最新打书最PDF
最新打书最PDF
最新打书最PDF
最新打书最PDF

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

非盗品，仅供非商业用途或交流学习使用
严禁网络传播本PDF，侵害责任自负！
版权所有，严禁传播，违者必追法律责任！
澳洲学习指南
深入浅出
Python机器学习
版权所有，严禁传播，违者必追法律责任！
非盗品，仅供非商业用途或交流学习使用
严禁网络传播本PDF，侵害责任自负！
版权所有，严禁传播，违者必追法律责任！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权相关注意事项：
1、书籍版权证意味出版社所有
2、本PDF来自于个人的收藏学习，经过整理加工
3、本PDF仅提供学习使用或用于非商业用途或个人交流研究学习使用
4、本PDF获得是不用在互联网上以任何形式进行传播，流程是通过迅雷等任何软件，流程者自负
5、如果您非常喜欢，请一定购买正版实体书，多多支持版权拥有者的作者和版权的出版社！当然，如果您有内容不满入目，请退款下，你也可以选择提前退款服务PDF
6、技术有书籍是来源网络知识，不是来自收藏的，你得到了书籍不非常你得到了知识，所以请不要用你书籍出版社发现法律责任，要知其然知其所然！经常阅读
7、请子下载PDF是24小时内所获资源有效期限！

版权所有，严禁传播，违者自负法律责任！

目 录

第1章 概 述

1.1 什么是机器学习——从一个小故事开始 / 002
1.2 机器学习的一些应用场景——蝙蝠公司的业务单元 / 003
1.3 机器学习应该如何入门——世上无难事 / 005
1.4 有监督学习与无监督学习 / 007
1.5 机器学习中的分类与回归 / 008
1.6 模型的泛化、过拟合与欠拟合 / 008
1.7 小结 / 009

第2章 基于Python语言的环境配置

2.1 Python的下载和安装 / 012
2.2 Jupyter Notebook的安装与使用方法 / 013
2.2.1 使用pip进行Jupyter Notebook的下载和安装 / 013
2.2.2 运行Jupyter Notebook / 014
2.2.3 Jupyter Notebook的使用方法 / 015
2.3 一些必需库的安装及功能简介 / 017
2.3.1 Numpy——基础科学计算库 / 017
2.3.2 Scipy——强大的科学计算工具集 / 018
2.3.3 pandas——数据分析的利器 / 019
2.3.4 matplotlib——画出优美的图形 / 020

2.4 scikit-learn——非常流行的 Python 机器学习库 / 021
2.5 小结 / 022

第 3 章 K 最近邻算法——近朱者赤，近墨者黑

3.1 K 最近邻算法的原理 / 024
3.2 K 最近邻算法的用法 / 025
3.2.1 K最近邻算法在分类任务中的应用 / 025
3.2.2 K最近邻算法处理多元分类任务 / 029
3.2.3 K最近邻算法用于回归分析 / 031
3.3 K 最近邻算法项目实战——酒的分类 / 034
3.3.1 对数据集进行分析 / 034
3.3.2 生成训练数据集和测试数据集 / 036
3.3.3 使用K最近邻算法进行建模 / 038
3.3.4 使用模型对新样本的分类进行预测 / 039
3.4 小结 / 041

第 4 章 广义线性模型——“耿直”的算法模型

4.1 线性模型的基本概念 / 044
4.1.1 线性模型的一般公式 / 044
4.1.2 线性模型的图形表示 / 045
4.1.3 线性模型的特点 / 049
4.2 最基本的线性模型——线性回归 / 050
4.2.1 线性回归的基本原理 / 050
4.2.2 线性回归的性能表现 / 051
4.3 使用 L2 正则化的线性模型——岭回归 / 053
4.3.1 岭回归的原理 / 053
4.3.2 岭回归的参数调节 / 054
4.4 使用 L1 正则化的线性模型——套索回归 / 058
4.4.1 套索回归的原理 / 058
4.4.2 套索回归的参数调节 / 059
4.4.3 套索回归与岭回归的对比 / 060

4.5　小结 / 062

第 5 章　朴素贝叶斯——打雷啦，收衣服啊

5.1　朴素贝叶斯基本概念 / 064

5.1.1　贝叶斯定理 / 064

5.1.2　朴素贝叶斯的简单应用 / 064

5.2　朴素贝叶斯算法的不同方法 / 068

5.2.1　贝努利朴素贝叶斯 / 068

5.2.2　高斯朴素贝叶斯 / 071

5.2.3　多项式朴素贝叶斯 / 072

5.3　朴素贝叶斯实战——判断肿瘤是良性还是恶性 / 075

5.3.1　对数据集进行分析 / 076

5.3.2　使用高斯朴素贝叶斯进行建模 / 077

5.3.3　高斯朴素贝叶斯的学习曲线 / 078

5.4　小结 / 080

第 6 章　决策树与随机森林——会玩读心术的算法

6.1　决策树 / 082

6.1.1　决策树基本原理 / 082

6.1.2　决策树的构建 / 082

6.1.3　决策树的优势和不足 / 088

6.2　随机森林 / 088

6.2.1　随机森林的基本概念 / 089

6.2.2　随机森林的构建 / 089

6.2.3　随机森林的优势和不足 / 092

6.3　随机森林实例——要不要和相亲对象进一步发展 / 093

6.3.1　数据集的准备 / 093

6.3.2　用get_dummies处理数据 / 094

6.3.3　用决策树建模并做出预测 / 096

6.4　小结 / 098

第 7 章　支持向量机 SVM——专治线性不可分

7.1　支持向量机 SVM 基本概念　/　100
7.1.1　支持向量机SVM的原理　/　100
7.1.2　支持向量机SVM的核函数　/　102
7.2　SVM 的核函数与参数选择　/　104
7.2.1　不同核函数的SVM对比　/　104
7.2.2　支持向量机的gamma参数调节　/　106
7.2.3　SVM算法的优势与不足　/　108
7.3　SVM 实例——波士顿房价回归分析　/　108
7.3.1　初步了解数据集　/　109
7.3.2　使用SVR进行建模　/　110
7.4　小结　/　114

第 8 章　神经网络——曾入“冷宫”，如今得宠

8.1　神经网络的前世今生　/　116
8.1.1　神经网络的起源　/　116
8.1.2　第一个感知器学习法则　/　116
8.1.3　神经网络之父——杰弗瑞·欣顿　/　117
8.2　神经网络的原理及使用　/　118
8.2.1　神经网络的原理　/　118
8.2.2　神经网络中的非线性矫正　/　119
8.2.3　神经网络的参数设置　/　121
8.3　神经网络实例——手写识别　/　127
8.3.1　使用MNIST数据集　/　128
8.3.2　训练MLP神经网络　/　129
8.3.3　使用模型进行数字识别　/　130
8.4　小结　/　131

试读结束：需要全本请在线购买： www.ertongbook.com

第 9 章 数据预处理、降维、特征提取及聚类——快刀斩乱麻

9.1 数据预处理 / 134
9.1.1 使用StandardScaler进行数据预处理 / 134
9.1.2 使用MinMaxScaler进行数据预处理 / 135
9.1.3 使用RobustScaler进行数据预处理 / 136
9.1.4 使用Normalizer进行数据预处理 / 137
9.1.5 通过数据预处理提高模型准确率 / 138
9.2 数据降维 / 140
9.2.1 PCA主成分分析原理 / 140
9.2.2 对数据降维以便于进行可视化 / 142
9.2.3 原始特征与PCA主成分之间的关系 / 143
9.3 特征提取 / 144
9.3.1 PCA主成分分析法用于特征提取 / 145
9.3.2 非负矩阵分解用于特征提取 / 148
9.4 聚类算法 / 149
9.4.1 *K*均值聚类算法 / 150
9.4.2 凝聚聚类算法 / 153
9.4.3 DBSCAN算法 / 154
9.5 小结 / 157

第 10 章 数据表达与特征工程——锦上再添花

10.1 数据表达 / 160
10.1.1 使用哑变量转化类型特征 / 160
10.1.2 对数据进行装箱处理 / 162
10.2 数据“升维” / 166
10.2.1 向数据集添加交互式特征 / 166
10.2.2 向数据集添加多项式特征 / 170
10.3 自动特征选择 / 173
10.3.1 使用单一变量法进行特征选择 / 173

10.3.2 基于模型的特征选择 / 178
10.3.3 迭代式特征选择 / 180
10.4 小结 / 182

第 11 章 模型评估与优化——只有更好，没有最好

11.1 使用交叉验证进行模型评估 / 184
11.1.1 scikit-learn中的交叉验证法 / 184
11.1.2 随机拆分和“挨个儿试试” / 186
11.1.3 为什么要使用交叉验证法 / 188
11.2 使用网格搜索优化模型参数 / 188
11.2.1 简单网格搜索 / 189
11.2.2 与交叉验证结合的网格搜索 / 191
11.3 分类模型的可信度评估 / 193
11.3.1 分类模型中的预测准确率 / 194
11.3.2 分类模型中的决定系数 / 197
11.4 小结 / 198

第 12 章 建立算法的管道模型——团结就是力量

12.1 管道模型的概念及用法 / 202
12.1.1 管道模型的基本概念 / 202
12.1.2 使用管道模型进行网格搜索 / 206
12.2 使用管道模型对股票涨幅进行回归分析 / 209
12.2.1 数据集准备 / 209
12.2.2 建立包含预处理和MLP模型的管道模型 / 213
12.2.3 向管道模型添加特征选择步骤 / 214
12.3 使用管道模型进行模型选择和参数调优 / 216
12.3.1 使用管道模型进行模型选择 / 216
12.3.2 使用管道模型寻找更优参数 / 217
12.4 小结 / 220

第 13 章　文本数据处理——亲，见字如“数”

13.1　文本数据的特征提取、中文分词及词袋模型　/　222

13.1.1　使用CountVectorizer对文本进行特征提取　/　222

13.1.2　使用分词工具对中文文本进行分词　/　223

13.1.3　使用词袋模型将文本数据转为数组　/　224

13.2　对文本数据进一步进行优化处理　/　226

13.2.1　使用*n*-Gram改善词袋模型　/　226

13.2.2　使用tf-idf模型对文本数据进行处理　/　228

13.2.3　删除文本中的停用词　/　234

13.3　小结　/　236

第 14 章　从数据获取到话题提取——从“研究员”到“段子手”

14.1　简单页面的爬取　/　238

14.1.1　准备Requests库和User Agent　/　238

14.1.2　确定一个目标网站并分析其结构　/　240

14.1.3　进行爬取并保存为本地文件　/　241

14.2　稍微复杂一点的爬取　/　244

14.2.1　确定目标页面并进行分析　/　245

14.2.2　Python中的正则表达式　/　247

14.2.3　使用BeautifulSoup进行HTML解析　/　251

14.2.4　对目标页面进行爬取并保存到本地　/　256

14.3　对文本数据进行话题提取　/　258

14.3.1　寻找目标网站并分析结构　/　259

14.3.2　编写爬虫进行内容爬取　/　261

14.3.3　使用潜在狄利克雷分布进行话题提取　/　263

14.4　小结　/　265

第 15 章　人才需求现状与未来学习方向——你是不是下一个“大牛”

15.1　人才需求现状　/　268

15.1.1　全球AI从业者达190万，人才需求3年翻8倍　/　268

15.1.2　AI人才需求集中于一线城市，七成从业者月薪过万　/　269

15.1.3　人才困境仍难缓解，政策支援亟不可待　/　269

15.2　未来学习方向　/　270

15.2.1　用于大数据分析的计算引擎　/　270

15.2.2　深度学习开源框架　/　271

15.2.3　使用概率模型进行推理　/　272

15.3　技能磨炼与实际应用　/　272

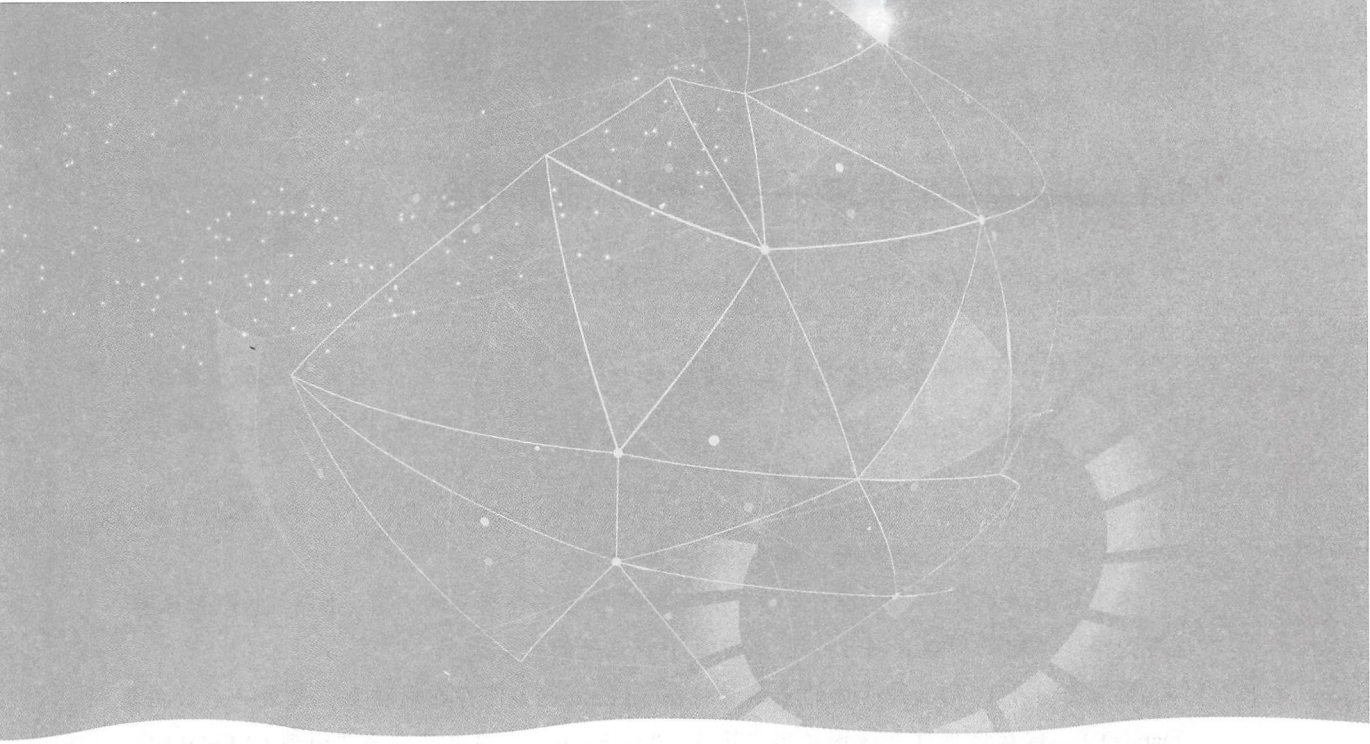
15.3.1　Kaggle算法大赛平台和OpenML平台　/　272

15.3.2　在工业级场景中的应用　/　273

15.3.3　对算法模型进行A/B测试　/　273

15.4　小结　/　274

参考文献　/　275



第1章 概 述

近年来，全球新一代信息技术创新浪潮迭起。作为全球信息领域产业竞争的新一轮焦点，人工智能的发展迎来了第三次浪潮，它正在推动工业发展进入新的阶段，掀起第四次工业革命的序幕。而作为人工智能的重要组成部分，机器学习也成了炙手可热的概念。本章将向读者介绍机器学习的基础知识，为后面的学习打好基础。

本章主要涉及的知识点有：

- 什么是机器学习
- 机器学习的主要应用场景
- 机器学习应该如何入门
- 有监督学习和无监督学习的概念
- 分类、回归、泛化、过拟合与欠拟合等概念

1.1 什么是机器学习——从一个小故事开始

要搞清楚什么是机器学习，我们可以从一个小故事开始。

小 C 是一个即将毕业的大学生、单身的小伙子，他一直在暗地里喜欢隔壁班的女神，可是又苦于没有机会接近她，于是在很长一段时间里，小 C 只能保持这种暗恋的状态。

突然有一天，在一个很偶然的机会下，小 C 得到了女神的微信号，并且添加了她。然后开始密切关注她的朋友圈，观察她的一举一动。

不久小 C 就有了重大发现，女神在朋友圈经常发三种类型的内容：书籍、电影和旅游。这可是个了不起的发现，对于小 C 来说，千载难逢的机会来了。

接下来，小 C 把女神喜欢的书名和特征（Features）保存在电脑上，做成一个数据集（Dataset），然后根据这个数据集用“算法（Algorithm）”建立了一个“模型（Model）”，并且通过这个模型预测出了女神会喜欢哪一本新书，之后小 C 买下了模型预测出来的书，作为礼物送给了女神。

收到新书的女神很开心，也对小 C 产生了好感。

后来小 C 又用同样的方法预测出了女神喜欢的电影，并买票请女神去看。不出所料，每次女神的观影体验都棒极了，两个人的关系也越来越近。

再后来，小 C 又预测了女神会喜欢的旅游地点，订好机票和酒店，对女神发出了邀请。当然，女神不会拒绝小 C 了，因为这次旅游的目的地可是她一直想去的地方呢！

整个旅途愉快极了，小 C 总能像手术刀一样精准地切到女神最感兴趣的话题上。女神觉得太不可思议了，她问小 C：“为什么你会这么了解我呢？”小 C 按捺住内心的喜悦，故作镇定地说道：“这是机器学习的力量。”

“什么是机器学习啊？”女神不解。

是时候让小 C 展现出扎实的学术底蕴了，他抬头 45° 仰望星空，深沉地说道：

“机器学习，最早是由一位人工智能领域的先驱，Arthur Samuel（见图 1-1），在 1959 年提出来的。本意指的是一种让计算机在不经过明显编程的情况下，对数据进行学习，并且做出预测的方法，属于计算机科学领域的一个子集。公认的世界上第一个自我学习项目，就是 Samuel 跳棋游戏。而我也是通过机器学习的方法，通过你在社交媒体的数据预测出你的喜好的。”

试读结束：需要全本请在线购买： www.ertongbook.com

图 1-1 Arthur Samuel 和他的跳棋游戏

毫无悬念地，女神对小 C 产生了深深的崇拜感，并且芳心暗许。从此以后，两个人走在了一起，并过上了幸福的生活。

对于一部童话来说，故事到这里就可以结束了。可是对于一本机器学习的入门书来说，我们才刚刚开始。

有了女朋友的小 C 也要背负起自己的责任了，他需要一份工作，才能为两个人的生活提供经济来源。很幸运的是，他通过校园招聘进入了国内最大的互联网公司——蝙蝠公司，成为一名机器学习工程师，从此开始了他的职业生涯。

1.2 机器学习的一些应用场景——蝙蝠公司的业务单元

小 C 入职的蝙蝠公司，作为国内互联网行业的龙头企业，其业务覆盖面十分广泛，包括电子商务、社交网络、互联网金融以及新闻资讯等。每一个方向在内部都被称为一个 BU（业务单元）。每个 BU 相对独立运作，有自己完善的体系。但机器学习技术，在每个 BU 都有非常深入的应用。下面我们来大致了解一下。

1. 电子商务中的智能推荐

蝙蝠公司的电子商务 BU 是国内最大的在线零售平台，其用户接近 5 亿，每天在线商品数超过 8 亿，平均每分钟会售出 4.8 万件商品。正因此，电子商务 BU 拥有海量的用户和商品数据。当然，为了让平台的成交总额（Gross merchandise Volume，GMV）不断提高，电子商务 BU 必须精确地为用户提供商品优惠信息。和小 C 预测女神的喜好类似，电子商务 BU 要通过机器学习，来对用户的行为进行预测。但在如此海量的数据下，模型要比小 C 的模型复杂很多。

比如某个男性用户的浏览记录和购买记录中有大量的数码产品，而且系统识别出该

用户访问平台时使用的设备是 iPhone 7，则算法很有可能会给该用户推荐 iPhone X 的购买链接。而另外一个女性用户浏览和购买最多的是化妆品和奢侈品，那么机器就会把最新款的 Hermès 或者 Chanel 产品推荐给她。

2. 社交网络中的效果广告

蝙蝠公司旗下的社交网络平台目前有超过 9 亿的月活跃用户（Monthly Active Users，MAU），其主要盈利模式是通过在社交网络中投放效果广告，从而为商家提供精准营销的服务。在这种盈利模式下，该 BU 需要保证广告的投放尽可能精准地到达目标受众，并转化为销售。因此需要机器学习算法来预测用户可能感兴趣的广告内容，并且将符合要求的内容展示给用户。

比如用户经常转发或点赞和汽车有关的信息，那么系统就会把某品牌新车上市的广告展示给用户，而如果用户经常转发或点赞的是和时尚相关的信息，那么系统推荐的广告就会是新一季的服装搭配潮流等。

3. 互联网金融中的风控系统

蝙蝠公司旗下另外一个业务单元——互联网金融事业部，主要是为用户提供小额贷款和投资理财服务。目前该 BU 拥有 4.5 亿用户，具有每秒处理近 9 万笔支付的能力。而且它的资产损失比率仅有 0.001%，这是一个非常恐怖的数字！要知道全球最老牌的在线支付工具资产损失比率也要 0.2%。要达到如此低的资产损失比率，必须要有强大的风控系统作为支撑，而风控系统背后，就是机器学习算法的应用。

例如，在这个场景中，风控系统要能够收集已知的用户欺诈行为，并对欺诈者的行为数据进行分析，再建立模型，在类似的欺诈行为再度发生之前就把它们扼杀在摇篮里，从而降低平台的资产损失比率。

4. 新闻资讯中的内容审查

蝙蝠公司旗下的新闻资讯业务单元的表现也同样让人眼前一亮。这个 BU 的产品主要是新闻客户端 APP，据称其激活用户数已经超过 6 亿，而平均每个用户使用时长达到了 76 分钟。而令人咂舌的是，这个业务单元下据说没有编辑人员，所有的内容处理都是通过机器学习算法自动完成的。

比如该产品的“精准辟谣”功能，就是主要依赖机器学习的算法，对内容进行识别。如果判断为是虚假信息，则会提交给审核团队，审核属实之后，虚假信息就会被系统屏蔽，不会给用户进行推送。

当然，蝙蝠公司的业务单元远远不止上述这几个，同时机器学习在这些业务单元中的应用也远远不止上述这几个方面。限于篇幅，本书就不再一一罗列了。

5. 机器学习在蝙蝠公司之外的应用

蝙蝠公司代表的是互联网行业，然而在互联网行业之外，机器学习也被广泛的应用。例如在医疗行业中所使用的专家系统，典型的案例就是诞生于20世纪70年代的MYCIN系统，该系统由斯坦福大学研制，它可以用患者的病史、症状和化验结果等作为原始数据，运用医疗专家的知识进行逆向推理，找出导致感染的细菌。若是多种细菌，则用0到1的数字给出每种细菌的可能性，并在上述基础上给出针对这些可能的细菌的药方。

此外，还有诸如智能物流、智能家居、无人驾驶等领域。可以说机器学习，已经非常深切地融入了我们的生活与工作当中。

6. 一些炫酷的“黑科技”

除了上述我们提到的已经广泛应用的领域，还有一些代表着未来发展趋势的案例，例如：2016年，Google旗下的AI程序AlphaGo首次战胜了人类围棋世界冠军。2017年，埃隆·马斯克创办的OpenAI公司开发的人工智能程序在电子竞技游戏DOTA中战胜了人类世界冠军Dendi。除此之外，还有很多诸如AI写新闻稿、画插画、写诗词等消息充斥着各大新闻网站的首页，仿佛用不了多久，AI就会在各个方面全面替代人类，进而统治世界了。

本书并不想争论AI究竟会让人类生活得更美好，还是会成为地球的主宰者奴役我们。在这里只想和大家一起探究一下这些案例背后的原理。不管是AlphaGo还是OpenAI，其背后的原理都是机器学习中的深度学习，它们的崛起在全球范围内掀起了一阵人工智能和深度学习的热潮。

实际上，人工智能这个概念并不是最近几年才出现的。早在20世纪60年代，人工智能就被提出，并且分为诸多学派。其中联结学派就是神经网络，或者说深度学习的代表。但受限于当时的计算能力，人工智能的发展也出现了停滞。

而随着时代的发展，现在的芯片计算能力越来越强，同时用户的数据量也越来越大，为人工智能的进一步发展提供了必要的先决条件，而机器学习、深度学习、神经网络等概念也随之火爆起来。同时在人才市场上，机器学习工程师、算法工程师、数据分析师等职位也呈现出了供不应求的场面，因此有更多的人开始投身到机器学习的研究当中。

1.3 机器学习应该如何入门——世上无难事

相信在看了上面的内容之后，一些读者朋友也已经动心，想要加入机器学习的领域当中了。保不齐能像故事中的小C一样，既能抱得美人归，又能找到一份心仪的工作。

但另外，又会担心自己基础薄弱，不知道从何入手。

不用担心！只要你肯多动脑，勤动手，相信很快就可以入门的。下面是我们给大家的一点学习方面的建议。

1. 从一种编程语言开始

如果你之前完全没有编程的基础，那么我们建议先从一门编程语言开始。目前市面上常用的编程语言有很多种，如 C++、Java、Python、R 等。那么该选择哪一种呢？不必纠结，编程语言并没有绝对的“好”和“不好”的区别，只是它们各自有各自的特点而已。而且如果你掌握了其中的一种，再学习其他的编程语言时，上手会快得多。

本书使用的语言是 Python，主要原因是：在数据科学领域，Python 已经成为了一门通用的编程语言。它既有通用编程语言的强大能力，同时还具有诸如 MATLAB 或者 R 之类针对某个专门领域语言的易用性。同时丰富和强大的库，让 Python 在数据挖掘、数据可视化、图像处理、自然语言处理等领域都有非常不俗的表现。

Python 还被称为“胶水语言”，因为它能够把用其他语言编写的各种模块轻松连接在一起。而它简洁清晰的语法和强制缩进的特点，都让 Python 对初学者非常友好。此外，它还是完全开源的，用户完全不需要支付任何费用。

由于 Python 语言的简洁性、易读性以及可扩展性，在国外用 Python 做科学计算的研究机构日益增多，一些知名大学已经采用 Python 来教授程序设计课程。众多开源的科学计算软件包都提供了 Python 的调用接口，如著名的计算机视觉库 OpenCV、三维可视化库 VTK、医学图像处理库 ITK。Google 的深度学习框架 TensorFlow 兼容得最好的语言之一，也是 Python。

2017 年 7 月 20 日，IEEE 发布 2017 年编程语言排行榜：Python 高居首位。

还有一个重要的原因，对于用户来说，Python 的学习成本是非常低的。哪怕是完全零基础的读者，在一个月左右的努力学习之后，也可以大致掌握它的基本语法和主要的功能模块。

因此我们推荐读者使用 Python 进行机器学习方面的研究与开发，在后面的章节我们会带大家配置基于 Python 的开发环境。

2. 熟悉机器学习中的基本概念

在对编程语言有了基本的掌握之后，读者朋友需要熟悉机器学习中的一些基本概念，比如什么是“有监督学习”，什么是“无监督学习”，它们之间的区别是什么，在应用方面有什么不同。另外，对机器学习的“分类”和“回归”有基本认知，清楚在什么场景下使用分类算法，在什么场景下使用回归算法。最后理解模型的“泛化”，明白在什

么情况下模型会出现“过拟合”的现象，在什么情况下会出现“欠拟合”的现象。

3. 了解机器学习中最常见的算法

在了解了基本概念之后，读者朋友就可以开始了解机器学习中最常用的一些算法了。比如K最近邻算法、线性模型、朴素贝叶斯、决策树、随机森林、SVMs、神经网络等。

在这个过程中，读者需要了解每种算法的基本原理和用途，它们的特性分别是什么，在不同的数据集中表现如何，如何使用它们建模，模型的参数如何调整等。

4. 掌握对数据进行处理的技巧

读者朋友可根据前述内容，对小数据集进行建模并且做出一些预测。但是在真实世界中，数据往往比我们拿来实验的小数据集复杂很多倍。它们的特征变量会大很多，也就是说数据的维度会高很多，同时可能完全没有训练数据集供你使用，这时候读者就需要掌握一些数据处理的技能，比如如何对数据进行降维，或者聚类，从而让数据更容易被理解，并从中找到关键点，为建模奠定基础。

5. 学会让模型更好地工作

学会用算法建模和对数据进行处理之后，读者朋友要做的是如何让模型更好地工作。例如，怎样做可以让算法的效率更高，怎样找到最适合的模型，模型最优的参数是什么，以及如何打造一个流水线，让几个模型在其中共同协作，以解决你的问题等。

6. 动手，一定要动手操作

学习一门知识最好的办法就是使用它，因此建议读者朋友一定要自己动手实操。不要嫌麻烦，尽可能把本书中的代码全部自己敲一下这样才能对内容有更加深刻的理解。如果觉得不够过瘾，还可以到知名的Kaggle大赛平台，或者“天池”算法大赛平台上，使用那些来自真实世界的数据来磨炼自己的技能。

当然，还有个更好的方法，那就是去企业中寻找一个机器学习工程师或是算法工程师的职位，在工作中学习，效果是最好的了。

1.4 有监督学习与无监督学习

在机器学习领域，有监督学习和无监督学习是两种常用的方法。有监督学习是通过现有训练数据集进行建模，再用模型对新的数据样本进行分类或者回归分析的机器学习方法。在监督式学习中，训练数据集一般包含样本特征变量及分类标签，机器使用不同的算法通过这些数据推断出分类的方法，并用于新的样本中。目前有监督学习算法已经比较成熟，并且在很多领域都有很好的表现。

而无监督学习，或者说非监督式学习，则是在没有训练数据集的情况下，对没有标签的数据进行分析并建立合适的模型，以便给出问题解决方案的方法。在无监督学习当中，常见的两种任务类型是数据转换和聚类分析。

其中数据转换的目的是，把本来非常复杂的数据集通过非监督式学习算法进行转换，使其变得更容易理解。常见的数据转换方法之一便是数据降维，即通过对特征变量较多的数据集进行分析，将无关紧要的特征变量去除，保留关键特征变量（例如，把数据集降至二维，方便进行数据可视化处理）。

而聚类算法则是通过把样本划归到不同分组的算法，每个分组中的元素都具有比较接近的特征。目前，聚类算法主要应用在统计数据分析、图像分析、计算机视觉等领域。

1.5 机器学习中的分类与回归

分类和回归是有监督学习中两个最常见的方法。对于分类来说，机器学习的目标是对样本的类标签进行预测，判断样本属于哪一个分类，结果是离散的数值。而对于回归分析来说，其目标是要预测一个连续的数值或者是范围。

这样讲可能会有一点抽象，我们还是用小 C 的例子来理解一下这两个概念。

比如，小 C 在使用算法模型预测女神的电影喜好时，他可以将电影分为“女神喜欢的”和“女神不喜欢的”两种类型，这就是二元分类，如果他要把电影分为“女神特别喜欢的”“女神有点喜欢的”“女神不怎么喜欢的”以及“女神讨厌的”四种类型，那么这就属于多元分类。

但如果小 C 要使用算法模型预测女神对某部电影的评分，例如，女神会给“速度与激情 8”打多少分，从 0 到 100，分数越高说明女神越喜欢，最终模型预测女神会给这部电影打 88 分，这个过程就称为回归。小 C 需要将女神给其他电影的评分和相对应的电影特征作为训练数据集，通过建立回归模型，来给“速度与激情 8”打分。

1.6 模型的泛化、过拟合与欠拟合

在有监督学习中，我们会在训练数据集上建立一个模型，之后会把这个模型用于新的，之前从未见过的数据中，这个过程称为模型的泛化（generalization）。当然我们希望模型对于新数据的预测能够尽可能准确，这样才能说模型泛化的准确度比较高。

那么我们用什么样的标准来判断一个模型的泛化是比较好的，还是比较差的呢？

试读结束：需要全本请在线购买： www.ertongbook.com