第 1章 数据科学概述 1

1.1　挑战　2

1.1.1　工程实现的挑战　2

1.1.2　模型搭建的挑战　3

1.2　机器学习　5

1.2.1　机器学习与传统编程　5

1.2.2　监督式学习和非监督式学习　8

1.3　统计模型　8

1.4　关于本书　10

第　2章 Python安装指南与简介：告别空谈　12

2.1　Python简介　13

2.1.1　什么是Python　15

2.1.2　Python在数据科学中的地位　16

2.1.3　不可能绕过的第三方库　17

2.2　Python安装　17

2.2.1　Windows下的安装　18

2.2.2　Mac下的安装　21

2.2.3　Linux下的安装　24

2.3　Python上手实践　26

2.3.1　Python shell　26

2.3.2　第 一个Python程序：Word Count　28

2.3.3　Python编程基础　30

2.3.4　Python的工程结构　34

2.4　本章小结　35

第3章　数学基础：恼人但又不可或缺的知识　36

3.1　矩阵和向量空间　37

3.1.1　标量、向量与矩阵　37

3.1.2　特殊矩阵　39

3.1.3　矩阵运算　39

3.1.4　代码实现　42

3.1.5　向量空间　44

3.2　概率：量化随机　46

3.2.1　定义概率：事件和概率空间　47

3.2.2　条件概率：信息的价值　48

3.2.3　随机变量：两种不同的随机　50

3.2.4　正态分布：殊途同归　52

3.2.5　P-value：自信的猜测　53

3.3　微积分　55

3.3.1　导数和积分：位置、速度　55

3.3.2　极限：变化的终点　57

3.3.3　复合函数：链式法则　58

3.3.4　多元函数：偏导数　59

3.3.5　极值与最值：最优选择　59

3.4　本章小结　61

第4章　线性回归：模型之母　62

4.1　一个简单的例子　64

4.1.1　从机器学习的角度看这个问题　66

4.1.2　从统计学的角度看这个问题　69

4.2　上手实践：模型实现　73

4.2.1　机器学习代码实现　74

4.2.2　统计方法代码实现　77

4.3　模型陷阱　82

4.3.1　过度拟合：模型越复杂越好吗　84

4.3.2　模型幻觉之统计学方案：假设检验　87

4.3.3　模型幻觉之机器学习方案：惩罚项　89

4.3.4　比较两种方案　92

4.4　模型持久化　92

4.4.1　模型的生命周期　93

4.4.2　保存模型　93

4.5　本章小结　96

第5章　逻辑回归：隐藏因子　97

5.1　二元分类问题：是与否　98

5.1.1　线性回归：为何失效　98

5.1.2　窗口效应：看不见的才是关键　100

5.1.3　逻辑分布：胜者生存　102

5.1.4　参数估计之似然函数：统计学角度　104

5.1.5　参数估计之损失函数：机器学习角度　104

5.1.6　参数估计之最终预测：从概率到选择　106

5.1.7　空间变换：非线性到线性　106

5.2　上手实践：模型实现　108

5.2.1　初步分析数据：直观印象　108

5.2.2　搭建模型　113

5.2.3　理解模型结果　116

5.3　评估模型效果：孰优孰劣　118

5.3.1　查准率与查全率　119

5.3.2　ROC曲线与AUC　123

5.4　多元分类问题：超越是与否　127

5.4.1　多元逻辑回归：逻辑分布的威力　128

5.4.2　One-vs.-all：从二元到多元　129

5.4.3　模型实现　130

5.5　非均衡数据集　132

5.5.1　准确度悖论　132

5.5.2　一个例子　133

5.5.3　解决方法　135

5.6　本章小结　136

第6章　工程实现：计算机是怎么算的　138

6.1　算法思路：模拟滚动　139

6.2　数值求解：梯度下降法　141

6.3　上手实践：代码实现　142

6.3.1　TensorFlow基础　143

6.3.2　定义模型　148

6.3.3　梯度下降　149

6.3.4　分析运行细节　150

6.4　更优化的算法：随机梯度下降法　153

6.4.1　算法细节　153

6.4.2　代码实现　154

6.4.3　两种算法比较　156

6.5　本章小结　158

第7章　计量经济学的启示：他山之石　159

7.1　定量与定性：变量的数学运算合理吗　161

7.2　定性变量的处理　162

7.2.1　虚拟变量　162

7.2.2　上手实践：代码实现　164

7.2.3　从定性变量到定量变量　168

7.3　定量变量的处理　170

7.3.1　定量变量转换为定性变量　171

7.3.2　上手实践：代码实现　171

7.3.3　基于卡方检验的方法　173

7.4　显著性　175

7.5　多重共线性：多变量的烦恼　176

7.5.1　多重共线性效应　176

7.5.2　检测多重共线性　180

7.5.3　解决方法　185

7.5.4　虚拟变量陷阱　188

7.6　内生性：变化来自何处　191

7.6.1　来源　192

7.6.2　内生性效应　193

7.6.3　工具变量　195

7.6.4　逻辑回归的内生性　198

7.6.5　模型的联结　200

7.7　本章小结　201

第8章　监督式学习： 目标明确　202

8.1　支持向量学习机　203

8.1.1　直观例子　204

8.1.2　用数学理解直观　205

8.1.3　从几何直观到最优化问题　207

8.1.4　损失项　209

8.1.5　损失函数与惩罚项　210

8.1.6　Hard margin 与soft margin比较　211

8.1.7　支持向量学习机与逻辑回归：隐藏的假设　213

8.2　核函数　216

8.2.1　空间变换：从非线性到线性　216

8.2.2　拉格朗日对偶　218

8.2.3　支持向量　220

8.2.4　核函数的定义：优化运算　221

8.2.5　常用的核函数　222

8.2.6　Scale variant　225

8.3　决策树　227

8.3.1　决策规则　227

8.3.2　评判标准　229

8.3.3　代码实现　231

8.3.4　决策树预测算法以及模型的联结　231

8.3.5　剪枝　235

8.4　树的集成　238

8.4.1　随机森林　238

8.4.2　Random forest embedding　239

8.4.3　GBTs之梯度提升　241

8.4.4　GBTs之算法细节　242

8.5　本章小结　244

第9章　生成式模型：量化信息的价值　246

9.1　贝叶斯框架　248

9.1.1　蒙提霍尔问题　248

9.1.2　条件概率　249

9.1.3　先验概率与后验概率　251

9.1.4　参数估计与预测公式　251

9.1.5　贝叶斯学派与频率学派　252

9.2　朴素贝叶斯　254

9.2.1　特征提取：文字到数字　254

9.2.2　伯努利模型　256

9.2.3　多项式模型　258

9.2.4　TF-IDF　259

9.2.5　文本分类的代码实现　260

9.2.6　模型的联结　265

9.3　判别分析　266

9.3.1　线性判别分析　267

9.3.2　线性判别分析与逻辑回归比较　269

9.3.3　数据降维　270

9.3.4　代码实现　273

9.3.5　二次判别分析　275

9.4　隐马尔可夫模型　276

9.4.1　一个简单的例子　276

9.4.2　马尔可夫链　278

9.4.3　模型架构　279

9.4.4　中文分词：监督式学习　280

9.4.5　中文分词之代码实现　282

9.4.6　股票市场：非监督式学习　284

9.4.7　股票市场之代码实现　286

9.5　本章小结　289

第　10章 非监督式学习：聚类与降维　290

10.1　K-means　292

10.1.1　模型原理　292

10.1.2　收敛过程　293

10.1.3　如何选择聚类个数　295

10.1.4　应用示例　297

10.2　其他聚类模型　298

10.2.1　混合高斯之模型原理　299

10.2.2　混合高斯之模型实现　300

10.2.3　谱聚类之聚类结果　303

10.2.4　谱聚类之模型原理　304

10.2.5　谱聚类之图片分割　307

10.3　Pipeline　308

10.4　主成分分析　309

10.4.1　模型原理　310

10.4.2　模型实现　312

10.4.3　核函数　313

10.4.4　Kernel PCA的数学原理　315

10.4.5　应用示例　316

10.5　奇异值分解　317

10.5.1　定义　317

10.5.2　截断奇异值分解　317

10.5.3　潜在语义分析　318

10.5.4　大型推荐系统　320

10.6　本章小结　323

第　11章 分布式机器学习：集体力量　325

11.1　Spark简介　327

11.1.1　Spark安装　328

11.1.2　从MapReduce到Spark　333

11.1.3　运行Spark　335

11.1.4　Spark DataFrame　336

11.1.5　Spark的运行架构　339

11.2　最优化问题的分布式解法　341

11.2.1　分布式机器学习的原理　341

11.2.2　一个简单的例子　342

11.3　大数据模型的两个维度　344

11.3.1　数据量维度　344

11.3.2　模型数量维度　346

11.4　开源工具的另一面　348

11.4.1　一个简单的例子　349

11.4.2　开源工具的阿喀琉斯之踵　351

11.5　本章小结　351

第　12章 神经网络：模拟人的大脑　353

12.1　神经元　355

12.1.1　神经元模型　355

12.1.2　Sigmoid神经元与二元逻辑回归　356

12.1.3　Softmax函数与多元逻辑回归　358

12.2　神经网络　360

12.2.1　图形表示　360

12.2.2　数学基础　361

12.2.3　分类例子　363

12.2.4　代码实现　365

12.2.5　模型的联结　369

12.3　反向传播算法　370

12.3.1　随机梯度下降法回顾　370

12.3.2　数学推导　371

12.3.3　算法步骤　373

12.4　提高神经网络的学习效率　373

12.4.1　学习的原理　373

12.4.2　激活函数的改进　375

12.4.3　参数初始化　378

12.4.4　不稳定的梯度　380

12.5　本章小结　381

第　13章 深度学习：继续探索　383

13.1　利用神经网络识别数字　384

13.1.1　搭建模型　384

13.1.2　防止过拟合之惩罚项　386

13.1.3　防止过拟合之dropout　387

13.1.4　代码实现　389

13.2　卷积神经网络　394

13.2.1　模型结构之卷积层　395

13.2.2　模型结构之池化层　397

13.2.3　模型结构之完整结构　399

13.2.4　代码实现　400

13.2.5　结构真的那么重要吗　405

13.3　其他深度学习模型　406

13.3.1　递归神经网络　406

13.3.2　长短期记忆　407

13.3.3　非监督式学习　409

13.4　本章小结　411