



深度学习 与R语言

程显毅 施 仨 / 编 著



提供源程序代码

下载网址为 www.cmpbook.com



机械工业出版社
CHINA MACHINE PRESS

深度学习与 R 语言

程显毅 施 仨 编著



机械工业出版社

近年来,深度学习可谓是机器学习方向的明星概念,不同的深度学习模型分别在图像处理与自然语言处理等任务中取得了前所未有的好成绩。

在许多场合都有这样的需求“如何对感兴趣的领域快速理解和使用深度学习技术?”答案涉及复杂的数学、编程语言(如C、C++和Java)。但随着R的兴起,现在使用深度学习技术比以往更容易。因为R易学易用,不要求很扎实的编程基础,它被广泛地应用于机器学习实践和教学中。即使对R语言不是很了解的用户也可以通过一些包来搭建深度学习网络。

全书11章,分为原理篇(第1~8章)和应用篇(第9~11章)。原理篇按照深度学习的发展过程,主要讨论了浅层神经网络、深度神经网络、卷积神经网络、递归神经网络、自编码网络、受限玻耳兹曼机和深度置信网。应用篇讨论R环境部署深度学习环境的一些策略,包括:MXNetR、H2O和其他深度学习R包以及一些典型的应用。

本书可用作本科高年级机器学习课程参考书或数据科学课程教材,也可供对人工智能、机器学习感兴趣的读者参考阅读。

图书在版编目(CIP)数据

深度学习与R语言/程显毅,施佺编著. —北京:机械工业出版社,2017.6
ISBN 978-7-111-57073-8

I. ①深… II. ①程… ②施… III. ①学习(人工智能) ②程序语言-程序设计 IV. ①TP18 ②TP312

中国版本图书馆CIP数据核字(2017)第125518号

机械工业出版社(北京市百万庄大街22号 邮政编码 100037)

策划编辑:汤枫 责任编辑:汤枫

责任校对:张艳霞 责任印制:李飞

北京振兴源印务有限公司印刷

2017年6月第1版·第1次印刷

184mm×260mm·13.5印张·318千字

0001-3000册

标准书号:ISBN 978-7-111-57073-8

定价:49.00元

凡购本书,如有缺页、倒页、脱页,由本社发行部调换

电话服务

网络服务

服务咨询热线:(010) 88361066

机工官网:www.cmpbook.com

读者购书热线:(010) 68326294

机工官博:weibo.com/cmp1952

(010) 88379203

教育服务网:www.cmpedu.com

封面无防伪标均为盗版

金书网:www.golden-book.com

前 言

深度学习是机器学习领域一个新的研究方向，近年来在语音识别、计算机视觉等多类应用中取得突破性的进展，其动机在于建立模型模拟人类大脑的神经连接结构，进而给出数据的解释。

深度学习之所以被称为“深度”，是相对支持向量机（Support Vector Machine, SVM）、提升方法（Boosting）、最大熵方法等“浅层学习”方法而言的。浅层学习依靠人工经验抽取样本特征，网络模型学习后获得的是没有层次结构的单层特征；而深度学习通过对原始信号进行逐层特征变换，将样本在原空间的特征表示变换到新的特征空间，自动地学习得到层次化的特征表示，从而更有利于分类或特征的可视化。

本书的目的是把强大的深度学习技术传递到想实践深度学习的读者手中，而不是让读者理解深度学习的理论细节。因此，内容重点是数据分析和建模，注意力完全集中在能有效工作的深度学习技术、理念和策略上，这样可以用最少的时间快速消化和部署深度学习应用。

本书具有以下特点：

1) 让读者清楚如何在 R 中使用深度学习。书中给出了大量的深度学习应用案例，这些例子可以直接输入到 R 环境中运行，指导读者一步一步构建和部署深度学习模型。

2) 深度学习不需要很深的数学基础作为前提。无论你是谁？无论你来自哪里？无论你的受教育背景如何？都有能力使用这本书中论述的方法。

3) 每一章都提供了进一步学习的详细参考资料，并且大部分是免费的。

图的上半部分给出了本书的学习路线，第 1、5、7 章相对独立，是学习深度神经网络的基础，虚线表示分类，实线表示支持，核心是第 3、4、8 章，每一章下面的英文表示依赖的 R 包。图的下半部分是深度学习 R 包与各章的关系。

本书主要参考了 N. D. Lewis 所著的《Deep Learning Made Easy with R——A Gentle Introduction for Data Science》，在此表示感谢。感谢研究生谢璐、胡海涛、周春瑜、姚泽峰和沈佳杰在材料整理方面所做的工作。感谢胡彬、陈晓勇、李跃华老师

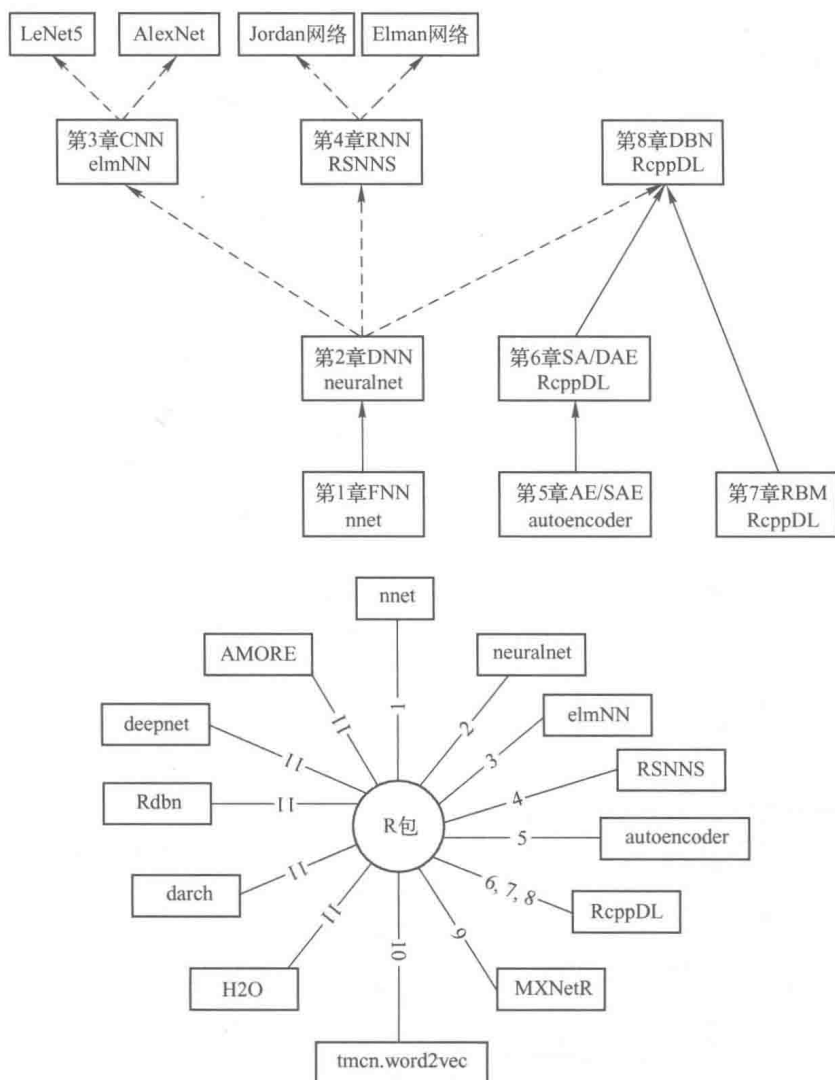


图 深度学习 R 包与各章的关系

对本书提出的宝贵意见。

本书的出版得到南通大学－南通智能信息技术联合研究中心开放课题项目的资助和南通大学学术著作出版基金资助。

深度学习领域发展迅猛，对许多问题作者并未做深入研究，一些有价值的新内容也来不及收入本书。加上作者知识水平和实践经验有限，书中难免存在不足之处，敬请读者批评指正。

编 者

目 录

前言

第1章 引言	1
1.1 关于深度学习	1
1.1.1 深度学习兴起的渊源	1
1.1.2 深度学习总体框架	3
1.1.3 深度学习本质	4
1.1.4 深度学习应用	5
1.2 前向反馈神经网络 FNN	5
1.2.1 多层感知器	5
1.2.2 神经元的作用	6
1.2.3 激活函数	8
1.2.4 学习算法	9
1.3 R 语言基础	11
1.3.1 入门	11
1.3.2 基本语法	13
1.3.3 数据	14
1.3.4 绘图	19
1.3.5 数据准备	21
1.3.6 基本运算	23
1.4 FNN 的 R 实现	23
1.5 学习指南	27
第2章 深度神经网络 DNN	28
2.1 DNN 原理	28
2.2 DNN 应用	30
2.2.1 提高雾天视觉能见度	31
2.2.2 打击黑客和网络犯罪	31

2.2.3 图像压缩	33
2.2.4 函数逼近	34
2.3 DNN 应用需要注意的一些问题	38
2.3.1 神经元数量	38
2.3.2 最佳层数的选择	39
2.3.3 训练时间过长	39
2.3.4 过拟合	40
2.4 DNN 应用技巧	40
2.5 单响应变量 DNN 的 R 实现	42
2.6 多响应变量 DNN 的 R 实现	46
2.7 学习指南	51
第 3 章 卷积神经网络 CNN	52
3.1 CNN 原理	52
3.1.1 局部感知	53
3.1.2 权值共享	54
3.1.3 多卷积核	54
3.1.4 池化	56
3.2 多层卷积	57
3.2.1 ImageNet-2010 网络结构	57
3.2.2 DeepID 网络结构	58
3.3 CNN 的 R 实现	58
3.4 学习指南	59
第 4 章 递归神经网络 RNN	61
4.1 RNN 原理	61
4.2 Elman 网络	61
4.2.1 承接层神经元的作用	62
4.2.2 信息流动	62
4.2.3 Elman 网络应用	63
4.3 Jordan 网络	65
4.3.1 Jordan 网络结构	65
4.3.2 Jordan 网络应用	65

4.4	RNN 的 R 实现	66
4.5	学习指南	77
第 5 章	自编码网络 AE	79
5.1	无监督学习过程	79
5.2	AE 基本结构	80
5.2.1	降维问题	81
5.2.2	特征抽取	82
5.3	稀疏自动编码网络 SAE	83
5.3.1	Kullback - Leibler 散度	84
5.3.2	使用 SAE 注意事项	84
5.4	SAE 的 R 实现	85
5.5	学习指南	94
第 6 章	堆栈自编码网络 SA	95
6.1	SA 原理	96
6.2	SA 的 R 实现	97
6.3	降噪自编码网络 DAE	99
6.3.1	随机掩蔽的椒盐噪声	100
6.3.2	DAE 基本任务	100
6.3.3	标准化堆栈降噪自编码网络	100
6.4	DAE 的 R 实现	101
6.5	学习指南	105
第 7 章	受限玻耳兹曼机 RBM	107
7.1	RBM 原理	107
7.1.1	玻耳兹曼机的四类知识	107
7.1.2	能量和概率的作用	108
7.1.3	联合概率分布表示的自编码网络	109
7.1.4	模型学习的目标	109
7.2	训练技巧	110
7.2.1	技巧 1: Gibbs 采样	110
7.2.2	技巧 2: 最小化 KL 距离	110
7.2.3	技巧 3: 使用 RLU 激活函数	111

7.2.4 技巧4: 模拟退火	111
7.3 对深度学习的质疑	112
7.4 RBM 应用	112
7.4.1 肝癌分类的 RBM	113
7.4.2 麻醉镇定作用预测的 RBM	114
7.5 RBM 的 R 实现	115
7.6 学习指南	118
第8章 深度置信网络 DBN	120
8.1 DBN 原理	120
8.2 应用案例	121
8.3 DBN 的 R 实现	123
8.4 学习指南	126
第9章 MXNetR	128
9.1 MXNet 技术特性	128
9.2 MXNetR 安装	129
9.2.1 安装 MXNet 基本需求	129
9.2.2 MXNet 云设置	130
9.2.3 MXNet 安装方法	130
9.2.4 MXNetR 安装方法	131
9.2.5 常见的安装问题	132
9.3 MXNetR 在深度学习中的应用	133
9.3.1 二分类模型	133
9.3.2 回归模型与自定义神经网络	135
9.3.3 手写数字竞赛	137
9.3.4 图像识别应用	141
9.4 学习指南	143
第10章 word2vec 的 R 语言实现	144
10.1 word2vec 词向量由来	144
10.1.1 统计语言模型	144
10.1.2 神经网络概率语言模型	145
10.2 word2vec——词向量特征提取模型	145

10.2.1 词向量	145
10.2.2 CBOW 的分层网络结构——HCBOV	146
10.2.3 word2vec 流程	150
10.3 word2vec 的 R 实现	151
10.3.1 tmcn.word2vec 包	151
10.3.2 word2vec 自编译函数	153
10.3.3 使用 tmcn.word2vec 和 word2vec 注意的问题	154
10.4 学习指南	155
第 11 章 R 语言其他深度学习包	156
11.1 darch 包	156
11.2 Rdbn 包	161
11.2.1 Rdbn 原理	161
11.2.2 Rdbn 安装	161
11.2.3 Rdbn 应用	162
11.3 H2O 包	164
11.3.1 H2O 原理	164
11.3.2 H2O 应用	167
11.4 deepnet 包	169
11.5 mbench 包	171
11.6 AMORE 包	175
11.7 学习指南	177
附录	178
附录 A 深度学习发展史	178
附录 B 深度学习的未来——GAN	195
附录 C R 包分类	198
参考文献	201
后记	204

第1章 引言

1.1 关于深度学习

机器学习是人工智能的一个分支，在很多时候几乎成为人工智能的代名词。简单来说，机器学习就是通过算法使得机器能从大量历史数据中学习规律，从而对新的样本做智能识别或对未来做预测。从20世纪80年代末期以来，机器学习的发展大致经历了两次浪潮：浅层学习(Shallow Learning)和深度学习(Deep Learning)。

1.1.1 深度学习兴起的渊源

深度学习起源于对神经网络的研究，20世纪60年代，受神经科学对人脑结构研究的启发，为了让机器也具有类似人一样的智能，人工神经网络被提出用于模拟人脑处理数据的流程。最著名的学习算法称为感知机。但随后人们发现，两层结构的感知机模型不包含隐层单元，输入是人工预先选择好的特征，输出是预测的分类结果，因此只能用于学习固定特征的线性函数，而无法处理非线性分类问题。Minsky等指出了感知机的这一局限，由于当时其他人工智能研究学派的抵触等原因，使得对神经网络的研究遭受到巨大的打击，陷入低谷。直到20世纪80年代中期，反向传播(Back Propagation, BP)算法的提出，提供了一条如何学习含有多隐层结构的神经网络模型的途径，让神经网络研究得以复苏。

由于增加了隐层单元，多层神经网络比感知机具有更灵活且更丰富的表达力，可以用于建立更复杂的数学模型，但同时也增加了模型学习的难度，特别是当包含的隐层数量增加的时候，使用BP算法训练网络模型时，常常会陷入局部最小值，而在计算每层结点梯度时，在网络低层方向会出现梯度衰竭的现象。因此，训练含有许多隐层的深度神经网络一直存在困难，导致神经网络模型的深度受到限制，制约了其性能。

2006 年之前,大多数机器学习仍然在探索浅层结构 (Shallow - structured), 这种结构包含了一层典型的非线性特征变换的单层, 而缺乏自适应非线性特征的多层结构。如隐马尔可夫模型 (HMM)、线性或非线性动态系统、条件随机域 (CRFs)、最大熵 (Max - entropy) 模型、支持向量机 (SVM)、逻辑回归、内核回归和具有单层隐藏层的多层感知器 (MLP) 神经网络。这些浅层学习模型的共性是由仅有的单层组成的简单架构负责转换原始输入信号或输入特征为特定问题特征空间时, 其过程不可观察。以支持向量机为例, 它是一种浅层线性独立模型, 当使用核技巧时, SVM 具有一个特征转换层, 否则特征转换层个数为 0。浅层架构在许多简单或受限问题中, 早已被证明卓有成效, 但是由于它们的有限建模与表现能力, 导致在处理涉及自然信号如人的讲话、自然的声音和语言、自然的图像和视觉场景等更为复杂的现实应用时, 产生了困难。

在实际应用中, 如对象分类问题 (对象可以是文档、图像、音频等), 人们不得不面对的一个问题是如何用数据来表示这个对象, 当然这里的数据并非初始的像素或者文字, 也就是这些数据是比初始数据具有更为高层的含义, 这里的数据往往指的是对象的特征。例如人们常常将文档、网页等数据用词的集合来表示, 根据文档的词集合表示到一个词组短语的向量空间 (Vector Space Model, VSM) 中, 然后才能根据不同的学习方法设计出适用的分类器来对目标对象进行分类。因此, 选取什么特征或者用什么特征来表示某一对象对于解决一个实际问题非常得重要。然而, 人为地选取特征的时间代价是非常昂贵的。而所谓的启发式算法得到的结果往往不稳定, 结果的好坏经常是依靠经验和运气。于是, 人们考虑到利用自动学习来完成特征抽取这一任务。深度学习的产生就是缘于此任务, 它又被称为无监督的特征学习, 从这个名称就可以知道这是一个没有人为参与的特征选取方法。

深度结构学习, 或者通常更多人称之为深度学习, 从 2006 年开始作为一个新兴的领域出现在机器学习的研究中。深度学习的概念是 2006 年由 Geoffrey Hinton 等人在《Science》上发表的一篇文章 *Reducing the dimensionality of data with neural networks* 提出来的, 开启了深度学习在学术界和工业界的浪潮。这篇文章有两个主要观点: ①多隐层的人工神经网络具有优异的特征学习能力, 学习得到的特征对数据有更本质的刻画, 从而有利于可视化或分类; ②深度神经网络在训练上的难度, 可以通过“逐层初始化”来有效克服, 在这篇文章中, 逐层初始化是通过无监督学习实现的。2006 年的另外 3 篇论文^[2-4]改变了训练深度架构失败的状况, 由 Hinton 革命

性地在深度置信网络（Deep Belief Networks, DBN）上的工作所引领。

1.1.2 深度学习总体框架

随着计算机处理器速度越来越快、存储器容量越来越大以及数据表达的形式多样化，各类大小事务都能使用深度学习进行实时数据分析。

深度学习是利用多层感知机学习模型来对数据进行有监督或者无监督学习。模型中的层是由多段非线性数据变换构成的，层次越高，对数据特征的表达就越抽象。

图 1.1 说明了深度学习的金字塔结构。在底部是数据学习的两种核心形式：有监督学习和无监督学习。核心要素非线性变换位于金字塔结构的中心，金字塔上部是各种各样的神经网络。

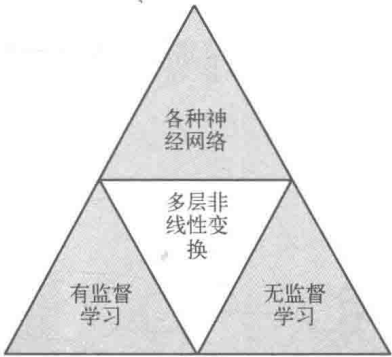


图 1.1 深度学习金字塔

- 1) 有监督学习：训练数据包含了已知结果，根据这些结果来训练模型。
- 2) 无监督学习：训练数据不含任何已知结果，在这种情况下，算法需要自身发现数据间的关系。

当使用 R 来实现深度学习时，采用的总体方法如图 1.2 所示。无论开发了什么特定的模型，最终还是会回到这个基本框架上来。数据输入到模型中，并且被多个非线性层过滤，最后一层由分类器构成，这个分类器决定了感兴趣的目标所属的类别。



图 1.2 深度学习总体框架

深度学习的目的在于预测一个响应变量（分类变量）。这在某种程度上类似于线性回归做的事情。在线性回归中，一个线性模型使用一组独立变量（亦称属性或者特征）来预测响应变量。尽管如此，传统的线性回归模型并不被认为是深度的。

其他流行的学习方法，如决策树、随机森林以及支持向量机，都不是深度结构。决策树和随机森林使用的都是原始输入数据，对这些数据它们不做变换也不产生新的特征；支持向量是由一个核和一个线性变换构成的模型。

1.1.3 深度学习本质

深度学习能够以适当数目、并行、非线性步骤对非线性数据进行分类或者预测，它的威力正是源于此。从原始输入数据一直到数据分类结果，一个深度学习模型对输入数据进行分层次的学习，每一层从前一层的输出中提取特征。

深度学习模型是多隐藏层的神经网络。最简单的深度神经网络如图 1.3 所示，它包含了至少两个隐藏层。在这个神经网络中，每一层都把前一层的输出看成是本层的输入。

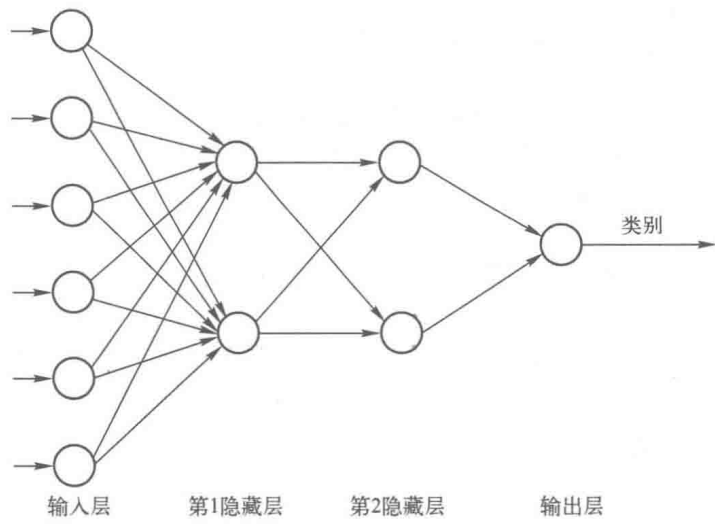


图 1.3 两个隐藏层的前馈神经网络

深度多层神经网络包含了许多非线性变换，这些变换使得深度多层神经网络能够简洁地近似表达复杂的非线性函数。深度多层神经网络善于分辨出数据中的复杂模式，它们已经被用来解决实际问题，如计算机视觉、自然语言处理、语音识别等。

1.1.4 深度学习应用

深度学习技术已经商业化了，应用于健康护理产业和医学图像处理，自然语言处理已经被用于广告业来改善点击率。微软、谷歌、IBM、雅虎、推特、百度、Paypal 以及 Facebook 都正在利用深度学习来理解用户的行为，目的是能够有针对性地推荐服务与产品。深度学习技术处处存在，很难想出在商业活动的哪个领域不需要利用深度学习，以下列出深度学习商业化的一些领域：

- 1) 过程建模与控制。
- 2) 健康诊断。
- 3) 投资证券管理。
- 4) 军事目标识别。
- 5) 核磁共振图像与 X 射线图像分析。
- 6) 银行以及其他财务机构对个人的信用评价。
- 7) 市场推广。
- 8) 语音识别。
- 9) 股票市场预测。
- 10) 文本搜索。
- 11) 财务欺诈检测。
- 12) 光学字符识别。

1.2 前向反馈神经网络 FNN

1.2.1 多层感知器

前向反馈神经网络（Feed - forward Neural Networks, FNN）的引入显著改善了预测的准确率，并且随着 FNN 训练方法的逐渐完善使得商业和科学研究持续受益。

对模拟人脑的生理结构以及功能的需求催生了 FNN。尽管这样的需求从来没有具体化，但人们很快发现 FNN 在分类和预测任务中表现得相当好。

FNN 能被用来解决许多分类问题。这是因为从理论上讲，它能逼近任何可以计

算的函数。在实践中，有些问题对错误比较宽容，有些问题不能简单地应用严格的规则，神经网络在应对这些问题时特别有效。

一个 FNN 是由一些相互连接的神经元构建而成。这些神经元通常被安置到一些层中。一个典型的前馈神经网络至少有一个输入层、一个隐藏层和一个输出层。输入层的神经元数对应于输入神经网络的特征或者属性的数目。这类似于在线性回归模型中使用的自变量。输出神经元的数目对应于分类数或者预测的类别数。隐藏层结点大致上是用来执行对原始输入特征的非线性变换。

最简单形式的前馈神经网络通过网络来传播特征信息，用来做出预测。前馈神经网络的输出要么是连续的（回归），要么是离散的（分类）。图 1.4 说明了一种典型的前馈神经网络拓扑结构。它有 7 个输入神经元、一个具有 5 个神经元的隐藏层，以及 3 个输出神经元。信息从输入特征前馈到隐藏层，并且接着再前馈到做出分类或者回归预测的输出层。它之所以叫前馈神经网络是因为信息是向前流动通过整个网络的。

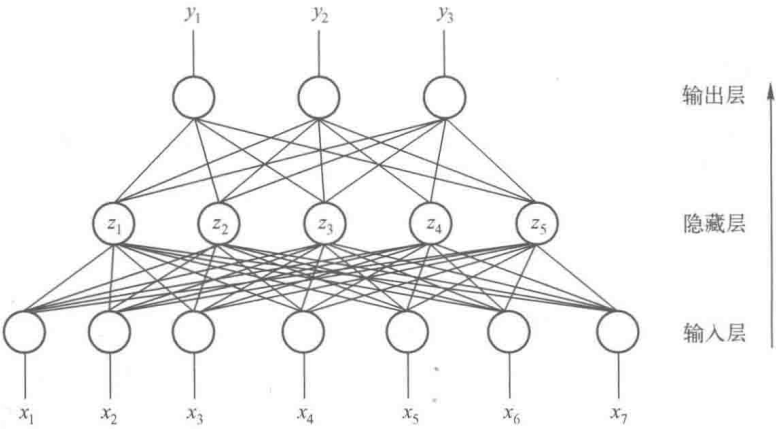


图 1.4 一种基本的 FNN

1.2.2 神经元的作用

图 1.5 说明了一个生物神经元是如何工作的。生物神经元通过电信号在相互之间传递信号或者消息。相邻的神经元通过它们的树突接收这些信号。信息从树突流向称为胞体的主细胞体，再通过轴突到轴突末端。实质上，生物神经元就是彼此之间相互传递与各种生物功能相关信息的计算机器。

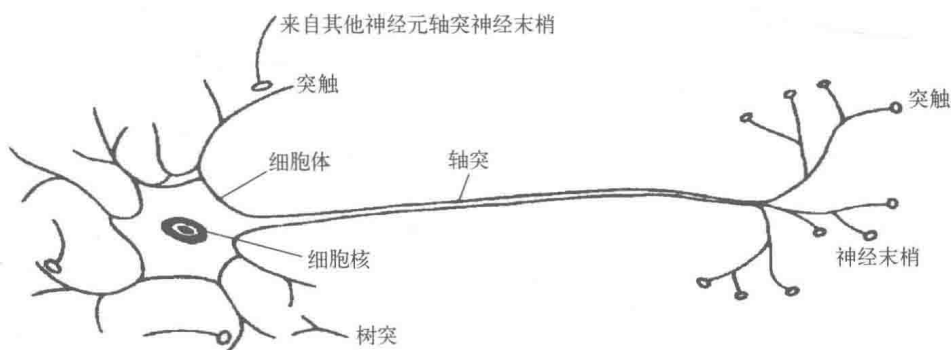


图 1.5 生物神经元

神经元是人工神经网络的核心，它是最基本的处理要素。输入层神经元接收进入网络的信息，这些信息通过一个数学函数来处理，接着被传递到隐藏层神经元。这种信息经过隐藏层神经元处理传递到输出层神经元。信息通过一个激活函数进行处理，这是神经元工作的关键。激活函数模拟了大脑神经元，大脑神经元是否工作取决于输入信号的强度。

第一个感知器模型是 1958 年在康奈尔航空实验室中开发的。它由无反馈的三层所构成：

- 1) 传递输入数据到第二层（视网膜）。
- 2) 把加权输入和阈值阶梯函数组合起来（神经元连接）。
- 3) 输出层。

处理的结果接着被加权并且被传递到下一层的神经元中。实质上，神经元通过加权和相互激活，根据处理信息的权重来确定两个神经元之间的联系强度大小。

每一个神经元都包含一个激活函数和一个阈值。阈值是输入要激活神经元所必须的最小值。因此，神经元的任务就是，在传递输出到下一层之前，求输入信号的加权和，以及执行一个激活函数。

综上，输入层对输入数据进行求和；中间层神经元对由输入层神经元传输而来的加权信息进行求和；并且输出层对由中间层神经元传输而来的加权信息进行求和。

图 1.6 说明了神经元是如何工作的。给定一个输入样本的特征 $\{x_1, \dots, x_n\}$ 和一个权重 W_i ；接着计算神经元输入的加权和，公式如下：

$$y = \sigma(\text{net}) = \sigma \left(\sum_{i=1}^k (W_i x_i + b_i) \right)$$