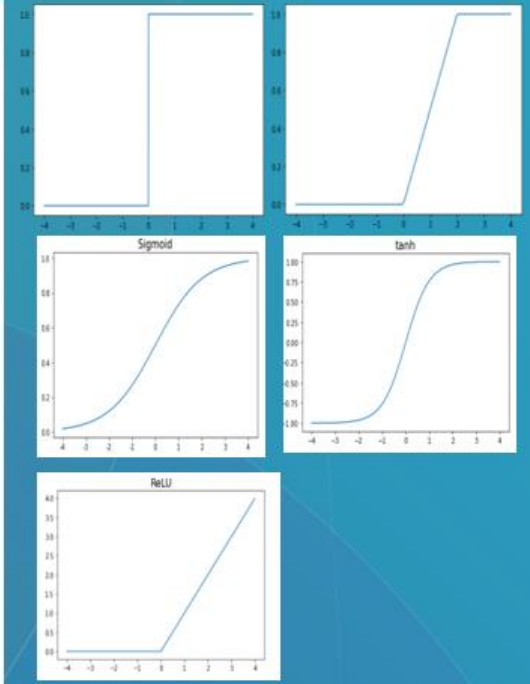
1. 什么是人工神经网络

- 人工神经网络（Artificial Neural Network，ANN）

人工神经网络

更为复杂的非线性问题需要依靠多层感知机结构的人工神经网络（增加了隐含层）





在神经网络中，神经元的不同主要在于其采用了不同的激活函数，激活函数具有反映神经元输出与其激活状态之间关系的功能。

2. R语言与神经网络

R语言中已经有许多用于神经网络的package。
例如nnet、AMORE、neuralnet。

- nnet提供了最常见的前馈反向传播神经网络算法。
- AMORE包则更进一步提供了更为丰富的控制参数，并可以增加多个隐藏层。
- neuralnet包的改进在于提供了弹性反向传播算法和更多的激活函数形式。
- Stuttgart Neural Network Simulator（SNNS）是德国斯图加特大学开发的优秀神经网络仿真软件，为国外的神经网络研究者所广泛采用。其手册内容极为丰富，同时支持友好的 Linux 平台。而RSNNS则是连接R和SNNS的工具，在R中即可直接调用SNNS的函数命令。

http://www.360doc.com/content/18/0426/18/7378868_748973845.shtml

3.构建BP神经网络预测模型

【问题】 airdata.csv是预测空气质量可能的相关因素，数据已经进行了归一化处理。其中，f1_f10分别代表气压、气温、相对湿度、云量、日照时数、风速、SO2浓度、 NO2浓度、PM2.5浓度、PM10浓度。

如何利用这1758条数据，预测PM2.5的浓度?

3.构建BP神经网络预测模型

人工神经网络可以用在数据的**分类、预测**，甚至是无监督的模式识别。其中BP （Back-propagation，反向传播）神经网络广泛使用。

BP神经网络可以逼近任意连续函数，具有很强的非线性映射能力，而且网络的中间层数、各层的处理单元数及网络的学习系数等参数可根据具体情况设定，灵活性大。

BP网络的参数优化就是选择神经网络的各权重系数，使得期望值$y0_j$与实际输出值y_j之差的误差最小（损失函数）。**反向传播算法**就是在模拟过程中（训练神经网络时）收集系统所产生的误差，之后用这些误差来调整神经元的权重，这样生成一个可以模拟出原始问题的人工神经网络系统。

3.构建BP神经网络预测模型

算法训练步骤:

① 定义参数x(输入向量),w(权值向量),b(偏置),y(实际输出),d(期望输出),a(学习率参数)

② 初始化，n=0,w=0

③ 输入训练样本，对每个训练样本指定其期望输出

④ 计算实际输出y=f(w*x+b)

⑤ 更新权值向量w(n+1)=w(n)+a[d-y(n)]*x(n),0

⑥ 判断，若满足收敛条件，算法结束，否则返回3

注意，其中学习率a为了权值的稳定性不应过大，为了体现误差对权值的修正不应过小（凭经验）。

3.构建BP神经网络预测模型

预测结果评价：

- MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

- RMSE：　对MSE的开根号

- MAE

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

- MAPE

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right|$$

3.构建BP神经网络预测模型

预测结果评价：

- R^2

$$R^2 = 1 - \frac{\sum (Y\_actual - Y\_predict)^2}{\sum (Y\_actual - Y\_mean)^2}$$

越接近1，表明方程的变量对y的解释能力越强，这个模型对数据拟合的也较好。越接近0，表明模型拟合的越差。

经验值：>0.4， 拟合效果好

缺点：

数据集的样本越大，R^2越大，因此，不同数据集的模型结果比较会有一定的误差。

- 校正R^2

$$R^2\_adjusted = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1}$$

n为样本数量，p为特征数量

消除了样本数量和特征数量的影响

3.构建BP神经网络预测模型

【步骤】

① 设置工作目录

```
setwd("D:/R")
```

② 安装neuralnet包，neuralnet函数用于神经网络建模

```
install.packages("neuralnet")
library("neuralnet")
```

③ 运用read.csv函数读取数据

```
data<-read.csv("airdata.csv")
class(data)   #data是一个数据框
```

3.构建BP神经网络预测模型

④ 拆分训练用数据（训练集）和测试数据（测试集）

```
trainingdata <- data[1:1700,1:10]  #前1700个数据
testdatax <- data[1701:1758,
                  c('f1','f2','f3','f4','f5','f6','f7','f8','f10')]
testdatay <- data[1701:1758,9] #后58条数据
```

3.构建BP神经网络预测模型

⑤ 建立BP模型(f9 ~ f1+f2+f3+f4+f5+f6+f7+f8+f10 ）

```
BPnet <- neuralnet(f9 ~
f1+f2+f3+f4+f5+f6+f7+f8+f10 ,trainingdata,hidden=2)
```

Hidden：隐含层神经元个数

Threshold：设定训练停止的一个条件

act.fct：设置激活（传递）函数

.....

请参考help

⑥ 画出网络结构图查看模型参数等

```
plot(BPnet)
```

3.构建BP神经网络预测模型

⑦ 对测试数据集进行预测，求解预测值与目标值的偏差

```
BPnet.y <- compute(BPnet, testdatax)
print(BPnet.y$net.result-testdatay)
# 让结果更直观些
output <- cbind(c(1:58),testdatay,BPnet.y$net.result)
colnames(output) <- c("DataID","Expected Output","BP Output")
print(output)
write.csv(output, 'BPresult.csv')
```

3.构建BP神经网络预测模型

⑧ 对结果作图

```
library('ggplot2')
output<-as.data.frame(output) #转换成数据框
p_line=ggplot(output)+
geom_line(aes(x=output[,1],y=output[,2]),color="red")+
geom_line(aes(x=output[,1],y=output[,3]),color="green")
p_line
```

⑨ 预测结果评价

MSE、RMSE、MAE、MAPE、R^2、adjusted- R^2

4.拓展练习

【请思考】

1、上面的工作能实现真正的预测吗?

提示：昨天预测今天，合并列函数cbind()

2、如何进行随机采样?

提示： tdata<-data[sample(nrow(data),1700, replace = FALSE, prob = NULL),]

3、如何将预测的标准化数据还原真实值?

提示：airdata数据采用的标准化方法是：

y=(x-MinValue)/(MaxValue-MinValue)

假设：Max=480μg/m^3 ， Min=15μg/m^3

4、如何用一个网络同时预测PM2.5、PM10、SO2和NOX?

5.拓展学习——词频统计

词频统计：文本信息，如新闻、微博、书籍以及病历等，我们可以采用分词提取关键要素进行统计。可应用于：

- **应试** 如：通过高频词汇，知道考试中常常考到单词和短语。
  - CET4，CET6等等。
  - 面试
- **热点**：通过词频统计政府可以研究相关热点走势等等。如：
  - 通过微博看新冠病毒发展
- **科学研究** 如：
  - 基于计算机的词频统计研究 ————考证《红楼梦》作者是否唯一
- **政府** 更好更快地利用大数据资源，加速和提升我国经济社会发展速度及其质量的对策建议。
- ……

5.拓展学习——词频统计

jieba中文分词包由Qin Wenfeng开发，支持最大概率法、隐式马尔科夫模型、索引模型、混合模型四种分词模式，以及标记词性、提取关键词、计算Simhash和海明距离，加载用户自定义词库(包括停用词词库)，繁体中文分词。基本步骤：

首先安装和载入分词包library(jiebaR)

然后构建分词器worker()

最后通过segment()进行分词

具体参考“R语言—jieba分词”

https://mp.weixin.qq.com/s?src=11&timestamp=1607850763&ver=2764&signature=lLcxQBLf-n8*6YnnyVzXTfOj-8drv-GcPP4kluAzuMpk0sxzwjeSmSH6xW9QtMO1A68Xkb-TOCfZMvLOsfO2Oc*tH1lmZ*nfxFkKtLOhXepLO0GRNLm6o1mK*5kUg0Cb&new=1

6.课下作业

你和你的小伙伴完成你们的Project。

下周开始展示。

提示：大家共同评出两个优秀成果，有优秀课程学习证明和小礼物!

荣誉证书

郑迪亚、孙婧瑶、张成龙 同学：

在 2018-2019(1)计算机基础（理）课程的问题求解实践项目中表现优异，特此证明。

南开大学公共计算机基础教学部

2018 年 12 月 31 日

Certificate of Excellence

This is to certify that:

【Diya Zheng】

【Jingyao Sun】

【Chenglong Zhang】

are honored with the Best Course Project Award in cooperative problem solving in Fundamentals of Computer for Science and Engineering 2018-2019(1).

Hong Zhao

Department of General Computer,

Nankai University

Dec. 31th, 2018

南開大學 NANKAI UNIVERSITY, P.R.CHINA

允公允能 日新月异