

# 机器学习

主编：严晓东

副主编：陈华 王国长 赵旋

## 编 者

- 第一章：赵旋
- 第二章：胡琴琴
- 第三章：张梅
- 第四章：国忠金
- 第五章：王重仁
- 第六章：王国长
- 第七章：周峰
- 第八——十二章：严晓东
- 第十三章：句媛媛
- 第十四章：梁虎
- 第十五章：陈华
- 编辑秘书：呼亚楠；崔文海；吉晓婷；陈熙；连蓓蓓；封文丽；吴淑琦

# 序　　言

随着新一代科学技术的发展，人工智能（artificial intelligence, AI）技术已经在各领域，包括科学、社会、经济、管理等方面，体现出解决各类复杂问题的优势，得到广泛认可，已经成为创新驱动发展的核心驱动力之一。大部分人工智能技术在应用落地方面得到很好的应用，尤其是 AI 中的机器学习日益成为一门通用的技能。最近有一些统计学家逐渐从大数据统计方法研究转入到前沿机器学习的科学问题的挖掘，例如基本原理、数学理论等存在的诸多瓶颈，影响着“可解释、可通用”的下一代机器学习的开发。华为创始人兼 CEO 任正非在接受央视《面对面》采访时特别强调，“人工智能就是统计学，计算机与统计学就是人工智能”，也间接强调了统计学在研究机器学习技术原理与理论上的重要地位。每一个学习者不能因为原理不懂而放弃使用机器学习方法的机会，这也正是写作本书意义和出发点。尽量使用统计语言介绍机器学习与大数据分析方法，做到详尽的数学推导，直观的文字解释，便捷的程序实现，给读者愉悦的学习体验，迎合普通高等学校理工类以及经济管理类等高年级本科生和研究生的基础知识储备。

本书的主要应用价值体现在机器学习方法与程序操作相结合。学习机器学习的最主要目的之一就是应用，应用离不开软件操作。同时，软件操作又有助于对机器学习方法的理解。R 语言为统计学家的母语，擅长于统计计算，Python 语言更偏于计算机及工程类专业使用的软件，因此 R 与 Python 语言是机器学习的两大常用语言。为此，本书提供“双语”服务，在介绍每个机器学习方法之后，最后章节详细介绍相应的 R 与 Python 语言实操。

目前市场上已出版了不少机器学习的书籍，再增加这一本书主要原因在于，计算机科学家的机器学习教材强调算法，使用“伪代码”（pseudo codes）来介绍方法，从实际工程（engineering）问题的角度讲解方法的使用。统计学家的机器学习教材一般缺少大数据统计方法的介绍，例如解决超高维数据的特征筛选、流数据的在线学习以及模型平均的集成学习等方法讲解。另外，本教材在一些概念上进行了混用，比如统计学里的变量、响应变量、协变量、离散响应变量等分别对应机器学习里的特征、因变量、自变量、标签等，但是并不影响我们的理解。

本教材依据“数据中国”的国家战略需求，针对“统计学”，“数据科学与大数据技术”等专业对《机器学习》教学资源的需求，依靠山东省大数据研究会大数据专业建设委员会搭建了大数据教材编写团队，形成了《机器学习》的编写队伍。山东国家应用数学中心、山东大学金融研究院等十多个教学科研平台也成为本教材编写的重要支撑。

严晓东  
2022 年 06 月 06 日

## 内容提要

本书主要围绕机器学习以及大数据分析的核心方法，进行深入而详细的讲解，特别关注方法理论推导以及程序实现，特色在于机器学习和大数据分析方法的统计语言介绍，并配有较多的插图、例子来直观地解释机器学习的原理。同时，结合流行的 R 与 Python 语言，介绍方法的程序实现，为读者提供“双语”服务。本书的机器学习和大数据分析方法配有详尽的数学符号与公式，并辅以直观的文字解释、细致的数学推导解读复杂的机器学习原理，做到删繁就简，娓娓道来，逻辑清晰。本书适合普通高等学校理工类以及经济管理类等高年级本科生和研究生使用。先修课包括数学分析、高等数学、线性代数、概率论与数理统计、统计学等以及 R 语言与 Python 语言的基本操作。

本书主要介绍常用机器学习方法，具体包括无监督学习、监督学习、集成学习、在线学习、机器学习预测以及大数据分析。

### 无监督学习

包括第一章与第二章，主要介绍两种无监督机器学习算法：聚类分析与主成分分析。

第一章：聚类分析。详细介绍了几种较为常用的聚类算法及其 python 语言与 R 语言代码。

第二章：主成分分析。介绍了线性主成分分析算法与非线性主成分分析算法，并利用 python 语言与 R 语言进行主成分分析实践。

### 监督学习

包括第三章，第四章与第五章。此部分主要介绍了监督学习中常见算法：回归分析，支持向量机与决策树算法。

第三章：回归分析。此章节主要介绍单响应变量的线性回归模型、单响应变量的广义线性模型 (GLM) 和多元响应变量协方差广义线性模型 (McGLM)。并给出了基于 Python 与 R 语言代码的相应案例实践。

第四章：支持向量机。本章分别介绍了线性可分 SVM、线性 SVM、非线性 SVM 处理分类问题以及回归问题的相关理论与方法。最后给出了基于 Python 语言与 R 语言的实现代码。

第五章：决策树。介绍的内容主要包括：决策树的基本概念、回归树和分类树的建模过程、防止决策树过拟合的方法及实现决策树相关的 Python 和 R 语言程序。

### 集成学习

包括第六章，第七章与第八章。这一部分主要介绍了集成学习的三种不同方法。

第六章：随机森林。主要介绍了基本概念，详细介绍了随机森林方法，并给出了基于 Python 语言与 R 语言的实现代码。

第七章：Boosting 方法。本章主要介绍 Boosting 方法及其代表性算法 AdaBoost，并给出了程序实现。

第八章：模型平均。本章节主要介绍了频率和贝叶斯模型平均方法及其理论依据，最后也给出了模型平均方法的程序实现。

### 在线学习

第十一章：在线学习。此章节主要介绍均值模型、线性模型、岭回归以及高维模型的在线学习算法和程序算法。

### 深度学习

包括第十三章与第十四章。主要介绍了深度学习中基本概念以及较为常见的模型。

第十三章：人工神经网络。此章主要讨论了前馈神经网络的数学模型、正向和反向计算方法以及 RBF 神经网络的数学模型和计算方法，对于其它神经网络形式做了简单介绍。同时也给出了基于 R 语言与 Python 语言的人工神经网络模型实践分析。

第十四章：深度学习。此章节针对图像数据，讨论了卷积神经网络、循环神经网络、长短时记忆网络、自编码器等网络的结构和计算过程及其基于 Python 语言与 R 语言的实现代码。

### 机器学习预测

第十二章：保形预测。该章节主要介绍了保形预测的基本思想及保形预测的几类常见算法，并给出了基于 Python 语言和 R 语言的实现代码。

### 数据分析

包括第八章，第十章与第十五章。这一部分主要介绍的是针对高维以及大规模数据集需要的技术。

第八章：模型选择。主要介绍了两种常用的模型选择的方法：子集选择法和压缩估计法及基于 Python 语言与 R 语言的实现代码。

第十章：特征筛选。此章节主要介绍了针对超高维数据进行特征筛选的几种常用方法，并给出了基于 Python 语言与 R 语言的实现代码。

第十五章：并行计算。此章节主要基于 Python 和 R 编程环境，给出了基于 CPU 和基于 GPU 的并行计算的相关概念、原理和计算流程。

# 检查分工

一元变量:  $X$ ; 一元样本:  $X_i$

多元变量:  $\mathbf{X} = (X_1, \dots, X_p)^T$ ; 多元样本:  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$

样本矩阵:  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$  是  $n \times p$  的设计矩阵

注意: 设  $\mathbf{X}_j$  为  $\mathbf{X}$  的第  $j$  列, 则  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ 。

随机误差: 令  $\varepsilon$  是一个一般随机误差,  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$

**崔文海: 检查数学符号、公式推导:**

- (1) 检查数学符号: 包括正体、斜体、粗体等
- (2) 向量、矩阵表示等是否正确, 以及转置符号
- (3) 公式、以及相关推导是否正确, 讲解是否清楚。

**吉晓婷: 检查语言文字**

- (1) 错别字
- (2) 标点符号与括号。中文文字处标点符号用中文格式, 美元符号中的标点符号都用英文; 括号也是一样, 也要区分中文还是英文。
- (3) 正体与斜体。如果是变量必须要用斜体 (例如  $K, n$  等); 如果是一些符号简写则用正体 (例如 AIC BIC RSS 等)。  
期望符号 E 用正体 E, 方差 var, 协方差 cov
- (4) 认真阅读每一句, 是否通顺, 表达是否清楚。
- (5) 概念定义处用蓝体字加强
- (6) 参考文献要添加上, 放在最后 (不是章节最后), 并且统一格式, 不知道的问我。

**陈熙: 检查程序与图片、表格**

- (1) 若涉及到真实数据, 建立按章节的子文件夹进行数据保存。如果是 R 或者 python 自带数据集就不需要了。
- (2) R 与 python code 是否可以运行, 并保证运行结果正确 (因此稍微熟悉一下方法)
- (3) 检查代码的中文解释是否正确, 并进行一定的修正。代码中文解释的文字不要太多, 简单、清楚、正确即可。没有中文解释的你需要适度的补充上。
- (4) 代码框中的代码一定要整齐、紧凑; 要保证代码框之间的连贯性, 做好过渡语句; 输出结果需要的解释的做一些简单的解释, 但是也不要太多。
- (5) 程序单独放在一小章节中, 例如第二章。
- (6) 图片是否清晰 (包括正文); 图片表格的名命以及引用是否冲突 (根据章节名命不会出错)

# 目 录

<b>第一章 聚类分析</b> .....	<b>1</b>
1.1 简介 .....	1
1.2 相似度 .....	1
1.2.1 数据对象间相似度 .....	1
1.2.2 簇间相似度 .....	3
1.3 K 均值聚类 .....	4
1.3.1 原理 .....	4
1.3.2 特点 .....	5
1.4 模糊 C 均值聚类 .....	7
1.5 高斯混合聚类 .....	8
1.6 层次聚类 .....	10
1.7 DBSCAN 聚类 .....	11
1.8 其他类型聚类方法 .....	14
1.8.1 混合型数据聚类方法 .....	14
1.8.2 双向聚类方法 .....	15
1.9 聚类实践 .....	16
1.9.1 R 语言实践 .....	17
1.9.2 Python 语言实践 .....	19
1.10 总结 .....	26
1.11 习题 .....	27
<b>第二章 主成分分析</b> .....	<b>29</b>
2.1 简介 .....	29
2.2 总体的主成分 .....	29
2.2.1 总体主成分的定义 .....	29
2.2.2 总体主成分的求法 .....	30
2.2.3 总体主成分的性质 .....	32
2.2.4 标准化变量的主成分 .....	34
2.3 样本主成分 .....	36

2.4 非线性主成分分析 ······	37
2.4.1 核主成分分析 ······	38
2.4.2 $t$ -SNE 非线性降维算法 ······	39
2.5 主成分分析实践 ······	42
2.5.1 R 语言实践 ······	42
2.5.2 Python 语言实践 ······	46
2.6 总结 ······	47
2.7 习题 ······	48
<b>第三章 回归分析 ······</b>	<b>49</b>
3.1 简介 ······	49
3.2 单响应变量的线性回归模型 ······	49
3.2.1 线性回归模型的原理 ······	50
3.2.2 多重共线性 ······	55
3.2.3 岭回归 ······	56
3.3 广义线性模型 ······	57
3.3.1 指数族分布 ······	57
3.3.2 连接函数 ······	58
3.3.3 广义线性模型 ······	59
3.4 多元响应变量协方差广义线性模型 McGLM ······	60
3.4.1 McGLM 模型的原理 ······	60
3.4.2 参数估计 ······	63
3.5 回归分析实践 ······	64
3.5.1 R 语言实践 ······	64
3.5.2 Python 语言实践 ······	68
3.6 习题 ······	73
<b>第四章 支持向量机 ······</b>	<b>77</b>
4.1 简介 ······	77
4.2 SVM 算法 ······	77
4.2.1 SVM 的基本内容 ······	77
4.2.2 线性可分 SVM ······	78
4.2.3 软间隔与线性 SVM ······	83
4.2.4 核函数与非线性 SVM ······	85
4.3 SVM 与 logistic 回归的关系 ······	89
4.4 支持向量回归 ······	89

4.5 SVM 实践 ······	90
4.5.1 R 语言实践 ······	90
4.5.2 Python 语言实践 ······	91
4.6 总结 ······	91
4.7 习题 ······	92
<b>第五章 决策树 ······</b>	<b>93</b>
5.1 简介 ······	93
5.2 决策树的基本原理 ······	93
5.3 分类树与回归树 ······	95
5.3.1 分类树 ······	95
5.3.2 回归树 ······	96
5.4 分支条件 ······	96
5.4.1 信息熵 ······	97
5.4.2 信息增益 ······	98
5.4.3 增益率 ······	100
5.4.4 基尼指数 ······	101
5.4.5 分类误差 ······	102
5.4.6 均方误差 ······	102
5.4.7 算法总结 ······	103
5.5 剪枝 ······	104
5.5.1 预剪枝 ······	105
5.5.2 后剪枝 ······	105
5.6 决策树实践 ······	107
5.6.1 R 语言实践 ······	107
5.6.2 Python 语言实践 ······	108
5.7 本章小节 ······	111
5.8 习题 ······	112
<b>第六章 随机森林 ······</b>	<b>113</b>
6.1 简介 ······	113
6.2 随机森林基本概况 ······	113
6.3 随机森林基本理论 ······	115
6.3.1 回归树基本理论 ······	115
6.3.2 分类树基本理论 ······	116

6.4 随机森林实践 ······	119
6.4.1 R 语言实践 ······	119
6.4.2 Python 语言实践 ······	124
6.5 小结 ······	129
6.6 习题 ······	130

## 第七章 Boosting 方法 ······ 132

7.1 简介 ······	132
7.1.1 Boosting 方法起源 ······	133
7.1.2 AdaBoost 算法 ······	134
7.1.3 AdaBoost 实例 ······	137
7.2 AdaBoost 算法的误差分析 ······	139
7.2.1 AdaBoost 算法的训练误差 ······	139
7.2.2 AdaBoost 算法的泛化误差 ······	142
7.3 AdaBoost 算法原理探析 ······	144
7.3.1 损失函数最小化视域 ······	144
7.3.2 向前逐段可加视域 ······	148
7.4 Boosting 算法的演化 ······	149
7.4.1 回归问题的 Boosting 算法 ······	150
7.4.2 梯度 Boosting 方法 ······	152
7.5 AdaBoost 算法实践 ······	155
7.5.1 R 语言实践 ······	156
7.5.2 Python 语言实践 ······	159
7.6 总结 ······	161
7.7 习题 ······	162

## 第八章 模型选择 ······ 164

8.1 简介 ······	164
8.2 基于检验的方法 ······	165
8.2.1 最优子集法 ······	165
8.2.2 逐步选择法 ······	166
8.3 基于准则的方法 ······	169
8.3.1 各种准则 ······	169
8.3.2 交叉验证 ······	171
8.4 正则化方法 ······	173
8.4.1 Lasso 回归 ······	176

8.4.2 非凸惩罚函数回归——SCAD 和 MCP . . . . .	177
8.4.3 群组变量选择方法 . . . . .	179
8.4.4 双层变量选择方法 . . . . .	181
8.5 模型选择实践 . . . . .	182
8.5.1 R 语言实现 . . . . .	182
8.5.2 Python 语言实现 . . . . .	188
8.6 练习 . . . . .	193
<b>第九章 模型平均 . . . . .</b>	<b>196</b>
9.1 简介 . . . . .	196
9.1.1 模型不确定性 . . . . .	196
9.1.2 模型选择与模型平均 . . . . .	196
9.2 贝叶斯模型平均 . . . . .	197
9.3 频率模型平均 . . . . .	199
9.4 权重选择方法 . . . . .	199
9.4.1 基于信息准则 . . . . .	200
9.4.2 基于 Mallows 准则 . . . . .	200
9.4.3 基于 Jackknife 准则 . . . . .	201
9.5 Python 语言实现 . . . . .	202
9.6 练习 . . . . .	211
<b>第十章 特征筛选 . . . . .</b>	<b>213</b>
10.1 简介 . . . . .	213
10.2 基于边际模型的特征筛选 . . . . .	213
10.2.1 边际最小二乘 . . . . .	213
10.2.2 边际极大似然 . . . . .	214
10.2.3 边际非参估计 . . . . .	215
10.3 基于边际相关系数的特征筛选 . . . . .	216
10.3.1 广义和秩相关系数 . . . . .	216
10.3.2 确定独立秩筛选 . . . . .	218
10.3.3 距离相关系数 . . . . .	218
10.4 高维分类数据的特征筛选 . . . . .	220
10.4.1 Kolmogorov-Smirnov 统计量 . . . . .	220
10.4.2 均值-方差统计量 . . . . .	220
10.4.3 类别自适应筛选统计量 . . . . .	221

10.5 R 语言实现 ······	222
10.5.1 基于边际模型拟合的特征筛选 ······	222
10.5.2 基于边际相关系数的特征筛选 ······	225
10.5.3 高维分类数据的特征筛选 ······	226
10.6 Python 语言实现 ······	229
10.6.1 基于边际模型拟合的特征筛选 ······	229
10.6.2 基于边际相关系数的特征筛选 ······	231
10.6.3 高维分类数据的特征筛选 ······	233
10.7 练习 ······	236
<b>第十一章 在线学习 ······</b>	<b>238</b>
11.1 简介 ······	238
11.2 累积统计量在线学习 ······	238
11.2.1 均值模型 ······	239
11.2.2 线性模型 ······	240
11.3 在线梯度下降 ······	241
11.3.1 OGD 算法一般形式 ······	241
11.3.2 OGD 算法收敛性分析 ······	242
11.3.3 线性模型的 OGD 算法 ······	243
11.3.4 岭回归模型的 OGD 算法 ······	244
11.4 基于正则化的在线梯度下降 ······	245
11.4.1 FTL 算法 ······	245
11.4.2 FTRL 算法 ······	246
11.4.3 FTRL-Proximal 算法 ······	247
11.5 在线学习实践 ······	248
11.5.1 R 语言实践 ······	249
11.5.2 Python 语言实践 ······	255
11.6 练习 ······	262
<b>第十二章 保形预测 ······</b>	<b>263</b>
12.1 简介 ······	263
12.1.1 简单运用 ······	264
12.1.2 预测集的有效性 ······	266
12.1.3 效率 ······	267
12.2 保形回归 ······	267
12.2.1 保形预测基本思想 ······	267
12.2.2 完全保形预测 ······	268

12.3 保形方法 ······	270
12.3.1 分裂保形预测 ······	270
12.3.2 Jackknife 保形预测 ······	271
12.3.3 局部加权保形预测 ······	272
12.4 保形分类 ······	273
12.4.1 Softmax 法 ······	273
12.4.2 最近邻法 ······	275
12.5 R 语言实现 ······	275
12.5.1 分裂保形预测的简单代码 ······	276
12.5.2 局部加权保形预测的简单代码 ······	277
12.6 Python 语言实现 ······	280
12.6.1 完全保形预测的简单代码 ······	280
12.6.2 保形分类和保形回归的简单代码 ······	281
<b>第十三章 人工神经网络 ······</b>	<b>284</b>
13.1 人工神经元 ······	284
13.2 前馈神经网络 ······	285
13.2.1 单层感知器 ······	286
13.2.2 多层感知器 ······	287
13.3 神经网络的正向与反向传播算法 ······	288
13.3.1 神经网络的正向传播 ······	288
13.3.2 神经网络的损失函数 ······	291
13.3.3 反向传播算法 ······	292
13.3.4 全局最小与局部极小 ······	293
13.4 径向基网络 ······	295
13.5 其它常见的神经网络 ······	298
13.6 神经网络实践 ······	301
13.6.1 R 语言实践 ······	301
13.6.2 Python 语言实践 ······	301
13.7 总结 ······	304
13.8 练习 ······	304
<b>第十四章 深度学习 ······</b>	<b>305</b>
14.1 卷积神经网络 ······	306
14.2 简单循环神经网络 ······	319
14.3 长短时记忆神经网络 ······	320

14.4	自编码器 ······	323
14.5	玻尔兹曼机 ······	327
14.5.1	随机神经网络 ······	327
14.5.2	模拟退火算法 ······	328
14.5.3	BM ······	328
14.6	深度学习实践 ······	331
14.6.1	R 语言实践 ······	331
14.6.2	Python 语言实践 ······	332
14.7	总结 ······	335
14.8	练习 ······	335
<b>第十五章 并行计算 ······</b>		<b>337</b>
15.1	并行计算相关概念 ······	337
15.1.1	进程 ······	337
15.1.2	线程 ······	339
15.1.3	并行计算与分布式计算 ······	341
15.1.4	同步与异步 ······	341
15.1.5	通信 ······	342
15.1.6	加速比 ······	343
15.2	基于 CPU 线程的并行计算 ······	345
15.2.1	创建线程 ······	345
15.2.2	同步 ······	348
15.3	基于 CPU 进程的并行计算 ······	352
15.3.1	创建进程 ······	352
15.3.2	进程间通信 ······	355
15.3.3	同步 ······	358
15.4	基于 GPU 线程的并行计算 ······	358
15.4.1	CUDA 基本概念 ······	358
15.4.2	CUDA 线程组织 ······	360
15.4.3	CUDA 内存组织 ······	363
15.4.4	PyCUDA ······	364
15.4.5	TensorFlow ······	365
15.5	并行计算实践 ······	365
15.5.1	R 语言实践 ······	365
15.5.2	Python 语言实践 ······	372
15.6	总结 ······	375

15.7 练习 ······	375
<b>参考文献 ······</b>	<b>377</b>



# 第一章 聚类分析

## 1.1 简介

当数据集有明确的标签时，可利用标签信息进行监督学习，发现数据内在特性，然而，在一些实际问题中，标签数据有时较难获取或者无法获取，此时，通常使用无监督学习对数据进行贴签，聚类分析是无监督学习的典型代表。

所谓聚类是指根据某种规则，将样本集中具有相似特征的样本进行分组的过程，每个分组组成一个“簇”，不同簇之间互不相交，所有簇的并集构成整个样本集，以簇中心表示簇内所有样本的特性。聚类过程使得同一簇内样本的相似度尽可能高，不同簇间样本的相似度尽可能低，从而达到“物以类聚”的目的。

目前，聚类算法被广泛应用于实际问题中，例如：企业对客户的价值分析，由于数据量较大，通常很难采用人工方式定义某一客户是重点客户或者一般客户，此时可通过聚类算法将客户分为不同簇，每个簇表示一个类别，然后通过领域专家判断每个类别的特性。

本章将介绍几种较为常用的聚类算法，在此之前，首先介绍聚类算法中的重要概念—相似度。

## 1.2 相似度

相似度在聚类算法中至关重要，是聚类算法的核心问题，具体表现在两个层面：（1）不同的聚类算法其相似度计算方法也不同；（2）同一聚类算法采用不同的相似度计算将得到不同的聚类效果。本节主要介绍数据对象间相似度和簇间相似度。

### 1.2.1 数据对象间相似度

聚类是对所有数据对象或样本进行一定的处理，达到“物以类聚”的目标。此过程涉及不同样本之间相似度的刻画，选取不同的相似度将产生不同的聚类结果，本节介绍几种样本之间相似度的度量方法。

### 1、闵可夫斯基距离 (Minkowski distance)

每一个样本可以看作是一个向量，所有样本则是向量空间中点的集合，可以采用向量之间的距离度量刻画样本之间的相似度。假设  $\mathbf{X} = (X_1, \dots, X_p)^T$ 、 $\mathbf{Y} = (Y_1, \dots, Y_p)^T$ 、 $\mathbf{Z} = (Z_1, \dots, Z_p)^T$  表示三个样本，其中  $p$  表示样本特征的维数，给定函数  $d(\cdot, \cdot)$ ，如果其满足下面四个性质，则称其为**距离度量**：

- (1) **非负性**:  $d(\mathbf{X}, \mathbf{Y}) \geq 0$ ;
- (2) **同一性**:  $d(\mathbf{X}, \mathbf{Y}) = 0$  当且仅当  $\mathbf{X} = \mathbf{Y}$ ;
- (3) **对称性**:  $d(\mathbf{X}, \mathbf{Y}) = d(\mathbf{Y}, \mathbf{X})$ ;
- (4) **直递性**:  $d(\mathbf{X}, \mathbf{Y}) \leq d(\mathbf{X}, \mathbf{Z}) + d(\mathbf{Z}, \mathbf{Y})$ 。

闵可夫斯基距离是常用的距离度量，其计算公式如下：

$$d(\mathbf{X}, \mathbf{Y}) = \left( \sum_{k=1}^p |X_k - Y_k|^t \right)^{\frac{1}{t}} \quad (1.2.1)$$

上式中  $t \geq 1$ 。下面介绍闵可夫斯基距离的三种特殊形式。

当  $t = 1$  时，称为**曼哈顿距离**，定义为：

$$d(\mathbf{X}, \mathbf{Y}) = \sum_{k=1}^p |X_k - Y_k| \quad (1.2.2)$$

当  $t = 2$  时，称为**欧氏距离**，定义为：

$$d(\mathbf{X}, \mathbf{Y}) = \left( \sum_{k=1}^p |X_k - Y_k|^2 \right)^{\frac{1}{2}} \quad (1.2.3)$$

当  $t = \infty$  时，称为**切比雪夫距离**，定义为：

$$d(\mathbf{X}, \mathbf{Y}) = \max_{1 \leq k \leq p} |X_k - Y_k| \quad (1.2.4)$$

### 2、马氏距离 (Mahalanobis distance)

马氏距离可视为欧式距离的一种泛化形式，一定程度上可以克服欧式距离中由于不同特征量纲不一致导致的影响距离计算的问题。

若样本  $\mathbf{X} = (X_1, \dots, X_p)^T$ 、 $\mathbf{Y} = (Y_1, \dots, Y_p)^T$  服从同一类型分布且具有相同的协方差阵，其协方差矩阵为  $\Sigma$ ，则马氏距离可以表示为：

$$d(\mathbf{X}, \mathbf{Y}) = \sqrt{(\mathbf{X} - \mathbf{Y}) \Sigma^{-1} (\mathbf{X} - \mathbf{Y})^T} \quad (1.2.5)$$

上式中，当协方差矩阵  $\Sigma$  是单位矩阵时，即各个维度是独立同分布的，则马氏距离等同于欧式距离。

### 3、夹角余弦

两个向量之间的夹角余弦可用于衡量两者之间的相似性，当把样本看作向量时，若两者之间的夹角余弦是 0，即角度为 90 度，那么说明两者之间是独立的；若夹角

余弦是 1，说明两者指向相同方向；若夹角余弦是 -1，说明两者指向相反方向。夹角余弦越靠近 1，表明样本之间的相似度越高。

因此  $\mathbf{X}$  与  $\mathbf{Y}$  的夹角余弦定义为：

$$d(\mathbf{X}, \mathbf{Y}) = \frac{\sum_{k=1}^p X_k Y_k}{[(\sum_{k=1}^p X_k^2)(\sum_{k=1}^p Y_k^2)]^{\frac{1}{2}}} \quad (1.2.6)$$

#### 4、相关系数

相关系数介于 [0,1] 之间，可以衡量变量之间线性相关的程度，在聚类算法中可用于样本间相似度的度量，其绝对值越接近 1，样本之间的相似度越高；越接近 0，样本之间相似度越低。从两类样本集  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$  和  $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)^T$  中任取两个样本  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ 、 $\mathbf{Y}_j = (Y_{j1}, \dots, Y_{jp})^T$ ，两者之间的相关系数定义为：

$$d(\mathbf{X}_i, \mathbf{Y}_j) = \frac{(\mathbf{X}_i - \bar{\mathbf{X}})^T (\mathbf{Y}_j - \bar{\mathbf{Y}})}{\|\mathbf{X}_i - \bar{\mathbf{X}}\|_2 \|\mathbf{Y}_j - \bar{\mathbf{Y}}\|_2} \quad (1.2.7)$$

其中  $\|\cdot\|_2$  表示  $L_2$  范数（欧几里得范数）， $\bar{\mathbf{X}}$  和  $\bar{\mathbf{Y}}$  分别表示样本集  $\mathbf{X}$  和  $\mathbf{Y}$  的样本均值，

$$\bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i, \quad \bar{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i$$

### 1.2.2 簇间相似度

聚类过程将样本集划分为互不相交的簇，簇间相似度可通过计算簇间距离得到，簇间距离可以通过样本间的距离定义。假设  $G_1$ 、 $G_2$  是两个不同的簇，两者之间的距离可通过以下方式定义。

#### 1、最小距离

最小距离由簇间距离最近的两个样本决定，定义为：

$$D_{12} = \min_{\mathbf{x}_i \in G_1, \mathbf{x}_j \in G_2} d_{ij} \quad (1.2.8)$$

其中  $d_{ij}$  表示样本  $\mathbf{x}_i$  与  $\mathbf{x}_j$  之间的距离。

#### 2、最大距离

最大距离由簇间距离最远的两个样本决定，定义为：

$$D_{12} = \max_{\mathbf{x}_i \in G_1, \mathbf{x}_j \in G_2} d_{ij} \quad (1.2.9)$$

#### 3、平均距离

平均距离等于两簇之间所有样本间的距离的平均值，定义为：

$$D_{12} = \frac{1}{|G_1||G_2|} \sum_{\mathbf{x}_i \in G_1, \mathbf{x}_j \in G_2} d_{ij} \quad (1.2.10)$$

其中  $|G_1|$  表示簇  $G_1$  中包含的样本个数,  $|G_2|$  表示簇  $G_2$  中包含的样本个数。

#### 4、中心距离

两个不同簇中心之间距离定义为中心距离, 假设  $\bar{\mathbf{X}}_1$  表示簇  $G_1$  的中心,  $\bar{\mathbf{X}}_2$  表示簇  $G_2$  的中心, 则中心距离定义为:

$$D_{12} = d_{\bar{\mathbf{X}}_1 \bar{\mathbf{X}}_2} \quad (1.2.11)$$

其中,

$$\bar{\mathbf{X}}_1 = \frac{1}{|G_1|} \sum_{\mathbf{X}_i \in G_1} \mathbf{X}_i,$$

$$\bar{\mathbf{X}}_2 = \frac{1}{|G_2|} \sum_{\mathbf{X}_j \in G_2} \mathbf{X}_j$$

## 1.3 K 均值聚类

### 1.3.1 原理

K 均值也称为 K-means, 是一种原理简单、应用广泛的聚类算法。其目标是将包含  $n$  个样本的数据集, 按照某种规则划分到  $K$  个互不相交的子集中, 称为  $K$  个簇, 相同簇中样本的相似度较高, 不同簇之间样本的相似度较低。

假设样本集合为  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ , 其中  $\mathbf{X}_i$  表示第  $i$  个样本。令  $\mu_1, \mu_2, \dots, \mu_K$  是  $K$  个向量, 表示  $K$  个簇中心, 其中  $\mu_k$  代表第  $k$  个簇的中心。K 均值聚类采用欧氏距离计算样本与簇中心之间的距离, 然后将样本置入距离最近的簇中心所代表的簇中, 这样形成  $K$  个互不相交的集合(簇)  $G_1, G_2, \dots, G_K$ , 当  $j \neq k$  时满足  $G_j \cap G_k = \emptyset$ , 并且  $\cup_{k=1}^K G_k = \mathbf{X}$ 。

基于欧式距离, K 均值聚类算法的目标函数为:

$$L(\Theta) = \sum_{k=1}^K \sum_{\mathbf{X}_i \in G_k} \|\mathbf{X}_i - \mu_k\|^2 \quad (1.3.1)$$

其中,  $\mu_k = \frac{1}{|G_k|} \sum_{\mathbf{X}_i \in G_k} \mathbf{X}_i$ 。

计算上述目标函数的最优解并非易事, K 均值聚类算法采用迭代更新策略得到优化问题的近似解。首先选取初始簇中心  $\mu_k, k = 1, 2, \dots, K$ , 然后按照欧氏距离最小原则将样本划入不同簇中, 最后更新簇中心, 以便得到更小的  $L(\Theta)$  值。将此过程不断迭代下去, 直到聚类结果基本保持不变, 停止计算。K 均值聚类算法的流程如下:

- 1、从样本集  $\mathbf{X}$  中随机选取  $K$  个样本向量  $\{\mu_1, \mu_2, \dots, \mu_K\}$  作为初始簇中心。
- 2、计算每一个样本到  $K$  个簇中心的距离，并将样本划入到距离最小的簇中心所代表的簇中。
- 3、重新计算每个簇的中心向量  $\{\mu'_1, \mu'_2, \dots, \mu'_K\}$ 。如果  $\mu'_k \neq \mu_k$ ，则将  $\mu_k$  更新为  $\mu'_k$ 。
- 4、不断迭代第 2 步和第 3 步，直到所有的簇中心不再更新为止。
- 5、算法结束，输出划分结果。

如图 1.1，展示了 K 均值聚类的整个过程，每张图代表每一次的聚类结果。首先从原始数据集中随机选取三个样本作为初始聚类中心，分别用红色、绿色和黑色的三角形表示，如图 1.1(a) 所示。之后计算每个样本与三个聚类中心的距离，根据距离最近原则将样本划分为颜色各异的三个簇，如图 1.1(b)。随后，重新计算每个簇的样本均值，作为新的三个聚类中心，如图 1.1(c)，此时簇中心已经发生改变，计算样本到新的簇中心之间的距离，得到新的簇划分。不断迭代此过程，直至聚类中心不再变化。

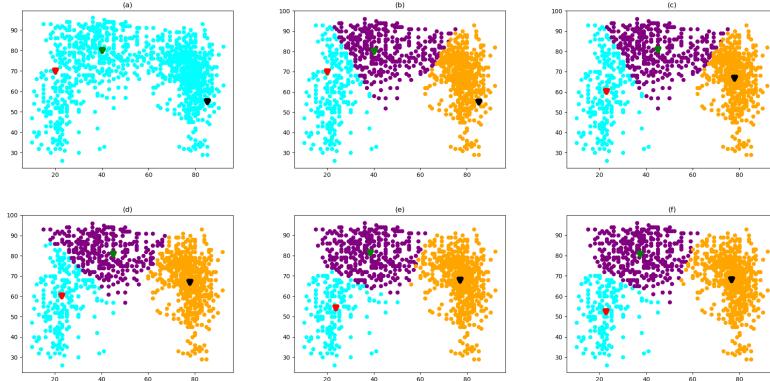


图 1.1 K 均值聚类过程

### 1.3.2 特点

K 均值聚类是机器学习中的常用无监督学习算法，其优点是易于实现、拥有较好的性能，并且相比于其他机器学习算法具有较少的超参数。然而， $K$  的取值以及初始簇中心的选择对算法有较大的影响。 $K$  的取值关乎聚类输出的效果以及可解释性，应当选取合适的  $K$  值作为聚类簇数。初始簇中心的选择影响算法的迭代效果，好的初始簇中心能够加快算法的收敛，反之将影响收敛速度。本节将从这两个方面对 K 均值算法进行分析。

#### 1、 $K$ 值的确定

在 K 均值算法中，簇个数  $K$  的选择对算法最终的输出结果影响较大，不同的  $K$  值将产生不同的聚类效果，在实际应用中， $K$  的最优值通常较难获取，本节介绍三种  $K$  值确定方法。

(1) 根据数据的先验知识确定  $K$  值，或者可通过简单数据分析得到  $K$  的可能取值。

(2) **指示函数确定法**：也可称为肘部确定法。定义一个指示函数，如图 1.2，横轴表示  $K$  的取值，纵轴表示指示函数的取值。函数值随着  $K$  值的变化而变化，对于选定的某个临界值，如果  $K$  大于该值时，函数的变化趋于平缓，反之小于该值时，函数值的变化较为剧烈，则可选择该值作为最优  $K$  值。其中指示函数可以选择误差平方和、平均直径、平均半径等。

(3) **平均轮廓系数法**：轮廓系数的计算公式如下：

$$R(\mathbf{X}_i) = \frac{B(\mathbf{X}_i) - I(\mathbf{X}_i)}{\max[B(\mathbf{X}_i), I(\mathbf{X}_i)]} \quad (1.3.2)$$

上式中， $\mathbf{X}_i$  表示第  $i$  个样本向量， $I(\mathbf{X}_i)$  表示样本  $\mathbf{X}_i$  与其所在簇中其他样本之间的平均距离， $B(\mathbf{X}_i)$  表示样本  $\mathbf{X}_i$  与其他簇中所有样本之间的平均距离。由式 (1.3.2) 可计算出每个样本的轮廓系数，然后再求平均值，得到平均轮廓系数。一般情形下，平均轮廓系数越接近 1 聚类效果越好。

## 2、初始簇中心的选择

实际问题中，合适的初始簇中心对算法至关重要。本部分介绍四种选择初始簇中心的方法。

- (1) 根据对数据集的先验知识，由领域专家设定初始簇中心。
- (2) 采用不同的簇中心，多次运行算法，选取最优的初始簇中心。
- (3) 选择尽可能远离的  $K$  个点：首先从样本集合中选择一个样本，然后选择距离此样本最远的样本作为第二个簇中心，依次进行下去，直到选出  $K$  个初始簇中心。
- (4) 采用层次聚类等算法进行初始聚类，利用这些类的中心作为 K 均值算法的初始簇中心。

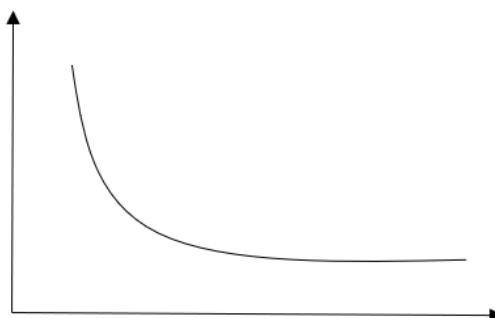


图 1.2 指示函数与  $K$  值关系图

## 1.4 模糊 C 均值聚类

K 均值聚类算法将样本划分到不同簇，其结果具有非黑即白的性质，即非 0 即 1，所有样本对象均被硬性划分到某一簇类别，在很多实际问题中可以达到较好效果，如是否下雨、是否旅行等；然而，这种硬聚类的方式不适用于类别界限不明确的问题，如健康程度、冷热程度等。**模糊 C 均值聚类**(也称为 FCM) 采用软聚类思想，用**隶属度**描述样本与簇之间的关系，隶属度表示样本属于某一簇的确定程度。

与 K 均值聚类相似，FCM 算法也是通过最小化目标函数，不断迭代更新聚类簇中心，当目标函数收敛时，算法停止迭代；不同的是，FCM 算法在 K 均值聚类算法目标函数的基础上，增加了隶属度度量，其表达式为：

$$L(\Theta) = \sum_{i=1}^n \sum_{j=1}^C \omega_{i,j}^m \| \mathbf{X}_i - \boldsymbol{\mu}_j \|^2, \quad 1 \leq m < \infty \quad (1.4.1)$$

其中， $n$  为样本总个数， $C$  为聚类中心数， $m$  为隶属度参数， $\mathbf{X}_i$  表示第  $i$  个样本， $\boldsymbol{\mu}_j$  表示第  $j$  个聚类中心， $\omega_{i,j}^m$  用于刻画隶属度并且满足，

$$\sum_{j=1}^C \omega_{i,j}^m = 1$$

通过拉格朗日乘子法可以得到：

$$\begin{aligned} \boldsymbol{\mu}_j &= \frac{\sum_{i=1}^n \omega_{i,j}^m \mathbf{X}_i}{\sum_{i=1}^n \omega_{i,j}^m} \\ \omega_{i,j}^m &= \frac{1}{\sum_{k=1}^C \frac{\| \mathbf{X}_i - \boldsymbol{\mu}_j \|^{\frac{2}{m-1}}}{\| \mathbf{X}_i - \boldsymbol{\mu}_k \|^{\frac{2}{m-1}}}} \end{aligned}$$

FCM 算法通过不断迭代更新  $\omega_{i,j}^m$  与  $\boldsymbol{\mu}_j$  的取值，最终得到聚类结果，迭代终止条件可设置为：

$$\max_{i,j} \{ |\omega_{i,j}^{m,s+1} - \omega_{i,j}^{m,s}| \} < \epsilon$$

其中， $s$  表示迭代次数， $\epsilon$  为给定的误差阈值。其含义为，当两次迭代之间的最大隶属度变化量小于给定阈值时，说明继续迭代隶属度的变化量也会非常小，此时停止迭代，得到最终的聚类结果。

FCM 算法的算法流程为：

- 1、超参数确定：选择合适的聚类簇数  $C$ 、隶属度参数  $m$  以及误差阈值  $\epsilon$ ；
- 2、初始化隶属度矩阵  $\omega^{m,0}$ ；
- 3、计算并更新聚类中心  $\boldsymbol{\mu}_j$ ；
- 4、计算并更新隶属度矩阵  $\omega^{m,s}$ ；

5、比较更新前后隶属度矩阵的最大变化量，如果小于  $\epsilon$ ，算法停止，否则返回第 3 步。

## 1.5 高斯混合聚类

**高斯混合聚类**是在概率框架下实施聚类过程，采用软聚类的思想计算样本被划分到不同簇的概率，选取概率最大的簇作为最终的划分结果。

在阐述高斯混合聚类算法之前，首先给出**多变量高斯分布概率密度函数**的定义：如果样本空间中的  $p$  维随机向量  $\mathbf{X} = (X_1, \dots, X_p)^T$  服从高斯分布，则其概率密度函数为

$$\Pr(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu})}. \quad (1.5.1)$$

其中， $\boldsymbol{\mu}$  表示均值向量， $\boldsymbol{\Sigma}$  表示协方差矩阵，显然其概率密度函数由  $\boldsymbol{\mu}$  和  $\boldsymbol{\Sigma}$  唯一确定。下面以二维高斯分布为例，给出其概率密度函数图像（如图 1.3）。

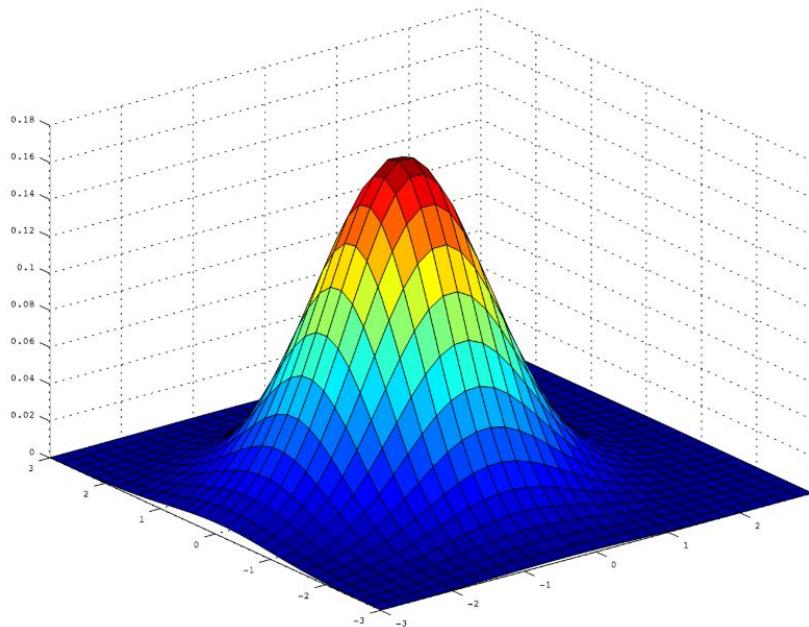


图 1.3 二维高斯分布概率密度函数

高斯混合聚类模型的算法可分为三个环节：

(1) 假设样本集是由多个高斯分布生成，每个分布的概率密度函数为式 (1.5.1)，不同高斯分布的  $\boldsymbol{\mu}$  和  $\boldsymbol{\Sigma}$  值也不同；

(2) 采用 EM 算法不断迭代更新  $\mu$  和  $\Sigma$  的值拟合样本数据，直到算法收敛或者达到最大迭代次数为止；

(3) 利用所计算出的概率密度函数对样本集进行聚类。其中第 (2) 步是整个算法的核心，下面将详细推导参数的更新策略。

假设高斯混合聚类模型由  $K$  个高斯分布组成，每个高斯分布称为一个成分，每个成分对应一个簇划分，对高斯分布进行线性组合得到高斯混合聚类的概率密度函数：

$$\Pr(\mathbf{X}) = \sum_{k=1}^K \lambda_k \Pr(\mathbf{X} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (1.5.2)$$

其中， $\boldsymbol{\mu}_k$  和  $\boldsymbol{\Sigma}_k$  分别表示第  $k$  个混合成分的均值向量与协方差矩阵。 $\lambda_k$  是第  $k$  个混合成分系数（权重），并且满足

$$\sum_{k=1}^K \lambda_k = 1$$

可以采用极大似然估计求解模型参数  $\lambda_k$ 、 $\boldsymbol{\mu}_k$  以及  $\boldsymbol{\Sigma}_k$ 。假设样本集  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ ，对数似然函数可以定义为：

$$\begin{aligned} L(\Theta) &= \ln \left( \prod_{i=1}^n \Pr(\mathbf{X}_i) \right) \\ &= \sum_{i=1}^n \ln \left( \sum_{k=1}^K \lambda_k \Pr(\mathbf{X}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \end{aligned} \quad (1.5.3)$$

此时的目标转化为：求使得式 (1.5.3) 最大化的  $\lambda_k$ 、 $\boldsymbol{\mu}_k$  和  $\boldsymbol{\Sigma}_k$ ，其中  $1 \leq k \leq K$ 。对于  $\boldsymbol{\mu}_k$  和  $\boldsymbol{\Sigma}_k$  的求解，可分别对其求导，并令导数等于 0，即

$$\frac{\partial L(\Theta)}{\partial \boldsymbol{\mu}_k} = 0 \quad (1.5.4)$$

$$\frac{\partial L(\Theta)}{\partial \boldsymbol{\Sigma}_k} = 0 \quad (1.5.5)$$

求解式 (1.5.4)，我们可以得到，

$$\boldsymbol{\mu}_k = \frac{\sum_{i=1}^n \theta_{ik} \mathbf{X}_i}{\sum_{i=1}^n \theta_{ik}} \quad (1.5.6)$$

其中  $\theta_{ik}$  定义为，

$$\theta_{ik} = \frac{\lambda_k \Pr(\mathbf{X}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^K \lambda_k \Pr(\mathbf{X}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}$$

$\theta_{ik}$  表示第  $i$  个样本  $\mathbf{X}_i$  被划入第  $k$  个混合成分的概率, 高斯混合聚类选取  $\max(\theta_{ik}), k = 1, 2, \dots, K$  作为  $\mathbf{X}_i$  的簇划分。同理求解式 (1.5.5), 我们可以得到

$$\boldsymbol{\Sigma}_k = \frac{\sum_{i=1}^n \theta_{ik} (\mathbf{X}_i - \boldsymbol{\mu}_k)(\mathbf{X}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^n \theta_{ik}} \quad (1.5.7)$$

下面我们考虑  $\lambda_k$  的求解方法。由于  $\lambda_k$  满足条件

$$\sum_{k=1}^K \lambda_k = 1, \lambda_k \geq 0$$

因此需求解带约束的极值优化问题, 采用拉格朗日乘子法容易得到:

$$\lambda_k = \frac{1}{n} \sum_{i=1}^n \theta_{ik} \quad (1.5.8)$$

通过上述推导过程可以看出, 高斯混合聚类可以采用 EM 算法对参数进行更新, 具体可分为两个步骤: (1) 根据前一个迭代步 (或初始值) 的参数值计算  $\theta_{ik}$ ; (2) 利用第 (1) 步求得的  $\theta_{ik}$  更新  $\lambda_k$ 、 $\boldsymbol{\mu}_k$  和  $\boldsymbol{\Sigma}_k$ 。

高斯混合聚类算法的流程可以概括为:

- 1、参数初始化, 包括  $\lambda_k$ 、 $\boldsymbol{\mu}_k$ 、 $\boldsymbol{\Sigma}_k$  以及混合成分个数  $K$ 。
- 2、根据前一个迭代步 (或初始值) 的参数值计算  $\theta_{ik}$ 。
- 3、利用式 (1.5.6)、(1.5.7)、(1.5.8) 更新模型参数  $\boldsymbol{\mu}_k$ 、 $\boldsymbol{\Sigma}_k$ 、 $\lambda_k$ 。如果满足停止条件, 则停止迭代; 否则返回第 2 步, 继续迭代计算。
- 4、根据所求得到  $\theta_{ik}$ , 将  $\mathbf{X}_i$  划入到相应簇划分  $C_k$ ,  $C_k = C_k \cup \mathbf{X}_i$ 。
- 5、得到最终的簇划分  $C = \{C_1, C_2, \dots, C_K\}$ , 算法结束。

## 1.6 层次聚类

**层次聚类**顾名思义是一个逐层进行聚类的过程, 通过计算相似度形成一种树形结构。按照形成树形结构方式的不同可以分为**聚合聚类**和**分裂聚类**。

**聚合聚类**是自下而上进行聚类, 初始状态将每个样本当作一个簇, 然后按照簇间距离最近原则合并两个簇, 组成一个新簇, 不断迭代此过程, 直到满足条件为止; **分裂聚类**是自上而下进行聚类, 初始状态是将样本集看作一个簇, 然后按照簇间距离最远的原则将样本划分到两个新的簇, 不断迭代此过程, 直到满足条件为止。

本节以聚合聚类为例, 介绍层次聚类的算法思想。假设样本集为  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ , 样本个数为  $n$ , 特征维数为  $p$ , 将样本集划分为  $K$  个簇, 则聚合聚类的算法流程为:

- 1、将  $n$  个样本分为  $n$  个簇, 即每个簇中只包含一个样本。
- 2、对于给定数据集, 通过计算各簇之间的距离得到距离矩阵  $D = (d_{ij})_{n \times n}$ 。其中  $d_{ij}$  表示第  $i$  个簇与第  $j$  个簇之间的距离。

- 3、将距离最近的两个簇合并为一个新簇。
- 4、计算新簇与其他簇之间的距离，并更新距离矩阵。
- 5、不断迭代 3、4 两个过程，直到满足条件停止。

在聚合聚类算法中，样本之间的距离（相似度）计算可以采用本章1.2.1节中的方法，簇间距离的计算可采用1.2.2节中的方法。通常根据距离最小的原则将相似簇进行合并，算法的停止条件可以是达到设定的簇的个数  $K$  或者达到某个指示函数的阈值。

下面通过例题1.1阐述聚合聚类的算法流程。

**例 1.1** 假设某数据集中有 4 个样本，通过计算样本之间的欧氏距离得到如下距离矩阵  $D$ ，请利用聚合聚类算法将 4 个样本进行聚类。

$$D = \begin{pmatrix} 0 & 2 & 9 & 4 \\ 2 & 0 & 3 & 5 \\ 9 & 3 & 0 & 6 \\ 4 & 5 & 6 & 0 \end{pmatrix}$$

解：(1) 将 4 个样本分为 4 个簇  $G_1, G_2, G_3, G_4$ ，每个簇中只有一个样本，则簇间距离矩阵为  $D$ 。

(2) 根据  $D$ ，挑选出距离最近的两个簇  $G_1, G_2$ ，将这两个簇合并为一个新簇  $G_5 = \{G_1, G_2\}$ ，此时已划分为 3 个簇  $G_3, G_4, G_5$ 。

(3) 分别计算新簇  $G_5$  与  $G_3, G_4$  之间的最短距离，得到最短距离大小是  $D_{35} = 3$ ,  $D_{45} = 4$ 。

(4) 可以看出，距离最近的 2 两个簇是  $G_3, G_5$ ，合并这两个簇形成一个新簇  $G_6 = \{G_3, G_5\} = \{G_1, G_2, G_3\}$ 。此时已划分为 2 个簇  $G_4, G_6$ 。

(5) 将  $G_4, G_6$  合并为新簇  $G_7 = \{G_4, G_6\} = \{G_1, G_2, G_3, G_4\}$ ，算法结束。聚类过程可以表示为图 1.4。

层次聚类可以达到“分层”的效果，通过选取不同的聚类簇数，得到不用层次的聚类结果。然而算法实现过程的复杂度较高，计算量通常高于 K 均值算法。

## 1.7 DBSCAN 聚类

通常情形下，K 均值聚类一般适用于样本集具有凸形簇状结构（形似“椭球”的簇结构），对非凸样本集效果不够理想。本节介绍一种既可用于凸样本集，也可用于非凸样本集的DBSCAN (Density-Based Spatial Clustering of Applications with Noise) 方法。

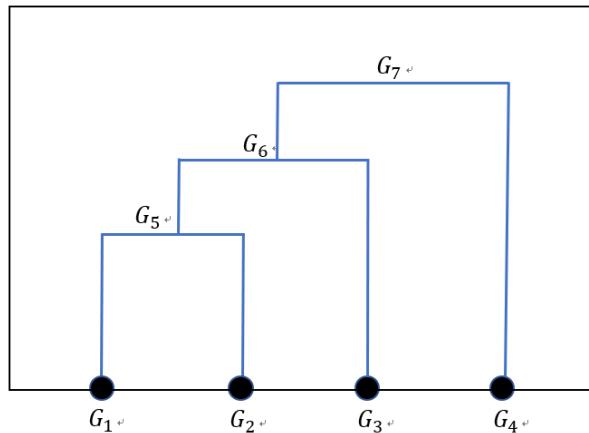


图 1.4 层次聚类过程图

DBSCAN 是一种基于密度的聚类方法，算法假设样本的类别划分可由样本之间分布的紧密程度决定，紧密程度较高的样本被划分到相同簇，反之，被划分到不同簇，最终，通过样本之间的紧密程度得到聚类结果。由此可见，紧密程度的刻画是 DBSCAN 算法的核心。

DBSCAN 算法采用邻域描述样本集分布的紧密程度，邻域是由  $(\epsilon, M)$  表示，其含义是某一样本  $\epsilon$  邻域内的样本个数不超过  $M$ 。下面给出与描述样本集  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$  紧密程度相关的一些概念：

**$\epsilon$ -邻域**: 对于  $\mathbf{X}$  中的任一样本  $\mathbf{X}_j$ ，其  $\epsilon$  邻域  $N_\epsilon(\mathbf{X}_j)$  表示与该样本距离不大于  $\epsilon$  的所有样本组成的集合，即  $N_\epsilon(\mathbf{X}_j) = \{\mathbf{X}_i \in \mathbf{X} | d(\mathbf{X}_i, \mathbf{X}_j) \leq \epsilon\}$ ，其中  $d(\mathbf{X}_i, \mathbf{X}_j)$  表示  $\mathbf{X}_i$  与  $\mathbf{X}_j$  之间的距离，例如闵可夫斯基距离。

**核心对象**: 对于  $\mathbf{X}$  中的任一样本  $\mathbf{X}_j$ ，若其所对应的  $N_\epsilon(\mathbf{X}_j)$  中样本个数大于等于  $M$ ，则称  $\mathbf{X}_j$  是核心对象。

**密度直达**: 若  $\mathbf{X}_i \in N_\epsilon(\mathbf{X}_j)$ ，并且  $\mathbf{X}_j$  是一个核心对象，则称样本  $\mathbf{X}_i$  由样本  $\mathbf{X}_j$  密度直达。

**密度可达**: 对于  $\mathbf{X}$  中的任意两个样本  $\mathbf{X}_i$  和  $\mathbf{X}_j$ ，如果存在样本序列  $q_1, q_2, \dots, q_t$  满足  $q_1 = \mathbf{X}_j$ ,  $q_t = \mathbf{X}_i$ ，并且  $q_{l+1}$  由  $q_l$  密度直达，其中  $1 \leq l \leq t-1$ ，则称样本  $\mathbf{X}_i$  由样本  $\mathbf{X}_j$  密度可达。

**密度相连**: 对于  $\mathbf{X}$  中的任意两个样本  $\mathbf{X}_i$  和  $\mathbf{X}_j$ ，如果存在一个样本  $\mathbf{X}_l$  使得  $\mathbf{X}_i$  和  $\mathbf{X}_j$  均可由  $\mathbf{X}_l$  密度可达，则称  $\mathbf{X}_i$  和  $\mathbf{X}_j$  密度相连。

上图1.5描述了上述基本概念，如图所示，样本  $A$ 、 $B$  和  $O$  均是核心对象， $Y$  由  $A$  密度直达， $A$  和  $B$  由  $O$  密度直达， $X$  由  $B$  密度直达； $X$  和  $Y$  均由  $O$  密度可达； $X$  与  $Y$  密度相连。

基于上述基本概念，可以给出 DBSCAN 算法中簇的定义：由密度可达关系导出的最大密度相连的样本集合，即为聚类结果的一个簇。

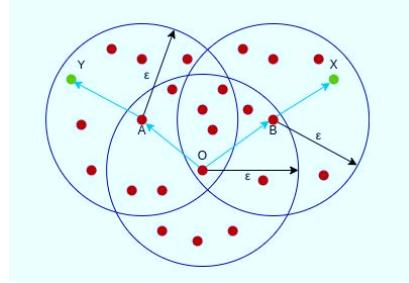


图 1.5 基本概念示意图

DBSCAN 算法的目标是从样本集中找出所有满足簇定义的簇划分，由此可见 DBSCAN 不需要预先给定簇划分的个数。在寻找簇划分时，DBSCAN 从样本集中选择一个没有被划分的核心对象作为“种子”，寻找由该核心对象密度可达的所有样本构成一个样本子集，即为一个簇划分；然后选择另外一个没有被划分的核心对象作为“种子”，依次进行下去，直到所有核心对象均被划分为止。

给定样本集  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ , DBSCAN 算法的流程可以描述为：

1、初始化  $\epsilon$ 、 $M$ 、核心对象集  $\Omega = \emptyset$ 、聚类簇数  $k=0$  以及未被访问的样本集  $\Phi = \mathbf{X}$ ；

2、遍历样本集中的所有样本，找出核心对象，将其添加到  $\Omega$  中，此时  $\Omega$  存储了所有的核心对象；

3、若  $\Omega = \emptyset$ ，则算法结束，否则进行第 4 步；

4、记录当前未被访问的样本集合： $\Phi^* = \Phi$ ，在  $\Omega$  中随机挑选一个核心对象  $o$ ，初始化队列  $\Omega^* = \langle o \rangle$ ，更新未被访问的样本集  $\Phi = \Phi - \{o\}$ ；

5、在队列  $\Omega^*$  中随机取出一个核心对象  $o$ ，此时  $\Omega^* = \Omega^* - \langle o \rangle$ ，并找出其  $\epsilon$ -邻域  $N_\epsilon(o)$ ，令  $\Delta = N_\epsilon(o) \cap \Phi$ ，更新  $\Phi = \Phi - \Delta$ ，更新  $\Omega^* = \Omega^* \cup (\Delta \cap \Omega)$ ；

6、如果  $\Omega^* = \emptyset$ ，进行第 7 步，否则返回第 5 步；

7、令  $k = k + 1$ ，更新当前簇  $C_k = \Phi^* - \Phi$ ，更新  $\Omega = \Omega - C_k$ ，返回第 3 步；

8、输出最终的簇划分  $C = \{C_1, C_2, \dots, C_k\}$ 。

注：若队列  $\Omega^*$  中的元素在此前循环中被访问过，那么  $\Omega^*$  中将不再包含此元素。

DBSCAN 算法从样本集紧密程度的角度进行簇划分，与 K 均值、高斯混合聚类等算法相比，DBSCAN 算法不需要提前确定聚类簇数，并且可应用于非凸数据集的聚类，在得到聚类结果的同时还可以识别出异常点。然而，当样本集规模较大时，算法收敛时间较长；并且算法调优涉及  $\epsilon$  和  $M$  两个参数，不同参数组合对聚类的结果影响较大，调参工作量较为复杂。

## 1.8 其他类型聚类方法

聚类方法多种多样，针对不同的应用场景可能会产生比传统聚类方法更好的算法，因此在使用聚类方法解决实际问题时，应当根据问题本身的特点选择合适的聚类方法。本节将介绍其他几种常用的聚类方法。

### 1.8.1 混合型数据聚类方法

在实际应用中，数据集通常是由数值型和分类型特征组成，尤其是当数据集是由多源数据集成而得时，如医疗数据（包括病人个人信息：姓名、学历、职业、收入情况等；医疗监测数据：血常规、血脂等）、客户数据（包括客户的年龄、性别、种族、消费记录、消费频次、消费金额等）等。针对这种混合型数据进行聚类分析时需进行特殊处理，本节介绍一种常用的混合型数据聚类方法：[K-prototypes](#)。

[K-prototypes](#) 是基于 K 均值方法的一种可以处理混合数据的聚类方法。与 K 均值聚类方法相比，主要存在两点不同：(1) 簇中心的选取；(2) 距离的计算。下面针对这两点进行详细阐述。

在 K 均值聚类中，每个簇选取簇内样本的均值作为当前簇中心，这种计算方式仅适用于特征都是数值型数据，当某些特征是分类型时将无法计算。因此 [K-prototypes](#) 方法在计算簇中心时将所有特征分为两部分：数值型、分类型。对于数值型特征依然计算均值，对于分类型特征选择众数，然后将两部分合并起来组成当前簇的中心。令样本集为  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ ，不失一般性，假设任意一个样本  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$  包含  $p$  个特征，其中  $q$  个数值型特征、 $p - q$  个分类型特征。令  $\{\mu_1, \mu_2, \dots, \mu_K\}$  表示  $K$  个簇中心。

K 均值常采用欧氏距离计算样本与簇中心之间的距离，在 [K-prototypes](#) 中数值型特征依然采用欧氏距离，分类型特征采用[汉明距离](#)，然后得到样本点与簇中心的距离。[K-prototypes](#) 的算法流程可以描述为：

- 1、初始化聚类簇数  $K$  以及簇中心  $\{\mu_1, \mu_2, \dots, \mu_K\}$ ；
- 2、计算数据集中样本点  $\mathbf{X}_i$  到  $K$  个簇中心的距离，将所有样本划入距离最近的簇中；
- 3、重新计算每个簇的中心；
- 4、重复第 2 步、第 3 步，直到满足停止条件（簇中心变化很小或者达到迭代的最大次数）；

注：对于两个数字来说，汉明距离就是转成二进制后，对应的位置值不相同的个数。例如，假设有两个十进制数  $a=93$  和  $b=73$ ，如果将这两个数用二进制表示的话，有  $a=1011101$ 、 $b=1001001$ ，可以看出，二者的从右往左数的第 3 位、第 5 位不同（从 1 开始数），因此， $a$  和  $b$  的汉明距离是 2。

## 1.8.2 双向聚类方法

传统的聚类方法可大致分为两种模式：一种是对样本进行聚类（如根据销售数据对客户类型聚类分析）；另一种是对特征或者变量进行聚类（如分析影响青少年成长的因素，分为积极因素、消极因素等）。这两种模式均是对单一方向（样本或者特征）进行聚类，通常被称为单向聚类。

尽管单向聚类可以解决多数实际问题，仍然存在一些特殊场景并不适合采用单向聚类进行分析，例如利用患某种疾病的病人基因数据对病人进行划分，通常只有少量基因对该病产生影响，并且不同的基因组合对疾病的影响也不尽相同，由于需要同时考虑样本和特征之间的关系，捕捉局部性质，单向聚类在利用所有基因进行聚类分析时效果不佳，因此利用双向聚类方法解决此类问题得到较多的研究。近年来双向聚类方法被广泛应用于基因表达分析、文本数据挖掘等领域，下面介绍一种双向聚类方法——稀疏双向 K 均值聚类。

稀疏双向聚类的目标是从原始数据集中根据矩阵元素的相似性提取具有特定结构的子矩阵，下面介绍几种常见的类型：

**全局常数型**：子矩阵中的所有元素是同一常数，即对于子矩阵  $B = (b_{tr})$ ，其中  $1 \leq t \leq T, 1 \leq r \leq R$ ，满足  $b_{tr} = c$ ，其中  $c$  是常数，如图1.6(a)。实际场景中由于噪声的存在，通常要求子矩阵在阈值  $\epsilon$  范围内满足元素为常数。

**行(列)常数型**：子矩阵的每一行(或列)的元素为常数。以行常数为例，对于子矩阵  $B = (b_{tr})$ ，其中  $1 \leq t \leq T, 1 \leq r \leq R$ ，满足  $b_{t1} = b_{t2} = \dots = b_{tR} = c_t$ ，其中  $c_t$  是常数，如图1.6(b)。

**线性关系型**：子矩阵的每行(列)的元素是其他任意一行(列)的线性组合。以行满足线性关系为例，对于子矩阵  $B = (b_{tr})$ ，其中  $1 \leq t \leq T, 1 \leq r \leq R$ ，满足  $B_{t_1} = c_{t_2} \times B_{t_2} + \theta_{t_2}$ ，其中  $c_{t_2}, \theta_{t_2}$  是常数， $B_{t_1}$  表示矩阵第  $t_1$  行。列满足线性关系同理可得。当  $c_{t_2} = 1$  时，被称为加法模型，如图1.6(c)；当  $\theta_{t_2} = 0$  时，被称为乘法模型，如图1.6(d)。

稀疏双向 K 均值聚类是将 K 均值聚类拓展到稀疏双向聚类中的一种算法，求解过程与 K 均值算法类似。现给定原始数据集矩阵为  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ ，其中  $n$  表示样本数， $p$  表示特征数且  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ 。假设  $n$  个样本可以划分为  $K_1$  个类别，记为  $G_1, G_2, \dots, G_{K_1}$ ， $p$  个特征可以划分为  $K_2$  个类别，记为  $G_1^*, G_2^*, \dots, G_{K_2}^*$ 。并且假设样本  $\mathbf{X}_i$  中的元素  $X_{ij}$  的期望  $E(X_{ij}) = \mu_{ls}$ ，其中  $\mu_{ls}$  表示  $X_{ij}$  属于第  $l$  个样本类别、第  $s$  个特征类别所对应子矩阵中的元素均值。双向聚类的目标函数为

1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1

1	1	1	1
2	2	2	2
3	3	3	3
4	4	4	4

1	2	3	4
2	3	4	5
3	4	5	6
4	5	6	7

(a) 全局常数型                          (b) 行常数型                          (c) 加法模型

1	2	3	4
2	4	6	8
3	6	9	12
4	8	12	16

(d) 乘法模型

图 1.6 双向聚类结构类型.

最小化下式：

$$\min \sum_{l=1}^{K_1} \sum_{s=1}^{K_2} \sum_{i \in G_l} \sum_{j \in G_s} (X_{ij} - \mu_{ls})^2 \quad (1.8.1)$$

由式 (1.8.1) 可以看出，当  $K_1 = 1$  时，转化为关于特征的 K 均值聚类，当  $K_2 = 1$  时，转化为关于样本的 K 均值聚类。

为了增强聚类结果的可解释性，并且减少类别方差，可对式 (1.8.1) 中的  $\mu_{ls}$  施加 Lasso 惩罚，此时目标函数可变为：

$$\min \left\{ \frac{1}{2} \sum_{l=1}^{K_1} \sum_{s=1}^{K_2} \sum_{i \in G_l} \sum_{j \in G_s} (X_{ij} - \mu_{ls})^2 + \lambda \sum_{l=1}^{K_1} \sum_{s=1}^{K_2} |\mu_{ls}| \right\}. \quad (1.8.2)$$

其中  $\lambda$  是非负参数。

显然稀疏双向 K 均值聚类是 K 均值聚类的一种泛化形式，在利用稀疏双向 K 均值聚类算法时，通常先将原始数据集矩阵  $\mathbf{X}$  进行中心化处理，然后再进入算法流程。算法参数的求解过程与 K 均值算法相似，可以利用迭代求解的思想不断更新参数的取值，直到参数在某个阈值  $\epsilon$  范围内不再变化或者达到最大迭代次数时停止计算。

## 1.9 聚类实践

### 1.9.1 R 语言实践

#### K 均值聚类算法

R 语言中有 K 均值聚类算法库 kmeans，本小节将基于该库函数，选取莺尾花数据集中的 4 个属性为原始数据，阐述 K 均值聚类算法的 R 语言实现过程。

莺尾花数据集共包含 5 个属性，最后一个属性为类别，本实验不予采用，仅使用前 4 个属性进行 K 均值聚类分析，属性含义分别为萼片长度 (Sepal.Length)、萼片宽度 (Sepal.Width)、花瓣长度 (Petal.Length)、花瓣宽度 (Petal.Width)。实验代码如下：

```
df<-iris[,1:4]
#调用 kmeans 函数, centers 参数为簇数量; nstart 参数表示随机起始分区数
KMmodel<-kmeans(df,centers=3,nstart=8)
KMmodel$centers#输出簇中心
KMmodel$cluster#输出每个样本所属的簇划分
```

#### 模糊 C 聚类算法

本小节将基于 R 语言 e1071 包中的 cmeans 函数实现 FCM 算法。本实验将随机生成 50 个均值为 0、标准差为 0.2 以及 50 个均值为 1、标准差为 0.3 的数据，构成包含 50 个样本的二维数据集合作为样本集。采用 cmeans 方法对该样本集进行聚类分析，代码如下：

```
set.seed(5)
cdata<-rbind(matrix(rnorm(50,mean=0,sd=0.2),ncol=2),
matrix(rnorm(50,mean=1,sd=0.3),ncol=2))#生成样本集
library(e1071)
#iter.max 参数为最大迭代次数; method 参数默认为 cmeans, 表示模糊C均值聚类
CMmodel<-cmeans(cdata,centers=2,iter.max=30,verbose=TRUE,method="cmeans")
CMmodel$centers#输出聚类中心
CMmodel$size#每个簇中数据点的个数
```

#### 高斯混合聚类算法

R 语言 mclust 包中的 Mclust 函数可以实现高斯混合聚类算法，本节将随机生成样本集，调用 Mclust 函数阐述高斯混合聚类算法的实现过程。

使用上一部分生成的样本集，并且调用 summary 函数、plot 函数等输出聚类结果，效果图如图1.7。

```
Gmodel<-Mclust(cdata)
summary(Gmodel)
plot(Gmodel,what="classification")
```

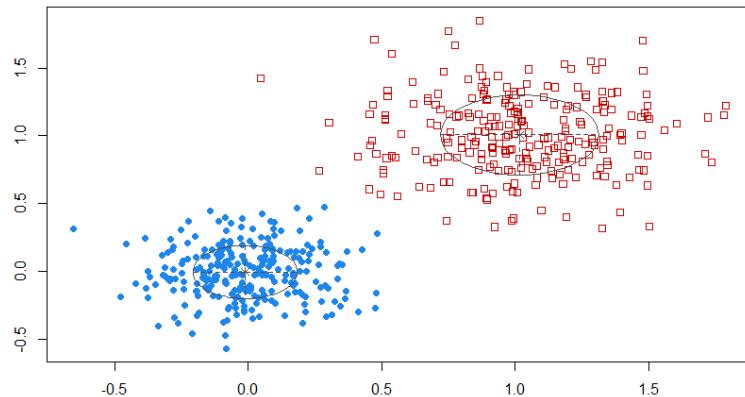


图 1.7 聚类效果图

## 层次聚类算法

R 语言中 hclust 函数可以实现层次聚类算法，本小节将随机生成样本集，调用 hclust 函数阐述层次聚类算法的实现过程。

```
set.seed(5)
Hdata<-rbind(matrix(rnorm(10,mean=0,sd=0.2),ncol=2),matrix(rnorm(10,mean=1,sd=0.3),
ncol=2),matrix(rnorm(10,mean=2,sd=0.3),ncol=2))#生成样本集
#调用 dist 函数计算样本之间的欧氏距离（或者其他形式的距离），得到距离矩阵
HDDist<-dist(Hdata,method="euclidean")
Hmodel=hclust(HDDist,method="single")#调用 hclust 函数采用最短距离法进行聚类分析
plot(Hmodel)#调用 plot 函数画出层次聚类效果图
rect.hclust(Hmodel,k=3)#调用 rect.hclust 函数用矩形框出
```

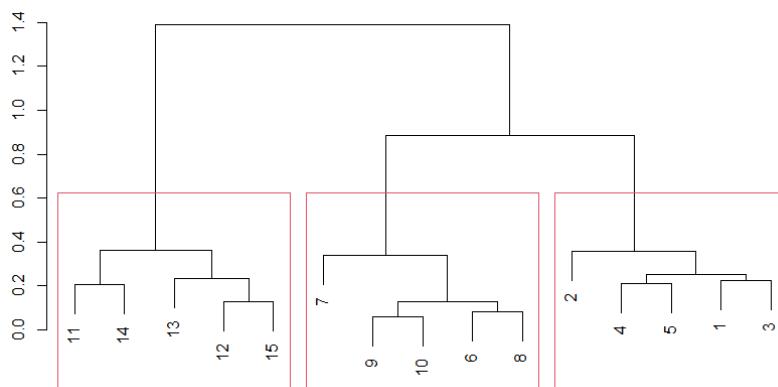


图 1.8 层次聚类效果图

## DBSCAN 聚类算法

R 语言中的 fpc 库和 dbscan 库均可实现 DBSCAN 算法，由于 dbscan 库中的 dbscan 函数具有更快的计算效率并且可以处理更大规模的数据集，因此本节将基于 dbscan 函数实现。此次实验同样选取莺尾花数据集的前 4 个属性作为实验数据。

```
DBdata<-iris[,1:4]
library(dbscan)
DBmodel<-dbscan(DBdata,eps=0.7,MinPts=5) #调用 dbscan 函数：选取 eps 为 0.7, MinPts 为 5
print(DBmodel$cluster) #结果输出
plot(DBdata[,1:2],col=DBmodel$cluster)
```

## 双向聚类算法

本小节以 Bimax 算法为例，利用 biclust 库中的函数实现算法整体流程，实现代码如下：

```
library(biclust)
#生成一个由 5000 个满足正态分布的随机数构成的 100×50 的矩阵
test <- matrix(rnorm(5000), 100, 50)
#生成由均值是 3、标准差是 0.1 的 100 个随机数构成的 10×10 矩阵
test[11:20,11:20] <- rnorm(100, 3, 0.1)
loma <- binarize(test,2) #进行预处理
#使用 bcbimax 函数进行双聚类并将结果保存在 res 中
res <- biclust(x=loma, method=BCBimax(), minr=4, minc=4, number=10)
```

输出结果如下：

```
An object of class Biclust
call:
biclust(x = loma, method = BCBimax(), minr = 4, minc = 4, number = 10)
There was one cluster found with
10 Rows and 10 columns #输出结果显示有一个 10×10 的簇
```

## 1.9.2 Python 语言实践

### K 均值聚类算法

K 均值聚类算法原理简单、容易实现，并且有较多资料库可以直接调用。本小节将基于 sklearn 库实现 K 均值算法。代码实现如下：

```
from sklearn.datasets import make_blobs
from matplotlib import pyplot as plt
from sklearn.cluster import KMeans
#调用 sklearn 库中的 make_blobs 函数建立数据集，特征维数设置为 2，样本数设置为 200
X, y = make_blobs(n_features=2, n_samples=200, centers=4,
cluster_std=0.80, random_state=0)
```

```
#利用KMeans.fit函数对构建的数据集进行K均值聚类，选取聚类簇数K=4
kmeans = KMeans(n_clusters=4).fit(X)
y_pre = kmeans.predict(X)#predict函数返回每个样本所属的簇
para = kmeans.get_params()#get_params函数返回模型参数
plt.scatter(X[:, 0], X[:, 1], c=y_pre)
centers = kmeans.cluster_centers_#cluster_centers_函数返回聚类算法的四个簇中心
plt.scatter(centers[:,0], centers[:, 1], c='black', s=100, marker='+')
plt.show()#绘制聚类效果图
```

聚类效果图如图 1.9 所示，不同颜色表示样本属于不同簇。

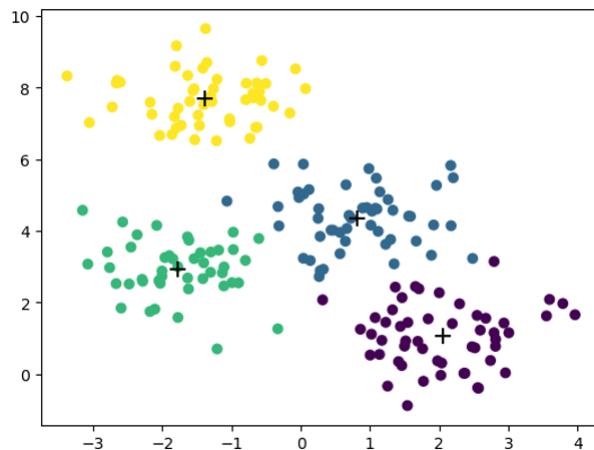


图 1.9 聚类效果图

## 模糊 C 聚类算法

本小节调用 skfuzzy.cluster 库中的 cmeans 函数实现 FCM 算法，代码实现如下：

```
import numpy as np
from skfuzzy.cluster import cmeans
data = np.random.uniform(1, 100, (2, 100))#在1到100中随机生成包含2个特征的100个样本
#调用cmeans函数实现FCM聚类，参数c为聚类的簇数；error为误差阈；maxiter为最大迭代次数
center, U, U0, dist, mj, knum, evalute = cmeans(data, m=2, c=3,
error=0.001, maxiter=10000)
```

其中 cmeans 函数输出参数：center 为聚类中心；U 为最终隶属矩阵；U0 为初始隶属矩阵；dist 为样本点到聚类中心的距离矩阵；mj 为目标函数值；knum 为迭代次数；evalute 为评价指标，越接近 1 效果越好。

```
import matplotlib.pyplot as plt
for i in U:
    label = np.argmax(U, axis=0) #根据输出结果，将样本点划分到不同簇
for i in range(np.shape(data)[1]):
    if label[i] == 0:
```

```

plt.scatter(data[0][i], data[1][i], c='b')
elif label[i] == 1:
    plt.scatter(data[0][i], data[1][i], c='g')
elif label[i] == 2:
    plt.scatter(data[0][i], data[1][i], c='r')
plt.show() #绘制聚类结果图

```

FCM 聚类结果图，如图 1.10 所示。

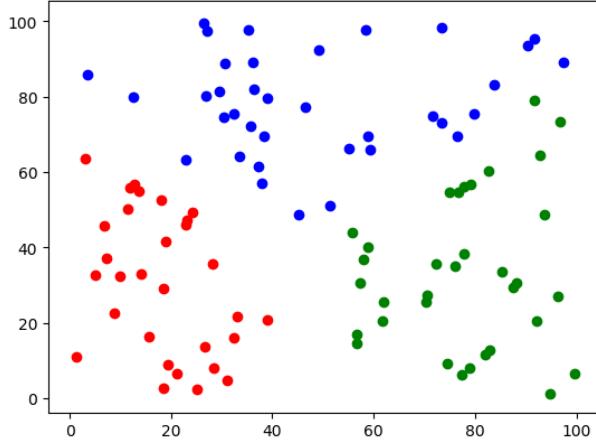


图 1.10 FCM 聚类结果图

## 高斯混合聚类算法

本小节将利用 sklearn 实现高斯混合聚类算法，实现代码如下：

```

from sklearn.datasets import make_blobs
from sklearn.mixture import GaussianMixture
from matplotlib import pyplot as plt
#调用sklearn库中的make_blobs函数建立数据集。特征维数设置为2，样本数设置为200
X, y = make_blobs(n_features=2, n_samples=200, centers=4,
cluster_std=0.80, random_state=0)
#GaussianMixture函数实现高斯混合聚类算法。混合成分k取2，采用fit方法作用于原始样本集
gmm = GaussianMixture(n_components=2, random_state=0).fit(X)
para = gmm.get_params()
y_pre = gmm.predict(X)
plt.scatter(X[:, 0], X[:, 1], c=y_pre)
plt.show() #绘制聚类散点图

```

高斯混合聚类效果图，如图 1.11 所示。

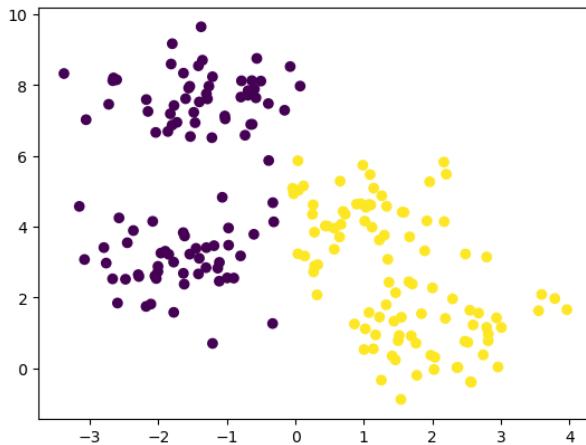


图 1.11 聚类效果图

## 层次聚类算法

本小节以聚合聚类算法为例，利用 sklearn 库中 AgglomerativeClustering 函数实现算法整体流程，实现代码如下：首先导入相应模块，建立数据集

```
from sklearn.datasets import make_blobs
from matplotlib import pyplot as plt
from sklearn.cluster import AgglomerativeClustering as Agc
#基于make_blobs函数建立数据集。设置特征维数为2，数据集中样本数为200
X, y = make_blobs(n_features=2, n_samples=200, centers=5,
cluster_std=1, random_state=0)
```

再采用 AgglomerativeClustering.fit 函数进行聚合聚类

```
#聚类簇数设置为5
model = Agc(n_clusters=5).fit(X)
para = model.get_params()#返回模型参数
y_pred = model.labels_#返回样本所属的簇
plt.scatter(X[:, 0], X[:, 1], c=y_pred)
plt.show()#绘制聚合聚类效果图
```

聚合聚类效果图，如图 1.12 所示，不同颜色表示样本属于不同簇。

## DBSCAN 聚类算法

本节我们采用 sklearn.cluster 中的 DBSCAN 方法，展示 DBSCAN 算法的实现过程，并与 K 均值算法进行对比。实现过程如下：

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.datasets import make_circles, make_blobs
```

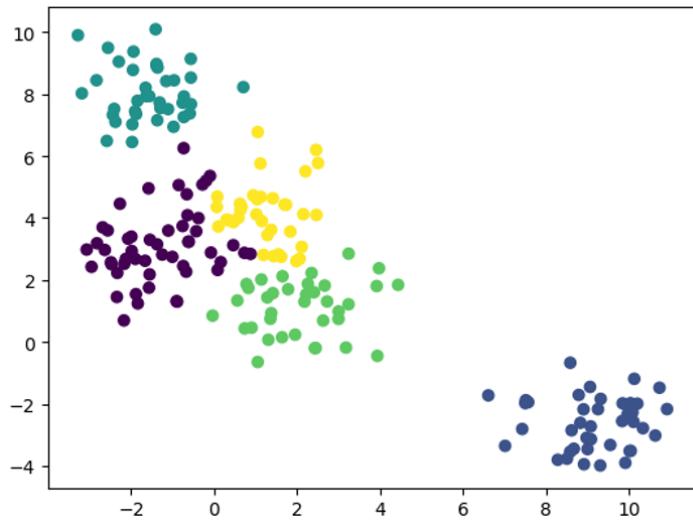


图 1.12 聚合聚类效果图

```
#调用make_circles方法生成包含1000个点的环形数据子集
X1, y1 = make_circles(n_samples=1000, factor=0.5, noise=0.03)
#调用make_blobs方法生成包含1000个点并以(1.3, 1.3)为中心的数据子集
X2, y2 = make_blobs(n_samples=1000, centers=[[1.3, 1.3]], cluster_std=[[.04]], random_state=6)
X = np.concatenate((X1, X2))#将两个数据子集进行拼接得到我们要分析的数据集合
plt.scatter(X[:, 0], X[:, 1], marker='o')
plt.show()#绘制数据集散点图
```

数据集散点图如图1.13所示。

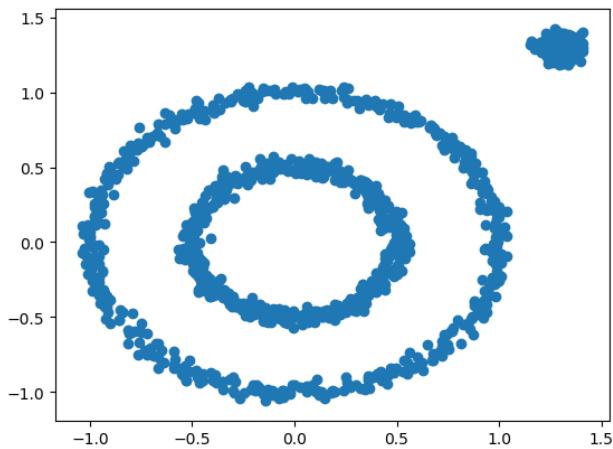


图 1.13 数据集散点图

首先采用 K 均值算法对上述数据集进行聚类分析，得到聚类结果如图1.14，可以看出 K 均值聚类不能较好地捕捉此数据集的内在特征。

```
from sklearn.cluster import KMeans
from sklearn.cluster import DBSCAN
y_pred = KMeans(n_clusters=3).fit_predict(X)#聚类簇数k取3
plt.scatter(X[:, 0], X[:, 1], c=y_pred)
plt.show()
```

再采用 DBSCAN 算法对上述数据集进行聚类，得到聚类结果图如图1.15，可以看出 DBSCAN 算法将数据集分为三个簇，较好地捕捉到数据的特征。

```
y_pred = DBSCAN(eps = 0.1, min_samples = 10).fit_predict(X)#选取eps为0.1, PNum为10
plt.scatter(X[:, 0], X[:, 1], c=y_pred)
plt.show()
```

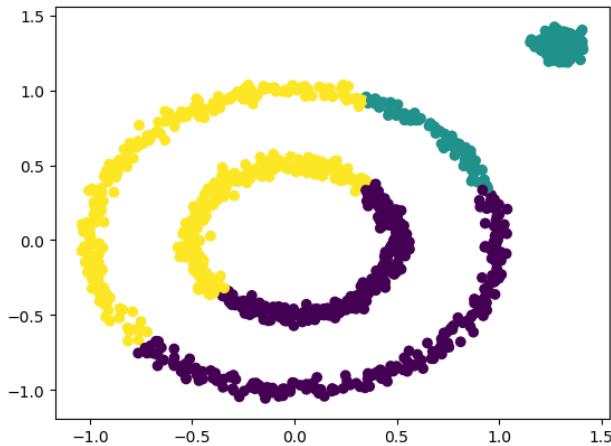


图 1.14 K 均值聚类结果图

## 双向聚类算法

本小节采用 sklearn.cluster 中的 SpectralCoclustering 方法，展示双向聚类算法的实现过程。首先建立原始数据集，实现代码如下：

```
from matplotlib import pyplot as plt
from sklearn.datasets import make_biclusters
from sklearn.cluster import SpectralCoclustering
from sklearn.metrics import consensus_score
data, rows, columns = make_biclusters(shape=(300, 300), n_clusters=5,
noise=5, shuffle=False, random_state=0)
plt.matshow(data, cmap=plt.cm.Blues)#构建原始数据集
plt.title("Original dataset")#绘制原始数据集
```

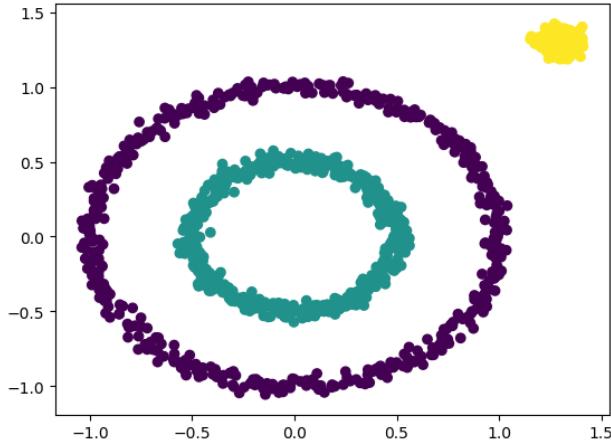


图 1.15 DBSCAN 聚类结果图

绘制出原始数据集如图1.16所示。

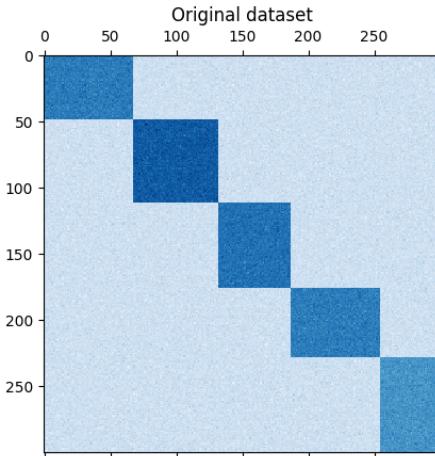


图 1.16 原始数据集

然后将原始数据集打乱：

```
rng = np.random.RandomState(0)
row_idx = rng.permutation(data.shape[0])
col_idx = rng.permutation(data.shape[1])
data = data[row_idx][:, col_idx] # 打乱原始数据集
plt.matshow(data, cmap=plt.cm.Blues)
plt.title("Shuffled dataset") # 绘制打乱后的数据集
```

打乱后的数据集图像1.17所示。下一步利用双向聚类算法对打乱后的数据集重

新排列：

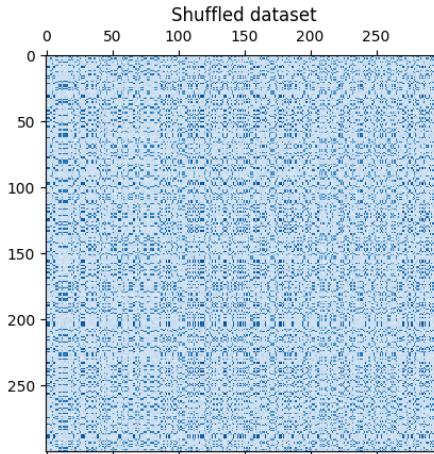


图 1.17 打乱后的数据集

```
model = SpectralCoclustering(n_clusters=5, random_state=0).fit(data)
#调用SpectralCoclustering函数重新排列打乱后的数据集
score = consensus_score(model.biclusters_, (rows[:, row_idx], columns[:, col_idx]))
#计算前后两种双向聚类的相似性
print("consensus score: {:.3f}".format(score))#输出结果consensus score: 1.000
fit_data = data[np.argsort(model.row_labels_), np.argsort(model.column_labels_)]
plt.matshow(fit_data, cmap=plt.cm.Blues)
plt.title("After biclustering; rearranged to show biclusters")
plt.show()#绘制利用双向聚类算法后的数据集
```

重排后的数据集如图1.18所示。可以看出双向聚类算法在本例中有较好的应用效果。

## 1.10 总结

本章介绍了多种常用的聚类算法。K 均值聚类算法采用距离最近原则将样本集划分到不同簇中，并计算每个簇中样本的均值作为簇中心，基于 K 均值的思想，发展了较多变种算法，例如：处理分类属性的 K-modes 算法 [1]、可加快迭代过程的 X-Means 算法 [2] 等。

相比于 K 均值聚类非黑即白的硬聚类过程，FCM 和高斯混合聚类算法采用软聚类的思想，通过计算样本属于某一簇划分的可能性，并且根据可能性大小进行簇划分。这类聚类算法在一些场景中得到较好的聚类效果，被广泛应用于实际问题。

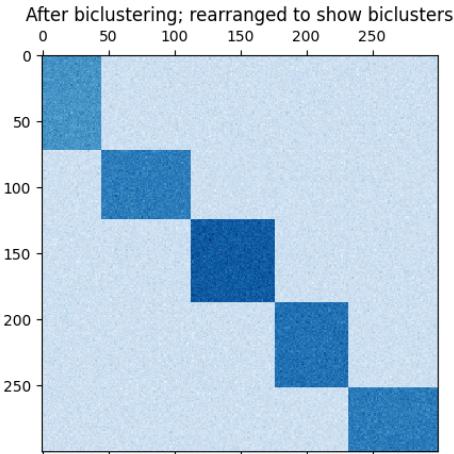


图 1.18 双向聚类算法效果

层次聚类算法通过对簇进行分裂或者合并，达到聚类的目标，本章主要介绍了聚合聚类的算法流程，分裂聚类与之相反，算法是由一个簇生成多个簇的过程。相关改进算法有：BIRCH [3] 等。

DBSCAN 算法从样本集分布紧密程度的角度出发进行簇划分，分布越紧密越有可能被划入同一簇。此方法可应用于非凸样本的聚类分析。更多基于密度的聚类可参考：ST-DBSCAN [4]、OPTICS [5]。

可处理混合型数据的聚类方法，除本章介绍的 K-prototypes 之外，其他方法可参考：ClustMD [6]、KAMILA [7] 等；更多双向聚类方法可参考：谱双向聚类 [8]、信息双向聚类 [9]、凸双向聚类 [10] 等。

## 1.11 习题

- 1、证明：当  $t < 1$  时，式 (1.2.1) 不是距离度量。
- 2、参考聚合聚类算法，试写出分裂聚类的算法流程，并采用分裂聚类对例题1.1进行分析。
- 3、在求解高斯混合模型参数时，我们用到了 EM 算法。下面请叙述 EM 算法一般化步骤并给出相关证明。
- 4、在 K 均值聚类方法中，可采用指示函数法确定最优  $K$  值，请写出误差平方和、平均直径、平均半径的指示函数形式。
- 5、分别采用 K 均值聚类、模糊 C 均值聚类、高斯混合聚类、层次聚类、DBSCAN 聚类分析 2019 年中国主要城市平均气温，并对结果进行分析与对比。（数据来源：

<http://www.stats.gov.cn/tjsj/ndsj/2020/indexch.htm>。数据表名称：8-5 主要城市平均气温（2019 年）。

6、编写程序实现 K-prototypes 算法，作用于 UCI 数据集 German Credit Data ([https://archive.ics.uci.edu/ml/datasets/Statlog+\(German+Credit+Data\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)))，并对结果进行分析。（注意：数据集最后一列为类别标签，聚类过程不予采用）

## 第二章 主成分分析

### 2.1 简介

在实际问题的研究中，往往涉及多变量。而且不同变量之间有一定的相关性。当变量较多时，将增加分析问题的复杂性。由于变量较多且变量之间存在相关性，使得观测到的数据在一定程度上有所重叠。因此利用变量间信息的重叠，通过变量的降维，可以使得复杂问题得到简化。

主成分分析 (principal component analysis, PCA) 是利用降维的思想，在尽量减少信息损失的前提下，将多变量转化为少数几个综合变量的一种机器学习方法。例如，在商品经济中，用主成分分析可以将复杂的经济数据综合成几个商业指数，如物价指数、生活费用指数以及商业活动指数等。主成分分析是由 Pearson [11] 提出，后来被 Hotelling [12] 发展起来的。通常将生成的综合变量称为主成分，这些主成分保留原始变量的绝大部分信息，都是原始变量的线性组合，而且各个主成分之间互不相关。在研究复杂问题时，通过主成分分析，可以从事物错综复杂的关系中找出一些主要成分，揭示事物内部变量之间的规律，简化问题，提高分析效率。

### 2.2 总体的主成分

#### 2.2.1 总体主成分的定义

主成分分析的目标是找到原始变量的一个能够按照“重要性”排序并且信息不重复的线性组合。具体地，假设  $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$  为  $p$  维随机向量，其均值为  $\boldsymbol{\mu}$ ，协方差矩阵为  $\boldsymbol{\Sigma}$ 。对  $\mathbf{X}$  进行线性变换，可以形成新的综合变量，记为  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_p)^T$ ，即

$$\left\{ \begin{array}{l} Y_1 = a_{11}X_1 + a_{21}X_2 + \dots + a_{p1}X_p = \mathbf{a}_1^T \mathbf{X} \\ Y_2 = a_{12}X_1 + a_{22}X_2 + \dots + a_{p2}X_p = \mathbf{a}_2^T \mathbf{X} \\ \vdots \\ Y_p = a_{1p}X_1 + a_{2p}X_2 + \dots + a_{pp}X_p = \mathbf{a}_p^T \mathbf{X} \end{array} \right.$$

首先用一个综合变量  $Y_1$  来替代原始的  $p$  个变量，为了使得  $Y_1$  在  $X_1, X_2, \dots, X_p$  的所有线性组合中最具代表性，应使其方差最大化，以最大程度地保留原始变量的方差和协方差结构信息。由于

$$\text{var}(Y_1) = \text{var}(\mathbf{a}_1^T \mathbf{X}) = \mathbf{a}_1^T \boldsymbol{\Sigma} \mathbf{a}_1$$

若对  $\mathbf{a}_1$  不加以约束，可使得  $Y_1$  的方差任意增大，那么方差最大化就变得没有意义。因此主成分分析限制  $\mathbf{a}_1$  为单位向量，即  $\mathbf{a}_1^T \mathbf{a}_1 = 1$ ，寻求向量  $\mathbf{a}_1$ ，使得  $\text{var}(Y_1) = \text{var}(\mathbf{a}_1^T \mathbf{X})$  达到最大， $Y_1$  就称为第一主成分。如果第一主成分所含信息不够多，不足以代表原始数据的  $p$  个变量，则需要考虑  $Y_2$ ，为了使得  $Y_2$  中所含信息与  $Y_1$  不重叠，则应该要求

$$\text{cov}(Y_1, Y_2) = 0$$

即  $Y_1$  和  $Y_2$  不相关。因此主成分分析在上述约束和条件  $\mathbf{a}_2^T \mathbf{a}_2 = 1$  的条件下寻求  $\mathbf{a}_2$ ，使得  $\text{var}(Y_2) = \text{var}(\mathbf{a}_2^T \mathbf{X})$  达到最大，所求的  $Y_2$  称为第二主成分。类似地，可以定义第三主成分， $\dots$ ，第  $p$  主成分。各主成分在总方差中所占比重依次递减。实际应用中，通常只需挑选前几个主成分，达到简化问题，抓住问题本质的目的。

### 2.2.2 总体主成分的求法

本节将阐述求解  $\mathbf{X}$  主成分的计算过程。假设  $\boldsymbol{\Sigma}$  是  $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$  的协方差矩阵， $\boldsymbol{\Sigma}$  的特征值及其相应的正交单位化特征向量分别为  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > \lambda_{r+1} = \dots = \lambda_p = 0$  及  $\mathbf{a}_1, \dots, \mathbf{a}_p$ ，其中  $r = \text{rank}(\boldsymbol{\Sigma})$ 。首先求出  $\mathbf{X}$  的第一个主成分  $Y_1 = \mathbf{a}_1^T \mathbf{X}$ 。由于第一主成分的系数  $\mathbf{a}_1$  应在条件  $\mathbf{a}_1^T \mathbf{a}_1 = 1$  下，使得  $\mathbf{X}$  的所有线性变换中方差

$$\text{var}(\mathbf{a}_1^T \mathbf{X}) = \mathbf{a}_1^T \boldsymbol{\Sigma} \mathbf{a}_1$$

最大化。因此，求第一主成分就转换为求解以下约束最优化问题：

$$\max_{\mathbf{a}_1} \mathbf{a}_1^T \boldsymbol{\Sigma} \mathbf{a}_1$$

$$\text{s.t.} \quad \mathbf{a}_1^T \mathbf{a}_1 = 1$$

根据拉格朗日乘子法，定义拉格朗日函数

$$L(\mathbf{a}_1, \lambda) = \mathbf{a}_1^T \boldsymbol{\Sigma} \mathbf{a}_1 - \lambda(\mathbf{a}_1^T \mathbf{a}_1 - 1)$$

其中  $\lambda$  为拉格朗日乘子。将拉格朗日函数  $L(\mathbf{a}_1, \lambda)$  分别对参数  $\mathbf{a}_1, \lambda$  求导，令其为 0，即得：

$$\begin{cases} \boldsymbol{\Sigma} \mathbf{a}_1 - \lambda \mathbf{a}_1 = 0 \\ \mathbf{a}_1^T \mathbf{a}_1 = 1 \end{cases}$$

因此,  $\lambda$  是协方差矩阵  $\Sigma$  的特征值,  $\mathbf{a}_1$  是其对应的单位特征向量。可得目标函数

$$\mathbf{a}_1^T \Sigma \mathbf{a}_1 = \mathbf{a}_1^T \lambda \mathbf{a}_1 = \lambda \mathbf{a}_1^T \mathbf{a}_1 = \lambda$$

因此如果  $\mathbf{a}_1$  是  $\Sigma$  的最大特征值  $\lambda_1$  对应的单位特征向量, 则  $\mathbf{a}_1$  与  $\lambda_1$  是上述最优化问题的解。即可得第一个主成分  $Y_1 = \mathbf{a}_1^T \mathbf{X}$ , 其方差为协方差矩阵  $\Sigma$  的最大特征值  $\lambda_1$ , 其系数  $\mathbf{a}_1$  是  $\lambda_1$  对应的单位特征向量。

下面求解  $\mathbf{X}$  的第二个主成分  $Y_2 = \mathbf{a}_2^T \mathbf{X}$ 。由于第二个主成分的系数  $\mathbf{a}_2$  应满足以下条件: 单位向量  $\mathbf{a}_2^T \mathbf{a}_2 = 1$ , 且  $Y_2 = \mathbf{a}_2^T \mathbf{X}$  与  $Y_1 = \mathbf{a}_1^T \mathbf{X}$  不相关, 并使得  $\mathbf{X}$  的所有线性变换中方差

$$\text{var}(\mathbf{a}_2^T \mathbf{X}) = \mathbf{a}_2^T \Sigma \mathbf{a}_2$$

达到最大。因此, 求第二主成分就转换为求解以下约束最优化问题:

$$\begin{aligned} & \max_{\mathbf{a}_2} \mathbf{a}_2^T \Sigma \mathbf{a}_2 \\ \text{s.t. } & \mathbf{a}_2^T \mathbf{a}_2 = 1, \quad \mathbf{a}_1^T \Sigma \mathbf{a}_2 = 0, \quad \mathbf{a}_2^T \Sigma \mathbf{a}_1 = 0 \end{aligned}$$

由于

$$\mathbf{a}_1^T \Sigma \mathbf{a}_2 = \mathbf{a}_2^T \Sigma \mathbf{a}_1 = \mathbf{a}_2^T \lambda_1 \mathbf{a}_1 = \lambda_1 \mathbf{a}_2^T \mathbf{a}_1 = \lambda_1 \mathbf{a}_1^T \mathbf{a}_2$$

则有

$$\mathbf{a}_2^T \mathbf{a}_1 = 0, \quad \mathbf{a}_1^T \mathbf{a}_2 = 0$$

定义拉格朗日函数

$$L(\mathbf{a}_2, \lambda, \phi) = \mathbf{a}_2^T \Sigma \mathbf{a}_1 - \lambda(\mathbf{a}_2^T \mathbf{a}_2 - 1) - \phi \mathbf{a}_2^T \mathbf{a}_1$$

其中  $\lambda, \phi$  为拉格朗日乘子。将拉格朗日函数  $L(\mathbf{a}_2, \lambda, \phi)$  分别对参数  $\mathbf{a}_2, \lambda, \phi$  求导, 令其为 0, 即得:

$$\left\{ \begin{array}{l} 2\Sigma \mathbf{a}_2 - 2\lambda \mathbf{a}_2 - \phi \mathbf{a}_1 = \mathbf{0} \\ \mathbf{a}_2^T \mathbf{a}_2 = 1 \\ \mathbf{a}_2^T \mathbf{a}_1 = 0 \end{array} \right.$$

由于

$$2\mathbf{a}_1^T \Sigma \mathbf{a}_2 - 2\lambda \mathbf{a}_1^T \mathbf{a}_2 - \phi \mathbf{a}_1^T \mathbf{a}_1 = 0$$

则  $\phi = 0$ , 进而可以得到

$$\Sigma \mathbf{a}_2 - \lambda \mathbf{a}_2 = \mathbf{0}$$

显然,  $\lambda$  是协方差矩阵  $\Sigma$  的特征值,  $\mathbf{a}_2$  是其对应的单位特征向量。此时, 目标函数可表示为

$$\mathbf{a}_2^T \Sigma \mathbf{a}_2 = \mathbf{a}_2^T \lambda \mathbf{a}_2 = \lambda \mathbf{a}_2^T \mathbf{a}_2 = \lambda$$

因此如果  $\mathbf{a}_2$  是  $\Sigma$  的第二大特征值  $\lambda_2$  对应的单位特征向量，则  $\mathbf{a}_2$  与  $\lambda_2$  是上述最优化问题的解。即可得第二个主成分  $Y_2 = \mathbf{a}_2^T \mathbf{X}$ ，其方差为协方差矩阵  $\Sigma$  的第二大特征值  $\lambda_2$ ，系数向量  $\mathbf{a}_2$  是  $\lambda_2$  对应的单位特征向量。

以此类推，可知第  $k$  个主成分  $Y_k = \mathbf{a}_k^T \mathbf{X}$ ，其方差为协方差矩阵  $\Sigma$  的第  $k$  大特征值  $\lambda_k$ ，系数向量  $\mathbf{a}_k$  是  $\lambda_k$  对应的单位特征向量。

因此，假设  $\mathbf{X}$  的第  $i$  个主成分为

$$Y_i = \mathbf{a}_i^T \mathbf{X} = a_{1i} X_1 + a_{2i} X_2 + \cdots + a_{pi} X_p$$

其中  $\mathbf{a}_i = (a_{1i}, a_{2i}, \dots, a_{pi})^T$ 。显然有：

$$\begin{cases} \text{var}(Y_i) = \mathbf{a}_i^T \Sigma \mathbf{a}_i = \lambda_i \mathbf{a}_i^T \mathbf{a}_i = \lambda_i, i = 1, \dots, p \\ \text{cov}(Y_i, Y_k) = \mathbf{a}_i^T \Sigma \mathbf{a}_k = \lambda_i \mathbf{a}_i^T \mathbf{a}_k = 0, i \neq k \end{cases}$$

即令  $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_p)$ ，则  $\mathbf{A}$  是一个正交矩阵，且  $\mathbf{A}^T \Sigma \mathbf{A} = \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ ，其中  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$  表示对角矩阵。因此，求主成分问题就转化成了求协方差矩阵的特征值和特征向量。

### 2.2.3 总体主成分的性质

1、主成分的协方差矩阵  $\text{var}(\mathbf{Y}) = \Lambda$ 。即  $\text{var}(Y_i) = \lambda_i, i = 1, \dots, p$  且  $Y_1, Y_2, \dots, Y_p$  互不相关。

2、假设  $\Sigma = (\sigma_{ij})_{p \times p}$  表示  $\mathbf{X}$  的协方差矩阵，则总体主成分的方差之和可表示为：

$$\sum_{i=1}^p \lambda_i = \sum_{i=1}^p \sigma_{ii}$$

事实上，由于  $\Sigma = \mathbf{A} \Lambda \mathbf{A}^T$ ，则

$$\sum_{i=1}^p \lambda_i = \text{tr}(\Lambda) = \text{tr}(\Lambda \mathbf{A}^T \mathbf{A}) = \text{tr}(\mathbf{A} \Lambda \mathbf{A}^T) = \text{tr}(\Sigma) = \sum_{i=1}^p \sigma_{ii}$$

由此可知，主成分分析是把  $p$  个随机变量  $X_1, X_2, \dots, X_p$  的总方差分解为  $p$  个不相关的随机变量  $Y_1, Y_2, \dots, Y_p$  的方差之和。

在主成分分析中，令  $\eta_k$  表示第  $k$  个主成分的方差贡献率，定义为：

$$\eta_k = \frac{\lambda_k}{\sum_{i=1}^p \lambda_i}, \quad k = 1, \dots, p$$

其含义是第  $k$  个主成分  $Y_k$  所提取的信息占总信息的比例。根据主成分分析的算法原理，第一主成分的贡献率最大，意味着  $Y_1$  综合原始变量  $X_1, X_2, \dots, X_p$  所含的信息能力最强，而  $Y_2, Y_3, \dots, Y_p$  的综合能力依次减弱。

前  $m$  个主成分  $Y_1, \dots, Y_m$  的方差贡献率之和定义为:

$$\sum_{i=1}^m \eta_i = \frac{\sum_{i=1}^m \lambda_i}{\sum_{i=1}^p \lambda_i}$$

表示主成分  $Y_1, \dots, Y_m$  的累积贡献率, 其含义是前  $m$  个主成分综合提供原始变量信息的能力。在实际应用中, 通常选取  $m < p$ , 使得前  $m$  个主成分的累积贡献率达到较高的比例(例如, 大于 85%)。此时, 用  $Y_1, \dots, Y_m$  代替原始随机变量  $X_1, X_2, \dots, X_p$  不但使得变量的维数降低, 而且也不损失太多的信息。

3、设矩阵  $\mathbf{A}$  的第  $i$  行第  $j$  列元素为  $A_{ij}$ 。由于  $\mathbf{Y} = \mathbf{A}^T \mathbf{X}$ , 则有  $\mathbf{X} = \mathbf{A}\mathbf{Y}$ , 故而有  $X_j = A_{j1}Y_1 + A_{j2}Y_2 + \dots + A_{jp}Y_p$ ,  $\text{cov}(Y_i, X_j) = \lambda_i A_{ji}$ , 则可得主成分  $Y_i$  与原始变量  $X_j$  的相关系数为

$$\rho_{Y_i, X_j} = \frac{\text{cov}(Y_i, X_j)}{\sqrt{\text{var}(Y_i)} \sqrt{\text{var}(X_j)}} = \frac{\lambda_i A_{ji}}{\sqrt{\lambda_i} \sqrt{\sigma_{jj}}} = \frac{\sqrt{\lambda_i}}{\sqrt{\sigma_{jj}}} A_{ji}$$

它给出了主成分  $Y_i$  与原始变量  $X_j$  的线性关联性的度量, 也称为因子负荷量或因子载荷量。

4、之前所提到的累积贡献率度量了前  $m$  个主成分  $Y_1, \dots, Y_m$  综合提供原始变量  $X_1, X_2, \dots, X_p$  所含的信息能力, 那么前  $m$  个主成分中包含原始变量  $X_j$  有多少信息应该如何度量呢? 这个指标为前  $m$  个主成分  $Y_1, \dots, Y_m$  与原始变量  $X_j$  的相关系数的平方和, 我们称之为  $Y_1, \dots, Y_m$  对原始变量  $X_j$  的贡献率。

下面我们通过一个例子阐述总体主成分的计算方法。

**例 2.1** 设随机变量  $\mathbf{X} = (X_1, X_2, X_3)^T$  的协方差矩阵为

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

则知  $\boldsymbol{\Sigma}$  的特征值为  $\lambda_1 = 3 + \sqrt{8}$ ,  $\lambda_2 = 2$ ,  $\lambda_3 = 3 - \sqrt{8}$ , 相应的单位正交特征向量为

$$\mathbf{a}_1 = \begin{bmatrix} 0.383 \\ -0.924 \\ 0.000 \end{bmatrix}, \quad \mathbf{a}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{a}_3 = \begin{bmatrix} 0.924 \\ 0.383 \\ 0.000 \end{bmatrix}$$

因此, 主成分为

$$Y_1 = 0.383X_1 - 0.924X_2$$

$$Y_2 = X_3$$

$$Y_3 = 0.924X_1 + 0.383X_2$$

取  $m = 1$  时,  $Y_1$  的累积贡献率为  $\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = \frac{3 + \sqrt{8}}{8} = 72.8\%$ ; 取  $m = 2$  时,  $Y_1, Y_2$  的累积贡献率为 97.85%。下表列出  $m$  个主成分对变量  $X_i$  的贡献率。

$i$	$\rho_{Y_1, X_i}$	$\rho_{Y_1, X_i}^2$	$\rho_{Y_2, X_i}$	$\rho_{Y_2, X_i}^2$
1	0.925	0.855	0.000	0.000
2	-0.998	0.996	0.000	0.000
3	0.000	0.000	1.000	1.000

由此可见, 当  $m = 1$  时,  $Y_1$  的累积贡献率已达 72.8%, 但是  $Y_1$  对  $X_3$  的贡献率为 0, 这是因为在  $Y_1$  中没有包含  $X_3$  的任何信息, 因此仅取  $m = 1$  不够, 故而取  $m = 2$ , 这时  $Y_1, Y_2$  的累积贡献率为 97.85%, 且  $Y_1, Y_2$  对  $X_i (i = 1, 2, 3)$  的贡献率也比较高。

#### 2.2.4 标准化变量的主成分

在实际问题中, 通常有两种情形不适合直接从协方差矩阵  $\Sigma$  出发求主成分。一种是各变量的单位不全相同, 对同样的变量使用不同的单位直接进行主成分分析, 其结果一般是不一样的, 甚至差异较大, 这样做出来的分析也没有意义。另一种是各变量的单位虽相同, 但是其变量方差的差异较大, 以至于主成分分析的结果往往倾向于方差大的变量, 而方差小的变量几乎被忽略。因此, 对这两种情形, 通常先将各原始变量做标准化处理, 然后从标准化变量的协方差矩阵出发求主成分。常用的标准化变换为:

$$X_i^* = \frac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, i = 1, \dots, p$$

其中  $\mu_i = E(X_i), \sigma_{ii} = \text{var}(X_i)$ 。此时  $\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*)^T$  的协方差矩阵为原始变量  $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$  的相关系数矩阵  $\boldsymbol{\rho} = (\rho_{ij})_{p \times p}$ , 其中

$$\rho_{ij} = \frac{\text{cov}(X_i, X_j)}{\sqrt{\text{var}(X_i)} \sqrt{\text{var}(X_j)}}$$

因此只需直接从相关系数矩阵  $\boldsymbol{\rho}$  出发求主成分, 此时的主成分分析将均等地对待每一个原始变量。从相关系数矩阵出发求的主成分与从协方差矩阵出发是完全类似的, 并且主成分的一些性质具有更简单的数学形式。

设  $\boldsymbol{\rho}$  的特征值  $\lambda_1^* \geq \lambda_2^* \geq \dots \geq \lambda_r^* > \lambda_{r+1}^* = \dots = \lambda_p^* = 0$ , 其中  $r = \text{rank}(\boldsymbol{\rho})$ ,  $\boldsymbol{\rho}$  的  $p$  个单位特征向量为  $\mathbf{a}_1^*, \dots, \mathbf{a}_p^*$ , 且相互正交, 则  $p$  个主成分为:  $Y_1^* = \mathbf{a}_1^{*\top} \mathbf{X}^*, Y_2^* = \mathbf{a}_2^{*\top} \mathbf{X}^*, \dots, Y_p^* = \mathbf{a}_p^{*\top} \mathbf{X}^*$ 。记  $\mathbf{Y}^* = (Y_1^*, Y_2^*, \dots, Y_p^*)^T$ ,  $\mathbf{A}^* = (\mathbf{a}_1^*, \mathbf{a}_2^*, \dots, \mathbf{a}_p^*)$ , 则有  $\mathbf{Y}^* = \mathbf{A}^{*\top} \mathbf{X}^*$ 。上述主成分具有的性质可以概括如下:

1、 $E(\mathbf{Y}^*) = \mathbf{0}, \text{var}(\mathbf{Y}^*) = \mathbf{\Lambda}^* = \text{diag}(\lambda_1^*, \dots, \lambda_p^*)$ 。

- 2、 $\sum_{i=1}^p \text{var}(Y_i^*) = \sum_{i=1}^p \lambda_i^* = \sum_{i=1}^p \text{var}(X_i^*) = p$ 。
- 3、第  $k$  个主成分  $Y_k^*$  的贡献率为  $\frac{\lambda_k^*}{p}$ ，前  $m$  个主成分  $Y_1^*, \dots, Y_m^*$  的累积贡献率为  $\frac{\sum_{i=1}^m \lambda_i^*}{p}$ 。
- 4、主成分  $Y_i^*$  与  $X_j^*$  的相关系数为  $\rho_{Y_i^*, X_j^*} = \sqrt{\lambda_i^*} A_{ji}^*$ 。

下面通过一个例子说明分别从协方差矩阵和相关系数矩阵出发求主成分的差异。

**例 2.2** 随机变量  $\mathbf{X} = (X_1, X_2, X_3)^T$  的协方差矩阵为

$$\boldsymbol{\Sigma} = \begin{bmatrix} 16 & 2 & 30 \\ 2 & 1 & 4 \\ 30 & 4 & 100 \end{bmatrix}$$

其相关系数矩阵为

$$\boldsymbol{\rho} = \begin{bmatrix} 1.00 & 0.50 & 0.75 \\ 0.50 & 1.00 & 0.40 \\ 0.75 & 0.40 & 1.00 \end{bmatrix}$$

经计算可知  $\boldsymbol{\Sigma}$  的特征值为  $\lambda_1 = 109.793, \lambda_2 = 6.494, \lambda_3 = 0.738$ ，相应的单位正交特征向量为

$$\mathbf{a}_1 = \begin{bmatrix} 0.305 \\ 0.041 \\ 0.951 \end{bmatrix}, \quad \mathbf{a}_2 = \begin{bmatrix} 0.944 \\ 0.120 \\ -0.308 \end{bmatrix}, \quad \mathbf{a}_3 = \begin{bmatrix} -0.127 \\ 0.992 \\ -0.002 \end{bmatrix}$$

因此，主成分为

$$Y_1 = 0.305X_1 + 0.041X_2 + 0.951X_3$$

$$Y_2 = 0.944X_1 + 0.120X_2 - 0.308X_3$$

$$Y_3 = -0.127X_1 + 0.992X_2 - 0.002X_3$$

$Y_1$  的贡献率为  $\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3} = 0.938$ 。如此高的贡献率归因于  $X_1$  的大方差，以及  $X_1, X_2, X_3$  之间存在一定的相关性。

相关系数矩阵  $\boldsymbol{\rho}$  的特征值为  $\lambda_1^* = 2.114, \lambda_2^* = 0.646, \lambda_3^* = 0.240$ ，相应的单位正交特征向量为

$$\mathbf{a}_1^* = \begin{bmatrix} 0.627 \\ 0.497 \\ 0.600 \end{bmatrix}, \quad \mathbf{a}_2^* = \begin{bmatrix} -0.241 \\ 0.856 \\ -0.457 \end{bmatrix}, \quad \mathbf{a}_3^* = \begin{bmatrix} -0.741 \\ 0.142 \\ 0.656 \end{bmatrix}$$

因此，主成分为

$$Y_1^* = 0.627X_1^* + 0.497X_2^* + 0.600X_3^*$$

$$Y_2^* = -0.241X_1^* + 0.856X_2^* - 0.457X_3^*$$

$$Y_3^* = -0.741X_1^* + 0.142X_2^* + 0.656X_3^*$$

$Y_1^*$  的贡献率为  $\frac{\lambda_1^*}{3} = 0.705$ ,  $Y_1^*$  和  $Y_2^*$  的累积贡献率为  $\frac{\lambda_1^* + \lambda_2^*}{3} = 0.920$ 。比较从  $\Sigma$  出发和从  $\rho$  出发的主成分分析结果。可知从  $\rho$  出发的  $Y_1^*$  的贡献率 0.705 明显小于从  $\Sigma$  出发的  $Y_1$  的贡献率 0.938, 事实上, 原始变量方差之间的差异越大, 这一点往往越明显。此例也说明标准化后的结论可能会发生很大的变化, 因此标准化并不是无关紧要的。

## 2.3 样本主成分

在上一节, 我们可以从协方差矩阵  $\Sigma$  或相关系数矩阵  $\rho$  出发求主成分。但是在实际问题中,  $\Sigma$  和  $\rho$  一般都是未知的, 需要通过样本来进行估计得到。设

$$\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T, i = 1, \dots, n$$

为取自样本数据矩阵  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$  的一个简单随机样本, 其中  $p$  表示特征维数,  $n$  表示样本数。样本的协方差矩阵和相关系数矩阵分别为

$$\begin{aligned} \mathbf{S} &= (s_{ij})_{p \times p} = \frac{1}{n-1} \sum_{k=1}^n (\mathbf{X}_k - \bar{\mathbf{X}})(\mathbf{X}_k - \bar{\mathbf{X}})^T \\ \mathbf{R} &= (r_{ij})_{p \times p} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}} \end{aligned}$$

其中

$$\begin{aligned} \bar{\mathbf{X}} &= (\bar{X}_1, \dots, \bar{X}_p)^T, \quad \bar{X}_i = \frac{1}{n} \sum_{j=1}^n X_{ji}, \quad i = 1, \dots, p, \\ s_{ij} &= \frac{1}{n-1} \sum_{k=1}^n (X_{ki} - \bar{X}_i)(X_{kj} - \bar{X}_j) \quad i, j = 1, \dots, p \end{aligned}$$

分别以  $\mathbf{S}, \mathbf{R}$  作为  $\Sigma, \rho$  的估计, 再按照总体主成分的方法求得的主成分称之为样本主成分。

设  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_r > \hat{\lambda}_{r+1} = \dots = \hat{\lambda}_p = 0$  为样本协方差矩阵  $\mathbf{S}$  的特征值,  $\hat{\mathbf{a}}_1, \dots, \hat{\mathbf{a}}_p$  为对应的正交单位化特征向量, 其中  $\hat{\mathbf{a}}_i = (\hat{a}_{1i}, \dots, \hat{a}_{pi})^T$ , 则第  $m$  个样本的第  $i$  个主成分可表示为  $Y_i = \hat{\mathbf{a}}_i^T \mathbf{X}_m$ 。此时, 可以得到

1、 $Y_i$  的样本方差为

$$\text{var}(Y_i) = \hat{\mathbf{a}}_i^T \mathbf{S} \hat{\mathbf{a}}_i = \hat{\lambda}_i, \quad i = 1, \dots, p$$

2、 $Y_i$  与  $Y_j$  的样本协方差为

$$\text{cov}(Y_i, Y_j) = \hat{\mathbf{a}}_i^T \mathbf{S} \hat{\mathbf{a}}_j = 0$$

3、样本总方差为

$$\sum_{i=1}^p s_{ii} = \sum_{i=1}^p \hat{\lambda}_i$$

第  $i$  个主成分的贡献率为

$$\hat{\eta}_i = \frac{\hat{\lambda}_i}{\sum_{i=1}^p \hat{\lambda}_i}, \quad i = 1, \dots, p$$

前  $k$  个主成分累积贡献率为

$$\sum_{i=1}^k \hat{\eta}_i = \frac{\sum_{i=1}^k \hat{\lambda}_i}{\sum_{i=1}^p \hat{\lambda}_i}$$

类似的，为了避免单位不统一或者变量之间差异性较大产生的影响，可以对样本进行标准化处理，即令

$$\mathbf{X}_j^* = \left( \frac{X_{j1} - \bar{X}_1}{\sqrt{s_{11}}}, \frac{X_{j2} - \bar{X}_2}{\sqrt{s_{22}}}, \dots, \frac{X_{jp} - \bar{X}_p}{\sqrt{s_{pp}}} \right)^T, j = 1, \dots, n$$

标准化后数据的样本协方差矩阵即为原始数据的样本相关系数矩阵  $\mathbf{R}$ 。由  $\mathbf{R}$  出发所求的样本主成分称为标准化样本主成分。计算  $\mathbf{R}$  的特征值及相应的正交单位化特征向量即可求得标准化样本主成分。此时，标准化的样本总方差为  $p$ .

选取前  $m$  个样本主成分，使其累积贡献率达到一定的要求（例如：大于 85%），用这  $m$  个样本主成分代替原始数据进行分析，可以达到在保留大部分信息的前提下，降低原始数据维数的目的。

## 2.4 非线性主成分分析

传统主成分分析一般适应于线性降维，其对于非线性数据往往不能达到较好的效果，例如，不同人之间的人脸图像存在非线性关系，用传统的线性主成分分析结果不尽人意。下面介绍几种非线性主成分分析算法。

### 2.4.1 核主成分分析

**核主成分分析** (Kernel Principal Component Analysis, KPCA) 是对传统主成分分析(PCA)算法的非线性拓展。简单地说,通过将非线性不可分问题映射到维度更高的特征空间,使其在新的特征空间上线性可分。设样本数据矩阵  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ , 其中  $p$  表示特征维数,  $n$  表示样本数。 $\mathbf{X} \in \mathcal{R}^p$  为取自  $\mathbf{X}$  的一个简单随机样本,为了将其映射到维度更高的  $k$  维子空间, 定义如下非线性映射函数  $\phi$ :

$$\phi : \mathcal{R}^p \rightarrow \mathcal{R}^k (p \ll k)$$

换句话说, 利用 KPCA, 可以通过非线性变换将数据映射到一个高维空间, 然后在此高维空间中使用标准 PCA 将其映射到另外一个低维空间中, 并通过线性分类器进行划分。但是, 由于协方差矩阵每个元素都是向量的内积, 因此映射到高维度空间后, 向量维度增加导致计算量大幅度增大。故而, 可以利用核函数忽略映射函数的具体形式, 直接得到低维数据映射到高维后的内积。

假设  $\mathbf{X}$  表示标准化后的样本数据矩阵, 则  $\phi(\mathbf{X})$  是一个映射后的矩阵, 维数是  $n \times k$ , 可以计算得到协方差矩阵为

$$\Sigma = \frac{1}{n} \phi(\mathbf{X})^T \phi(\mathbf{X})$$

特征向量可表示为  $\nu = \phi(\mathbf{X})^T \mathbf{a}$ , 代入  $\Sigma \nu = \lambda \nu$ , 则可得

$$\frac{1}{n} \phi(\mathbf{X})^T \phi(\mathbf{X}) \phi(\mathbf{X})^T \mathbf{a} = \lambda \phi(\mathbf{X})^T \mathbf{a}$$

两边乘以  $\phi(\mathbf{X})$ , 则可得

$$\frac{1}{n} \phi(\mathbf{X}) \phi(\mathbf{X})^T \mathbf{a} = \lambda \mathbf{a}$$

即

$$\frac{1}{n} \mathbf{K} \mathbf{a} = \lambda \mathbf{a} \quad (2.4.1)$$

其中  $\mathbf{K}$  为**核矩阵**:  $\mathbf{K} = \phi(\mathbf{X}) \phi(\mathbf{X})^T$ 。显然,  $\mathbf{K}$  由高维空间中的内积决定, 为了避免由此带来的复杂计算, 可通过定义核函数表示样本点在高维空间中的内积, 令  $k(\mathbf{X}, \mathbf{Y})$  表示核函数,  $\mathbf{X} = (X_1, \dots, X_p)^T$ 、 $\mathbf{Y} = (Y_1, \dots, Y_p)^T$  表示样本。此时矩阵  $\mathbf{K}$  的第  $i$  行, 第  $j$  列元素  $\mathbf{K}_{ij} = k(\mathbf{X}_i, \mathbf{Y}_j)$ 。由单位特征向量的假定知  $\nu^T \nu = 1$ , 推出  $\mathbf{a}^T \mathbf{K} \mathbf{a} = 1$ , 因此得到条件

$$n \lambda \mathbf{a}^T \mathbf{a} = 1 \quad (2.4.2)$$

利用 (2.4.1) 和条件 (2.4.2) 可以求解出未知向量  $\mathbf{a}$ , 以及对应的特征值和特征向量。接下来, 对于一个新的样本  $\mathbf{X}$ , 我们可以得到他的第一主成分是

$$\phi(\mathbf{X})^T \nu_1 = \sum_{i=1}^n a_i k(\mathbf{X}, \mathbf{X}_i)$$

其中  $\nu_1$  是最大特征值对应的特征向量。

常用的核函数有：

1、**线性核函数**：

$$k(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^T \mathbf{Y} + c$$

其中  $c$  为参数。

2、**多项式核函数**：

$$k(\mathbf{X}, \mathbf{Y}) = (a\mathbf{X}^T \mathbf{Y} + c)^d$$

其中  $a, b, c$  为参数。

3、**高斯核函数**：

$$k(\mathbf{X}, \mathbf{Y}) = \exp\left(-\frac{\|\mathbf{X} - \mathbf{Y}\|^2}{2\sigma^2}\right)$$

其中  $\sigma$  为参数，高斯核函数是径向基函数核的一个典型代表。

4、**指数核函数**：

$$k(\mathbf{X}, \mathbf{Y}) = \exp\left(-\frac{\|\mathbf{X} - \mathbf{Y}\|}{2\sigma}\right)$$

其中  $\sigma$  为参数。指数核函数也是径向基函数核代表，与高斯核函数很像，只是将  $L_2$  范数变成  $L_1$  范数。

5、**拉普拉斯核函数**：

$$k(\mathbf{X}, \mathbf{Y}) = \exp\left(-\frac{\|\mathbf{X} - \mathbf{Y}\|}{\sigma}\right)$$

拉普拉斯核函数完全等价于指数核，区别在于前者对参数的敏感性降低，也是一种径向基函数核。

值得注意的是，上述理论均是在  $\phi(\mathbf{X})$  已经中心化的前提下完成的。在实际应用中，应首先将矩阵  $\phi(\mathbf{X})$  中心化，即

$$\tilde{\phi}(\mathbf{X}) = \phi(\mathbf{X}) - \mathbf{1}_n \cdot \phi(\mathbf{X})$$

其中  $\mathbf{1}_n = \frac{1}{n} \mathbf{1}_{n \times 1} \mathbf{1}_{n \times 1}^T$ ，为  $n \times n$  矩阵，其每个元素为  $\frac{1}{n}$ 。由上述可知，不需要显示计算  $\tilde{\phi}(\mathbf{X})$ ，只需得到中心化后的核矩阵即可：

$$\widetilde{\mathbf{K}} = \tilde{\phi}(\mathbf{X}) \tilde{\phi}(\mathbf{X})^T = \mathbf{K} - \mathbf{K} \cdot \mathbf{1}_n - \mathbf{1}_n \cdot \mathbf{K} + \mathbf{1}_n \cdot \mathbf{K} \cdot \mathbf{1}_n$$

因此上面介绍 KPCA 使用的  $\phi(\mathbf{X})$  和  $\mathbf{K}$  本质上就是这里的  $\tilde{\phi}(\mathbf{X})$  和  $\widetilde{\mathbf{K}}$ 。

## 2.4.2 t-SNE 非线性降维算法

**t 分布随机邻域嵌入** (*t*-distributed stochastic neighbor embedding *t*-SNE) 是一种针对高维数据的非线性降维算法，在 2008 年由 Laurens van der Maaten 和

Geoffrey Hinton [13] 提出。传统主成分分析是一种线性算法，不能解释特征之间的复杂多项式关系。而 t-SNE 是基于邻域图上随机游走的概率分布寻找数据内的结构，将数据点之间的相似度转化为条件概率，原始空间中数据点的相似度由正态分布表示，嵌入空间中数据点的相似度由  $t$  分布表示。通过原始空间和嵌入空间的联合概率分布的KL 散度（用于评估两个分布的相似度的指标）来评估嵌入效果的好坏。

### 1、SNE 算法

$t$ -SNE 算法是从 SNE 改进而来，所以先介绍 SNE。给定一组高维数据  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ ,  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$ 。目标是将这组数据降维到二维，SNE 的基本思想是若两个数据在高维空间中是相似的，那么降维到二维空间时距离应当较近。

随机邻近嵌入 (SNE) 首先通过将数据点之间的高维欧几里得距离转换为相似性的条件概率来描述两个数据之间的相似性。

假设高维空间中的两个点  $\mathbf{X}_i, \mathbf{X}_j$ ，以点  $\mathbf{X}_i$  为中心构建方差为  $\sigma_i$  的高斯分布。用  $P_{j|i}$  表示  $\mathbf{X}_j$  在  $\mathbf{X}_i$  邻域的概率，若  $\mathbf{X}_j$  与  $\mathbf{X}_i$  相距很近，那么  $P_{j|i}$  很大；反之， $P_{j|i}$  很小， $P_{j|i}$  可以表示为

$$P_{j|i} = \frac{\exp(-\|\mathbf{X}_j - \mathbf{X}_i\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{X}_k - \mathbf{X}_i\|^2 / 2\sigma_i^2)}, \quad P_{i|i} = 0,$$

其中  $\|\mathbf{X}_i - \mathbf{X}_j\|$  表示点  $\mathbf{X}_i, \mathbf{X}_j$  的欧式距离，这里只关注不同数据点间的距离是因为我们设置  $P_{i|i} = 0$ 。高斯核的带宽  $\sigma_i$  是条件概率中所涉及到的范围。有些特征点是稀疏的，有些比较紧密，因此带宽大小也是不同的。一般来说，数据密度高的区域带宽要小于数据密度低的区域。每个数据点的高斯核带宽的最优值可以通过简单的二进制搜索 [14] 等得到。

当把数据映射到低维空间后，高维数据点之间的相似性也应该在低维空间的数据点上体现出来。假设  $\mathbf{X}_i, \mathbf{X}_j$  映射到低维空间后对应  $\mathbf{Y}_i, \mathbf{Y}_j$ ，那么  $\mathbf{Y}_i, \mathbf{Y}_j$  邻域的条件概率为  $Q_{j|i}$ ：

$$Q_{j|i} = \frac{\exp(-\|\mathbf{Y}_j - \mathbf{Y}_i\|^2)}{\sum_{k \neq i} \exp(-\|\mathbf{Y}_k - \mathbf{Y}_i\|^2)}$$

低维空间中的方差直接设置为  $\sigma_i = \frac{1}{\sqrt{2}}$ 。同样  $Q_{i|i} = 0$ 。

若条件概率  $Q_{j|i}$  反映了高维数据点  $\mathbf{X}_i, \mathbf{X}_j$  之间的关系，那么我们希望条件概率  $P_{j|i}$  与  $Q_{j|i}$  应该完全相等。若给定  $\mathbf{X}_i$  与其他所有点之间的条件概率，则可构成一个条件概率分布  $\mathcal{P}_i$ 。同理在低维空间存在一个条件概率分布  $\mathcal{Q}_i$  与之对应，那么我们希望条件概率分布  $\mathcal{Q}_i$  与  $\mathcal{P}_i$  完全一样。为了衡量两个分布之间的相似性，采用 Kullback-Leibler (KL) 散度最小化低维与高维下两个分布条件概率的差异，SNE 最终目标就是对所有数据点最小化 KL 距离，可以使用梯度下降算法最小化如下代价函数：

$$C = \sum_i \text{KL}(\mathcal{P}_i \| \mathcal{Q}_i) = \sum_i \sum_j P_{j|i} \log \frac{P_{j|i}}{Q_{j|i}}$$

但由于 KL 距离是一个非对称的度量。这意味着当  $P_{j|i}$  较大,  $Q_{j|i}$  较小时, 代价较高; 而  $P_{j|i}$  较小,  $Q_{j|i}$  较大时, 代价较低。即高维空间中两个数据点距离较近时, 若映射到低维空间后距离较远, 那么将得到一个很高的惩罚, 这符合我们的初衷。反之, 高维空间中两个数据点距离较远时, 若映射到低维空间距离较近, 将得到一个很低的惩罚值, 我们的初衷是这里也应得到一个较高的惩罚。即 SNE 的代价函数更关注局部结构, 而忽视了全局结构。

## 2、*t*-SNE 算法

在 SNE 中, 高维空间中条件概率  $P_{j|i}$  不等于  $P_{i|j}$ , 低维空间中  $Q_{j|i}$  不等于  $Q_{i|j}$ , 于是为简化计算提出对称 SNE, 采用联合概率分布代替原始的条件概率, 使得  $P_{ij} = P_{ji}$ ,  $Q_{ij} = Q_{ji}$  优化  $P_{i|j}$  和  $Q_{i|j}$  的 KL 散度的一种替换思路是, 使用联合概率分布来替换条件概率分布, 即  $\mathcal{P}$  和  $\mathcal{Q}$  分别是高维空间和低维空间里各个点的联合概率分布, 此时目标函数为:

$$C = \text{KL}(\mathcal{P} \| \mathcal{Q}) = \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}}$$

其中, 在高维空间中定义  $P_{ij}$ :

$$P_{ij} = \frac{P_{i|j} + P_{j|i}}{2}$$

为了使得高维度下的中高等距离在映射后距离更大, 以减轻拥挤问题。在低维空间中使用更加偏重长尾分布的方式将距离转换为概率分布。因此, 对于高维数据点  $\mathbf{X}_i$  和  $\mathbf{X}_j$  的低维对应点  $\mathbf{Y}_i$  和  $\mathbf{Y}_j$ , 采用自由度为 1 的归一化的  $t$  核:

$$Q_{ij} = \frac{(1 + \|\mathbf{Y}_i - \mathbf{Y}_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|\mathbf{Y}_k - \mathbf{Y}_l\|^2)^{-1}}, \quad Q_{ii} = 0$$

因此  $t$ -SNE 是通过仿射变换将数据点映射到概率分布, 将两个数据点之间的距离转换为以一个点为中心一定范围内另一个点出现的条件概率。由于在嵌入空间中的目标函数是非凸的, 因此, 可以用梯度下降最小化目标函数, 此时

$$\frac{\partial C}{\partial \mathbf{Y}_i} = 4 \sum_{i \neq j} (P_{ij} - Q_{ij}) Q_{ij} Z (\mathbf{Y}_i - \mathbf{Y}_j)$$

其中  $Z = \sum_{k \neq l} (1 + \|\mathbf{Y}_k - \mathbf{Y}_l\|^2)^{-1}$ 。

因此,  $t$ -SNE 的主要优势在于通过  $t$  分布与正态分布的差异, 解决了样本分布拥挤、边界不明显的问题, 使得相似的样本能够聚集在一起, 而差异大的样本能够有效地分开。但它同时也有计算复杂度高, 目标函数非凸, 容易得到局部最优解; 对参数敏感、结果具有随机性等缺陷。

$t$ -SNE 非线性降维算法在实际问题中可用于处理高维数据集, 并应用于自然语言处理、图像处理、基因组数据和语音处理等。

## 2.5 主成分分析实践

### 2.5.1 R 语言实践

#### 主成分分析

R 语言中可用于进行主成分分析的函数有: prcomp() 和 princomp(), 以及 psych 包对数据进行主成分分析, 其中, principal 函数进行主成分分析, fa.parallel 函数生成碎石图等。也可根据 PCA 的原理, 利用自编函数的方法。下面将利用 R 中自带的范例数据集 iris 进行示范。

1、标准化数据: 为了统一数据的量纲并对数据进行中心化, 在主成分分析之前往往需要对原始数据进行标准化。

```
data<-iris
irisda<-as.matrix(scale(data[,1:4]))#对原数据进行z-score归一化
head(irisda)
```

2、计算相关系数 (协方差) 矩阵并求解特征值和相应的特征向量, 提取各主成分方差, 计算方差贡献率和累积贡献率, 利用碎石函数 screeplot() 绘制碎石图。碎石图是可以用于确定主成分合适个数的有用的可视化工具。其将特征值从大到小排列, 选取一个拐点对应的序号, 序号后的特征值全部较小且大小差不多, 因此此值可作为主成分的个数。

```
covm<-cor(irisda)#归一化后相关系数等于协方差
eign<-eigen(covm)
val <- eign$values#提取结果中的特征值, 即各主成分的方差
(Standard_deviation <- sqrt(val))#换算成标准差(Standard deviation);
(Proportion_of_variance <- val/sum(val))
(Cumulative_Proportion <- cumsum(Proportion_of_variance))#计算方差贡献率和累积贡献率
par(mar=c(6,6,2,2))
plot(eign$values,type="b", cex=2, cex.lab=2, cex.axis=2, lty=2, lwd=2,
xlab = "主成分编号", ylab="特征值")#碎石图绘制
```

#### 3、计算主成分得分

```
(U<-as.matrix(eign$vectors))#提取结果中的特征向量
PC <-irisda %*% U#进行矩阵乘法, 获得PC score;
colnames(PC) <- c("PC1","PC2","PC3","PC4")
head(PC)
```

#### 4、绘制主成分散点图

```
df<-data.frame(PC,iris$Species)#与iris数据集的第5列数据合并
head(df)
library(ggplot2)
```

```
#提取主成分的方差贡献率，生成坐标轴标题
xlab<-paste0("PC1(",round(Proportion_of_variance[1]*100,2),"%)")
ylab<-paste0("PC2(",round(Proportion_of_variance[2]*100,2),"%)")
p1<-ggplot(data = df,aes(x=PC1,y=PC2,color=iris.Species))+ 
stat_ellipse(aes(fill=iris.Species),
type ="norm", geom ="polygon",alpha=0.2,color=NA)+ 
geom_point() + labs(x=xlab,y=ylab,color="")+
guides(fill="none")#绘制散点图并添加置信椭圆
p1
```

本节只展示直观性强的碎石图2.1和主成分散点图2.2。

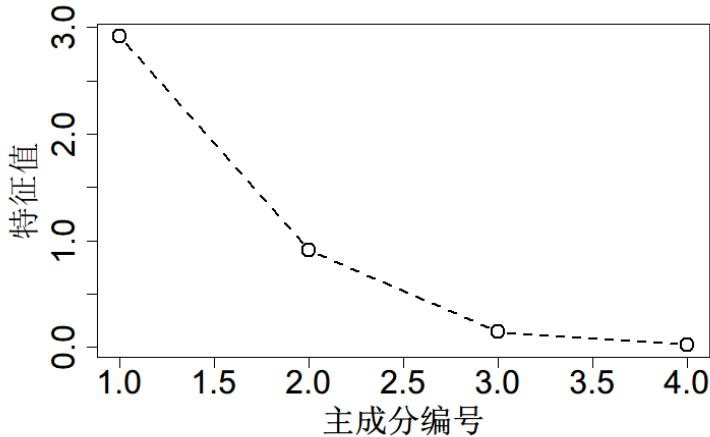


图 2.1 screeplot



图 2.2 PCA

在理解了主成分分析的具体步骤后，也可用 R 语言中常见的 PCA 函数验证以

上分析过程的正确性。在实际操作中，可以根据具体情况，按需选择函数。

### (1) prcomp() 函数

```
pca1 <- prcomp(data[,1:4], center = TRUE, scale. = TRUE)
summary(pca1)
```

(2) princomp() 函数：如果使用 princomp() 函数，需要先做归一化，princomp() 函数并无数据标准化相关的参数，且默认使用 covariance matrix，得到的结果与使用相关性矩阵有细微差异（如下）。原因是：根据相关系数公式可知，归一化后的相关性系数近乎等于协方差。

```
pca2 <- princomp(irisda, cor = T)
summary(pca2)
pca3 <- princomp(irisda)
summary(pca3)
```

通过比较分析得到 Standard deviation、Proportion of variance、Cumulative Proportion 发现，3 种主成分分析的方法得到的结果完全一致。

### (3) principal() 函数：

```
PC <- principal(irisda, nfactors=2, rotate = "none")
pc <- data.frame(PC$scores)
p <- ggplot(pc, aes(x=PC1, y=PC2, color="species")) +
  geom_point(size=4, alpha=0.5) +
  theme(axis.text = element_text(size=20)) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_line(color="steelblue"),
        panel.background = element_blank(),
        axis.line = element_line(colour = "black")) +
  stat_ellipse(lwd=1, level = 0.8)
p
```

其中 nfactors: 指定主成分个数，rotate: 指定模型旋转的方法，默認為最大方差法。最终可得主成分散点图2.3。

## 非线性主成分分析

(1) 核主成分分析 (KPCA)：在 R 中可以通过 kernlab 包中函数 kPCA() 实现基于核分析的主成分分析。以 R 中自带的范例数据集 iris 进行示范。

```
data(iris)
library(kernlab)
test <- sample(1:150, 20)
kpc <- kPCA(~., data=iris[-test, -5], kernel="rbfdot", kpar=list(sigma=0.2), features=2)
pcv(kpc)
#绘制数据图
plot(rotated(kpc), col=as.integer(iris[-test, 5]),
xlab="1st-Principal-Component", ylab="2nd-Principal-Component")
emb <- predict(kpc, iris[test, -5]) #嵌入余下的点
points(emb, col=as.integer(iris[test, 5]))
```

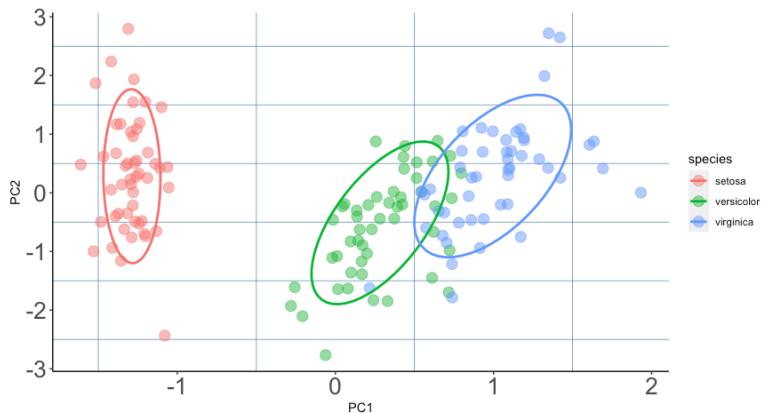


图 2.3 PCA-principal()

其可视化图为2.4。

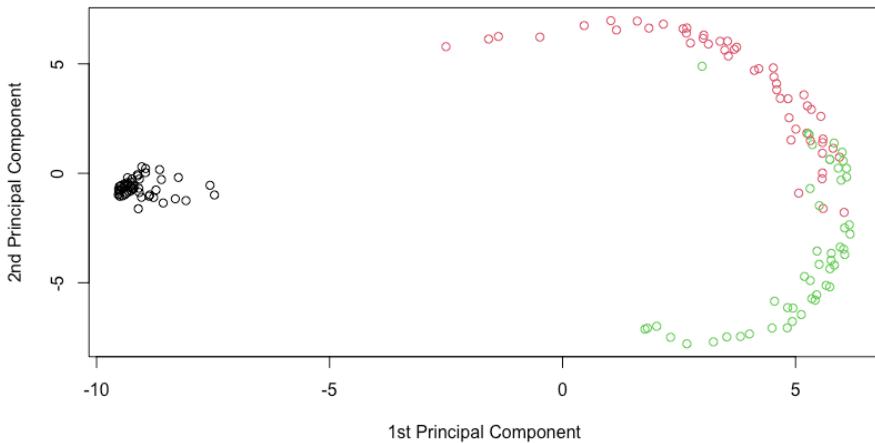


图 2.4 KPCA

(2) *t*-SNE 非线性降维算法：R 中可以通过 Rtsne 包可以实现 *t*-SNE 分析，所使用的函数为 Rtsne(X,...)，其中 X 为数据矩阵。以 R 中自带的范例数据集 iris 进行示范。

```
library(Rtsne)
data(iris)
iris.uni <- unique(iris)#去除重复数据
data <- iris.uni[,1:4]#获取性状信息
species <- iris.uni[,5]#获取品种信息
set.seed(321)
tsne_out <- Rtsne(data, dims = 2, pca = T, max_iter = 1000,
theta = 0.4, perplexity = 20, verbose = F)
library(ggplot2)
```

```
tsne_result <- as.data.frame(tsne_out$Y)
colnames(tsne_result) <- c("tSNE1", "tSNE2")
ggplot(tsne_result, aes(tSNE1, tSNE2, color=species)) + geom_point()
```

其可视化图为2.5。

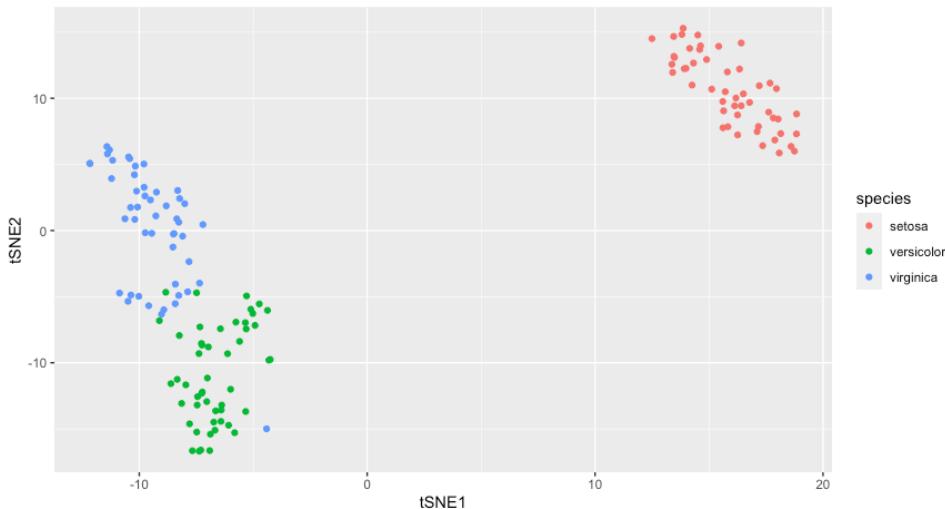


图 2.5 *t*-SNE

## 2.5.2 Python 语言实践

### 主成分分析

Python 语言中的 sklearn 库可实现主成分分析方法，本节将采用 PCA 对莺尾花数据集的特征进行分析，首先加载相应模块：

```
from sklearn.datasets import load_iris
from sklearn.decomposition import PCA
import numpy as np
```

再导入莺尾花数据集，利用 PCA 类进行分析

```
iris = load_iris()
X = iris.data
Pca_Model = PCA() # 导入PCA方法
Pca_Model = Pca_Model.fit(X) # 建立PCA模型
Pca_Result = Pca_Model.transform(X)
print(Pca_Result)
Ratio = Pca_Model.explained_variance_ratio_
plt.rcParams['font.sans-serif'] = ['SimHei']
plt.plot([1, 2, 3, 4], np.cumsum(Ratio))
plt.xticks([1, 2, 3, 4])
```

```
plt.title('累积贡献率')
plt.show()#基于各成分的贡献率，绘制累积贡献率折线图
```

累积贡献率折线图如图 2.6 所示，明显可以看出前两个成分的累积贡献率已经超过 97%，可选为主成分。

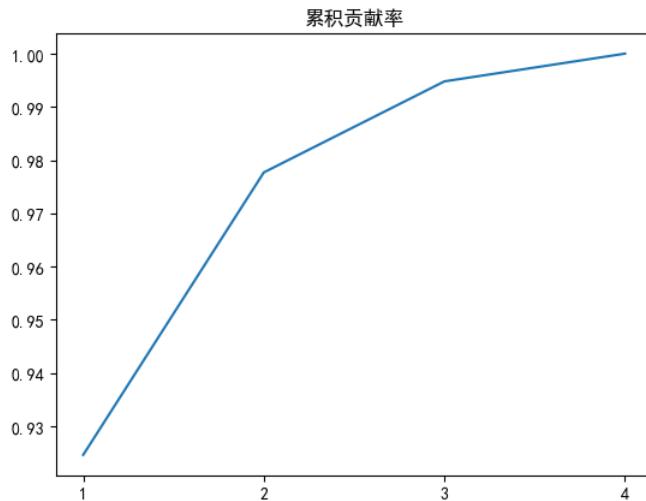


图 2.6 各成分累积贡献率

## 2.6 总结

主成分分析通过对原始高维特征进行映射变换得到少数几个主成分，得到的主要成分能够最大程度地表达原始数据，是一种处理高维数据的有效降维方法。本章主要介绍了总体主成分、样本主成分以及非线性主成分分析的算法理论，并通过实验案例阐述算法在实际问题中的应用。

除本章介绍的几种主成分分析方法之外，更多方法可参考：稀疏主成分分析 (SPCA) [15]，该方法不仅可以达到降维的效果，还能够增强算法可解释性；张量主成分分析 (TPCA) [16] [17]，其将张量与数据处理相结合，在高维信息压缩等领域得到广泛应用；鲁棒/稳健主成分分析 (RPCA) [18] 可提高算法对异常值、噪声的鲁棒性。

## 2.7 习题

1、设总体  $\mathbf{X} = (X_1, X_2)^T$  的协方差矩阵为

$$\boldsymbol{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

求  $\mathbf{X}$  的主成分  $Y_1, Y_2$ , 并计算第一主成分  $Y_1$  的贡献率。

2、总体  $\mathbf{X} = (X_1, \dots, X_p)^T$  的协方差矩阵为

$$\boldsymbol{\Sigma} = \text{diag}(\sigma_{11}, \sigma_{22}, \dots, \sigma_{pp})$$

其中  $\sigma_{11} \geq \sigma_{22} \geq \dots \geq \sigma_{pp}$ , 试求  $\mathbf{X}$  的主成分及其具有特征值。

3、试证明多元正态变量的主成分仍为正态变量且相互独立。

4、设总体  $\mathbf{X} = (X_1, X_2, X_3)^T$  的相关系数矩阵为

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}, (-1 < \rho < 1)$$

(1) 求  $\mathbf{X}$  的标准化变量的主成分及其主成分的贡献率和累积贡献率

(2) 将上述结果推广到  $p$  维情形。

5. 运用主成分分析对我国 2019 年分地区居民人均消费支出倾向进行分析。(数据来源: <http://www.stats.gov.cn/tjsj/ndsj/2020/indexch.htm>。数据表名称: 6-20 分地区居民人均消费支出 (2019 年))。

## 第三章 回归分析

### 3.1 简介

回归 (Regression) 这一概念最早由英国生物统计学家高尔顿 (F.Galton) 和皮尔逊 (K.Pearson) 在研究父母和子女的身高遗传特性时提出。回归分析 (Regression Analysis) 是处理多个变量间相关关系的一种数学方法，是机器学习中的监督学习方法。

回归模型是一种统计预测模型，需要预测的变量叫做因变量（或响应变量 (response variable)），用来解释因变量变化的变量叫做自变量（或协变量 (covariate)、特征 (feature)）。回归模型通过若干个自变量来预测响应变量。

从响应变量个数来看，回归模型中可以有一个响应变量，也可以有多个响应变量。从响应变量取值来看，回归模型中的响应变量可以是连续变量，也可以是离散变量。

一般的回归模型适用于单个响应变量、取值为连续型且一般服从正态分布的情形。广义线性模型 (generalized linear models, GLM) 适用于单个响应变量、取值是离散型且分布服从指数分布族的情形。多元响应变量协方差广义线性模型 (multivariate covariance generalized linear model, McGLM) 是广义线性模型的扩展，适用于服从非正态分布且不独立的多个响应变量的情形。

本章主要介绍单响应变量的线性回归模型、单响应变量的广义线性模型 (GLM) 和多元响应变量协方差广义线性模型 (McGLM)。

### 3.2 单响应变量的线性回归模型

本节主要介绍单响应变量的线性回归模型的原理、自变量的选择、多重共线性和岭回归。

### 3.2.1 线性回归模型的原理

下面介绍线性回归模型的一般形式、基本假设、未知参数的估计、显著性检验、预测等。

#### 一、线性回归模型的一般形式

**定义 3.1**  $p$  个自变量  $X_1, X_2, \dots, X_p$  对单个响应变量  $Y$  的线性回归模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon \quad (3.2.1)$$

式中,  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$  是  $p+1$  个未知参数。 $\beta_0$  称为**回归常数**,  $\beta_1, \beta_2, \dots, \beta_p$  为**回归系数**。当  $p=1$  时, 式 (3.2.1) 称为**一元线性回归模型**; 当  $p \geq 2$  时, 式 (3.2.1) 称为  $p$  元**线性回归模型**。式中  $\varepsilon$  是**随机误差**, 常假定  $E(\varepsilon) = 0, \text{var}(\varepsilon) = \sigma^2$ 。

若对单个响应变量的线性回归模型来说, 观测到样本数据为  $\{(X_{i1}, X_{i2}, \dots, X_{ip}; Y_i), i = 1, \dots, n\}$ , 则基于  $n$  个样本的线性回归模型可表示为

$$\left\{ \begin{array}{l} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \dots + \beta_p X_{1p} + \varepsilon_1 \\ Y_2 = \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \dots + \beta_p X_{2p} + \varepsilon_2 \\ \vdots \\ Y_n = \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \dots + \beta_p X_{np} + \varepsilon_n \end{array} \right.$$

写成矩阵形式为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3.2.2)$$

式 (3.2.2) 中  $\mathbf{X}$  为**设计矩阵**,

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

#### 二、线性回归模型的基本假设

为了方便地进行模型的参数估计及检验, 对线性回归模型进行如下基本假设:

1. 解释变量  $X_1, X_2, \dots, X_p$  是确定性变量, 而不是随机变量。且  $\text{rank}(\mathbf{X}) = p+1 < n$ , 即设计矩阵  $\mathbf{X}$  是列满秩矩阵, 其列之间不线性相关。
2. 随机误差项满足**高斯-马尔柯夫条件**, 即在给定自变量取值的情况下, 随机误差项满足零均值、同方差及序列不相关。

$$\begin{aligned} E(\varepsilon_i) &= 0, \quad i = 1, 2, \dots, n \\ \text{Cov}(\varepsilon_i, \varepsilon_j) &= \begin{cases} \sigma^2, & i = j \\ 0, & i \neq j \end{cases} \quad i, j = 1, 2, \dots, n \end{aligned} \quad (3.2.3)$$

3. 正态分布假定  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  相互独立, 且  $\varepsilon_i \sim N(0, \sigma^2)$ ,  $i = 1, 2, \dots, n$ 。  
矩阵形式为

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n) \quad (3.2.4)$$

其中  $\mathbf{I}_n$  是  $n \times n$  单位阵。在上述假定下, 因变量  $\mathbf{Y}$  服从  $n$  维正态分布

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n)$$

### 三、参数的估计

下面利用最小二乘法求线性回归模型  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$  中回归参数  $\boldsymbol{\beta}$  的最小二乘估计量。

**最小二乘法**(Ordinary Least Square Method, OLS)就是寻找未知参数  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$  的估计值  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ , 使得响应变量的离差平方和

$$Q(\beta_0, \beta_1, \beta_2, \dots, \beta_p) = \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})]^2$$

达到最小, 即满足

$$\begin{aligned} Q(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p) &= \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_p X_{ip})]^2 \\ &= \min_{\beta_0, \beta_1, \beta_2, \dots, \beta_p} \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})]^2 \end{aligned}$$

根据微积分求极值的原理, 只需求  $Q(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  的最小值点。将  $Q$  关于  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$  分别求偏导数, 并令其为零, 得到正规方程组

$$\left\{ \begin{array}{l} \frac{\partial Q}{\partial \beta_0} \Big|_{\beta_0=\hat{\beta}_0} = -2 \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_p X_{ip})] = 0 \\ \frac{\partial Q}{\partial \beta_1} \Big|_{\beta_1=\hat{\beta}_1} = -2 \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_p X_{ip})] X_{i1} = 0 \\ \dots \\ \frac{\partial Q}{\partial \beta_p} \Big|_{\beta_p=\hat{\beta}_p} = -2 \sum_{i=1}^n [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \dots + \hat{\beta}_p X_{ip})] X_{ip} = 0 \end{array} \right.$$

整理得矩阵形式的**正规方程组**

$$(\mathbf{X}^T \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{Y}$$

由  $\mathbf{X}$  的列满秩性可知,  $\mathbf{X}^T \mathbf{X}$  为满秩对称阵, 从而回归参数  $\boldsymbol{\beta}$  的**最小二乘估计**为

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (3.2.5)$$

可以证明，在高斯-马尔柯夫假定 (3.2.3) 下，最小二乘估计  $\hat{\beta}$  是  $\beta$  的最小方差线性无偏估计，且

$$\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1}) \quad (3.2.6)$$

证明可详见唐年胜，李会琼 [31]。

**定义 3.2** 对于线性回归模型  $Y = \mathbf{X}\beta + \varepsilon$ ，称

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p = \mathbf{X}^T \hat{\beta}$$

为**样本回归方程**，称  $\hat{Y}$  为响应变量的**拟合值**，称  $e = Y - \hat{Y}$  为**回归残差**。

**定义 3.3** 最小二乘法没有给出未知参数  $\sigma^2$  的估计。利用最大似然比原理可得  $\beta$  极大似然估计仍为  $\hat{\beta}$ ，同时给出  $\sigma^2$  的极大似然估计为

$$\frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta})}{n} = \frac{1}{n} \sum_{i=1}^n e_i^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (3.2.7)$$

因为上述估计有偏，通常取

$$\hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})^T (\mathbf{Y} - \mathbf{X}\hat{\beta})}{n-p-1} = \frac{1}{n-p-1} \sum_{i=1}^n e_i^2 = \frac{1}{n-p-1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (3.2.8)$$

作为  $\sigma^2$  的无偏估计，称  $\hat{\sigma}$  为**回归标准差**。

#### 四、显著性检验

下面介绍回归方程显著性的  $F$  检验、回归参数显著性的  $t$  检验以及衡量回归方程拟合程度的拟合优度检验。

##### 1、回归方程的显著性检验： $F$ 检验

###### (1) 平方和分解式

下面基于方差分析的思想，从数据出发来研究因变量数据变异的原因。

数据  $Y_1, Y_2, \dots, Y_n$  的波动大小可用**总离差平方和**  $SST = \sum_{i=1}^n (Y_i - \bar{Y})^2$  来度量，

其中  $\bar{Y} = \sum_{i=1}^n Y_i / n$ ，它反映响应变量  $Y$  的波动程度或不确定性。

称  $SSR = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$  为**回归平方和**，它是由回归方程确定的，是由自变量  $\mathbf{Y}$  的波动引起的，反映了自变量解释因变量波动的贡献。

称  $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$  为**残差平方和**，它是不能由自变量解释的波动，是由自变量之外的因素引起的。

易证  $SST = SSR + SSE$ ，称此式为**平方和分解式**。

平方和分解式表明，总离差平方和 SST 中能够由自变量解释的部分为 SSR，不能由自变量解释的部分为 SSE。因此，回归平方和越大，回归的效果就越好。

### (2) $F$ 检验统计量

在正态分布假设 (3.2.4) 下，需要检验自变量  $X_1, X_2, \dots, X_p$  从整体上对因变量  $Y$  是否有显著的影响，为此提出：

原假设

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

备择假设

$$H_1 : \beta_1, \beta_2, \dots, \beta_p \text{ 至少有一个不为零}$$

在原假设  $H_0$  下，构造检验统计量

$$F = \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)} \sim F(p, n-p-1) \quad (3.2.9)$$

给定显著性水平  $\alpha$  ( $0 < \alpha < 1$ )，右侧  $F$  检验的临界值为  $F_\alpha(p, n-p-1)$ 。当检验统计量  $F$  的观测值落在拒绝域，即  $F > F_\alpha(p, n-p-1)$  时，拒绝原假设，说明回归方程显著， $\mathbf{X}$  与  $Y$  有显著的线性关系；否则，只能接受原假设，认为回归方程不显著。

也可以利用  $P$  值法进行判断，其中  $P = \Pr(Z > F)$ ， $Z \sim F(p, n-p-1)$ ，当  $P < \alpha$  时，拒绝原假设；否则，只能接受原假设。

通常利用方差分析表 3.1 进行  $F$  检验。

### 2、回归系数的显著性检验： $t$ 检验

在多元线性回归分析中，回归方程显著并不意味着每个自变量对  $Y$  的影响都显著，所以需要对每个自变量进行显著性检验。为此，提出

原假设

$$H_{0j} : \beta_j = 0, \quad j = 1, \dots, p$$

备择假设

$$H_{1j} : \beta_j \neq 0, \quad j = 1, \dots, p$$

因  $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)^T \sim N_{p+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T \mathbf{X})^{-1})$ ，若记

$$(\mathbf{X}^T \mathbf{X})^{-1} = (c_{ij}) \stackrel{\Delta}{=} C, \quad i, j = 0, 1, 2, \dots, p \quad (3.2.10)$$

表 3.1 方差分析表

方差来源	自由度	平方和	均方	$F$ 值	$\Pr(Z > F)$
回归	$p$	SSR	$\text{SSR}/p$	$\frac{\text{SSR}/p}{\text{SSE}/(n-p-1)}$	
残差	$n-p-1$	SSE	$\text{SSE}/(n-p-1)$		
总和	$n-1$	SST			

则

$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2), \quad j = 1, 2, \dots, p$$

构造  $t$  统计量

$$t_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}\hat{\sigma}^2}} \sim t(n-p-1) \quad (3.2.11)$$

$$\text{上式中, } \hat{\sigma}^2 = \frac{(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{n-p-1}.$$

给定显著性水平  $\alpha$  ( $0 < \alpha < 1$ ), 双侧  $t$  检验的临界值为  $t_{\alpha/2}(n-p-1)$ 。当检验统计量的观测值落在拒绝域, 即  $|t_j| > t_{\alpha/2}(n-p-1)$  时, 拒绝原假设, 认为  $\beta_j$  显著不为 0,  $X_j$  对  $Y$  影响显著; 否则, 只能接受原假设。

还可以利用  $P$  值法进行判断, 其中  $P = \Pr(|Z| > |t_j|)$ ,  $Z \sim t(n-p-1)$ 。当  $P < \alpha$  时, 拒绝原假设; 否则, 只能接受原假设。

### 3、拟合优度检验

拟合优度检验用于检验回归方程对样本观测值的拟合程度。

#### 定义 3.4 样本决定系数为

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

样本决定系数  $R^2$  是一个回归直线与样本观测值拟合优度的相对指标, 反映了因变量的波动能被自变量解释的比例。其取值在 0 与 1 之间。 $R^2$  越接近 1, 拟合优度越好。

在应用过程中, 残差平方和往往随着解释变量个数的增加而减少。如果在模型中增加一个解释变量,  $R^2$  往往增大。而由增加解释变量个数引起的  $R^2$  的增大, 往往与拟合好坏无关, 因此样本决定系数需要调整。在样本容量一定的情况下, 增加解释变量必定使得自由度减少。故可将残差平方和与总离差平方和分别除以各自的自由度, 以剔除变量个数对拟合优度的影响。

#### 定义 3.5 记 $\bar{R}^2$ 为修正的决定系数, 则

$$\bar{R}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} \quad (3.2.12)$$

$\bar{R}^2$  越大, 对应的回归方程拟合程度越好。

## 五、预测

若基于观测数据  $\{(X_{i1}, X_{i2}, \dots, X_{ip}; Y_i), i = 1, \dots, n\}$  建立的回归方程

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p = \mathbf{X}^T \hat{\boldsymbol{\beta}}$$

通过了回归方程的显著性检验和回归系数的显著性检验，则对给定的新样本观测点  $\mathbf{X}_0 = (1, X_{01}, X_{02}, \dots, X_{0p})^T$ ，可以用  $\hat{Y}_0 = \mathbf{X}_0^T \hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 X_{01} + \dots + \hat{\beta}_p X_{0p}$  作为  $Y_0 = \beta_0 + \beta_1 X_{01} + \dots + \beta_p X_{0p} + \varepsilon_0$  的点预测。

因  $\hat{\boldsymbol{\beta}}$  与  $\varepsilon_0$  独立， $\hat{\boldsymbol{\beta}}$  与  $Y_0$  独立，从而  $\hat{Y}_0$  与  $Y_0$  独立。下面证明可详见唐年胜，李会琼 [31]，

$$\hat{Y}_0 - Y_0 \sim N(0, \sigma^2(1 + \mathbf{X}_0^T \mathbf{X})^{-1} \mathbf{X}_0)$$

从而

$$\frac{\hat{Y}_0 - Y_0}{\hat{\sigma} \sqrt{1 + \mathbf{X}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_0}} \sim t(n - p - 1)$$

由此得， $y_0$  的置信度  $1 - \alpha$  的置信区间为

$$\left( \hat{Y}_0 - t_{\alpha/2}(n - p - 1) \hat{\sigma} \sqrt{1 + \mathbf{X}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_0}, \hat{Y}_0 + t_{\alpha/2}(n - p - 1) \hat{\sigma} \sqrt{1 + \mathbf{X}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}_0} \right)$$

### 3.2.2 多重共线性

一般情况下，当回归方程的自变量之间存在很强的线性关系，回归方程的检验高度显著时，有些回归系数却不能通过显著性检验，甚至出现有的回归系数所带符号与实际经济意义不符，此时可认为变量间存在多重共线性。

我们已经知道多元线性回归模型 (3.2.2) 的未知参数  $\boldsymbol{\beta}$  的最小二乘估计为

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

这就要求  $\mathbf{X}$  列满秩，即  $\text{rank}(\mathbf{X}) = p + 1$ 。也就是要求  $\mathbf{X}$  的列向量之间线性无关。然而，往往  $\text{rank}(\mathbf{X}) < p + 1$ ，此时  $|\mathbf{X}^T \mathbf{X}| = 0$ ，从而  $(\mathbf{X}^T \mathbf{X})^{-1}$  不存在，也就无法得到参数的估计量。

在实际问题的研究中，经常见到近似多重共线性的情形。此时， $\text{rank}(\mathbf{X}) = p + 1$ ，但  $|\mathbf{X}^T \mathbf{X}| \approx 0$ ， $(\mathbf{X}^T \mathbf{X})^{-1}$  的对角线元素很大，从而  $\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$  的对角线元素很大，使得参数的估计精度很低。

多重共线性可以用方差膨胀因子进行判断。

**定义 3.6** 称式 (3.2.10) 中矩阵  $C = (c_{ij}) = (\mathbf{X}^T \mathbf{X})^{-1}$  中对角线元素  $c_{jj}$  为自变量  $X_j$  的方差膨胀因子 (variance inflation factor, VIF)，记作  $VIF_j$ 。

易证

$$c_{jj} = \frac{1}{1 - R_j^2}$$

其中， $R_j^2$  是以  $X_j$  为因变量对其余  $p - 1$  个自变量进行线性回归得到的样本决定系数，它度量了自变量  $X_j$  与其余  $p - 1$  个自变量的线性相关程度。这种相关程度越强，

说明自变量之间的多重共线性越严重。此时,  $R_j^2$  越接近 1,  $VIF_j$  越大。 $VIF_j$  的大小反映了自变量之间存在多重共线性的强弱。当  $VIF_j \geq 10$  时, 说明自变量  $X_j$  与其余自变量之间存在严重的多重共线性。

### 3.2.3 岭回归

针对自变量之间出现多重共线性时, 普通最小二乘法效果明显变差的问题, 霍尔 (Hoerl) [33] 于 1962 年提出一种改进最小二乘估计的方法。

当变量  $X_j$  之间相互独立时, 使用最小二乘法估计的参数  $\hat{\beta}$  是真实值的无偏估计, 即满足  $E(\hat{\beta}) = \beta$ , 并且在所有的无偏估计中具有最小的方差。从求解公式上看, 若要保证该回归系数有解, 必须确保  $\mathbf{X}^T \mathbf{X}$  矩阵是满秩的, 即  $\mathbf{X}^T \mathbf{X}$  可逆, 但在实际中, 若是  $X_j$  间存在高度多重共线性, 或者自变量个数大于等于观测个数, 不论哪种情况, 最终算出来的行列式都等于 0 或者是近似为 0, 此时  $\hat{\beta}$  的值表现不稳定, 即此时  $\hat{\beta}$  的方差很大 (虽然其在所有无偏估计类中最小, 但其本身仍很大), 导致其均方误差  $MSE(\hat{\beta}) = \text{var}(\hat{\beta}) + [\text{bias}(\hat{\beta})]^2$  很大。

当自变量之间存在多重共线性, 即  $|\mathbf{X}^T \mathbf{X}| \approx 0$  时, 如果给  $\mathbf{X}^T \mathbf{X}$  加上一个正的常数矩阵  $\lambda \mathbf{I}$  ( $\lambda > 0$ ), 那么  $\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}$  接近奇异的程度会比  $\mathbf{X}^T \mathbf{X}$  接近奇异的程度小得多, 从而使得行列式不再为零, 可以求逆。即得到岭估计:

$$\hat{\beta}(\lambda) \equiv (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{Y}$$

容易看出, 岭估计使用了单位矩阵乘以常数  $\lambda$ ,  $\lambda \mathbf{I}$  矩阵在 0 构成的平面上有一条 1 组成的“岭”, 这便是岭回归名称的由来。

**定义 3.7** 称  $\hat{\beta}(\lambda)$  为参数  $\beta$  的**岭估计**。其中,  $\lambda$  称为**岭参数**。由参数  $\beta$  的岭估计所建立的回归方程称为**岭回归方程**。

易证,  $\hat{\beta}(\lambda)$  是回归参数  $\beta$  的有偏估计。且  $\|\hat{\beta}(\lambda)\| < \|\hat{\beta}\|$ , 这表明  $\hat{\beta}(\lambda)$  可看成由  $\hat{\beta}$  进行某种向原点的压缩。在均方误差意义下,  $\hat{\beta}(\lambda)$  优于最小二乘估计  $\hat{\beta}$ 。当  $\lambda = 0$  时, 岭估计  $\hat{\beta}(0)$  即为普通最小二乘估计  $\hat{\beta}$ 。证明可详见唐年胜, 李会琼 [31]。

因为岭参数  $\lambda$  不是唯一确定的, 故岭回归估计  $\hat{\beta}(\lambda)$  是回归参数  $\beta$  的一个估计族。当岭参数  $\lambda$  在  $(0, \infty)$  内变化时,  $\hat{\beta}_j(\lambda)$  是  $\lambda$  的函数。在平面坐标系中, 把函数  $\hat{\beta}_j(\lambda)$  随  $\lambda > 0$  变化所描绘出来的曲线, 称为**岭迹**。

选择岭参数  $\lambda$  的值一般满足以下原则:

- 1、各回归系数的岭估计基本稳定;
- 2、用最小二乘估计所得符号不合理的回归系数, 其岭估计的符号变得合理;
- 3、回归系数没有不合乎经济意义的绝对值;
- 4、残差平方和增加不太多。

在实际应用中，可以利用  $\hat{\beta}(\lambda)$  的每一分量  $\hat{\beta}_j(\lambda)$  在同一图形上的岭迹的变化形状来确定适当的  $\lambda$ 。但是，岭迹法存在一定的主观性。此外，还可以用广义交叉验证 (generalized cross validation, GCV) 来确定岭参数。通常选取使得 GCV 最小的  $\lambda$  值。具体可详见 [34]。

### 3.3 广义线性模型

在本章的前面章节介绍了多元线性模型以及其假设，一般的线性模型主要适用于响应变量是连续型随机变量，并且其一般服从正态分布。但是我们会发现在实际应用中有些响应变量是离散的，而且并不服从正态分布。针对这种情况，Nelder 和 Wedderburn 将传统线性模型进行推广到广义线性模型 (generalized linear models)。广义线性模型将因变量的分布推广到指数分布族，从连续型变量拓展到离散型变量，通过连接函数将服从指数族分布的响应变量的期望与自变量的线性函数连接起来，下面开始介绍广义线性模型。在给出广义线性模型的定义前先介绍指数族分布与连接函数。

#### 3.3.1 指数族分布

指数族分布是以指数分布表示的一族分布，很多常见的分布如正态分布、Poisson 分布、二项分布等属于指数族分布。下面给出指数族分布的定义。

**定义 3.8** 若随机变量  $Y$  的概率质量 (离散型) 或概率密度 (连续型) 具有如下形式：

$$f(y; \theta, \phi) = \exp \left( \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right) \quad (3.3.1)$$

则称  $y$  服从指数族分布。其中  $a(\cdot)$ ,  $b(\cdot)$ ,  $c(\cdot, \cdot)$  为已知连续函数， $\theta$  和  $\phi$  为未知参数。通常， $\theta$  与  $E(Y)$  有关，是我们所感兴趣的参数，而  $\phi$  与  $\text{var}(Y)$  有关，称为多余参数 (nuisance parameter)。

下面以正态分布、泊松分布和二项分布为例说明这两种分布的概率密度或概率质量具有 (3.3.1) 式的形式，并求出对应的  $a(\cdot), b(\cdot), c(\cdot, \cdot)$  以及未知参数。

**例 3.1** 设  $Y \sim N(\mu, \sigma^2)$ ，则  $Y$  的概率密度可表为；

$$\begin{aligned} f(y; \theta, \phi) &= \frac{1}{\sqrt{2\pi}\sigma} \exp \left( -\frac{(y-\mu)^2}{2\sigma^2} \right) \\ &= \exp \left( \frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{1}{2} \left( \frac{y^2}{\sigma^2} + \ln(2\pi\sigma^2) \right) \right) \end{aligned}$$

对照 (3.3.1) 式有

$$\begin{aligned}\theta &= \mu, \phi = \sigma^2, a(\phi) = \phi \\ b(\theta) &= \frac{1}{2}\mu^2 = \frac{1}{2}\theta^2, c(y, \phi) = -\frac{1}{2} \left( \frac{y^2}{\phi} + \ln(2\pi\phi) \right)\end{aligned}$$

**例 3.2** 设  $Y \sim P(\mu)$  (参数为  $\mu$  的 Poisson 分布), 则  $Y$  的概率分布可表示为

$$P(Y = y) = \frac{\mu^y}{y!} \exp(-\mu) = \exp(y \ln \mu - \mu - \ln(y!)) \quad y = 0, 1, 2, \dots$$

对照 (3.3.1) 式有

$$\theta = \ln \mu, \phi = 1, a(\phi) = \phi, b(\theta) = \mu = \exp(\theta), c(y, \phi) = -\ln(y!)$$

**例 3.3** 设  $X \sim B(m, \mu)$  (二项分布), 其中  $0 < \mu < 1, m$  为正整数, 且二者均为未知参数。令  $Y = X/m$ , 则  $Y$  服从指数族分布。事实上,  $Y$  的所有可能的取值为  $y = 0, 1/m, 2/m, \dots, 1$ , 且

$$\begin{aligned}P(Y = y) &= P(X = my) \\ &= C_m^{my} \mu^{my} (1 - \mu)^{m-my} \\ &= \exp \left( \frac{y \ln \left( \frac{\mu}{1-\mu} \right) + \ln(1 - \mu)}{\frac{1}{m}} + \ln(C_m^{my}) \right)\end{aligned}$$

对照 (3.3.1) 式有

$$\begin{aligned}\theta &= \ln \left( \frac{\mu}{1 - \mu} \right), \phi = \frac{1}{m}, a(\phi) = \phi \\ b(\theta) &= -\ln(1 - \mu) = \ln(1 + \exp(\theta)), c(y, \phi) = \ln(C_m^{my})\end{aligned}$$

### 3.3.2 连接函数

通过前面的学习, 我们知“回归”一般是用于预测样本的值, 这个值通常是连续的。在分类问题的应用上效果往往不理想。为了保留线性回归“简单、效果佳”的特点, 又想让它能够进行分类, 需要对预测值再做一次处理。这个多出来的处理过程, 就是 GLM 所做的最主要的事。而处理这个过程的函数, 我们把它叫做连接函数。

**连接函数**  $g(\cdot)$  是将自变量第  $i$  组观测的线性组合与第  $i$  个观测值  $Y_i$  的期望连接起来的函数, 即

$$g(\mathbb{E}(Y_i)) = g(\mu_i) = \beta_0 + \sum_{j=1}^p X_{ij}\beta_j \quad i = 1, \dots, n \quad (3.3.2)$$

在线性回归模型中, 由于假定  $Y_i \sim N(\mu_i, \sigma^2)$ , 而  $E(Y_i) \in \mathcal{R}$ ,  $\beta_0 + \sum_{j=1}^p X_{ij}\beta_j$  一般也在  $\mathcal{R}$  中取值, 因此通常取  $g(E(Y_i)) = E(Y_i) = \beta_0 + \sum_{j=1}^p X_{ij}\beta_j$  ( $i = 1, \dots, n$ )。但是取  $g(E(Y_i)) = E(Y_i)$  并不总是合适的, 例如: 当  $Y_i$  服从泊松分布即  $Y_i \sim P(\mu_i)$  时, 则  $E(Y_i) = \mu_i > 0$  而  $\beta_0 + \sum_{j=1}^p X_{ij}\beta_j$  有可能取负值。因此直接建立模型

$$g(E(Y_i)) = E(Y_i) = \mu_i = \beta_0 + \sum_{j=1}^p X_{ij}\beta_j$$

是不合理的。

表3.2中介绍几种常见的连接函数。

表 3.2 常用的连接函数

连接函数名称	link function	连接函数公式
恒等函数	identity	$g(\mu) = \mu$
Logit 函数	logit	$g(\mu) = \ln(\mu/(1-\mu))$
Probit 函数	probit	$g(\mu) = \pi^{-1}(\mu)$
对数函数	log	$g(\mu) = \ln(\mu)$
逆函数	inverse	$g(\mu) = 1/\mu$
平方根函数	sqrt	$g(\mu) = \sqrt{\mu}$
逆高斯分布	$1/\mu^2$	$g(\mu) = 1/\mu^2$
重对数函数	loglog	$g(\mu) = \ln(-\ln(\mu))$
互补重对数函数	cloglog	$g(\mu) = \ln(-\ln(1-\mu))$
柯西函数	cauchit	$g(\mu) = \tan(\pi(\mu - 0.5))$

连接函数有很多种, 一般在广义线性模型的建模中我们会使用正则连接函数, 即

$$g(E(Y)) = g(\mu) = \theta$$

### 3.3.3 广义线性模型

**定义 3.9** 设  $(Y_i; X_{i1}, X_{i2}, \dots, X_{ip})$  ( $i = 1, 2, \dots, n$ ) 为因变量  $Y$  和自变量  $X_1, X_2, \dots, X_p$  的观测值, 若

(1)  $Y_1, Y_2, \dots, Y_n$  相互独立, 且对每个  $i$ ,  $Y_i$  服从指数族分布, 即

$$Y_i \sim f(y_i; \theta_i, \phi_i) = \exp \left( \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right)$$

$$(2) \quad g(\mu_i) = \beta_0 + \sum_{j=1}^p X_{ij}\beta_j, \text{ 其中 } \mu_i = E(Y_i) \quad (i = 1, 2, \dots, n)$$

则称  $Y$  与  $X_1, X_2, \dots, X_p$  服从广义线性模型。

表 3.3 常用的广义线性模型

分布	函数	模型
正态 (Gaussian)	$E(Y) = \mathbf{X}^T \boldsymbol{\beta}$	普通线性模型
二项 (Binomial)	$E(Y) = \frac{\exp(\mathbf{X}^T \boldsymbol{\beta})}{1 + \exp(\mathbf{X}^T \boldsymbol{\beta})}$	Logistic 模型
泊松 (Poisson)	$E(Y) = \exp(\mathbf{X}^T \boldsymbol{\beta})$	对数线性模型

表3.3是三种常见的广义线性模型。由于每一个模型都有相应的指数分布，因此我们可以采用极大似然方法进行参数估计。

## 3.4 多元响应变量协方差广义线性模型 McGLM

传统的线性模型 LM (Linear Model) 适用于处理单个响应变量的情形。广义线性模型 GLM (Generalized Linear Model, GLM) 扩展了线性模型，可用于处理独立非正态数据和单个响应变量情形，并且假设方差函数是已知的。Anderson [38] 和 Pourahmadi [39] 将 GLM 方法扩展到处理非独立非正态数据和单变量响应变量的情形，提出协方差广义线性模型 (Covariance Generalized Linear Model, cGLM)。Wagner Hugo Bonat 和 Bent Jørgensen [40] 将 cGLM 扩展到处理非独立非正态数据和多个响应变量的情形，提出多元响应变量协方差广义线性模型 (Multivariate Covariance Generalized Linear Model, McGLM)，它是广义线性模型及协方差广义线性模型的扩展。

本节介绍多元响应变量协方差广义线性模型的原理、参数估计和模型选择等。

### 3.4.1 McGLM 模型的原理

McGLM 旨在处理具有时间和空间相关结构的多元响应变量，通过采用传统广义线性模型的方差函数方法来考虑非正态性，采用连接函数和线性预测器建模均值结构，采用协方差连接函数和矩阵线性预测器建模协方差结构，为正态和非正态多元数据分析提供通用的统计建模框架。

McGLM 的每一个响应变量均可由连接函数、方差函数、协方差函数三种函数及线性预测器和矩阵线性预测器来构成。

### 定义 3.10 McGLM

设样本容量  $n$ , 响应变量个数  $R$ , 响应变量矩阵  $\mathbf{Y}_{n \times R} = (\mathbf{Y}_1, \dots, \mathbf{Y}_R)$ , 响应变量  $\mathbf{Y}$  的期望值矩阵为  $\mathbf{M}_{n \times R} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_R)$ 。 $n \times n$  阶矩阵  $\boldsymbol{\Sigma}_r$  表示每个响应变量  $\mathbf{Y}_r (r = 1, \dots, R)$  的协方差矩阵, 该矩阵描述了每个响应变量内的协方差结构。 $R \times R$  阶矩阵  $\boldsymbol{\Sigma}_b$  表示响应变量  $\mathbf{Y}_1, \dots, \mathbf{Y}_R$  间的相关系数矩阵。 $\mathbf{X}_r$  为设计矩阵的  $n \times k_r$  阶子阵。回归参数向量  $\boldsymbol{\beta}_r$  为  $k_r \times 1$  阶的回归参数子向量。[多元响应变量协方差广义线性模型](#) (McGLM) 满足:

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{M} = \{g_1^{-1}(\mathbf{X}_1 \boldsymbol{\beta}_1), \dots, g_R^{-1}(\mathbf{X}_R \boldsymbol{\beta}_R)\} \\ \text{var}(\mathbf{Y}) &= \mathbf{C} = \boldsymbol{\Sigma}_R \stackrel{G}{\otimes} \boldsymbol{\Sigma}_b \end{aligned} \quad (3.4.1)$$

其中  $\boldsymbol{\Sigma}_R \stackrel{G}{\otimes} \boldsymbol{\Sigma}_b = \text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \dots, \tilde{\boldsymbol{\Sigma}}_R)(\boldsymbol{\Sigma}_b \otimes \mathbf{I}_n)\text{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^T, \dots, \tilde{\boldsymbol{\Sigma}}_R^T)$  是  $\boldsymbol{\Sigma}_R$  与  $\boldsymbol{\Sigma}_b$  的[广义克罗内克积](#) [41]。它是  $nR \times nR$  阶矩阵, 表示所有响应变量的联合协方差矩阵。 $\tilde{\boldsymbol{\Sigma}}_r (r = 1, \dots, R)$  表示协方差矩阵  $\boldsymbol{\Sigma}_r$  的 Cholesky 分解的下三角矩阵, 是  $n \times n$  阶矩阵。运算符  $\text{Bdiag}(\cdot)$  表示分块对角矩阵。 $\mathbf{I}_n$  表示  $n \times n$  阶单位阵。 $\otimes$  表示克罗内克积。函数  $g_r(\cdot)$  为均值结构中的[连接函数](#) (link function)。

### 一、响应变量的协方差矩阵 $\boldsymbol{\Sigma}_r$

不同取值类型的响应变量  $\mathbf{Y}_r$  可取不同的协方差矩阵:

1、若  $\mathbf{Y}_r$  取值为连续型、二值型、二项式、比例或指数数据, 其协方差矩阵  $\boldsymbol{\Sigma}_r$  可定义为

$$\boldsymbol{\Sigma}_r = V(\boldsymbol{\mu}_r; \mathbf{p}_r)^{\frac{1}{2}} (\boldsymbol{\Omega}(\boldsymbol{\tau}_r)) V(\boldsymbol{\mu}_r; \mathbf{p}_r)^{\frac{1}{2}} \quad (3.4.2)$$

2、若  $\mathbf{Y}_r$  取值为计数数据, 其协方差矩阵  $\boldsymbol{\Sigma}_r$  可采用如下形式

$$\boldsymbol{\Sigma}_r = \text{diag}(\boldsymbol{\mu}_r) + V(\boldsymbol{\mu}_r; \mathbf{p}_r)^{\frac{1}{2}} (\boldsymbol{\Omega}(\boldsymbol{\tau}_r)) V(\boldsymbol{\mu}_r; \mathbf{p}_r)^{\frac{1}{2}} \quad (3.4.3)$$

式 (3.4.2) 与 (3.4.3) 中,  $V(\boldsymbol{\mu}_r; \mathbf{p}_r) = \text{diag}(\vartheta(\boldsymbol{\mu}_r; \mathbf{p}_r))$  是对角矩阵, 其主对角线元素由[方差函数](#) (variance function)  $\vartheta(\cdot; \mathbf{p}_r)$  按元素应用于向量  $\boldsymbol{\mu}_r$  而给出。方差函数  $\vartheta(\cdot; \mathbf{p}_r)$  中的参数  $\mathbf{p}_r$  为[幂参数](#) (power parameter)。

$\boldsymbol{\Omega}(\boldsymbol{\tau}_r)$  为[散度矩阵](#) (dispersion matrix), 描述了响应变量的协方差矩阵  $\boldsymbol{\Sigma}_r$  中不依赖于均值结构的协方差部分。 $\boldsymbol{\tau}_r$  为[散度参数](#) (dispersion parameter)。

### 二、方差函数 $\vartheta(\cdot; \mathbf{p}_r)$

方差函数  $\vartheta(\cdot; \mathbf{p}_r)$  在 McGLM 中起着重要作用。模型选择不同的方差函数, 则对应不同的响应变量边际分布。下面介绍 McGLM 常用的方差函数。

1、对于连续响应变量，方差函数采用幂函数  $\vartheta(\mu_r; p_r) = \mu_r^{p_r}$  来描述 Tweedie 分布族。Tweedie 分布族的特例有高斯分布 ( $p_r = 0$ )、伽马分布 ( $p_r = 2$ ) 和逆高斯分布 ( $p_r = 3$ )。

2、对于取值为二值型、二项式或比例数据的响应变量，方差函数常采用扩展二项式函数  $\vartheta(\mu_r; p_r) = \mu_r^{p_{r1}}(1 - \mu_r)^{p_{r2}}$ 。

3、对于计数数据的离散型响应变量，方差函数采用函数  $\vartheta(\mu_r; p_r) = \mu_r + \tau_r \mu_r^{p_r}$ ，(其中  $\tau_r$  是散度参数) 来描述 Poisson-Tweedie 分布族。由于散度参数仅出现在第二项中，因此协方差矩阵  $\Sigma_r$  采用式 (3.4.3) 的形式。Poisson-Tweedie 分布族的特例有 Hermite 分布 ( $p_r = 0$ )、Neyman 分布 ( $p_r = 1$ )、负二项分布 ( $p_r = 2$ ) 和泊松逆高斯 (Poisson-inverse Gaussian) 分布 ( $p_r = 3$ )。

### 三、幂参数 $p_r$

在 McGLM 二阶矩假设下，模型估计幂参数  $p_r$  的信息来自均值和方差的关系，因此估计幂参数  $p_r$  时需要均值向量有足够的变化，即要求线性预测器中存在显著的协变量。对于 McGLM 中所有的方差函数，幂参数  $p_r$  是区分重要分布的指标。模型中的算法允许估计幂参数  $p_r$ ，该参数可用于自动选择分布。

### 四、协方差连接函数 $h(\cdot)$ 与矩阵线性预测器

Anderson [38] 和 Pourahmadi [39]，Bonat 和 Jørgensen [40] 提出使用矩阵线性预测器结合协方差联接函数来对散度矩阵  $\Omega(\tau_r)$  进行建模，即根据已知矩阵的线性组合建立一个模型

$$h(\Omega(\tau_r)) = \tau_{r0}\mathbf{Z}_{r0} + \cdots + \tau_{rD}\mathbf{Z}_{rD} \quad (3.4.4)$$

这种结构与均值结构的线性预测器相似，称为矩阵线性预测器 (matrix linear predictor)。将矩阵线性预测器 (3.4.4) 代入式 (3.4.1) 中，即得到多元响应变量协方差广义线性模型。

式 (3.4.4) 中， $h(\cdot)$  为协方差连接函数 (covariance link function)； $\mathbf{Z}_{rd}$  ( $d = 0, \dots, D$ ) 是反映响应变量  $\mathbf{Y}_r$  内协方差结构的已知矩阵， $\tau_r = (\tau_{r0}, \dots, \tau_{rD})^T$  为  $D + 1$  维散度参数向量。

McGLM 中协方差连接函数常采用恒等函数 (identity)、逆函数 (inverse) 和指数矩阵函数 (exponential-matrix)。

实际上，很难定义散度参数向量  $\tau_r$  的参数空间。Bonat 和 Jørgensen 利用倒数似然算法，使用一个调整常数来控制算法的步长，并避免参数向量  $\tau_r$  的不合理值。

矩阵预测器中的协方差结构常采用协方差矩阵的线性模型，用于处理非高斯数据、纵向自相关数据 (如复合对称、移动平均和一阶自回归等)、空间数据及区域数据等。关于如何指定矩阵  $\mathbf{Z}_{rd}$ ，详见 [40]。

## 五、连接函数 $g(\cdot)$

McGLM 模型中，连接函数  $g(\cdot)$  将响应变量的期望与协变量联系起来，反映了模型的均值结构。常采用的连接函数也如表3.2。

### 3.4.2 参数估计

设参数向量  $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\lambda}^T)^T$ ，其中回归参数向量  $\boldsymbol{\beta}_{K \times 1} = (\boldsymbol{\beta}_1^T, \dots, \boldsymbol{\beta}_R^T)^T$ ，散度参数向量  $\boldsymbol{\lambda}_{Q \times 1} = (\rho_1, \dots, \rho_{R(R-1)/2}, p_1, \dots, p_R, \boldsymbol{\tau}_1^T, \dots, \boldsymbol{\tau}_R^T)^T$ 。

McGLM 模型采用基于二阶矩假设的估计函数法进行拟合。模型基于拟得分 (quasi-score) 函数和皮尔逊估计 (Pearson estimating) 函数的有效牛顿评分算法 (Newton scoring algorithm)，针对回归参数和散度参数，利用修正的追赶算法 (modified chaseralgorithm) 以及改进的倒数似然算法 (reciprocal likelihood algorithm)，通过调节常数  $\alpha$  来调节步长，从而进行参数估计及统计推断，进而拟合 McGLM 模型。

具体来讲，我们让  $\mathcal{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_R^T)^T$  和  $\mathcal{M} = (\boldsymbol{\mu}_1^T, \dots, \boldsymbol{\mu}_R^T)^T$  表示  $nR \times 1$  列向量，其中的元素分别由响应变量矩阵  $\mathbf{Y}_{n \times R}$  和期望值矩阵  $\mathbf{M}_{n \times R}$  按列排序得到。

为了进行参数估计，我们采取拟得分函数 [42]

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathbf{D}^T \mathbf{C}^{-1} (\mathcal{Y} - \mathcal{M}), \quad (3.4.5)$$

其中  $\mathbf{D} = \nabla_{\boldsymbol{\beta}} \mathcal{M}$  是一个  $nR \times K$  矩阵，并且  $\nabla_{\boldsymbol{\beta}}$  表示对相应参数求梯度。此外，可以得到  $\psi_{\boldsymbol{\beta}}$  的  $K \times K$  维敏感性矩阵和特异性矩阵

$$\mathbf{S}_{\boldsymbol{\beta}} = E(\nabla_{\boldsymbol{\beta}} \psi_{\boldsymbol{\beta}}) = -\mathbf{D}^T \mathbf{C}^{-1} \mathbf{D} \text{ 和 } V_{\boldsymbol{\beta}} = \text{var}(\psi_{\boldsymbol{\beta}}) = \mathbf{D}^T \mathbf{C}^{-1} \mathbf{D} \quad (3.4.6)$$

散度参数所采用的皮尔逊估计函数的各个部分定义如下

$$\psi_{\lambda_i}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \text{tr}(\mathbf{W}_{\lambda_i}(\mathbf{r}^T \mathbf{r} - \mathbf{C})), \quad i = 1, \dots, Q, \quad (3.4.7)$$

其中  $\mathbf{W}_{\lambda_i} = -\partial \mathbf{C}^{-1} / \partial \lambda_i$  并且  $\mathbf{r} = \mathcal{Y} - \mathcal{M}$ 。

$\psi_{\boldsymbol{\lambda}}$  的  $Q \times Q$  维度敏感性矩阵的第  $i$  第  $j$  列元素定义如下：

$$S_{\lambda_{ij}} = E\left(\frac{\partial}{\partial \lambda_i} \psi_{\lambda_j}\right) = -\text{tr}(\mathbf{W}_{\lambda_i} \mathbf{C} \mathbf{W}_{\lambda_j} \mathbf{C}). \quad (3.4.8)$$

$\psi_{\boldsymbol{\lambda}}$  的  $Q \times Q$  维特异性矩阵的第  $i$  第  $j$  列元素定义如下：

$$V_{\lambda_{ij}} = \text{cov}(\psi_{\lambda_i}, \psi_{\lambda_j}) = 2 \text{tr}(\mathbf{W}_{\lambda_i} \mathbf{C} \mathbf{W}_{\lambda_j} \mathbf{C}) + \sum_{l=1}^{nR} k_l^{(4)} (\mathbf{W}_{\lambda_i})_{ll} (\mathbf{W}_{\lambda_j})_{ll}, \quad (3.4.9)$$

$k_l^{(4)}$  是  $\mathcal{Y}_l$  的第四阶累积量 (the fourth cumulant)。2004 年, Jørgensen 和 Knudsen [43] 提出修正的追赶算法用来解  $\psi_{\beta} = 0$   $\psi_{\lambda} = 0$ 。具体更新公式如下:

$$\begin{aligned}\boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - \mathbf{S}_{\beta}^{-1} \psi_{\beta}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\lambda}^{(i)}) \\ \boldsymbol{\lambda}^{(i+1)} &= \boldsymbol{\lambda}^{(i)} - \alpha \mathbf{S}_{\lambda}^{-1} \psi_{\lambda}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)})\end{aligned}$$

2016 年, Bonat 和 Jørgensen [40] 提出倒数似然算法 (reciprocal likelihood algorithm), 通过调节常数  $\alpha$  来调节步长, 从而对参数  $\lambda$  估计。具体更新公式如下:

$$\boldsymbol{\lambda}^{(i+1)} = \boldsymbol{\lambda}^{(i)} - \left[ \alpha \psi_{\lambda}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)})^T \psi_{\lambda}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}) V_{\lambda}^{-1} S_{\lambda} + S_{\lambda} \right]^{-1} \psi_{\lambda}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\lambda}^{(i)}).$$

McGLM 中回归参数估计量对协方差结构的形式依赖性相对较小, 而回归参数估计量的标准误差则直接依赖于协方差结构的选择。McGLM 还可以进行回归参数的稳健和偏差校正标准误差、残差分析等。详见 [40]。

## 3.5 回归分析实践

### 3.5.1 R 语言实践

#### 单响应变量的线性回归模型

(1) 下面对影响城镇居民消费支出的影响因素进行回归分析。

用 R 语言对数据进行回归分析, 代码如下:

```
data=read.csv("D:/data.csv",header = T) #读取数据
rownames(data)=data[,1] #用地区名称对行命名
data=data[,-1] #删掉地区所在列
#y是城镇居民年消费支出
#剩余的变量x1,...,x9是自变量
lm=lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9,data=data) #建立回归方程
summary(lm) #输出回归结果及显著性检验结果
anova(lm) #方差分析表
```

R 输出结果此处省略。从回归参数估计结果可以看出,  $F$  检验对应的  $P$  值  $< 0.05$ , 即回归方程通过了显著性检验。拟合的回归方程为

$$\begin{aligned}\hat{Y} &= -7574 + 1.228X_1 + 1.838X_2 + 0.904X_3 + 0.988X_4 + 1.570X_5 + 0.021X_6 \\ &\quad + 0.008X_7 + 66.570X_8 - 29.490X_9\end{aligned}$$

回归参数估计结果显示，自变量  $X_1 \sim X_6$  对应  $t$  检验的  $P$  值  $< 0.05$ ，其回归系数显著不为零。 $X_1 \sim X_6$  对  $Y$  影响显著，且回归系数为正。但回归常数和自变量  $X_7 \sim X_9$  的显著性检验不能通过，还要进一步完善模型。

(2) 利用显著性自变量，使用 lm 函数对影响城镇居民消费支出作第二次回归，从而最终回归方程为

$$\hat{Y} = -736.4 + 1.23X_1 + 1.83X_2 + 0.99X_3 + 0.96X_4 + 1.6X_5 + 0.02X_6 \quad (3.5.1)$$

若自变量的新样本观测数据为

$$\mathbf{X}_0 = (8000, 2000, 6000, 2500, 3000, 9000, 8000, 103, 2.5)$$

则可利用最终回归方程 (3.5.1) 进行预测，由输出结果可知，因变量的点预测值为 26100.17，置信水平 95% 的预测区间为 (24443.79, 27756.55)。

```
newdata=data.frame(x1=8000,x2=2000,x3=6000,x4=2500, x5=3000,x6=9000,
x7=8000,x8=103,x9=2.5)
ypred=predict(lm_final,newdata,interval = "prediction",level = 0.95)
ypred#点预测及区间预测
  fit          lwr          upr
1 26100.17    24443.79    27756.55
```

(3) 对影响消费支出的各种因素进行岭回归分析。计算代码及运行结果如下：

```
datascale=data.frame(scale(data))#对数据进行标准化并转换为数据框格式存储
datascale=datascale[,1:7]#提取标准化后的因变量y及自变量x1-x6
library(MASS)#加载MASS包
ridge=lm.ridge(y~,data = datascale,lambda = seq(0,10,0.1))#岭回归
beta = ridge$coef#回归系数
lambda = ridge$lambda#岭参数
plot(lambda,lambda,type="n",xlab="岭参数",ylab="岭回归系数",ylim=c(0,0.6))
linetype=c(1:6); color=c(1:6)#创建没有点和线的图形区域
for(i in 1:6)
  lines(lambda,beta[i,],lty=linetype[i],cex=0.75,col=color[i])#绘制岭迹线
legend("topright",legend=c("x1","x2","x3","x4","x5","x6"),cex=0.8,
lty= linetype,col=color)#添加图例
select(ridge)#选择使得 GCV 达到最小值的岭参数 0.1
modified HKB estimator is 0.0409
modified L-W estimator is 0.0151
smallest value of GCV at 0.1
install.packages("ridge"); library(ridge)#下载并安装ridge包
#对选择的岭参数进行岭回归
Ridge_select=linearRidge(y~.-1,data = datascale,lambda = 0.1)
summary(Ridge_select)#输出结果
```

岭迹图3.3显示，随着岭参数  $\lambda$  的增大，岭迹逐渐趋于稳定。当  $\lambda \geq 0.1$  时，六条岭迹已较平稳了。使用 GCV 模型评估标准，可以选取  $\lambda = 0.1$  来建立岭回归方程。由输出结果可知，自变量  $X_1 \sim X_6$  均通过了显著性检验。

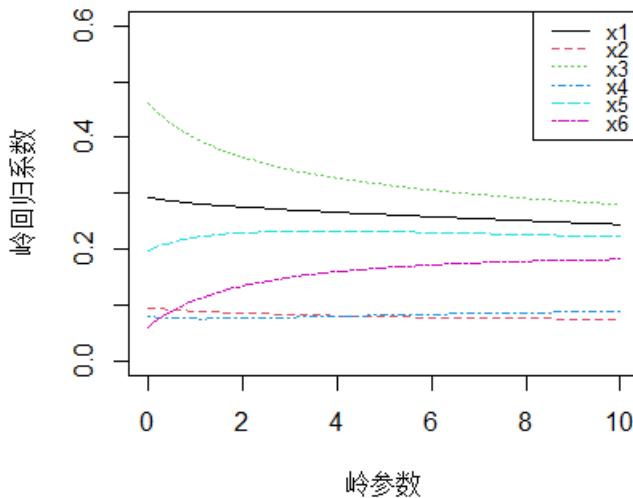


图 3.1 标准化岭迹图

## 广义线性模型

下面使用 `glm` 函数对数据建立广义线性模型，其中 `family` 参数可以根据因变量的数据结构作出选择，比如 `family` 可以包括正态分布的线性模型，两点分布的 Logistic 模型以及 Poisson 的回归模型。下面只展示 Logistic 建模过程。

R 代码如下所示：

```
#读取数据，数据是对45名驾驶员的调查结果
data=read.table("clipboard",header=T)
#x1视力状况（分类变量），x2年龄（数值型），x3驾车教育（分类变量）
#y去年是否出过事故（分类变量）
logit.glm<-glm(y~x1+x2+x3,family=binomial,data=data)#Logistic回归模型
summary(logit.glm)#Logistic回归模型结果
#输出结果如下：


|             | Estimate  | Std. Error | t value | Pr(> t ) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | 0.597610  | 0.894831   | 0.668   | 0.5042   |
| x1          | -1.496084 | 0.704861   | -2.123  | 0.0338*  |
| x2          | -0.001595 | 0.016758   | -0.095  | 0.9242   |
| x3          | 0.315865  | 0.701093   | 0.451   | 0.6523   |


```

由于  $X_1$  显著，重新运行仅仅基于  $X_1$  的 Logistic 模型，得到筛选后的的 Logistic 模型：

$$\Pr(Y = 1 | X_1) = \frac{\exp(0.6190 - 1.3728X_1)}{1 + \exp(0.6190 - 1.3728X_1)}$$

由结果可知，视力因素对事故率有影响。下面预测视力正常司机即  $X_1 = 1$  及视力有

问题司机，即  $X_1 = 0$  发生事故的概率。R 代码如下：

```
pre1<-predict(logit.step,data.frame(x1=1))#预测视力正常司机Logistic回归结果
p1<-exp(pre1)/(1+exp(pre1))#预测视力正常司机发生事故概率
pre2<-predict(logit.step,data.frame(x1=0))#预测视力有问题的司机Logistic回归结果
p2<-exp(pre2)/(1+exp(pre2))#预测视力有问题的司机发生事故概率
c(p1,p2)#结果显示
1           1
0.32       0.65
```

## McGLM 模型

mcglm 包是一个基于 Matrix 包的完整 R 实现。可以利用 mcglm 包的函数 mcglm() 来拟合 McGLM 模型。模型由一组列表指定，列表给出了线性和矩阵线性预测的符号描述。mcglm() 允许在连接、方差和协方差函数列表中进行选择，灵活地指定均值和协方差结构，并通过类似于 glm() 函数的公式界面来明确处理多元响应变量。mcglm() 将回归参数的拟得分函数和协方差参数的皮尔逊估计函数结合起来，采用估计函数法拟合 McGLM 模型。

本例采用澳大利亚健康调查数据集 ahs (Australian Health Survey)，数据包括 1987-1988 年澳大利亚有关卫生系统准入措施的 5190 个样本数据，包含 5 个计数响应变量和有关社会条件的 9 个协变量。数据集 ahs 中主要变量包括：Ndoc，离散响应变量，离散变量，咨询医生或专家的次数；Nndoc，离散响应变量，与卫生专业人员咨询的次数；age，连续自变量，受访者的年龄（以年为单位）除以 100；income，连续自变量；受访者的年收入（以澳元计）除以 1000。利用数据集 ahs，建立响应变量 Ndoc、Nndoc 关于自变量 income、age 的二元响应变量协方差广义线性模型。

```
library(mcglm)#加载mcglm包
data(ahs)#数据集ahs
form1 = Ndoc ~ income+age#第一个响应变量Ndoc的线性预测公式
form2 = Nndoc ~ income+age#第二个响应变量Nndoc的线性预测公式
Z0 = mc_id(ahs)#构建矩阵线性预测器的第一个分量矩阵
options(digits = 3)#设定保留小数点后3位
#拟合McGLM模型，其中参数：linear_pred指定线性预测器；matrix_pred指定矩阵线性预测器；link指定连接函数；variance指定方差函数；data指定数据集
fit.ahs = mcglm(linear_pred = c(form1,form2),matrix_pred = list(Z0,Z0),
link= c("log","log"),variancec("poisson_tweedie","poisson_tweedie"),data = ahs)
summary(fit.ahs)
coef(fit.ahs)#参数估计值
gof(fit.ahs)#输出拟合优度
confint(fit.ahs)#模型参数的区间估计
anova(fit.ahs)#拟合模型的方差分析表
vcov(fit.ahs)#参数估计量的方差协方差阵
fitted(fit.ahs)#输出拟合值
residuals(fit.ahs)#输出皮尔逊、原始和标准化残差
plot(fit.ahs)
#输出各自变量的得分信息准则SIC，用于选择线性预测器的分量
mc_sic(fit.ahs,scope = c("income","age"), data = ahs,response = 1)
```

```
#输出每个分量矩阵的SIC协方差值、自由度等
mc_sic_covariance(fit.ahs, scope =Z0, idx = 1, response = 1)
```

对响应变量 Ndoc、Nndoc 分别选取线性预测器 form1、form2，矩阵线性预测器的第一个分量矩阵均为单位矩阵，连接函数均为对数函数，方差函数均为 poisson\_tweedie，协方差连接函数均采用恒等变换。利用数据集 ahs 拟合 McGLM，运行结果表明，所有回归参数、散度参数及相关系数估计均显著。

### 3.5.2 Python 语言实践

#### 单响应变量的线性回归模型

在 Python 中，常使用 statsmodels 和 sklearn 库来做线性回归。首先选择来自 sklearn 库的波士顿房价数据集建立回归模型。

```
import pandas as pd
from matplotlib import pyplot as plt
from sklearn import datasets, linear_model
import statsmodels.api as sm
boston = datasets.load_boston() #导入数据集
X = pd.DataFrame(boston.data, columns=boston.feature_names)
y = pd.DataFrame(boston.target, columns=['MEDV'])
X = sm.add_constant(X) #由于此模型没有截距项，给训练集加上一列数值为1的特征
model = sm.OLS(y,X).fit()
print(model.summary())
```

输出结果显示，p 值大于 0.05 即认为与波士顿的房价波动变化关系不大，因此剔除 INDUS、AGE 选项，再次使用 statsmodel 中的最小二乘法：

```
X.drop(['AGE', 'INDUS'], axis=1, inplace=True)
model = sm.OLS(y,X).fit()
print(model.summary())
#拟合效果可视化
pre = model.predict(X.values)
plt.figure(figsize=(16,10))
plt.xticks(range(0, 506, 10), rotation=45)
plt.plot(range(len(X)), pre, 'b', label='predict')
plt.plot(range(len(y)), y.values, 'r', label='real')
plt.legend()
plt.show()
```

拟合效果如图3.2所示。

此外，用 sklearn 进行线性回归也是十分常见的作法，下面介绍用 Python 语言实现的，利用 sklearn 简单线性回归的通用方程拟合线性模型。这里我们将使用 SciKit-Learn 糖尿病数据集 (<https://scikit-learn.org/stable/datasets/index.html>)。

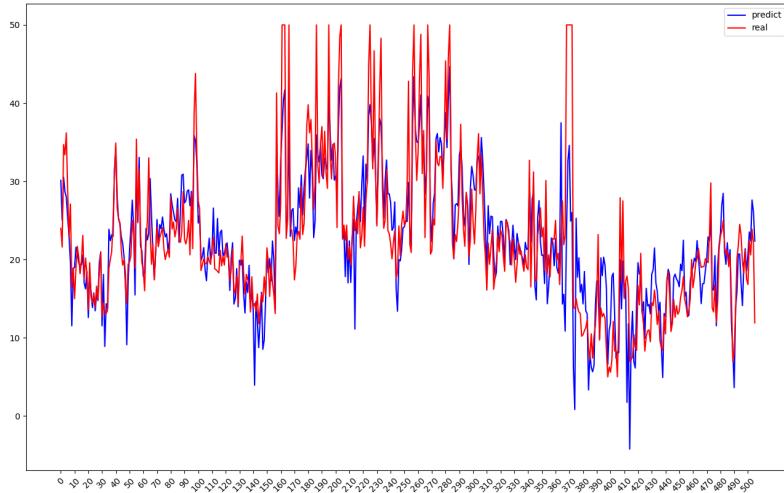


图 3.2 模型拟合效果

该数据集包含 10 个特征（均值中心化和缩放后）和一个目标值：衡量基本线后一年的疾病进展情况。下面是代码：

```
from sklearn.datasets import load_diabetes
from sklearn.model_selection import train_test_split
diabetes, target = load_diabetes(return_X_y=True) #载入数据集
diabetes = pd.DataFrame(diabetes) #准备模型数据
y = target #提取特征和目标
X = diabetes
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,
random_state=27) #设置测试集和训练集
```

用 sklearn 进行线性回归，得到  $R^2$  值 (CV Mean) 和标准偏差 (STD)。

```
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score
lr = LinearRegression().fit(X_train, y_train) #训练模型
#得到交叉验证评分
print('CV Mean: ', cross_val_score(lr, X_train, y_train).mean())
print('STD: ', cross_val_score(lr, X_train, y_train).std())
```

平均  $R^2$  值 (CV Mean) 为 0.48 和标准偏差 (STD) 为 0.14。这里  $R^2$  值较低表明模型不是很准确。标准偏差值 (STD) 为 0.14 表明可能对训练数据进行了过度拟合。解决过拟合的一个办法是简化模型。下面通过引入正则化来简化线性回归模型。

与其他线性模型一样，岭回归 (Ridge) 将在其拟合方法中接受数组  $X, y$ ，并将线性模型的系数  $w$  存储在其 *coefmember* 中。在岭回归中用正则化参数  $\alpha$  来控制模型的复杂性。 $\alpha$  的数值越高，就越能迫使系数向零移动，增加对模型的限制。这降

低了训练性能，但也增加了模型的普适性。但把  $\alpha$  设置得太高可能会导致模型过于简单，对数据的拟合不足。

下面使用 Scikit-Learn 的 Ridge 类来提高性能。

```
from sklearn.linear_model import Ridge
ridge = Ridge(alpha=1).fit(X_train, y_train) #选取 alpha=1 训练模型
print('CV Mean: ', cross_val_score(ridge, X_train, y_train).mean())
print('STD: ', cross_val_score(ridge, X_train, y_train).std()) #得到交叉验证评分
```

以上结果， $R^2$  值 (CV Mean) 为 0.48 为 0.38，这意味着岭回归模型只能解释百分之 38 的方差，与上面的线性回归相比，没有改进。但标准差减少，这表明不太可能出现过拟合。

我们使用了上述正则化参数  $\alpha$  的默认值，这可能不会带来最佳性能，下面通过调整正则化参数  $\alpha$  来提高  $R^2$  值 (CV Mean)，下面使用网格搜索来找到一个最佳的  $\alpha$  值。

```
from sklearn.model_selection import GridSearchCV
alpha = [0.001, 0.01, 0.1, 1, 10, 100, 1000]
param_grid = dict(alpha=alpha) #使用网格搜索来找到一个最佳 alpha
grid = GridSearchCV(estimator=ridge, param_grid=param_grid,
scoring='r2', verbose=1, n_jobs=-1)
grid_result = grid.fit(X_train, y_train)
print('Best Score:', grid_result.best_Score_)
print('Best Params:', grid_result.best_params_)
```

## 广义线性模型

此小节我们学习一个利用 Python 语言实现的广义线性模型中对数线性泊松回归。下面这个例子说明了对数线性泊松回归在法国汽车第三方责任索赔数据集 (<https://www.openml.org/d/41214>) 上的使用，具体可参见：A.Noll, R.Salzmann, ndM.V.Wuthrich, CaseStudy : FrenchMotorThird-PartyLiabilityClaims(November8, 2018).doi : 10.2139/ssrn.3164764

首先，绘制数据集，在这个数据集中，每个样本对应一个保险单。可用的特征包括：司机年龄、车龄、车辆功率等。我们的目标是给定投保人群体的历史数据，预测一个新的投保人在发生车祸后的预期索赔频率。

从 OpenML 下载汽车索赔数据集：<https://www.openml.org/d/41214>

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.datasets import fetch_openml#这一模块主要包含现实生活中的大型的数据集
df = fetch_openml(data_id=41214, as_frame=True).frame#导入汽车第三方责任索赔数据集
```

索赔数量 (ClaimNb) 是一个正整数，可以被建模为泊松分布。然后假定它是在一个给定的时间间隔内以恒定的速率发生的离散事件的数量 (曝光率，以年为单位)。

这里通过一个（按比例）泊松分布对  $X$  的频率  $y = ClaimNb/Exposure$  进行建模，并使用  $Exposure$  作为样本权重。

```
df["Frequency"] = df["ClaimNb"] / df["Exposure"]
# 使用 Exposure 作为样本权重，得到平均频率
print("Average Frequency= {}".format(np.average(df["Frequency"]),
weights=df["Exposure"])))
# 输出零索赔的比例
print("Fraction of exposure with zero claims={:.1%}".format(
df.loc[df["ClaimNb"] == 0, "Exposure"].sum() / df["Exposure"].sum()))
fig, (ax0, ax1, ax2) = plt.subplots(ncols=3, figsize=(16, 4))
# 利用 plt.subplots 函数绘制并列的三个图
ax0.set_title("Number of claims")
ax0 = df["ClaimNb"].hist(bins=30, log=True, ax=ax0) # 索赔数量
ax1.set_title("Exposure in years")
# 风险指某一保单的承保期限，以年为单位
ax1 = df["Exposure"].hist(bins=30, log=True, ax=ax1)
ax2.set_title("Frequency (number of claims per year)")
ax2 = df["Frequency"].hist(bins=30, log=True, ax=ax2) # 索赔频率（以每年索赔数量衡量）
```

运行产生以下结果和图像：

```
Average Frequency= 0.10070308464041304
Fraction of exposure with zero claims= 93.9%
```

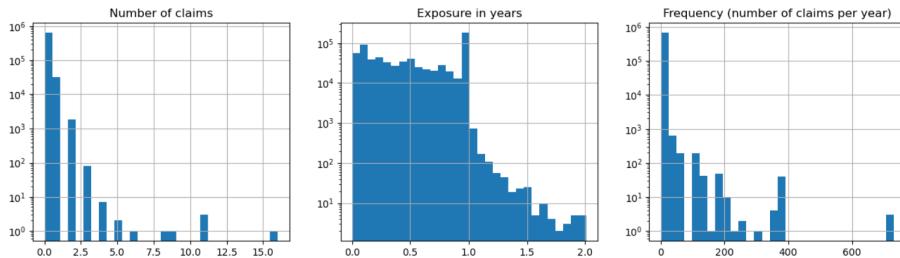


图 3.3 汽车索赔数据图

剩下的几栏可以用来预测索赔事件的频率。这些列是非常异质的，混合了不同尺度的分类和数字变量，可能分布非常不均匀。因此，为了用这些预测因子来拟合线性模型，可进行标准特征转换。

```
from sklearn.pipeline import make_pipeline
from sklearn.preprocessing import FunctionTransformer, OneHotEncoder
from sklearn.preprocessing import StandardScaler, KBinsDiscretizer
from sklearn.compose import ColumnTransformer
# 导入 pipeline 类，可以将多个处理步骤合并为单个 Scikit-Learn 学习器
log_scale_transformer = make_pipeline(FunctionTransformer(np.log, validate=False),
StandardScaler())
# 函数 ColumnTransformer，可以选择地进行数据转换和预处理。
linear_model_preprocessor=ColumnTransformer([('passthrough_numeric', "passthrough",
["BonusMalus"]),
("binned_numeric", KBinsDiscretizer(n_bins=10, encode='onehot-dense'))])
```

```
), ["VehAge", "DrivAge"]),
("log_scaled_numeric", log_scale_transformer, ["Density"]),
("onehot_categorical", OneHotEncoder(sparse=False), ["VehBrand", "VehPower",
"VehGas", "Region", "Area"]]))
```

首先用 (L2 惩罚) 最小二乘法线性回归模型 (即 Ridge 回归) 对目标变量进行建模。预计这样的线性模型在这样一个大的数据集上会出现拟合不足, 故使用较低的惩罚  $\alpha$ 。

```
from sklearn.linear_model import Ridge # 导入岭回归模型
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error, mean_absolute_error,
mean_poisson_deviance
# Pipeline 把若干个命令串起来, 前面命令的输出成为后面命令的输入
y, X = df.pop('Frequency'), df
y_train, y_test, X_train, X_test = train_test_split(y, X, random_state=0)
ridge_glm = Pipeline([("preprocessor", linear_model_preprocessor),
("regressor", Ridge(alpha=1e-6))]).fit(X_train, y_train, regressor__sample_weight=X_train["Exposure"])
# model.fit 用来拟合模型参数
```

泊松偏差不能用模型所预测的非正数值来计算。对于确实返回一些非阳性预测的模型 (如 Ridge), 我们忽略了相应的样本, 这意味着得到的泊松偏差是近似的。

```
pos = (y_test > 0) & (ridge_glm.predict(X_test) > 0)
print("Ridge evaluation: ")
print("MSE: ", mean_squared_error(y_test, ridge_glm.predict(X_test)))
print("MAE: ", mean_absolute_error(y_test, ridge_glm.predict(X_test)))
# 输出岭回归评价 MSE 和 MAE 及平均泊松偏差
print("平均泊松偏差: ", mean_poisson_deviance(y_test[pos], ridge_glm.predict(X_test)[pos]))
```

接下来在目标变量上拟合泊松回归。将正则化强度  $\alpha$  设置为大约  $1e-6$ , 超过样本数 (即  $1e-12$ ), 以模仿 Ridge 回归, 其 L2 惩罚项随样本数的变化而变化。由于泊松回归在内部对预期目标值的对数而不是直接对预期值进行建模 (对数与恒等连接函数),  $X$  和  $y$  之间的关系不再是完全线性的。因此, 泊松回归被称为广义线性模型 (GLM), 而不是像 Ridge 回归那样的普通线性模型。

```
from sklearn.linear_model import PoissonRegressor # 导入 PoissonRegressor 拟合泊松回归
from sklearn.pipeline import Pipeline
from sklearn.metrics import mean_squared_error, mean_absolute_error,
mean_poisson_deviance
n_samples = df_train.shape[0]
# Pipeline 函数将多个任务抽象为一个包括多个步骤的流水线式工作
y, X = df.pop('Frequency'), df
y_train, y_test, X_train, X_test = train_test_split(y, X, random_state=0)
poisson_glm = Pipeline([("preprocessor", linear_model_preprocessor),
("regressor", PoissonRegressor(alpha=1e-12, max_iter=300))])
poisson_glm.fit(X_train, y_train, regressor__sample_weight=X_train["Exposure"])
```

同样地, 输出输出泊松回归的 MSE 和 MAE 及平均泊松偏差

```

print("Ridge evaluation: ")
print("MSE: ", mean_squared_error(y_test, poisson_glm.predict(X_test)))
print("MAE: ", mean_absolute_error(y_test, poisson_glm.predict(X_test)))
print("平均泊松偏差: ", mean_poisson_deviance(y_test, poisson_glm.predict(X_test)))

```

## McGLM 模型

McGLM 方法可以很方便的通过 R 中 mcglm 包中函数 mcglm() 实现。但是如果你想在 python 中使用上述方法，并且调用相关结果的话，可以参考下述步骤：

- 1、以上述 R 语言代码为例，首先将 McGLM 方法以 R 语言形式保存至桌面且命名为“rcode.R”。
- 2、打开 python 编译器，并且安装“rpy2”库。
- 3、运行下述代码。

```

import rpy2.robjects as robjects #导入rpy2.robjects函数，用来调用R脚本。
robjects.r.source("C:/Users/Desktop/rkode.R")
#输入R脚本本地址后，函数robjects.r.source会直接调用R脚本，并且输出结果。

```

## 3.6 习题

1、研究货运总量  $Y$  (万吨) 与工业总产值  $X_1$  (亿元)、农业总产值  $X_2$  (亿元)、居民非商品支出  $X_3$  (亿元) 的关系，数据见表3.4。

表 3.4 习题 1 数据表

序号	1	2	3	4	5	6	7	8	9	10
$Y$	160	260	210	265	240	220	275	160	275	250
$X_1$	70	75	65	74	72	68	78	66	70	65
$X_2$	35	40	40	42	38	45	42	36	44	42
$X_3$	1	2.4	2	3	1.2	1.5	4	2	3.2	3

- (1) 计算  $Y, X_1, X_2, X_3$  的相关系数矩阵；

- (2) 求  $Y$  关于  $X_1, X_2, X_3$  的线性回归方程;
- (3) 对拟合的线性回归方程进行拟合优度检验、方程的显著性检验，对每个回归系数做显著性检验;
- (4) 若有回归系数未通过显著性检验，剔除，重新拟合回归方程并进行检验;
- (5) 求回归系数的置信度 95% 的置信区间;
- (6) 求标准化回归方程;
- (7) 将  $X_1, X_2, X_3$  分别取 75, 42, 3.1, 对因变量进行点预测及置信度 95% 的区间预测。

2、某种水泥在凝固时放出的热量  $y$  (卡/克, cal/g) 与水泥中的四种化学成分的含量 (%) 有关, 这四种化学成分分别是  $X_1$  铝酸三钙 ( $3\text{CaO}\cdot\text{Al}_2\text{O}_3$ ),  $X_2$  硅酸三钙 ( $3\text{CaO}\cdot\text{SiO}_2$ ),  $X_3$  铁铝酸四钙 ( $4\text{CaO}\cdot\text{Al}_2\text{O}_3\cdot\text{Fe}_2\text{O}_3$ ),  $X_4$  硅酸二钙 ( $2\text{CaO}\cdot\text{SiO}_2$ )。现观测到 13 组数据, 数据见表3.5。用逐步回归法做变量选择, 建立  $Y$  关于四种成分的线性回归方程。

表 3.5 习题 2 数据

序号	1	2	3	4	5	6	7	8	9	10	11	12	13
$Y$	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4
$X_1$	7	1	11	11	7	11	3	1	2	21	1	11	10
$X_2$	26	29	56	31	52	55	71	31	54	47	40	66	68
$X_3$	6	15	8	8	6	9	17	22	18	4	23	9	8
$X_4$	60	52	20	47	33	22	6	44	22	26	34	12	12

3、一家大型商业银行有多家分行, 近年来, 该银行的贷款额平稳增长, 但不良贷款额也有较大比例的提高。为弄清楚不良贷款形成的原因, 下面利用银行业务的有关数据进行定量分析, 找出控制不良贷款的办法。数据表3.6是该银行所属 25 家分行某年的有关业务数据。

- (1) 建立不良贷款与其余 4 个自变量的线性回归方程, 对模型进行检验;
- (2) 分析回归模型的共线性;
- (3) 采取逐步回归法选择变量, 对模型进行检验;
- (4) 建立不良贷款  $y$  对 4 个自变量的岭回归。

4、临床医学中为了研究麻醉剂用量与患者是否保持静止的关系, 对 30 名患者在手术前 15 分钟给予一定浓度的麻醉剂后的情况进行了记录。记录数据来自于 R 软件 DAAG 包中自带的 *anesthetic* 数据集, 见表3.7, 其中, 麻醉剂浓度为自变量  $X$ , 患者是否保持静止为因变量  $Y$ ,  $Y$  取 1 时表示患者静止,  $Y$  取 0 时表示患者有移动, 试建立  $Y$  关于  $X$  的 Logistic 回归模型。

5、R 中包 *mcglm* 自带数据集 *NewBorn* 是有关早产儿的呼吸物理治疗数据集, 旨在评估呼吸物理疗法对出生体重低于 1500g 的早产儿通气后心肺功能的影

表 3.6 习题 3 数据

分行编号	1	2	3	4	5	6	7	8	9	10	11	12	13
y	0.9	1.1	4.8	3.2	7.8	2.7	1.6	12.5	1	2.6	0.3	4	0.8
x1	67.3	111.3	173	80.8	199.7	16.2	107.4	185.4	96.1	72.8	64.2	132.2	58.6
x2	6.8	19.8	7.7	7.2	16.5	2.2	10.7	27.1	1.7	9.1	2.1	11.2	6
x3	5	16	17	10	19	1	17	18	10	14	11	23	14
x4	51.9	90.9	73.7	14.5	63.2	2.2	20.2	43.8	55.9	64.3	42.7	76.7	22.8
分行编号	14	15	16	17	18	19	20	21	22	23	24	25	
y	3.5	10.2	3	0.2	0.4	1	6.8	11.6	1.6	1.2	7.2	3.2	
x1	174.6	263.5	79.3	14.8	73.5	24.7	139.4	368.2	95.7	109.6	196.2	102.2	
x2	12.7	15.6	8.9	0.6	5.9	5	7.2	16.8	3.8	10.3	15.8	12	
x3	26	34	15	2	11	4	28	32	10	14	16	10	
x4	117.1	146.7	29.9	42.1	25.3	13.4	64.3	163.9	44.5	67.9	39.7	97.1	

表 3.7 习题 4 数据

序号	1	2	3	4	5	6	7	8	9	10
麻醉剂浓度 X	1	1.2	1.4	1.4	1.2	2.5	1.6	0.8	1.6	1.4
患者是否保持静止 Y	1	0	1	0	0	1	1	0	1	0
序号	11	12	13	14	15	16	17	18	19	20
麻醉剂浓度 X	0.8	1.6	2.5	1.4	1.6	1.4	1.4	0.8	0.8	1.2
患者是否保持静止 Y	0	1	1	1	1	1	1	0	1	1
序号	21	22	23	24	25	26	27	28	29	30
麻醉剂浓度 X	0.8	0.8	1	0.8	1	1.2	1	1.2	1	1.2
患者是否保持静止 Y	0	0	0	0	0	1	0	1	0	1

响。NewBorn 数据集由护理团队 Waldemar Monastier hospital, Campo Largo, PR, Brazil 收集。样本容量 270 个，变量 21 个。数据集中主要变量含义解释见表3.8。利用响应变量 SPO2 关于自变量 Sex、APGAR1M、APGAR5M、PRE、HD, SUR 建立 McGLM 模型。

6、R 中包 mcglm 自带数据集 soya 中有关豌豆的样本数据。数据集有 7 个变量的 75 个观察值。数据集中的试验是在一个种植大豆的温室中进行的。此试验有两株按地块种植的植物。数据集 soya 中变量含义见表3.9。为研究豌豆的产量、每株豌豆的种子数量及可生长发育的豌豆比率问题，建立三元独立响应变量 grain、seeds、viablepeasP 关于协变量 pot、water、block 的 McGLM 模型并对模型拟合结果进行分析。

表 3.8 习题 5 NewBorn 数据集的主要变量含义表

变量	描述
Sex	性别
APGAR1M	生命第一分钟的 APGAR 指数
APGAR5M	生命第五分钟的 APGAR 指数
PRE	因素, 两个类别 (过早: 是; 否)
HD	因子, 两个水平 (汉森病, 是; 否)
SUR	因子, 两级 (表面活性剂, 是; 否)
SPO2	氧饱和度 (有界)

表 3.9 习题 6 soya 数据集中变量含义表

变量	描述
pot	自变量, 钾肥, 五水平因子 (0,30,60,120,180)
water	自变量, 土壤中的水分, 三水平因子 (37.5,50,62.5)
block	自变量, 地块, 五水平因子 (I,II,III,IV,V)
grain	响应变量, 连续型数据, 单株粮食产量
seeds	响应变量, 计数型数据, 每株植物的种子数量
viablepeas	二项式型数据, -每株可存活豌豆的数量。
totalpeas	二项式型数据, -每株豌豆的总数量
viablepeasP	响应变量, 每株可存活豌豆比率, 值为 viablepeas/totalpeas

## 第四章 支持向量机

### 4.1 简介

对给定样本数据集进行分类处理通常有多种办法，如前面章节中介绍的：GLM 或者 McGLM 等，而支持向量机（Support Vector Machine, SVM）作为众多分类器中的一种，却有着其他分类器没有的优点，下面我们就来介绍一下 SVM。

支持向量机（Support Vector Machine, SVM）是一种基于统计学习理论的有监督新型学习机，是由前苏联教授 Vladimir N. Vapnik 等 [58] 于 1963 年在统计学习理论基础上提出的。支持向量机理论最初用来解决两类线性可分问题，之后逐步推广到多分类、非线性等问题。与传统学习方法不同，支持向量机是结构风险最小化方法的近似实现，是借助最优化方法来解决机器学习分类问题的新工具。在处理线性可分问题时，其“对样本依赖小”等优点使其成为了众多分类器当中的佼佼者。

### 4.2 SVM 算法

#### 4.2.1 SVM 的基本内容

最初 SVM 被提出时，是用来解决二分类问题的。其主要内容是：对于给定的包含两个类别的样本数据集  $\mathbf{Z} = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}, \mathbf{X}_i \in \mathcal{R}^p, Y_i \in \{-1, 1\}$ ，确定一个超平面，对数据集进行二分类处理，如图4.1。

其分类判别式为：

$$f(\mathbf{X}) = \text{sign}(\mathbf{w}^\top \mathbf{X} + b) \quad (4.2.1)$$

可以看到，不止有一个超平面能够把样本数据正确的分为两类，SVM 算法是在这众多超平面中挑选出鲁棒性最强的超平面  $\mathbf{w}^{*\top} \mathbf{X} + b^*$  来作为最终的分类器。

SVM 可分为三种不同的类别：[线性可分 SVM](#)、[线性 SVM](#)、[非线性 SVM](#)。下面分别对这三种类型的 SVM 进行介绍。

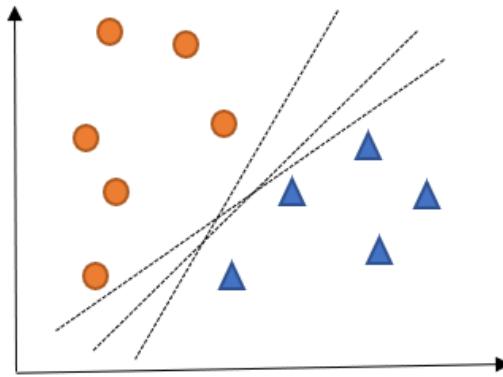


图 4.1 二分类超平面

### 4.2.2 线性可分 SVM

若至少存在一个超平面，可以将两类数据完全分开，此时的 SVM 被称为**线性可分 SVM**。线性可分 SVM 的原理就是要达到硬间隔最大化，又被称为**最大间隔分类器**。在样本点可以被超平面分为两类的基础上，线性可分 SVM 的目标是找到一个超平面，使得每个样本点到该超平面的距离都足够大。下面对算法理论进行详细阐述。对于给定样本数据集  $\mathbf{Z} = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}, \mathbf{X}_i \in \mathcal{R}^p, Y_i \in \{-1, 1\}$ 。在样本空间中，将划分超平面用如下线性方程来描述：

$$\mathbf{w}^T \mathbf{X} + b = 0$$

其中  $\mathbf{w} = (w_1, w_2, \dots, w_p)^T$  为法向量， $b$  为位移项。显然，超平面可以被法向量  $\mathbf{w}$  和位移  $b$  所确定，记为  $(\mathbf{w}, b)$ 。对于  $(\mathbf{X}_i, Y_i) \in \mathbf{Z}$ ，其到超平面  $(\mathbf{w}, b)$  的距离可表示为：

$$r_i = \frac{|\mathbf{w}^T \mathbf{X}_i + b|}{\|\mathbf{w}\|} \quad (4.2.2)$$

所有  $r_i$  ( $i = 1, \dots, n$ ) 的最小值称为**间隔**，可表示为式 (4.2.3)。而这些和超平面  $(\mathbf{w}, b)$  距离最小的样本点称作**支持向量**（如图 4.2 所示）

$$\text{margin} = \min_{1 \leq i \leq n} r_i \quad (4.2.3)$$

为了使这个超平面更具鲁棒性，SVM 算法的目标是通过调整参数  $\mathbf{w}$  和  $b$ ，找到以最大间隔把两类样本分开的超平面，也称之为最大间隔超平面。首先，我们希望

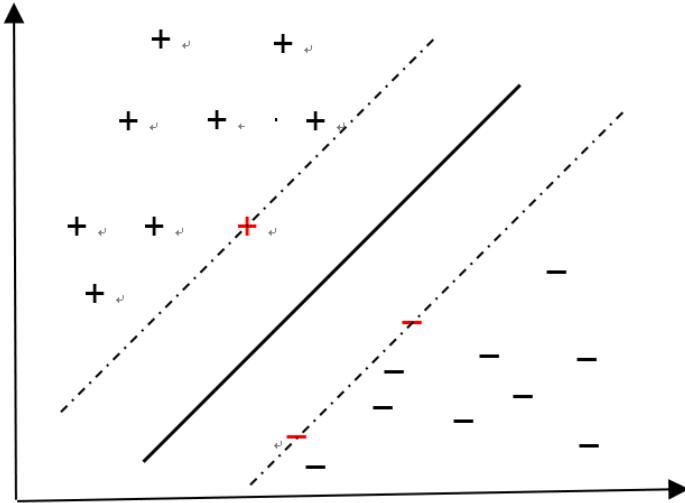


图 4.2 支持向量

两类样本分别分割在该超平面的两侧；其次，两侧距离超平面最近的样本点到超平面的距离被最大化。即：

$$\max_{\mathbf{w}, b} \min_{1 \leq i \leq n} r_i \quad (4.2.4)$$

假设样本点可以被超平面分为两个类别，显然对于  $(\mathbf{X}_i, Y_i) \in \mathbf{Z}$ ，满足下面约束条件：

$$\begin{cases} \mathbf{w}^T \mathbf{X}_i + b > 0, & Y_i = 1 \\ \mathbf{w}^T \mathbf{X}_i + b < 0, & Y_i = -1 \end{cases}$$

可将其化简为：

$$Y_i (\mathbf{w}^T \mathbf{X}_i + b) > 0 \quad (4.2.5)$$

这一约束条件意味着所有样本点被超平面正确分类。下面希望从能将样本数据集正确分类的超平面里选取特殊的一个超平面，使得两类样本数据集到这一超平面的最短距离最大化，即间隔最大化：

$$\begin{aligned} & \max_{\mathbf{w}, b} \min_{1 \leq i \leq n} r_i \\ \text{s.t. } & Y_i (\mathbf{w}^T \mathbf{X}_i + b) > 0 \end{aligned} \quad (4.2.6)$$

在该约束条件下，可以将  $r_i$  转化为：

$$r_i = \frac{Y_i (\mathbf{w}^T \mathbf{X}_i + b)}{\|\mathbf{w}\|}$$

优化目标函数转化为：

$$\max_{\mathbf{w}, b} \min_{1 \leq i \leq n} \frac{Y_i (\mathbf{w}^T \mathbf{X}_i + b)}{\|\mathbf{w}\|}$$

为了便于化简计算，对超平面  $(\mathbf{w}, b)$  进行放缩变换，使得  $|\mathbf{w}^T \mathbf{X}_i + b| \geq 1$ ，即  $Y_i (\mathbf{w}^T \mathbf{X}_i + b) \geq 1$ 。

因此，优化问题 (4.2.6) 转换为下面形式：

$$\begin{aligned} & \max_{\mathbf{w}, b} \frac{1}{\|\mathbf{w}\|} \\ & \text{s.t. } Y_i (\mathbf{w}^T \mathbf{X}_i + b) \geq 1 \quad i = 1, 2, \dots, n \end{aligned} \quad (4.2.7)$$

注意到目标函数中的  $\|\mathbf{w}\|$  在分母上，所以可以将求最大值的约束优化问题转换成求最小值的凸二次规划问题：

$$\begin{aligned} & \min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \\ & \text{s.t. } Y_i (\mathbf{w}^T \mathbf{X}_i + b) \geq 1 \quad i = 1, 2, \dots, n \end{aligned} \quad (4.2.8)$$

(注：此处的  $\frac{1}{2}$  并不影响最终的计算结果)

上式即为线性可分 SVM 的最优化问题。上述优化问题为不等式约束的优化问题，可利用[拉格朗日乘子法](#)将原问题转化为对偶问题。此外，这样做可以更自然的引入核函数，进而推广到非线性的分类问题。对于更一般化的约束优化问题来说，对偶问题可以将非凸问题转化为凸优化问题。

首先，引入拉格朗日乘子向量  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)^T$ ，并写出[拉格朗日函数](#)：

$$L(\mathbf{w}, b, \lambda) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \lambda_i (Y_i (\mathbf{w}^T \mathbf{X}_i + b) - 1), \quad \lambda_i \geq 0 \quad (4.2.9)$$

此时，可以得到原问题 (4.2.8) 等价于一个极大极小化问题：

$$\min_{\mathbf{w}, b} \max_{\lambda} L(\mathbf{w}, b, \lambda) \quad (4.2.10)$$

进而，可定义原问题 (4.2.8) 的[对偶问题](#)：

$$\max_{\lambda} \min_{\mathbf{w}, b} L(\mathbf{w}, b, \lambda) \quad (4.2.11)$$

根据凸优化理论，可以知道，当原问题的目标函数和不等式约束函数是凸函数时，在不等式约束函数严格可行情况下，原问题最优解  $(\mathbf{w}^*, b^*)$  与其对偶问题的最优解  $\lambda^*$  满足下面等式

$$L(\mathbf{w}^*, b^*, \lambda^*) = \min_{\mathbf{w}, b} \max_{\lambda} L(\mathbf{w}, b, \lambda) = \max_{\lambda} \min_{\mathbf{w}, b} L(\mathbf{w}, b, \lambda) \quad (4.2.12)$$

显然, 这里目标函数  $\frac{1}{2}\|\mathbf{w}\|^2$  及其约束函数  $1 - Y_i(\mathbf{w}^\top \mathbf{X}_i + b)$  ( $i = 1, 2, \dots, n$ ) 是凸函数, 那么上述等式在 SVM 中仍然成立。此时, 可以对原问题的对偶问题求最优解, 进而可以求出原问题的最优解。

根据 (4.2.11), 首先对  $\mathbf{w}$  和  $b$  求极小值。令  $L(\mathbf{w}, b, \lambda)$  对  $\mathbf{w}$  和  $b$  的偏导数为 0, 可以得到:

$$\begin{aligned}\frac{\partial L}{\partial b} &= 0 \Rightarrow \sum_{i=1}^n \lambda_i Y_i = 0 \\ \frac{\partial L}{\partial \mathbf{w}} &= 0 \Rightarrow \mathbf{w} = \sum_{i=1}^n \lambda_i Y_i \mathbf{X}_i\end{aligned}$$

将所得的关系带入  $L(\mathbf{w}, b, \lambda)$ , 式 (4.2.10) 转化为

$$\begin{aligned}\max_{\lambda} & \sum_{i=1}^n \lambda_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j Y_i Y_j \mathbf{X}_i^\top \mathbf{X}_j \\ \text{s.t. } & \sum_{i=1}^n \lambda_i Y_i = 0, \quad \lambda_i \geq 0.\end{aligned}\tag{4.2.13}$$

式 (4.2.13) 是一个含等式约束的优化问题。我们将其转化为求最小值问题, 上式等价于

$$\begin{aligned}\min_{\lambda} & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \lambda_i \lambda_j Y_i Y_j \mathbf{X}_i^\top \mathbf{X}_j - \sum_{i=1}^n \lambda_i \\ \text{s.t. } & \sum_{i=1}^n \lambda_i Y_i = 0 \quad \lambda_i \geq 0\end{aligned}\tag{4.2.14}$$

可以发现, 上述对偶问题是一个二次规划问题, 可以采用序列最小最优化 (sequential minimal optimization, SMO) 算法求解最优化问题 (4.2.14)。序列最小优化算法, 其核心思想非常简单: 每次只优化一个参数, 其他参数先固定住, 仅求当前这个优化参数的极值。

为提高求解效率, 其求解过程每次选择两个变量进行优化, 其他变量固定, 具体而言, 算法流程可描述为:

- (1) 选取两个待优化变量;
- (2) 固定其他变量的值, 求解优化问题 (4.2.14) (直接令目标函数关于待优化变量梯度等于零即可);
- (3) 不断重复第 (1)、(2) 两个步骤, 直到算法收敛。

SMO 算法一般选择违反 KKT 条件最严重的样本所对应的变量作为第一个优化变量, 第二个变量的选择应当使得目标函数有足够大的下降。一种启发式的选择方式为: 首先寻找与第一个变量所对应样本间隔最大的样本, 然后将该样本所对应的变量设置为第二个变量。

使用 SMO 算法可以求得对偶问题的最优解  $\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_n^*)$ 。根据式 (4.2.12)，对偶问题的最优解也是原问题的最优的拉格朗日乘子向量，接下来把最优的拉格朗日乘子向量带入原问题求解即可。

为了简化计算，我们引入 KKT 条件。已知 KKT 条件是判断等式约束和不等式约束的优化问题的必要条件。对于只含不等式约束的优化问题来说，如果目标函数和约束函数是凸函数，且不等式约束函数是可以满足的，那么 KKT 条件是这一不等式约束的优化问题的充要条件。在 SVM 中，原问题显然满足上述要求，那么我们可以得到原问题的充要条件（KKT 条件）即：

$$\left\{ \begin{array}{l} \frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \lambda_i Y_i = 0 \\ \frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^n \lambda_i Y_i X_i \\ \lambda_i \geq 0 \quad i = 1, \dots, n \\ Y_i (w^T X_i + b) - 1 \geq 0 \quad i = 1, \dots, n \\ \lambda_i (Y_i (w^T X_i + b) - 1) = 0 \quad i = 1, \dots, n \end{array} \right.$$

从而求解原问题等价于求解上述 KKT 条件，其中  $\lambda_i$  是拉格朗日乘数。

分析 KKT 条件我们还能发现更多关于样本的性质。对任意训练样本  $(X_i, Y_i)$  总有  $\lambda_i = 0$  或  $Y_i (w^T X_i + b) - 1 = 0$ 。若  $\lambda_i = 0$ ，则所对应的样本点不会出现在系数  $w$  中，即该样本点不影响最终模型；若  $\lambda_i > 0$ ，则必有  $Y_i (w^T X_i + b) - 1 = 0$ ，则所对应的样本点位于最大间隔边界上，是一个支持向量。这显示出支持向量机的一个重要性质：训练完成后，大部分的训练样本都不需要保留，最终模型仅与支持向量有关。

将对偶问题的最优解  $\lambda^* = (\lambda_1^*, \lambda_2^*, \dots, \lambda_n^*)$  代入上述 KKT 条件等式，可以得到  $w^*$ ，

$$w^* = \sum_{i=1}^n \lambda_i^* Y_i X_i \quad (4.2.15)$$

为了得到最优决策平面，还需求解  $b^*$ 。注意到对任意支持向量  $(X_k, Y_k)$ ，都有  $\lambda_k^* > 0$  且  $Y_k (w^{*T} X_k + b) - 1 = 0$ ；对任意非支持向量  $(X_l, Y_l)$ ，都有  $\lambda_l^* = 0$ 。

那么此时我们可以得到：

$$w^* = \sum_{k \in K} \lambda_k^* Y_k X_k \quad (4.2.16)$$

$K = \{i \mid \lambda_i^* > 0, i = 1, \dots, n\}$  为所有支持向量的下标集。

将  $w^*$  带入  $Y_s (w^{*T} X_s + b) - 1 = 0$ ，其中  $(X_s, Y_s)$  是一支持向量，进而可以得到：

$$Y_s \left( \sum_{k \in K} \lambda_k^* Y_k X_k^T X_s + b \right) = 1$$

由于  $Y_s^2 = 1$ , 等式两边同时乘以  $Y_s$  可以解得  $b^*$ 。理论上可选取任意支持向量通过上式解得  $b^*$ , 但在现实任务中, 常采用取所有支持向量的平均值的做法, 即:

$$b^* = \frac{1}{|K|} \sum_{s \in K} \left( Y_s - \sum_{k \in K} \lambda_k^* Y_k \mathbf{X}_k^T \mathbf{X}_s \right)$$

最终得到最优决策平面:

$$\mathbf{w}^{*T} \mathbf{X} + b^* = \sum_{i \in K} \lambda_i^* Y_i \mathbf{X}_i^T \mathbf{X} + \frac{1}{|K|} \sum_{s \in K} \left( Y_s - \sum_{k \in K} \lambda_k^* Y_k \mathbf{X}_k^T \mathbf{X}_s \right).$$

观察上述式子可以发现, 最优决策平面并不需要用所有的样本来计算, 而只需要用到支持向量。此外, 最优决策平面与  $\mathbf{X}_i^T \mathbf{X}$  有关, 这为将数据映射到高维空间, 引入核函数做好了铺垫。

### 4.2.3 软间隔与线性 SVM

在实际问题当中常存在一些“噪声点”(如图4.3所示), 采用线性可分SVM解决此类问题的过程中试图把这些“噪声点”也正确分类, 往往影响模型的泛化能力, 所以应当对模型放松要求, 允许其在分类时出一点错误, 这样会使模型的鲁棒性更好, 这也是线性SVM的基本出发点: 即划分超平面虽然不能完全分开所有样本, 但是可以使绝大多数样本正确被分类, 其目标是软间隔最大化。

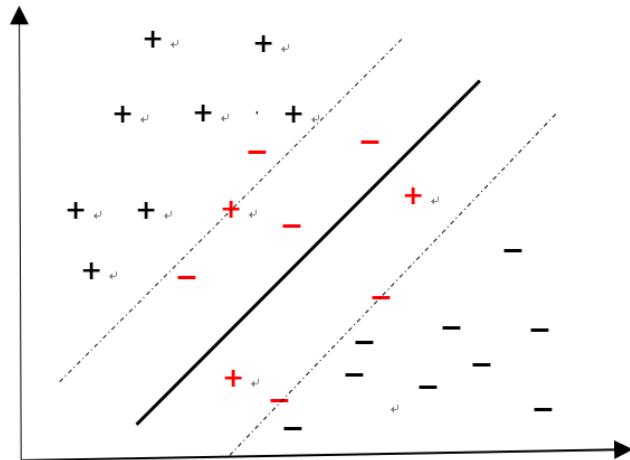


图 4.3 分类问题中的“噪声点”

假设划分超平面为：

$$f(\mathbf{X}) = w_1 X_1 + w_2 X_2 + \cdots + w_p X_p + b = \mathbf{w}^T \mathbf{X} + b = 0 \quad (4.2.17)$$

由于存在样本点不能满足式 (4.2.7) 中的约束条件，使得该约束不成立，因此在线性 SVM 中，对每一个样本引入松弛因子  $\zeta_i$ ，此时约束条件被放宽为：

$$Y_i f(\mathbf{X}_i) = Y_i (\mathbf{w}^T \mathbf{X}_i + b) \geq 1 - \zeta_i, \quad (i = 1, \dots, n)$$

上述约束条件说明当我们允许样本位于间隔区域内，也允许出现错误分类。为了使在最大化间隔的同时不满足约束的点尽可能的少，线性 SVM 优化问题可写成如下形式：

$$\begin{aligned} & \min_{\mathbf{w}, b, \zeta} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \zeta_i \\ & \text{s.t. } Y_i (\mathbf{w}^T \mathbf{X}_i + b) \geq 1 - \zeta_i, \quad (i = 1, \dots, n) \\ & \quad \zeta_i \geq 0, \quad (i = 1, \dots, n) \end{aligned} \quad (4.2.18)$$

其中  $C$  是可调节控制参数，在最小化目标函数处加了  $C \sum_{i=1}^n \zeta_i$ ，相当于施加了一个惩罚项。当  $C$  值越大时，对  $\zeta_i$  惩罚越大；反之， $C$  值越小，对  $\zeta_i$  惩罚越小。

最优化问题 (4.2.18) 是一个凸二次规划问题，可以通过拉格朗日乘子法进行求解，拉格朗日函数为：

$$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\|\mathbf{w}\|^2}{2} + C \sum_{i=1}^n \zeta_i - \sum_{i=1}^n \alpha_i (Y_i (\mathbf{w}^T \mathbf{X}_i + b) - 1 + \zeta_i) - \sum_{i=1}^n \beta_i \zeta_i \quad (4.2.19)$$

其中  $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)^T$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_n)^T$ 。分别对  $\mathbf{w}, b, \zeta_i$  求导数，并令值为零，可以得到：

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^n \alpha_i Y_i \mathbf{X}_i \\ \sum_{i=1}^n \alpha_i Y_i &= 0 \\ C - \alpha_i - \beta_i &= 0 \end{aligned} \quad (4.2.20)$$

将上面结果代入拉格朗日函数 (4.2.19)，可以得到如下对偶问题：

$$\begin{aligned} & \max_{\boldsymbol{\alpha}} \quad -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j Y_i Y_j \mathbf{X}_i^T \mathbf{X}_j + \sum_{i=1}^n \alpha_i \\ & \text{s.t. } \sum_{i=1}^n \alpha_i Y_i = 0 \\ & \quad 0 \leq \alpha_i \leq C, \quad (i = 1, \dots, n) \end{aligned} \quad (4.2.21)$$

对偶问题 (4.2.21) 等价于

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j Y_i Y_j \mathbf{X}_i^T \mathbf{X}_j - \sum_{i=1}^n \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i Y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad (i = 1, \dots, n) \end{aligned} \quad (4.2.22)$$

求解优化问题 (4.2.22)，得到最优解  $\boldsymbol{\alpha}^* = (\alpha_1^*, \alpha_2^*, \dots, \alpha_n^*)^T$ ，代入式 (4.2.20)，即可得到  $\mathbf{w}^*$ 。由于原始问题是凸二次规划问题，其解满足 KKT 条件，即

$$\beta_i \zeta_i = 0, \quad (i = 1, \dots, n) \quad (4.2.23)$$

$$\alpha_i (Y_i (\mathbf{w}^T \mathbf{X}_i + b) - 1 + \zeta_i) = 0, \quad (i = 1, \dots, n) \quad (4.2.24)$$

在最优解  $\boldsymbol{\alpha}^*$  中，取分量  $\alpha_k^*$  满足  $0 < \alpha_k^* < C$ ，由式  $C - \alpha_k^* - \beta_k^* = 0$  得  $0 < \beta_k^* < C$ ，再由式 (4.2.23) 和 (4.2.24) 易计算得到  $b^*$ ，进而得到划分超平面。

由  $\mathbf{w}$  的表达式可知，当  $\alpha_i^* = 0$  时，该样本不会对最终的模型产生影响，当  $0 < \alpha_i^* \leq C$  时，该样本是支持向量，包括以下几种类型：

- (1)  $0 < \alpha_i^* < C$ ，则  $\beta_i > 0$ ，进而由式 (4.2.23) 可以得到  $\zeta_i = 0$ ，此时  $\mathbf{X}_i$  位于最大间隔边界上；
- (2)  $\alpha_i^* = C$ ， $0 < \zeta_i < 1$ ，由式 (4.2.24) 知， $Y_i (\mathbf{w}^{*T} \mathbf{X}_i + b^*) = 1 - \zeta_i > 0$ 。那么，此时  $\mathbf{X}_i$  分类正确，并且位于划分超平面和间隔边界之间；
- (3)  $\alpha_i^* = C$ ， $\zeta_i = 1$ ，由式 (4.2.24) 知， $Y_i (\mathbf{w}^{*T} \mathbf{X}_i + b^*) = 1 - \zeta_i = 0$ 。此时  $\mathbf{X}_i$  位于划分超平面上；
- (4)  $\alpha_i^* = C$ ， $\zeta_i > 1$ ，由式 (4.2.24) 知， $Y_i (\mathbf{w}^{*T} \mathbf{X}_i + b^*) = 1 - \zeta_i < 0$ 。此时  $\mathbf{X}_i$  被分类错误。

#### 4.2.4 核函数与非线性 SVM

一般地，如果包含两个类别的样本集之间存在线性边界，那么建立线性 SVM 可以得到较好的分类效果。但是一些情况下，如果问题本身不是线性可分的，即边界是非线性的，那么线性 SVM 往往效果不佳，此时需建立非线性 SVM 对样本数据进行非线性分类。

首先看一个简单的例子，如图4.4（左）所示，一维空间中的样本点属于两个类别，分别用红色和蓝色表示不同类别。在这种情况下，无法用一个点（即一维空间的

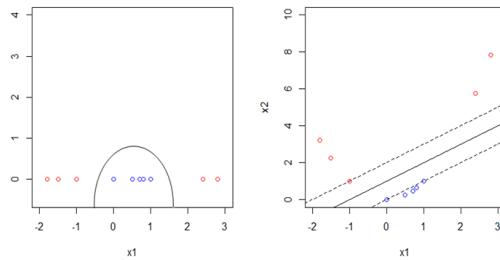


图 4.4 非线性 SVM 图例

超平面) 来将不同类别的样本分开, 但是可用一条复杂的曲线将它们分为两个类别。显然, 分类边界不是线性的。

为了解决上述非线性可分问题, 尝试将自变量的二次项添加到超平面中, 即此时的划分面是:

$$\{X : f(X) = w_1 X + w_2 X^2 + b = 0\} \quad (4.2.25)$$

式 (4.2.25) 中, 可以将  $X$  看作一个变量  $X_1$ ,  $X^2$  看作另一个变量  $X_2$ , 此时变成二维空间的线性问题。如图4.4 (右), 在构造出的二维空间中, 样本点可以用一个线性超平面 (即图中的黑色实线) 分为两个类别。因此, 在二维空间中, 线性 SVM 是有效的。同理, 可将上述问题扩展到  $p$  维空间中, 此外, 还可使用不同类型的多项式, 如三次、四次甚至是更高阶, 以及交叉项构造特征空间, 进而在新构造的特征空间中建立线性超平面。

总的来说, 对于线性不可分的数据 (如图 4.5), 可对样本进行变换, 将其映射到高维特征空间, 使得样本集在高维空间中线性可分, 如图 4.6 所示。

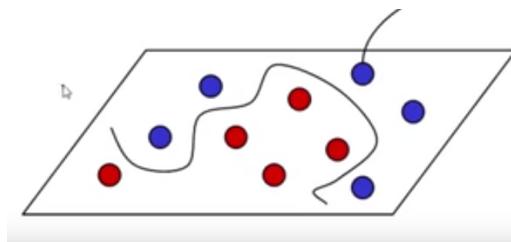


图 4.5 线性不可分的数据

如果可以找到合适的特征空间, 便可以将原问题通过特征空间中的线性超平面进行划分。然而实际问题中, 一方面, 构造合适的特征空间是非常困难的; 另一方面, 特征空间的构造方法可能并不唯一, 如果处理不当, 将会得到维数较大的特征空间, 此时的计算量将变得复杂。因此有必要寻找合适的构造特征空间的方法, 使得在新的特征空间中能有效求解得到线性超平面。

**核方法**通过**核函数** (kernel) 构造特征空间, 使得在新的特征空间中能有效求解线性超平面。在介绍核函数之前, 有必要先对内积的概念进行介绍。

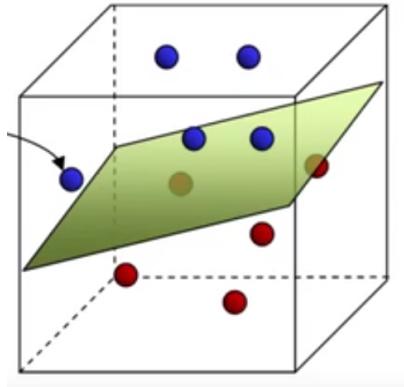


图 4.6 高维空间线性可分

两个样本  $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$  和  $\mathbf{X}_k = (X_{k1}, \dots, X_{kp})^T$  的内积定义为:

$$\langle \mathbf{X}_i, \mathbf{X}_k \rangle = \sum_{j=1}^p X_{ij} X_{kj} \quad (4.2.26)$$

通过前几小节的证明可以发现, 线性支持向量机的对偶问题可以描述成内积的形式, 即:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j Y_i Y_j \langle \mathbf{X}_i, \mathbf{X}_j \rangle - \sum_{i=1}^n \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i Y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad (i = 1, \dots, n). \end{aligned} \quad (4.2.27)$$

通过计算可得到对偶问题的拉格朗日乘子  $\alpha_i$ , 根据 KKT 条件, 可以计算最优决策平面且并将其描述为内积的形式:

$$f(\mathbf{X}) = \sum_{i=1}^n \alpha_i Y_i \langle \mathbf{X}, \mathbf{X}_i \rangle + b \quad (4.2.28)$$

上式中有  $n$  个参数  $\alpha_i, i = 1, \dots, n$ , 每个训练样本对应一个参数。为了得到  $f(\mathbf{X})$ , 需要计算  $\mathbf{X}$  与每个训练样本  $\mathbf{X}_i$  之间的内积, 但事实证明, 有且仅有支持向量所对应的  $\alpha_i$  是非零的。若用  $S$  表示支持向量样本点指标的集合, 那么式 (4.2.28) 可以改写成:

$$f(\mathbf{X}) = \sum_{i \in S} \alpha_i Y_i \langle \mathbf{X}, \mathbf{X}_i \rangle + b \quad (4.2.29)$$

式 (4.2.29) 的求和项比式 (4.2.28) 少得多。总而言之, 我们仅需知道内积便可以计算  $f(\mathbf{X})$ 。

对于非线性可分问题，我们的初衷是想找到这样一个映射  $\psi$ ，将样本点  $\mathbf{X}_i$  映射为新的高维特征空间的特征向量  $\psi(\mathbf{X}_i)$ ，进而通过其对偶问题求解相应拉格朗日乘子。最后，考虑原问题的 KKT 条件，求解最优决策平面。

在上述步骤中，可以发现在原空间内，对偶问题及原问题的 KKT 条件求解以内积的形式出现。自然的，在新的特征空间内，只需知道  $\langle \psi(\mathbf{X}_i), \psi(\mathbf{X}_j) \rangle$  便可以在新的特征空间中求解原问题的对偶问题，然后得到最优决策平面。高维空间中内积的计算较为复杂，可通过核技巧直接定义核函数  $K(\mathbf{X}_i, \mathbf{X}_j) = \langle \psi(\mathbf{X}_i), \psi(\mathbf{X}_j) \rangle$  来解决复杂计算问题，通俗的来说，就是将求解映射  $\psi$  的问题转化为选择合适核函数  $K(\mathbf{X}_i, \mathbf{X}_j)$  的问题。

对于非线性可分问题，考虑核技巧后，此时  $f(\mathbf{X})$  变为：

$$f(\mathbf{X}) = \sum_{i \in S} \alpha_i Y_i K(\mathbf{X}, \mathbf{X}_i) + b \quad (4.2.30)$$

其中  $K(\mathbf{X}_i, \mathbf{X}_j)$  被称为核函数。

非线性 SVM 的处理方法是构造核函数以代替高维空间中的内积计算，在高维特征空间中解决原始空间中线性不可分的问题。具体来说，在线性不可分的情况下，首先选择一个合适的核函数，核函数的作用是避免在高维空间中进行内积计算，然后在高维空间中执行线性可分的 SVM，最终计算出最优的划分超平面，从而达到把低维空间中线性不可分的数据集进行分类的目的。

核函数的选择至关重要，将影响 SVM 对数据集的分类效果。实际问题中，应当根据问题本身特征选择合适的核函数。表 4.1 给出了常用的几种核函数：

表 4.1 几种常用核函数

名称	表达式	参数
线性核	$K(\mathbf{X}_i, \mathbf{X}_j) = \mathbf{X}_i^T \mathbf{X}_j$	
多项式核	$K(\mathbf{X}_i, \mathbf{X}_j) = (\mathbf{X}_i^T \mathbf{X}_j)^d$	$d \geq 0$ 为多项式的次数
高斯核	$K(\mathbf{X}_i, \mathbf{X}_j) = \exp\left(-\frac{\ \mathbf{X}_i - \mathbf{X}_j\ ^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯的带宽 (width)
拉普拉斯核	$K(\mathbf{X}_i, \mathbf{X}_j) = \exp\left(-\frac{\ \mathbf{X}_i - \mathbf{X}_j\ }{\sigma}\right)$	$\sigma > 0$
Sigmoid 核	$K(\mathbf{X}_i, \mathbf{X}_j) = \tanh(\beta \mathbf{X}_i^T \mathbf{X}_j + \theta)$	$\tanh$ 为双曲正切函数, $\beta > 0, \theta > 0$

注意，关于核函数的选择一直以来都是支持向量机研究的热点，通常情况下，高斯核函数是使用最多的。此外，采用核函数而不是直接将样本集映射到特征空间的优势在于，无需明确指明映射函数，并且可以避免计算高维空间中的内积，大大降低计算量。

## 4.3 SVM 与 logistic 回归的关系

本节讨论 SVM 与 Logistic 回归的关系。首先，为了建立支持向量分类器

$$f(\mathbf{X}) = b + w_1 X_1 + \cdots + w_p X_p$$

对于最优化问题 (4.2.18)，可以将它改写成下面形式：

$$\min_{\mathbf{w}, b} \left\{ \frac{1}{n} \sum_{i=1}^n \max [0, 1 - Y_i f(\mathbf{X}_i)] + \gamma \sum_{j=1}^p w_j^2 \right\} \quad (4.3.1)$$

其中  $\gamma$  为调节控制参数， $\gamma \sum_{j=1}^p w_j^2$  是岭回归的惩罚项，这一项需要根据偏差的关系来确定。式 (4.3.1) 本质是如下的“损失函数 + 惩罚”的形式：

$$\min_{\mathbf{w}, b} \{L(\mathbf{w}, b | \mathbf{X}, \mathbf{Y}) + \gamma P(\mathbf{w})\}$$

这里，把具有形式

$$L(\mathbf{w}, b | \mathbf{X}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n \max [0, 1 - Y_i f(\mathbf{X}_i)]$$

的损失函数称为 hinge 损失。这类损失函数的特点是，对于边界外且分类正确的样本点，损失为零；对于边界上的样本点以及分类错误的样本点，损失是线性的。hinge 损失与 Logistic 回归使用的损失函数非常接近，只不过 Logistic 回归的损失函数在任何时候都是非零的，所以通常来说，SVM 和 Logistic 回归的结果也是非常接近的。

对于一个给定的问题，该如何选择是使用 SVM 还是 Logistic 回归呢？这个问题更多时候需要根据实际数据选择合适的算法，当然也有一些特殊情形：(1) 当类别区分度较高时，可以选择 SVM；(2) 如果想要得到估计的概率，那么就需要选择 Logistic 回归；(3) 对于决策边界是非线性的情况，核函数的 SVM 方法应用更加广泛。

## 4.4 支持向量回归

本节将介绍如何将支持向量机扩展到回归问题中，称为支持向量回归 (Support Vector Regression, SVR)，其思想与分类问题类似，不同之处在于支持向量回归的目的是要寻找一个超平面，在距离超平面  $\epsilon$  范围内尽可能的包含最多的样本点。

可将上述过程用数学公式来定义，考虑如下回归模型：

$$f(\mathbf{X}) = \sum_{j=1}^p w_j X_j + b \quad (4.4.1)$$

则支持向量回归超平面是如下最优化问题的解：

$$\min_{\mathbf{w}, b} H(\mathbf{w}, b \mid \mathbf{X}, \mathbf{Y}) = \min_{\mathbf{w}, b} \left\{ \frac{1}{n} \sum_{i=1}^n V_\epsilon(Y_i - f(\mathbf{X}_i)) + \lambda \sum_{j=1}^p w_j^2 \right\} \quad (4.4.2)$$

这里同样采用了“损失函数 + 惩罚”的形式，其中

$$V_\epsilon(r) = \begin{cases} 0, & \text{若 } |r| < \epsilon \\ |r| - \epsilon, & \text{其他} \end{cases}$$

称  $\epsilon$  为**不敏感损失**。当样本点与超平面的距离小于  $\epsilon$  时，损失定义为 0；反之，当样本点与超平面的距离大于等于  $\epsilon$  时，损失定义为线性的形式。因此，只有距离大于等于  $\epsilon$  的样本才会影响超平面的确定，与分类问题类似，称这些样本为**支持向量**。

同样地，与非线性回归模型一样，对  $f(\mathbf{X})$  我们可以采用非线性函数形式建立非线性支持向量回归，例如可以采用多项式、样条或者核方法进行逼近。

支持向量回归可视为是对普通线性回归的一种推广，相比于处理线性回归的最小二乘法而言，优势在于其避开了对回归中因变量分布的假设，也不再局限于线性模型。

## 4.5 SVM 实践

### 4.5.1 R 语言实践

R 语言中的 e1071 包可以实现 SVM 算法，算法实现流程如下：

```
library(e1071)
data(iris) # 加载 iris 数据集
attach(iris)
set.seed(1)
# 将数据集分为两部分，训练集占 2/3，测试集占 1/3
index <- sample(1:nrow(iris), round(nrow(iris)*2/3))
train_data <- iris[index,]
test_data <- iris[-index,]
train_data_x <- train_data[, 1:4]
train_data_y <- train_data[, 5]
test_data_x <- test_data[, 1:4]
test_data_y <- test_data[, 5]
# 构建 SVM 模型，type 为 C-classification，核函数为 radial
svmmodel <- svm(train_data_x, train_data_y, type='C-classification', kernel = 'radial')
# 检验模型在训练集上的拟合效果
train_data_pred <- predict(object=svmmodel, train_data_x)
train_data_result <- table(train_data_pred, train_data_y)
train_data_result
```

```

train_data_accuracy<-sum(diag(train_data_result))/sum(train_data_result)
train_data_accuracy
#评估模型的预测性能
test_data_pred<-predict(object=svmmmodel,test_data_x)
test_data_result<-table(test_data_pred,test_data_y)
test_data_result
test_data_accuracy<-sum(diag(test_data_result))/sum(test_data_result)
test_data_accuracy

```

### 4.5.2 Python 语言实践

本节将基于莺尾花数据集，采用 sklearn.svm 中的 SVC 函数训练一个 SVM 模型，算法实现流程如下：

```

from sklearn import datasets
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split
from sklearn import svm
from sklearn.metrics import classification_report
iris = datasets.load_iris()#导入样本集
x = iris.data[:,0:4]#截取前四列作为特征
y = iris.target#取出标签
#调用MinMaxScaler函数将数据集映射到0-1范围内，避免特征取值相差较大影响模型
ScalerModel = MinMaxScaler(feature_range=(0,1))
x = ScalerModel.fit_transform(x)
print(x)
#调用train_test_split函数将原数据集划分为训练集和测试集，训练集占2/3，测试集占1/3
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=1/3,
random_state=0, stratify=y)
#构建非线性SVM模型，核函数选择高斯核，参数C表示正则化参数，gamma为核函数的参数
model = svm.SVC(C=0.8, kernel='rbf', gamma=20, decision_function_shape='ovr')
model.fit(x_train, y_train)#拟合样本数据集
print('accuracy', model.score(x_train, y_train))#输出拟合评分
y_predict = model.predict(x_test)#将模型作用于测试集
ResultReport = classification_report(y_test, y_predict)
print(ResultReport)#输出评估结果

```

## 4.6 总结

支持向量机一直是机器学习领域的研究热点，本章分别介绍了线性可分 SVM、线性 SVM、非线性 SVM 处理分类问题以及支持向量回归处理回归问题的相关理论与方法。支持向量机以间隔为基本出发点，通常采用凸优化理论求解，借助 SMO 算

法实现快速计算，利用核函数解决线性不可分问题，使得该方法较为广泛地适用于实际生产。

除本章介绍的几种基本方法之外，还有较多的扩展模型，如基于粒度划分方法的粒度支持向量机（Granular Support Vector Machines, GSVM）[59,60]；结合模糊数学与支持向量机的模糊支持向量机（Fuzzy Support Vector Machines, FSVM）[61]，此类方法一定程度上可以克服噪声点对算法的影响；基于排序学习的排序支持向量机（Ranking Support Vector machines, RSVM）[62]；孪生支持向量机（Twin Support Vector Machines, TWSVM）[63–65]，该方法提升了经典 SVM 的泛化性能以及计算效率，对大规模数据集具有较好的处理效果。

支持向量机相关算法库可参考：LIBSVM、LIBLINEAR、e1071 等。

## 4.7 习题

- 1、请简要分析优化问题（4.2.18）转化成（4.3.1）中定义的损失函数的思路。
- 2、支持向量回归使用的目标函数是如何得到的？请给出简要分析思路。
- 3、假设样本集  $\mathbf{X}$  线性可分，采用线性可分 SVM 得到划分超平面  $\mathbf{w}^T \mathbf{X} + b = 0$ ，其中  $\mathbf{w} = (2, 1, -1)^T$ ,  $b = 2$ 。试求出间隔并判断下面三个样本是否为支持向量。
  - (1)  $\mathbf{X}_1 = (3, 1, -6)^T$
  - (2)  $\mathbf{X}_2 = (2, -2, 1)^T$
  - (3)  $\mathbf{X}_3 = (-2, 3, 2)^T$
- 4、对于同一样本集  $\mathbf{X}$ ，通过训练得到 SVM 和 Logistic 回归两个分类器。此时，若在训练集中新加入一个远离 SVM 决策边界的样本点，重新训练得到的两个新分类器是否发生变化？请用数学语言解释原因。
- 5、自选 UCI 数据集 (<https://archive.ics.uci.edu/ml/datasets.php>)，采用不同的核函数训练 SVM，对得到的模型以及实验结果进行对比。
- 6、自选 UCI 数据集 (<https://archive.ics.uci.edu/ml/datasets.php>)，训练一个支持向量回归模型。

# 第五章 决策树

## 5.1 简介

决策树方法最早产生于 20 世纪 60 年代，其中 CART 算法是决策树最经典和最主要的算法。[CART 算法](#) (Classification and Regression Tree) 是 Breiman 等 [67] 在 1984 年提出来的一种非参数方法，它既可以用于解决分类问题（预测定性变量，或者说当因变量是离散变量时），又可以用于回归问题（预测定量变量，或者说当因变量是连续变量时），分别称为分类树 (Classification tree) 和回归树 (Regression tree)。CART 算法的基本思想是一种二分递归分割方法，在计算过程中充分利用二叉树，在一定的分割规则下将当前样本集分割为两个子样本集，使得生成的决策树的每个非叶子节点都有两个分裂，这个过程又在子样本集上重复进行，直至无法再分成叶子节点为止。在本章中，我们介绍的内容主要包括：决策树的基本概念、回归树和分类树的建模过程、防止决策树过拟合的方法及实施决策树相关的 Python 和 R 语言程序。

## 5.2 决策树的基本原理

[决策树](#)(decision tree) 是一种常见的机器学习方法，决策树模型呈树形结构，采用自顶向下递归的方法。决策树包含三种节点，分别是：根节点、内部节点和叶子节点。刚开始构建模型时，全部样本组成的节点，没有入边，只有出边，称为[根节点](#) (root node)；不再继续分裂的节点称为树的[叶子节点](#) (leaf node)，没有出边，只有入边，叶子节点的个数决定了决策树的规模和复杂程度；根节点和叶子节点之外的节点都称作[内部节点](#) (internal node)，内部节点既有入边，又有出边。

决策树模型结构如图5.1所示，一棵决策树一般包含一个根节点（黄色圆形图标）、若干个内部节点（蓝色圆形图标）和若干个叶子节点（方形图标）。叶子节点对应于决策结果，每个内部节点对应于一个自变量，每个内部节点包含的样本集合根据自变量测试的结果被划分到它的子节点中，根节点包含全部训练样本。根节点在一定的分割规则下被分割成两个子节点，这个过程在子节点上重复进行，直至无法再分为叶子节点为止。从根节点到每个叶子节点的路径对应了一个判定测试序列。

决策树算法遵循自顶向下、分而治之的策略，**具体步骤**为：

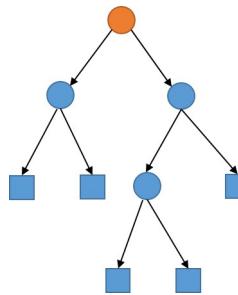


图 5.1 决策树模型结构

- (1) 选择最好的自变量作为测试自变量并创建树的根节点；
- (2) 为测试自变量每个可能的取值产生一个分支；
- (3) 将训练样本划分到适当的分支形成子节点；
- (4) 对每个子节点，重复上面的过程，直到所有的节点都是叶子节点。

若考虑自变量只有二维的情况，将决策树分类过程在二维的坐标轴中画出，如图5.2所示。则可以发现，决策树实际上就是对自变量空间的划分，且划分区域的数目就是叶节点的数目，如图5.2中的红线和绿线。

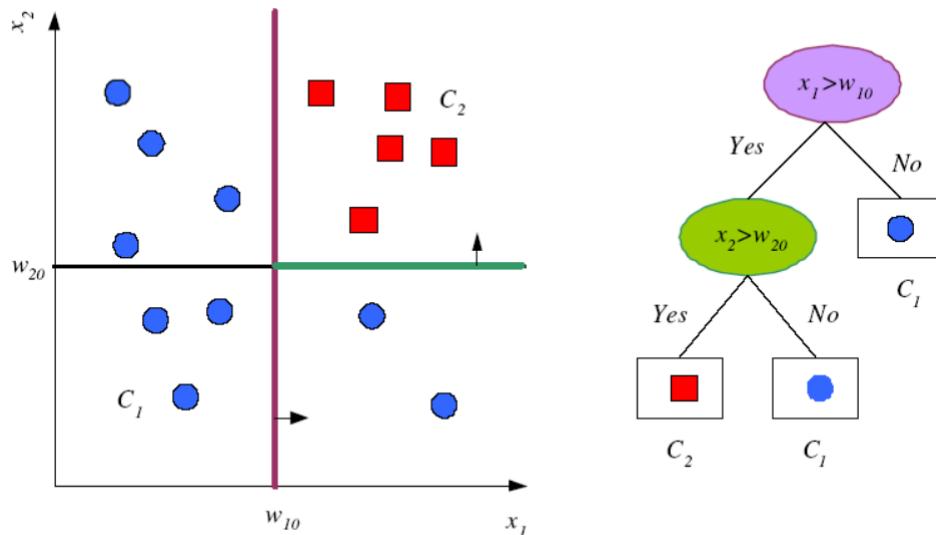


图 5.2 二维坐标轴中的决策树分类过程

决策树的建立过程可以概括为以下两个关键环节：

- (1) 将自变量空间（即  $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$  的所有可能取值构成的集合）分割成  $T$  个互不重叠的叶子节点  $R_1, \dots, R_T$ ；
- (2) 对落入节点  $R_t$  的每个观测，其预测值为  $R_t$  节点中训练集的因变量值的众

数（分类树），或者平均数（回归树）。

决策树生长是一个递归的过程，因此在算法中需要包含跳出递归的条件。直观的看，决策树停止生长的条件有以下三种情况，满足以下三个条件之一就能使决策树停止生长：

- (1) 当前节点包含的样本全部属于同一类别；
- (2) 当前自变量集为空，或所有样本在所有自变量上取值相同；
- (3) 当前节点包含的样本集合为空。

决策树算法有以下几条优点：

- (1) 计算量相对小，训练速度快；
- (2) 易理解、解释性强；
- (3) 不需要任何先验假设；
- (4) 可以处理连续变量和离散变量（性别）；
- (5) 可以处理缺失值和具有尺度不变性，即对自变量变量做一个单调变换，不改变树的生成结果。

决策树算法包括如下缺点：

- (1) 决策树算法可以创建很复杂的树，但容易导致过拟合，可以通过剪枝等手段缓解；
- (2) 决策树算法训练结果方差大、不稳定，数据很小的扰动可能得到完全不同的分裂结果，有可能是完全不同的决策树，这个缺点可以通过集成学习来解决。

## 5.3 分类树与回归树

对于决策树而言，根据预测变量的类型不同，可分为不同的决策树。若目标变量是离散的，则为分类树，若目标变量是连续的，则为回归树。

### 5.3.1 分类树

我们首先介绍分类树。当因变量为定性变量即离散型变量时，可建立分类树模型。分类树是一种特殊的分类模型，是一种直接以树的形式表征的非循环图。它的建模过程就是自动选择分裂变量，以及根据这个变量进行分裂的条件。

假设数据  $X_i = (X_{i1}, \dots, X_{ip})^T$  包含  $p$  个输入变量和一个离散型的因变量  $Y \in (1, \dots, K)$ ，样本量为  $n$ 。现在我们想把数据分成  $T$  个区域（或称为节点），第  $t$  个节点  $R_t$  的样本量为  $n_t (t = 1, \dots, T)$ 。

则建立分类树的过程可以用如下算法表示：

**1、分支。**采用递归二叉分裂法在训练集中生成一棵分类树。递归二叉分裂法是指是从树顶端开始依次分裂自变量空间，每个分裂点都产生两个新的分裂，并且每次分裂选取最优分裂方案。最优的分裂方案是使得  $T$  个节点的不纯度减少到最小，衡量节点不纯度的指标在第三节详细介绍。

**2、剪枝。**对生成的回归树进行剪枝，得到一系列最优子树，剪枝在第四节详细介绍。

**3、预测。**令  $\hat{P}_{tk} = \frac{1}{n_t} \sum_{i \in R_t} I(Y_i = k)$  表示在叶子节点  $t$  中第  $k$  类样本点的比例，则预测叶子节点  $t$  的类别为：

$$\hat{t}_m = \operatorname{argmax}_k \hat{P}_{tk}$$

即叶子节点  $t$  中类别最多的一类。

### 5.3.2 回归树

当因变量为连续型变量时，可建立回归树模型。[回归树](#)和分类树的思想类似，首先是在分支准则上有差异，即衡量节点不纯度的指标不一样，其次是预测准则上存在差异，对于分类树，将落在该叶节点的观测点的最大比例类别作为该叶节点预测值，而对于回归树，则是将落在该叶节点的观测点的平均值作为该叶节点预测值。

回归树的数据结构与分类树类似，唯一区别是回归树解决的是连续型的因变量  $Y$ ，并且回归树算法过程跟分类树类似：

**1、分支。**采用递归二叉分裂法在训练集中生成一棵回归树。最优的分裂方案是使得  $T$  个节点的不纯度减少到最小，其中，衡量节点不纯度的指标为样本均方误差。

**2、剪枝。**对生成的分类树进行剪枝，得到一系列最优子树，剪枝在第四节详细介绍。

**3、预测。**节点得到后，可以确定某一给定预测数据所属的节点，回归树用这一节点的训练集的因变量的平均值作为预测值：

$$\hat{Y}_t = \frac{1}{n_t} \sum_{i \in R_t} Y_i.$$

## 5.4 分支条件

我们现在再来讨论该如何构建节点，即对决策树进行分支。理论上，我们可以将节点所对应的区域形状作任意分割，但出于模型的简化和可解释性考虑，一般只将区域划分为高维矩形。不过，若要将自变量空间划分为矩形区域的所有可能性都进行考虑，这在计算上是不可行的。所以，我们对区域的划分一般采用一种自上而

下 (top-down)、贪婪 (greedy) 的方法: 递归二叉分裂 (recursive binary splitting)。自上而下指的是它从树顶端开始依次分裂自变量空间, 每个分裂点都产生两个新的分裂。贪婪指在建立树的每一步中, 最优分裂确定仅限于某一步进程, 而不是针对全局, 选择那些能够在未来进程中构建出更好的树的分裂点。对每一个节点重复以上过程, 寻找继续分割数据集的最优自变量和最优分裂点。此时被分割的不再是整个自变量空间, 而是之前确定的两个区域之一。这一过程不断持续, 直到符合某个分裂准则再停止, 譬如, 当叶节点包含的观测值个数低于某个最小值时, 分裂停止。

如何确定最优的分裂方案? 我们基于不纯度的减少来作为分裂准则, 即通过最小化节点不纯度来确定最优分裂变量和最优分裂点。对于分类树和回归树有不同的衡量节点不纯度的指标, 我们将在后面的分类树和回归树部分进行具体的描述。

决策树算法的关键是划分自变量的选择, 决策树自变量划分方法很多, 比较经典有: 适用于分类树的信息增益、增益率和基尼指数, 以及适用于回归树的均方误差。

### 5.4.1 信息熵

在介绍信息增益之前, 我们先引入信息量和信息熵的概念。

假设随机变量  $\mathbf{X}$  的概率分布为  $\Pr(\mathbf{x})$ , 则任意一事件  $\mathbf{X} = \mathbf{x}$  的信息量可以表示为  $h(\mathbf{x}) = -\log_2 \Pr(\mathbf{X} = \mathbf{x})$ , 其中  $\Pr(\mathbf{X} = \mathbf{x})$  表示该事件发生的概率。某事件发生的概率越小, 该事件的信息量越大。

信息熵是度量样本集合纯度最常用的一种指标, 也是一种不确定性的度量。1948年, 香农 (Shannon) [68] 在他著名的《通信的数学原理》论文中指出: “信息是用来消除随机不确定性的东西”, 并提出了“信息熵”的概念 (借用了热力学中熵的概念), 来解决信息的度量问题。

对于随机变量  $\mathbf{X}$  而言, 信息熵可表示为:  $H(\mathbf{X}) = - \sum_{\mathbf{X}_i \in \mathcal{X}} \Pr(\mathbf{X} = \mathbf{X}_i) \ln \Pr(\mathbf{X} = \mathbf{X}_i)$ , 其中  $\mathcal{X}$  表示所有可测事件的集合。特殊的, 若假设两点分布中某一点发生的概率是  $\theta$ , 则其信息熵:

$$H(\mathbf{X}) = - \sum_{\mathbf{X}_i \in \mathcal{X}} \Pr(\mathbf{X} = \mathbf{X}_i) \ln \Pr(\mathbf{X} = \mathbf{X}_i) = -\theta \ln \theta - (1-\theta) \ln(1-\theta)$$

若信息熵退化为定值则熵最小为 0; 若随机变量是连续型随机变量且取值在有限区间内 (即随机变量密度函数在有限区间内大于零), 则均匀分布的熵最大; 随机变量是连续型随机变量且取值在无限区间内, 通过拉格朗日乘子法可以推导正态分布的熵最大。

下面考虑响应变量是离散随机变量的情况介绍信息熵。即  $Y \in \{1, \dots, K\}$  的  $K$  分类问题。

对于离散响应变量集合  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$  来说，当前离散响应变量集合中第  $k$  类样本所占的比例为  $\hat{P}_k = \frac{1}{n} \sum_{i=1}^n I(Y_i = k)$ ，则信息熵的计算公式为：

$$H(\mathbf{Y}) = - \sum_{k=1}^K \hat{P}_k \cdot \log_2 \hat{P}_k \quad (5.4.1)$$

信息熵越大，意味着不确定性越大；信息熵越小，则不确定性越小。例如，对于一盏灯有两种可能的状态：开和关，假设样本集合中一半是开的样本，一半是关的样本，那么信息熵为：

$$H(\mathbf{Y}) = - \sum_{k=1}^K \hat{P}_k \cdot \log_2 \hat{P}_k = -\left(\frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{2} \cdot \log_2 \frac{1}{2}\right) = 1.$$

在此情况下，信息熵很大，说明灯的两种状态不确定性很大。

假如这盏灯出现故障，无法正常打开，那么此时信息熵为：

$$H(\mathbf{Y}) = - \sum_{k=1}^K \hat{P}_k \log_2 \hat{P}_k = -\left(\frac{1}{1} \cdot \log_2 \frac{1}{1}\right) = 0.$$

在此情况下，灯的状态是确定的，因此信息熵很小。

#### 5.4.2 信息增益

有了信息熵的概念，我们引入信息增益，信息增益是针对某一种自变量而言的，假设为  $X$ （一维）。

假设使用自变量  $X$  的样本集  $\mathbf{X} = (X_1, \dots, X_n)^T$  对分类树划分出  $M$  个区域  $R_1, \dots, R_M$ （在二叉分裂时， $M = 2$ ），其中叶子节点  $R_m$  中含有样本数为  $n_m$ ，简记为  $|R_m| = n_m$ 。假设节点  $R_m$  中类  $k$  的观测比例为： $\hat{P}_{mk} = \frac{1}{n_m} \sum_{i \in R_m} I(Y_i = k)$ 。

假设自变量是离散变量， $M$  个节点是由自变量  $X$  的  $M$  个可能的取值划分出来的。若自变量是连续变量，我们可以从连续变量取值范围上取几个点对样本进行划分，此时被划分的样本集可等价理解成离散变量样本。

接下来，计算出用离散变量样本集  $\mathbf{X}$  对样本集  $\mathbf{Y}$  进行再次划分所获得的信息增益，假设  $\mathbf{X}$  取之于集合  $\{1, \dots, M\}$ ， $\mathbf{Y}$  取之于集合  $\{1, \dots, K\}$ 。

**信息增益** (information gain) 表示自变量  $X$  使目标不确定性减少的程度。信息增益计算公式如下：

$$\Delta = H(\mathbf{Y}) - \sum_{m=1}^M \frac{n_m}{n} H(\mathbf{Y} | R_m) \quad (5.4.2)$$

其中  $H(\mathbf{Y})$  此时称为父节点的信息熵以及

$$H(\mathbf{Y} | R_m) = - \sum_{k=1}^K \hat{P}_{mk} \log_2 \hat{P}_{mk} \quad (5.4.3)$$

信息增益是自变量划分前的信息熵与自变量划分后加权信息熵的差值，即**信息增益 = 信息熵 - 条件熵**。

信息增益越大，则意味着使用自变量来进行划分所获得的“纯度提升”越大。下面以打网球实例来演示信息增益的具体计算过程，该实例是根据天气状况来决定是否打网球，影响打网球的自变量包括天气、温度、是否有风，数据集如表5.1所示，数据集中一共包括 10 个样本，构建决策树算法对是否适合打网球进行预测。首先计算划分前根节点的信息熵，根节点是否打网球两类的样本所占比例分别为  $\frac{6}{10}$  和  $\frac{4}{10}$ ，信息熵计算公式为：

$$H(\text{parent}) = - \sum_{k=1}^K \hat{P}_k \cdot \log_2 \hat{P}_k = -\left(\frac{6}{10} \cdot \log_2 \frac{6}{10} + \frac{4}{10} \cdot \log_2 \frac{4}{10}\right) = 0.971.$$

假设因变量  $Y$  “是否打球” 分别用 1 和 2 表示， $Y=1$  表示打球， $Y=2$  表示不打球。

表 5.1 打网球数据集

序号	weather	temperature	windy	play_tennis
1	晴	热	否	否
2	晴	热	是	否
3	阴	热	否	是
4	雨	温	否	是
5	雨	凉	是	否
6	雨	凉	否	是
7	阴	凉	是	是
8	晴	温	否	否
9	晴	凉	否	是
10	雨	温	否	是

在此基础上，首先计算天气的信息增益，对于天气，自变量取值包括：晴、阴、雨，分别对应节点  $R_1, R_2, R_3$ 。其中对于“晴”取值，共有 4 个样本，即  $n_1 = 4$ ，打球数量占比为  $\hat{P}_{11} = \frac{1}{n_1} \sum_{i \in R_1} I(Y_i = 1) = \frac{1}{4}$ ，不打球数量占比为  $\hat{P}_{12} = \frac{1}{n_1} \sum_{i \in R_1} I(Y_i = 2) = \frac{3}{4}$ ，即天气的其他取值采用同样的计算方式。首先根据公式(5.4.3)，用 1、2、3 分别代表

晴天、阴天和雨天，分别计算不同自变量取值的信息熵：

$$\begin{aligned} H(\mathbf{Y} | R_1) &= -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) = 0.811, \\ H(\mathbf{Y} | R_2) &= -\left(\frac{2}{2}\log_2\frac{2}{2}\right) = 0, \\ H(\mathbf{Y} | R_3) &= -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) = 0.811. \end{aligned}$$

信息增益：

$$\begin{aligned} \Delta_{\text{weather}} &= H(\text{parent}) - \sum_{m=1}^M \frac{n_m}{n} H(\mathbf{Y} | R_m), \\ &= 0.971 - ((4/10) * 0.811 + (2/10) * 0 + (4/10) * 0.811) = 0.971 - 0.649 = 0.322. \end{aligned}$$

同样可以求出温度和是否有风两个自变量的信息增益：

$$\begin{aligned} \Delta_{\text{temperature}} &= 0.971 - ((3/10) * 0.918 + (4/10) * 0.918 + (3/10) * 1) = 0.971 - 0.95 = 0.02 \\ \Delta_{\text{windy}} &= 0.971 - ((3/10) * 0.918 + (7/10) * 0.863) = 0.971 - 0.879 = 0.09. \end{aligned}$$

由信息增益的计算结果可以看出，天气的信息增益最大，因此首先选择天气作为划分自变量。

### 5.4.3 增益率

不同的离散自变量的离散值的取值个数是不同的，即自变量不同  $M$  不同。如果选取的某一自变量取值个数较多，即  $M$  较大，那么这一划分的信息增益越大。特别的，对于某一自变量，如果该自变量的每个取值内只含一个样本点，那么选取该自变量划分后的加权信息熵是 0，进而此划分的信息增益最大。因此，在以信息增益作为划分训练数据集的特征时，我们会偏向于选择取值较多的自变量。故而，我们引入信息增益率是为了避免自变量值个数对信息增益的影响。

**信息增益率** (information gain ratio) 在信息增益的基础上增加了惩罚项，惩罚项是特征的固有值，是避免上述情况而设计的。信息增益率的定义如下：

$$\text{GainR} = \frac{\Delta}{\text{IV}},$$

其中

$$\text{IV} = -\sum_{m=1}^M \frac{n_m}{n} \log_2 \frac{n_m}{n}$$

一般的话，某个自变量的取值数目越多，则 IV 值越大。

以上面打网球的例子，计算各自变量增益率。首先计算各自变量的 IV 值：

$$\begin{aligned} \text{IV}_{\text{weather}} &= -\frac{4}{10} * \log_2 \frac{4}{10} - \frac{2}{10} * \log_2 \frac{2}{10} - \frac{4}{10} * \log_2 \frac{4}{10} = 1.52, \\ \text{IV}_{\text{temperature}} &= -\frac{3}{10} * \log_2 \frac{3}{10} - \frac{3}{10} * \log_2 \frac{3}{10} - \frac{4}{10} * \log_2 \frac{4}{10} = 1.57, \\ \text{IV}_{\text{windy}} &= -\frac{3}{10} * \log_2 \frac{3}{10} - \frac{7}{10} * \log_2 \frac{7}{10} = 0.88. \end{aligned}$$

得到三个自变量的 IV 值，由结果可知，天气和温度的取值个数比是否有风多，则天气和温度的 IV 值高。接下来计算各个自变量的信息增益率：

$$\begin{aligned} \text{GainR}_{\text{weather}} &= \frac{\Delta_{\text{weather}}}{\text{IV}_{\text{weather}}} = \frac{0.322}{1.52} = 0.211 \\ \text{GainR}_{\text{temperature}} &= \frac{\Delta_{\text{temperature}}}{\text{IV}_{\text{temperature}}} = \frac{0.02}{1.57} = 0.012 \\ \text{GainR}_{\text{windy}} &= \frac{\Delta_{\text{windy}}}{\text{IV}_{\text{windy}}} = \frac{0.09}{0.88} = 0.103 \end{aligned}$$

求得三个自变量的增益率，发现天气的增益率较高，因此首先选择天气作为划分自变量。

#### 5.4.4 基尼指数

对于分类因变量样本集合  $\mathbf{Y}$  来说，假定当前样本集合  $\mathbf{Y}$  中第  $k$  类样本所占的比例为  $\hat{P}_k(k = 1, \dots, K)$ ，数据集的纯度可用基尼值（Gini）来度量：

$$\text{Gini}(\mathbf{Y}) = 1 - \sum_{k=1}^K \hat{P}_k^2. \quad (5.4.4)$$

$\text{Gini}(\mathbf{Y})$  反映了从数据集  $\mathbf{Y}$  中随机抽取两个样本，其类别标记不一致的概率。 $\text{Gini}(\mathbf{Y})$  越小，则数据集  $\mathbf{Y}$  的纯度越高。基于某个离散自变量划分出的第  $m$  区域上的基尼值表示为  $\text{Gini}(\mathbf{Y} | R_m) = 1 - \sum_{k=1}^K \hat{P}_{mk}^2$ 。

基于分类因变量的基尼指数（Gini Index）定义为：

$$\text{GiniI} = \sum_{m=1}^M \frac{n_m}{n} \text{Gini}(\mathbf{Y} | R_m) \quad (5.4.5)$$

基尼指数越小，则不纯度越低，自变量越好。

### 5.4.5 分类误差

令节点  $R_m$  中类  $k$  的观测比例为  $\hat{P}_{mk} = \frac{1}{n_m} \sum_{i \in R_m} I(Y_i = k)$ , 并且我们得到节点  $R_m$  中的所有样本的预测类别为:

$$\hat{k}_m = \operatorname{argmax}_k \hat{P}_{mk}$$

它是节点  $R_m$  上样本数最多的类。节点  $R_m$  上的分类误差表示为:

$$\text{CE}(\mathbf{Y} | R_m) = \frac{1}{n_m} \sum_{i \in R_m} I(Y_i \neq \hat{k}_m) = 1 - \hat{P}_{m\hat{k}_m}$$

最终也可以采用加权的思想, 得到基于  $M$  个节点的分类误差 (CE):

$$\text{CE} = \sum_{m=1}^M \frac{n_m}{n} \text{CE}(\mathbf{Y} | R_m) \quad (5.4.6)$$

### 5.4.6 均方误差

如果分裂变量  $X_j$  是连续的自变量以及因变量  $Y$  也是连续变量, 我们使用分裂点  $t$  去分裂自变量  $X_j$  的样本集  $\mathbf{X}_j = (X_{1j}, \dots, X_{nj})^\top$ 。若考虑二分裂, 使用分裂变量  $X_j$  和  $t$  可以定义两个子区域:

$$R_1(j, t) = \{i \mid X_{ij} < t\} \quad \text{和} \quad R_2(j, t) = \{i \mid X_{ij} \geq t\} \quad (5.4.7)$$

其中  $X_{ij}$  表示样本点  $\mathbf{X}_i$  的第  $j$  个特征。

对应分裂变量  $X_j$  和分裂点  $t$  的均方误差定义为:

$$\text{MSE}(j, t) = \frac{1}{n_1} \sum_{i \in R_1(j, t)} (Y_i - \hat{c}_1)^2 + \frac{1}{n_2} \sum_{i' \in R_2(j, t)} (Y_{i'} - \hat{c}_2)^2 \quad (5.4.8)$$

其中  $n_1 = |R_1(j, t)|$ ,  $n_2 = |R_2(j, t)|$  分别表示落入第一个区域和第二个区域的训练样本的数量, 且

$$\hat{c}_1 = \frac{1}{n_1} \sum_{i \in R_1(j, t)} Y_i \quad \text{和} \quad \hat{c}_2 = \frac{1}{n_2} \sum_{i' \in R_2(j, t)} Y_{i'},$$

### 5.4.7 算法总结

下面介绍三种决策树算法，其中 ID3 与 C4.5 算法主要用于分类树，而 CART 算法既可以用于分类树也可以是回归树。对决策树  $\mathcal{T}$  的优劣衡量我们可以用以下代价函数。

代价函数为  $\mathcal{C}(\mathcal{T}) = \sum_{t=1}^T n_t H(t)$ ,  $T$  表示所有叶子节点的个数,  $n_t$  是每个叶子节点中样本的个数，在这个公式中相当于权重，因为各叶节点包含的样本数目不同，可使用样本数加权求熵和。 $H(t)$  是每个叶子节点的损失度量准则，上面章节中已介绍。代价函数越小越好。对所有叶节点的熵求和，该值越小说明对样本的分类或者回归越精确。

**ID3 算法（信息增益）**是一种贪心算法，构造的决策树用于分类。于 1986 年由 Quinlan [69] 提出的 ID3 决策树学习算法就以信息增益为准则来选择划分自变量。ID3 算法起源于概念学习系统 (CLS)，以信息熵的下降速度为选取自变量的标准，即在每个节点选取还未被用来划分的具有最高信息增益的自变量作为划分标准，然后继续这个过程，直到生成的决策树能完美分类训练样例。

上述信息增益的定义是基于离散自变量，对于连续自变量，我们可以采用 (5.4.7) 给出的策略，将连续自变量进行离散化，即使用分裂点对连续自变量进行分裂产生区域（即节点）。此时使用的信息增益准则，不仅要选择自变量还要估计出分裂点。

**C4.5 算法（信息增益率）**采用信息增益率来选择最优划分自变量。不是直接选择信息增益率最大的候选划分自变量，而是先从候选划分自变量中找出信息增益高于平均水平的自变量，再从中选择信息增益率最高的。采用了信息增益率来选择特征，以减少信息增益容易选择分类取值多的自变量的问题。

同样的，对于连续自变量，我们可以采用 (5.4.7) 给出的策略使用分裂点对连续自变量进行分裂，然后使用的上述的信息增益率准则，选择自变量以及最优分裂点。

**CART 算法（回归：均方误差；分类：基尼指数）**包括 CART 分类树和 CART 回归树。

#### CART 回归树

**回归树**是决策树模型的一种，专门针对因变量是连续型随机变量或向量，自变量是实值随机向量的树的模型。为了便于理解，我们首先考虑一维连续随机变量  $Y$ ，二元自变量  $X_1, X_2$  的回归问题，并以图5.3为例来阐述基于树的回归方法。

首先，按  $X_2 \geq t_1$  和  $X_2 < t_1$  把整个样本空间  $N_0$  划分为两个子区域  $R_1$  和  $N_1$ ，为了得到更加精确的预测，我们可以继续对  $N_1$  子区域进行划分，得到两个子区域  $R_2$  和  $R_3$ 。接下来我们形象地定义一些树中的概念。为了记号方便，我们称图5.3用于分裂的变量  $X_2$  和  $X_1$  为分裂变量， $t_1$  和  $t_2$  为分裂点， $N_0$  是根节点（包含所有样本）， $N_1$  是中间节点（也是  $R_2$  和  $R_3$  的父节点）， $R_1, R_2$  和  $R_3$  为叶子节点。回归树需要解决以下 4 个关键问题：

1、如何选取分裂自变量？图5.3第一次选择的分裂特征为  $X_1$ ，第二次选择的分

裂特征为  $X_2$ ；

2、如何确定分裂点的值？如5.3中的  $t_1, t_2$ ；

3、如何确定停止分裂的准则，即为什么叶子节点  $R_1, R_2, R_3$  不再分裂？

由问题 1-3 可知，基于回归树模型的核心就是要依次找出分裂变量和分裂点，制定停止规则和对每一个叶子节点的样本进行预测。

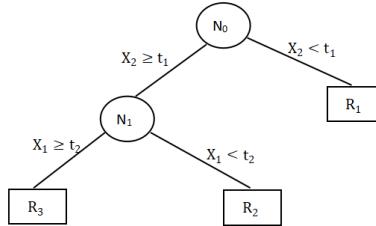


图 5.3 基于树模型定义的预测区域

接下来，我们将介绍如何逐步生成回归树。不失一般性，我们假设含有  $n$  个观测数据的数据集为： $N_0 = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ ，其中， $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$  是  $p$  维实值随机向量， $Y_i$  是一维实值随机变量。

具体的做法是：从所有数据出发，考虑分裂变量  $X_j$  和划分点  $t$  以及对应的划分成的两个子区域  $R_1(j, t)$  和  $R_2(j, t)$ ，根据表达式 (5.4.8)，通过最小化  $R_1(j, t)$  和  $R_2(j, t)$  的均方误差之和

$$\min_{j,t} \text{MSE}(j, t)$$

得到所对应的划分变量  $j$  和划分点  $t$ 。即通过搜索所有的自变量  $X_j, j, \dots, p$ ，以及对照分裂点  $t$  搜索所有的观测值  $\{X_{j1}, \dots, X_{jn}\}$ ，可以确定最好的  $(j, t)$ ，从而可以把整个样本分到两个区域。对每一个区域重复以上过程进行分裂。最后对所有的区域重复这一过程，直到达到最终的停止条件，关于停止条件及树修剪等方面详细知识可以参考 [70] [71] [67]。

**CART 分类树**类似回归树，我们只须把上式中的度量标准  $\text{MSE}(j, t)$  替换成基尼指数，序贯的选择最优变量去分裂样本，从而产生节点，最终生成分类树。

## 5.5 剪枝

决策树对训练数据有很好的分类和回归能力，虽然能在训练集中取得良好的预测效果，但对未知的测试数据未必有好的预测能力，即泛化能力弱，可能发生过拟合现象，导致在测试集上效果不佳。所以，为了防止决策树过度生长、出现过拟合现

象，我们的解决方案是对决策树进行剪枝或构建随机森林。在此处我们主要介绍剪枝，随机森林在后续章节进行介绍。

**剪枝**是决策树算法缓解过拟合的一种重要方法。可通过剪枝在一定程度上避免因决策分支过多，以致于把训练集自身的一些特点当做所有数据都具有的一般性质而导致的过拟合。剪枝主要分为两种：预剪枝（pre-prune）和后剪枝（post-prune）。

### 5.5.1 预剪枝

**预剪枝**在决策树构造时就进行剪枝。方法是在构造的过程中对节点进行评估，包括以下几种策略。

(1)、如果对某个节点进行划分，在验证集中不能带来准确性的提升，那么对这个节点进行划分就没有意义，这时就会把当前节点作为叶节点，不对其进行划分。

(2)、或者更为直接，预先设置每一个节点所包含的最小样本数目，例如 10，若该节点总样本数小于 10 时，则不再分；

(3)、或者预先指定树的高度或者深度，例如树的最大深度为 4；或者指定节点的熵小于某个值时就不再继续划分。

预剪枝的优点是可以降低训练模型的资源开销，缺点是存在欠拟合风险。这是因为有些分支的当前划分虽然不能提升性能，但在其基础上进行的后续划分却有可能导致性能显著提高。预剪枝提前把这些分支剪掉了，可能带来欠拟合风险。

### 5.5.2 后剪枝

预剪枝方法在建树过程中要求每个节点的分裂使得不纯度下降超过一定阈值。这种方法具有一定的短视，因为很有可能某一节点分裂不纯度的下降没超过阈值，但是其在后续节点分裂时不纯度会下降很多，而预剪枝法则在前一节点就已经停止分裂了。

**后剪枝**在生成决策树之后再进行剪枝，通常会从决策树的叶节点开始，逐层向上对每个节点进行评估。如果剪掉这个节点子树，与保留该节点子树在分类准确性上差别不大，或者剪掉该节点子树，能在验证集中带来准确性的提升，那么就可以把该节点子树进行剪枝。

后剪枝的优点是比预剪枝保留了更多的分支，欠拟合风险小，泛化性能往往优于预剪枝决策树，缺点是模型训练时会占用较多的资源。

在实际应用中，更多的是使用后剪枝法，其中的**代价复杂性剪枝**（cost complexity pruning）是最常用的方法。这种方法是先让树尽情生长，得到  $T_0$ ，然后再在  $T_0$  基础上进行修剪。设  $T$  表示子树  $T$  的叶节点数目， $n_t$  表示叶节点  $t$  的样本量，则代价

复杂性的剪枝法的损失函数为：

$$\mathcal{C}_\alpha(\mathcal{T}) = \sum_{t=1}^T n_t H(t) + \alpha T \quad (5.5.1)$$

其中  $\sum_{t=1}^T n_t H(t)$  是代价函数， $\alpha$  是参数。 $\mathcal{T}$  是任意一棵子树，它通过对  $\mathcal{T}_0$  进行剪枝得到，也就是减去  $\mathcal{T}_0$  某个中间节点的所有子节点，使其成为  $\mathcal{T}$  的叶节点。对于固定的  $\alpha$ ，一定存在这样一颗子树  $\mathcal{T}_\alpha$  使得损失函数  $\mathcal{C}_\alpha(\mathcal{T})$  达到最小。

调整参数  $\alpha$  控制着模型对数据的拟合与模型的复杂度（树的大小）之间的平衡。当  $\alpha = 0$  时，最优子树  $\mathcal{T}_\alpha$  等于原树  $\mathcal{T}_0$ 。随着  $\alpha$  取值增大，损失函数对叶节点数目  $T$  惩罚增大，那么此时我们偏向于选择叶节点数目少一点的树。

**后剪枝步骤**: 从  $\mathcal{T}_0$  开始，自下而上地考虑每个内部节点  $t$ ，考虑两种情况：

- (1)、以  $t$  为根节点的子树  $\mathcal{T}_t$ ，其损失为:  $\mathcal{C}_\alpha(\mathcal{T}_t) = \mathcal{C}(\mathcal{T}_t) + \alpha T_t$ ;
- (2)、对  $t$  进行剪枝，即将  $\mathcal{T}_t$  作为叶子节点，其损失为:  $\mathcal{C}_\alpha(t) = \mathcal{C}(t) + \alpha$ ;

我们通过比较  $\mathcal{C}_\alpha(\mathcal{T}_t)$  与  $\mathcal{C}_\alpha(t)$  大小来判断是否要对  $t$  进行剪枝：

当  $\alpha$  较小时，可以容忍模型有较高的复杂度，所以  $\mathcal{C}_\alpha(\mathcal{T}_t) < \mathcal{C}_\alpha(t)$ ，即这个时候不需要剪枝；

当  $\alpha$  逐渐增大到某一阈值时，这个时候需要考虑在训练数据上  $\mathcal{T}_t$  的损失增大，所以有  $\mathcal{C}_\alpha(\mathcal{T}_t) = \mathcal{C}_\alpha(t)$ ，即这个时候剪不剪枝都可以；

当  $\alpha$  继续增大，以至于大于上述阈值时，有  $\mathcal{C}_\alpha(\mathcal{T}_t) > \mathcal{C}_\alpha(t)$ ，这个时候剪枝带来的收益大于作为一棵子树  $\mathcal{T}_t$  所带来的收益，所以要剪枝。

从以上过程我们可以看出，对于树中的每个内部节点  $t$ ，都有一个特定的阈值  $g(t)$ 。当  $\alpha = g(t)$  时， $\mathcal{C}_\alpha(\mathcal{T}_t) = \mathcal{C}_\alpha(t)$ 。故而  $g(t)$  可以决定是否需要对其进行剪枝，且该阈值等于

$$g(t) = \frac{\mathcal{C}(t) - \mathcal{C}(\mathcal{T}_t)}{T_t - 1}.$$

上式  $g(t)$  可由等式  $\mathcal{C}_\alpha(\mathcal{T}_t), \mathcal{C}_\alpha(t)$  解出。

1、因此，在生成子树  $\mathcal{T}_1$  时，我们可以计算  $\mathcal{T}_0$  的每个内部节点的阈值  $g(t)$ ，选择其中最小的记为  $g(t_1)$ 。当  $\alpha$  大于该阈值  $g(t_1)$  时，这意味着我们要剪掉该节点  $t_1$  对应的子树  $\mathcal{T}_t$  得到新的树  $\mathcal{T}_1$  收益较大。当  $g(t_1) \leq \alpha < g(t_2)$  时， $\mathcal{T}_1$  使得损失函数最小。同理当  $g(t_2) < \alpha$  时，剪掉节点  $t_2$  的子树，使  $t_2$  为叶节点，得到子树  $\mathcal{T}_2$ 。同理子树  $\mathcal{T}_2$  在区间  $g(t_2) \leq \alpha < g(t_3)$  上使得损失函数最小。接着在  $\mathcal{T}_2$  的基础上持续剪枝，就可以得到最终的子树序列。

2、交叉验证在生成子树序列  $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \dots$  后，使用验证数据集，测试子树序列中每棵子树的损失，选择最小的子树作为剪枝后的决策树，这个时候也对应了一个  $\alpha_k$ 。

## 5.6 决策树实践

### 5.6.1 R 语言实践

R 语言中常用的决策树算法工具是 R 程序包 rpart (Recursive Partitioning and Regression Trees) 里的 rpart 函数，本小节以鸢尾花数据集为例，用决策树进行分类。具体实现过程代码如下：

```
library(rpart)
library(rpart.plot)
#表达式子Species~.表示Species因变量，其余都是自变量
#决策树类型为“class”分类树；使用信息增益为划分标准
fit <- rpart(Species~, data=iris,method="class", parms=list(split="information"))
rpart.plot(fit,sub="Classification tree")
```

决策树输出结果：

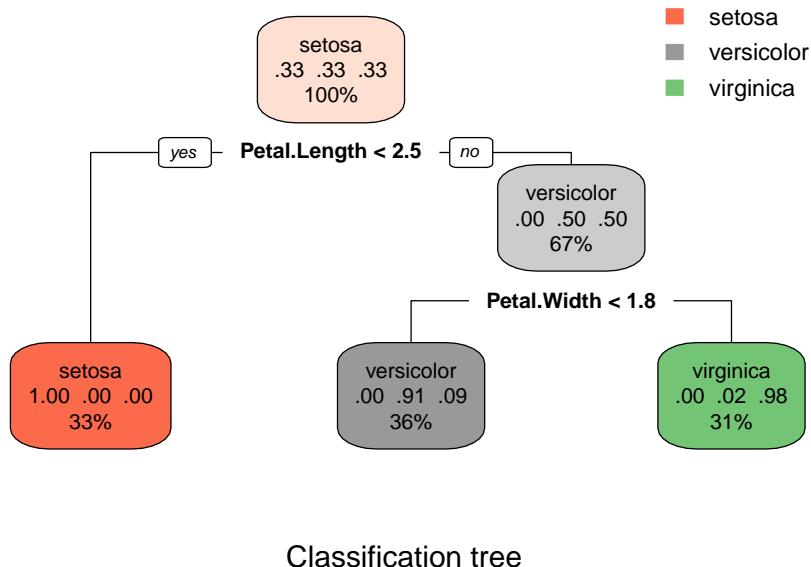


图 5.4 分类树树结构

## 5.6.2 Python 语言实践

决策树算法常用的 Python 工具是 scikit-learn（简称 sklearn），sklearn 是一个应用很广泛的 Python 机器学习库。sklearn 库中自带了决策树算法。

sklearn 库中的 DecisionTreeClassifier 函数可以实现分类树，DecisionTreeClassifier 将两个数组作为输入，第一个数组是用于保存训练样本的自变量数组  $X$ ，第二个数组是用于保存训练样本的类标签的整数值数组  $Y$ 。

如以下程序示例，从 sklearn 模块中导入分类树的函数，对数组  $X$  和  $Y$  赋值后进行拟合：

```
from sklearn import tree
X = [[0, 0], [1, 1]]
Y = [0, 1]
clf = tree.DecisionTreeClassifier()#建立分类树
clf = clf.fit(X, Y)#基于数据进行分类树拟合
clf.predict([[2., 2.]])#对新输入的数据进行分类预测
clf.predict_proba([[2., 2.]])#预测样本属于每个类的概率
```

使用 DecisionTreeRegressor 类，决策树也可以应用于回归问题。此时因变量  $Y$  具有浮点值而不是整数值。下面代码演示 DecisionTreeRegressor 类的使用：

```
from sklearn import tree
X = [[0, 0], [2, 2]]
Y = [0.5, 2.5]
clf = tree.DecisionTreeRegressor()#建立回归树
clf = clf.fit(X, Y)#基于数据进行回归树拟合
clf.predict([[1, 1]])#对新输入的数据进行回归预测
```

## 信息增益与增益率

本小节以前面介绍的预测是否打网球的实例，运用 Python 软件，演示信息增益和增益率的计算过程。

1、导入 math 库，自定义计算信息熵的函数并计算根节点的信息熵。

```
import math
def info(x,y):#定义计算特征熵的函数
    if x != y and x != 0:
        return -(x/y)*math.log2(x/y)-((y-x)/y)*math.log2((y-x)/y)#计算熵
    if x == y or x == 0:
        return 0#纯度最大，熵值为0
info_parent = info(6,10)#自变量划分前信息熵
print (info_parent)#计算根节点的信息熵
```

2、计算信息增益。首先计算天气 (weather)、温度 (temperature) 以及是否有风自变量 (windy) 的信息增益。代码如下：

```
#分别计算出三种天气情况下的信息熵，再结合根节点的信息熵计算出该自变量的信息增益
weather_sunny_entropy = info(3,4)
weather_cloudy_entropy = info(2,2)
weather_rain_entropy = info(1,4)
weather_entropy = (4/10) * weather_sunny_entropy + (2/10) * weather_cloudy_entropy +
(4/10) * weather_rain_entropy#计算天气特征信息熵
weather_info_gain = info_parent - weather_entropy#计算天气的信息增益
# 计算温度每种情况的熵
temperature_hot_entropy = info(1, 3)
temperature_warm_entropy = info(2, 3)
temperature_cold_entropy = info(2, 4)
#计算温度特征信息熵
temperature_entropy = (3/10) * temperature_hot_entropy + (3/10) *
temperature_warm_entropy + (4/10) * temperature_cold_entropy
temperature_info_gain = info_parent - temperature_entropy#计算温度信息增益
# 计算是否有风每种情况的熵
windy_yes_entropy = info(1,3)
windy_no_entropy = info(5,7)
#计算是否有风特征信息熵
windy_entropy = (3/10) * windy_yes_entropy + (7/10) * windy_no_entropy
windY_info_gain = info_parent - windy_entropy#计算是否有风信息增益
```

以是否有风自变量的信息熵与信息增益为例，将计算结果输出出来：

```
print ('windy_yes_entropy:', round(windy_yes_entropy, 4))
print ('windy_no_entropy:', round(windy_no_entropy, 4))
print('windy_entropy:', round(windy_entropy, 4))
print ('windY_info_gain:', round(windY_info_gain, 4))
```

3、接下来展示计算三种自变量的信息增益率。

```
#根据公式分别计算三种自变量对应的IV值：
iv_weather = -(4/10) * math.log2(4/10) - (2/10) * math.log2(2/10) - (4/10) *
math.log2(4/10)
iv_temperature = -(3/10) * math.log2(3/10) - (3/10) * math.log2(3/10) - (4/10) *
math.log2(4/10)
iv_windy = -(3/10) * math.log2(3/10) - (7/10) * math.log2(7/10)
#用各自变量的信息增益值除以对应的IV值得到三个自变量的信息增益率：
gain_ratio_weather = weather_info_gain / iv_weather
gain_ratio_temperature = temperature_info_gain / iv_temperature
gain_ratio_windy = windY_info_gain / iv_windy
print('gain_ratio_weather:', round(gain_ratio_weather, 4))
print('gain_ratio_temperature:', round(gain_ratio_temperature, 4))
print('gain_ratio_windy:', round(gain_ratio_windy, 4))
```

## 决策树算法

以下过程借助 sklearn 库完成，sklearn 库中自带了决策树算法，下面以打网球数据集，来演示决策树算法的应用。

1、导入 Pandas 库和 Sklearn 库，读入数据与数据预处理。

```

import pandas as pd
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics#导入相关函数
dataset=[ ['晴', '热', '否', '否'], ['晴', '热', '是', '否'], ['阴', '热', '否', '是'],
['雨', '温', '否', '是'], ['雨', '凉爽', '否', '是'], ['雨', '凉爽', '是', '否'],
['阴', '凉爽', '是', '是'], ['晴', '温', '否', '否'], ['晴', '凉爽', '否', '是'],
['雨', '温', '否', '是'] ]
#读入数据并命名
data = pd.DataFrame(dataset, columns=['weather', 'temperature', 'windy',
'play_tennis'])
#数据集中的四列数据都是字符串类型，因此字符串进行数值化，按类别进行赋值
labels = data['weather'].unique().tolist()
data['weather'] = data['weather'].apply(lambda n: labels.index(n))
labels = data['temperature'].unique().tolist()
data['temperature'] = data['temperature'].apply(lambda n: labels.index(n))
labels = data['windy'].unique().tolist()
data['windy'] = data['windy'].apply(lambda n: labels.index(n))
labels = data['play_tennis'].unique().tolist()
data['play_tennis'] = data['play_tennis'].apply(lambda n:labels.index(n))

```

2、决策树模型训练、预测与评估。首先对数据集进行拆分，前三列是特征自变量，最后一列是类别标签，分别记作  $x$  和  $y$ 。然后选择分类标准为基尼指数，对数据进行模型拟合。

```

x = data.drop(['play_tennis'],axis=1)#所有样本特征
y = data['play_tennis']#类别标签
clf = DecisionTreeClassifier(criterion='entropy')
clf.fit(x,y)#模型训练
DecisionTreeClassifier(criterion='entropy')#输出结果
y_pred = clf.predict(x) #模型预测
print('Accuracy:%.4f'%(metrics.accuracy_score(y, y_pred)))#模型评估

```

3、绘制决策树。导入绘图和可视化树形结构的库和函数，绘制树形图，保存图像并展示。结果如图5.5所示。

```

from IPython.display import Image
from sklearn import tree
import pydotplus
dot_data = tree.export_graphviz(clf, out_file=None, feature_names=data.columns[0:3],
class_names=data.columns[3], filled=True, rounded=True, special_characters=True)
graph = pydotplus.graph_from_dot_data(dot_data)
graph.write_png('example.png')#保存图像
Image(graph.create_png())

```

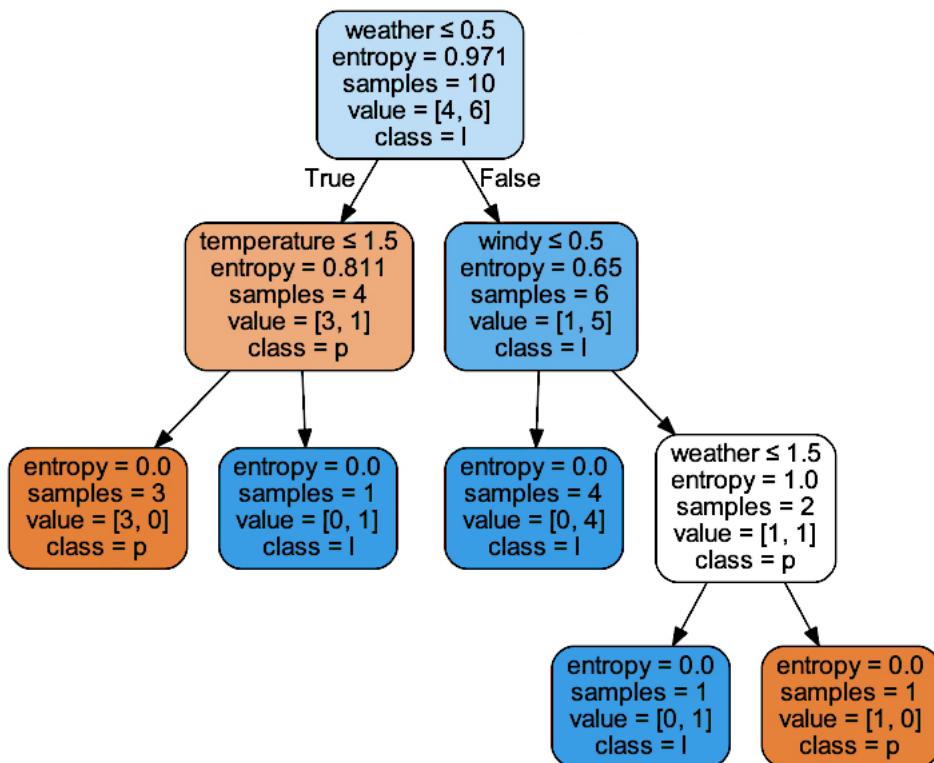


图 5.5 决策树树结构

## 5.7 本章小节

本章主要讨论了决策树中回归树和分类树的基本方法和算法，对于分类树主要采用投票准则来对新样本进行分类，而对于回归树则建议采用叶子节点里所有样本的样本均值来预测。本章介绍的基于树的方法具有以下优点：

- (1) 可以简单而自然的处理分类和有序变量；
- (2) 自动的逐步变量选择和降低复杂度；
- (3) 对任一单个有序变量的单调变换保持不变；
- (4) 对训练集中的异常点稳健；
- (5) 易于处理缺失值。

为了得到更加精确的预测，在叶子节点中样本量足够的情况下，也可以对叶子节点里的样本建立不同的预测模型而非简单的投票准则的分类树方法或者基于样本均值的回归树方法，不同的预测模型可以考虑线性或非线性模型等。其他关于基于回归树的集成算法包括 Bagged 树、Random Forests 和 Boosting 等集成算法和相关程序实现可以参考 Kuhn 和 Johnson 的书籍第八章（回归）和第十四章（分类）[71]。

## 5.8 习题

- 1、生成一模拟数据，来考察树方法在不同层次叶子节点的偏差。
- 2、Friedman [72] 介绍了如下的回归模型

$$Y = 10 \sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + \varepsilon$$

其中， $X_1, X_2, \dots, X_5 \sim U(0, 1)$ ， $\varepsilon \sim N(0, \sigma^2)$ ，我们也生成  $X_6, \dots, X_{10}$  5 个不相干的变量。

- (1) 设定样本量为 200，应用回归树方法拟合这一模拟数据；
- (2) 选取  $X_1, \dots, X_5$  中的一个进行单调变换，再次应用回归树模型拟合这一数据，与 (1) 中结果进行对比；
- (3) 将第一个样本扩大为原始值的 10 倍，应用回归树模型拟合这一数据，与 (1) 的结果进行对比。
- (4) 删除  $X_1, \dots, X_5$  中某一变量的部分观测值，然后应用回归树模型拟合这一数据，与 (1) (2) 的结果再次进行对比。
- 3、从鸢尾花数据集中任取 2 个自变量，应用分类树模型分析这一数据，并与书中例子进行对比。
- 4、思考一下对于回归树和分类树怎么评价变量的重要程度。
- 5、回归树和分类树都是对单一的变量进行划分，建立树模型，请考虑一种现代降维方法，先对自变量进行降维，然后再建立回归树，并应用习题 2 的模拟数据进行对比。
- 6、CART 分类树建模中，如果自变量是连续变量，如何结合基尼指数这一损失度量去选择最优变量，请论述。

# 第六章 随机森林

## 6.1 简介

回归树和分类树的优点是解释性强且易于计算，然而就像一个硬币具有两面，基于单棵树的回归或者分类模型也具有明显的缺陷：

- 1、由于模型的简单性及仅用单一的常数作为最终区域的预测值，从而使得单棵树的回归或分类模型很难具有最优的预测能力 [71]；
- 2、单棵树的预测是不稳定的 [73] [70]，数据有较小的变动就可能会导致完全不一样的分裂变量和分裂点；
- 3、单棵树具有容易遭受选择偏差影响的问题，即取值多的分类自变量比取值少的分类自变量更容易被选择为分裂变量 [74] [75] [76]。

为了克服单一统计模型（单棵树）稳定性差（方差大）的缺陷，集成学习方法得到了广泛的研究和发展。所谓集成学习，就是指分类（回归）器的集成。集成学习通过构建并结合多个弱学习器来完成学习任务，一般的方法是先产生一组个体学习器，再用某种策略将它们结合起来，常见的结合策略有：平均法、投票法和学习法等。例如Bagging方法利用 Bootstrap 方法（有放回抽样）对训练集进行抽样，得到一系列新的训练集，对每个训练集都构建一棵树，最后通过平均法、投票法组合所有预测器得到最终的预测模型。

后来，为了克服单棵树的缺陷及降低每次 Bootstrap 抽样之间的相关性，Breiman [77] 综合以往集成学习的优缺点提出了一种新的集成学习方法——随机森林（Random Forests）。接下来我们将详细介绍随机森林的基本思想、算法步骤及变量重要性评价等内容。

## 6.2 随机森林基本概况

在引入随机森林这一算法之前，我们先考虑与之相关的概念——Bagging。Bagging 是 Bootstrap Aggregating 的缩写。

特别的，我们考虑每个预测器都是决策树模型的 Bagging 算法。在 Bagging 算法中，我们注意到在对训练集  $N_0 = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$  进行 Bootstrap 抽样（样本量为  $n$ ）以获得  $M$  个新的训练集，记为  $\{N_m, m = 1, 2, \dots, M\}$ ，鉴于

Bootstrap 抽样的性质，可以证明新训练集  $N_b, b \in \{1, 2, \dots, M\}$  大约只包含原训练集  $N_0$  的三分之二（因为一个样本，经  $n$  次有放回抽样，仍未被抽中的概率是  $\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} \approx \frac{1}{3}$ ）。这些未被使用的观测值称为此树的袋外观测值（Out-Of-Bag, OOB）。袋外观测值构成的集合称为袋外示例。

可以将袋外示例作为对应训练集生成的树的测试集来评估训练的结果，即可以用所有将第  $i$  个观测值作为 OOB 的树来预测第  $i$  个观测值的响应值。我们可以对这些预测响应值求平均（回归情况下）或执行多数投票（分类情况下），已得到第  $i$  个观测值的一个 OOB 预测。用这种方法可以求出每个观测值的 OOB 预测，根据这些就可以计算总体的 OOB 均方误差（对回归问题）或分类误差（对分类问题）。由此得到的 OOB 误差是对 Bagging 模型测试误差的有效估计。我们将应用 OOB 误差来评估随机森林变量的重要性。

随机森林的基本思想是：为了降低单棵树的缺陷及各次抽样间的相关性，随机森林采用有放回抽样，每次抽取  $n$  个样本，再无放回随机抽取  $p$  个属性（自变量）中的  $k$ （一般取  $k$  大约为  $\sqrt{p}$ ）个属性，把这  $k$  个属性当成新的特征，并结合因变量和抽取的  $n$  个样本，生成一棵回归树或者分类树，重复这一过程  $M$  次得到  $M$  棵回归树或分类树，随机森林是通过集成上述  $M$  棵回归树或分类树而成。通过集成  $M$  棵树可以有效避免单棵树的不稳定性，而每一棵树只用  $k$  个属性代替  $p$  个属性来建模，不仅可以有效降低树之间的相关性，还能提高计算速度和节省计算机内存。

随机森林算法的步骤如下：

(1) 从数据集  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$  中进行 Bootstrap 抽样（有放回抽样），抽取  $n$  个样本，得到样本集  $N_m$ ；

(2) 利用  $N_m$  建立一棵决策树，对于树上的每个节点，重复以下步骤，直到节点的样本数达到指定的最小限定值  $n_{\min}$ ：

a) 从全部  $p$  个随机变量中随机取  $k (k < p)$  个；

b) 从这  $k$  个变量中选取最优分裂变量，将此节点分裂成两个子节点。

注：对于分类问题，构造每棵树时默认使用  $k = \sqrt{p}$  个随机变量，节点最小样本数为 1；对于回归问题，构造每棵树时默认使用  $k = \frac{p}{3}$  个随机变量，节点最小样本数为 5。

(3) 重复以上过程  $M$  次，得到  $M$  棵树构成一个随机森林。

(4) 当对新样本进行预测时，由每个决策树得到一个预测结果，再进行平均或“投票”得出最后的结果。a) 对于回归问题，最后的预测结果为所有决策树预测值的平均数；b) 对于分类问题，最终的预测结果为所有决策树预测结果中最多的那类，即采用“投票”得出最后的分类结果。

从随机森林的基本思想和算法可以看出，随机森林具有以下的特点：

(1) 与其他的集成学习如 Bagging 相比，由于每次只选取  $k (k < p)$  个预测，能够有效的降低树间的相关性，从而能最大程度的减少预测方差，提高预测的精度；

(2) 由于每次只用到  $k$  个自变量，因此能有效节省计算时间和计算机内存；

(3) 与 Bagging 相比, 随机森林最大的不同就在于自变量子集的规模  $k$ 。若取  $k = p$  建立随机森林, 则等同于建立 Bagging 树。因此, Bagging 是随机森林的特例。

**变量重要性评估:** 与所有集成学习方法一样, 随机森林很难得到自变量(特征)与因变量间的一个直接的显式表达关系。因而, 很难评估自变量的重要性。考虑到随机森林只用到了部分自变量, Breiman [77] 建议通过如下方式来度量某个特征  $X_j$  的重要性:

- (1) 根据未被抽取样本 OOB 计算随机森林中第  $i$  棵回归树的袋外误差  $e_i$ ;
- (2) 随机打乱训练集在变量  $X_j$  所在列的取值顺序, 并计算新的袋外误差  $e_i^j$ ;
- (3) 重复步骤直至计算出所有决策树的误差变化, 最后变量  $X_j$  预测误差的平均

变化, 即重要性指标:  $V(X_j) = \sum_{i=1}^M (e_i^j - e_i)^2 / M$ 。

这里袋外误差是指我们使用针对某一棵树的袋外数据得到的预测误差的均值。因为共有  $M$  棵树, 故而有  $M$  个袋外示例。由上述袋外示例会生成  $M$  个袋外误差  $e_i \quad i = 1, 2, \dots, M$ 。由此可知, 若特征变量  $X_j$  的变化引起重要性指标增加越大, 精度减少得越多, 则说明该变量越重要。

## 6.3 随机森林基本理论

在上一节我们已经详细介绍了随机森林算法的基本思想、算法及与其他集成学习相比的优缺点等, 这一节我们将详细介绍随机森林相关的基本理论。类似与 6.1.1 和 6.1.2 节, 我们将分别介绍回归树和分类树的基本理论。

### 6.3.1 回归树基本理论

假设随机森林是通过树的预测  $h(\mathbf{X}, \Theta)$  生成的, 其中,  $\Theta$  是  $q$  维随机向量,  $h(\mathbf{X}, \Theta)$  是  $R^{p+q} \rightarrow R$  上的实值函数。不妨假设  $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$  独立同分布地来自  $(\mathbf{X}, Y)$ , 并定义均方误差为:  $E_{\mathbf{X}, Y}(Y - h(\mathbf{X}, \Theta))^2$ 。

通过对  $M$  棵单一回归树  $h(\mathbf{X}, \Theta_i)$  取平均来生成随机森林的预测, 当  $M \rightarrow \infty$ , 我们可以得到定理 6.1 的结论, 即随机森林预测是均方收敛的。

**定理 6.1** 随着随机森林中树的数目趋向无穷, 则如下结论几乎处处成立:

$$\lim_{M \rightarrow \infty} E_{\mathbf{X}, Y} \left( Y - \frac{1}{M} \sum_{i=1}^M h(\mathbf{X}, \Theta_i) \right)^2 \rightarrow E_{\mathbf{X}, Y} (Y - E_{\Theta} h(\mathbf{X}, \Theta))^2$$

由定理6.1可知当树的数目  $M \rightarrow \infty$  时, 随机森林的预测误差趋向于总体预测误差。接下来我们将推导出随机森林预测误差的上界, 其结果将在如下的定理6.2中给出。

**定理 6.2** 假设对所有  $\Theta$ ,  $EY = E_{\mathbf{X}}(h(\mathbf{X}, \Theta))$ , 则如下结论成立:

$$E_{\mathbf{X}, Y} E_{\Theta} [Y - h(\mathbf{X}, \Theta)]^2 \leq \rho E_{\Theta} [E_{\mathbf{X}, Y} (Y - h(\mathbf{X}, \Theta))^2]$$

其中,  $\rho$  是  $Y - h(\mathbf{X}, \Theta)$  和  $Y - h(\mathbf{X}, \Theta')$  间的加权相关系数,  $\Theta$  和  $\Theta'$  相互独立。

### 证明

$$E_{\mathbf{X}, Y} [E_{\Theta} (Y - h(\mathbf{X}, \Theta))]^2 = E_{\Theta} E_{\Theta'} E_{\mathbf{X}, Y} (Y - h(\mathbf{X}, \Theta))(Y - h(\mathbf{X}, \Theta')) \quad (6.3.1)$$

式 (6.3.1) 的右边是一个协方差并且可以写成:

$$E_{\Theta} E_{\Theta'} (\rho(\Theta, \Theta') sd(\Theta) sd(\Theta')), \quad (6.3.2)$$

其中,  $sd(\Theta) = \sqrt{E_{\mathbf{X}, Y} (Y - h(\mathbf{X}, \Theta))^2}$ . 加权相关系数的定义为:

$$\rho = E_{\Theta} E_{\Theta'} (\rho(\Theta, \Theta') sd(\Theta) sd(\Theta')) / (E_{\Theta} sd(\Theta))^2. \quad (6.3.3)$$

因此,

$$\begin{aligned} E_{\mathbf{X}, Y} [E_{\Theta} (Y - h(\mathbf{X}, \Theta))]^2 &= \rho (E_{\Theta} sd(\Theta))^2 \\ &\leq \rho E_{\Theta} [E_{\mathbf{X}, Y} (Y - h(\mathbf{X}, \Theta))^2]. \end{aligned} \quad (6.3.4)$$

证毕。 □

由定理6.2可以看出随机森林预测的准确性取决于单棵树的预测能力及树之间相关性的强弱。

### 6.3.2 分类树基本理论

假设随机森林是树型分类器  $\{h(\mathbf{X}, \Theta_i), i = 1, \dots, M\}$  的集合, 其中,  $\mathbf{X}$  是预测向量;  $\Theta_i$  是独立同分布的随机向量, 决定了单棵分类树的生成过程; 元分类器  $h(\mathbf{X}, \Theta_i)$  是用 CART [67] 算法构建的无剪枝的分类决策树。则当  $M$  趋向无穷时, 对分类树也是收敛的, 即有如下结论:

**定理 6.3** 随着随机森林中树的数目趋向无穷, 则如下结论成立:

$$\lim_{M \rightarrow \infty} \Pr_{\mathbf{X}, Y} \left( \left[ \frac{1}{M} \sum_{i=1}^M I(h(\mathbf{X}, \Theta_i) = Y) - \max_{j \neq Y} \frac{1}{M} \sum_{i=1}^M I(h(\mathbf{X}, \Theta_i) = j) \right] < 0 \right)$$

$$\rightarrow \Pr_{\mathbf{X}, Y} \left( \left[ \Pr_{\Theta}(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} \Pr_{\Theta}(h(\mathbf{X}, \Theta) = j) \right] < 0 \right),$$

其中,  $I(\cdot)$  是示性函数。

**证明** 容易证明, 参数空间  $\Theta_1, \Theta_2, \dots, \Theta_M$  上存在一个零概率集合  $C$ , 在  $C$  之外, 对于所有的  $\mathbf{X}$ , 有下式成立

$$\frac{1}{M} \sum_{i=1}^M I(h(\mathbf{X}, \Theta_i) = g) \rightarrow \Pr_{\Theta}(h(\mathbf{X}, \Theta) = g)$$

在一个固定的训练集和参数空间  $\Theta$  上, 所有满足  $h(\Theta, \mathbf{X}) = g$  的  $\mathbf{X}$  构成的集合是一个超矩形单元。对于所有  $h(\Theta, \mathbf{X})$  只有有限的  $K$  个这种超矩阵单元, 记作  $S_1, \dots, S_K$ 。若  $\{\mathbf{X} : h(\Theta, \mathbf{X}) = g\} = S_g$ , 此时则定义  $\phi(\Theta) = g$ , 并令  $N_g$  为前  $N$  次试验中  $\phi(\Theta_n) = g$  的次数。那么有

$$\frac{1}{M} \sum_{i=1}^M I(h(\mathbf{X}, \Theta_i) = g) = \frac{1}{M} \sum_{g=1}^K N_g I(\mathbf{X} \in S_g)$$

再由大数定理可得

$$\frac{N_g}{M} = \frac{1}{M} \sum_{i=1}^M I(\phi(\Theta_i) = g)$$

会收敛到  $\Pr_{\Theta}(\phi(\Theta) = g)$ 。对于  $g$  的某个值, 所有集合的并集都不会发生收敛, 得到一个概率为零的集合  $C$ , 因此在  $C$  之外有

$$\frac{1}{M} \sum_{i=1}^M I(h(\mathbf{X}, \Theta_i) = g) \rightarrow \sum_{g=1}^K \Pr_{\Theta}(\phi(\Theta) = g) I(\mathbf{X} \in S_g).$$

上式右边即是  $\Pr_{\Theta}(h(\mathbf{X}, \Theta) = g)$ 。这样我们就证明了定理。  $\square$

由定理6.3可知, 随着随机森林中树的数量的增加, 模型的分类误差上限趋于一个固定值。即随机森林不会随着分类树数目的增加而产生过度拟合的问题, 将对未知实例预测提供较好的参考思路和应用性。类似定理6.2, 我们将给出随机森林分类误差的一个上界, 为叙述方便, 我们先给出如下定义:

给定一组分类器  $h(\mathbf{X}, \Theta_1), h(\mathbf{X}, \Theta_2), \dots, h(\mathbf{X}, \Theta_M)$ , 并使用从随机向量  $(\mathbf{X}, Y)$  的分布中随机抽取的训练集, 将**边际函数**定义为

$$\text{mg}(\mathbf{X}, Y) = \frac{1}{M} \sum_{i=1}^M I((h(\mathbf{X}, \Theta_i)) = Y) - \max_{j \neq Y} \frac{1}{M} \sum_{i=1}^M I((h(\mathbf{X}, \Theta_i)) = j)),$$

其中,  $I(\cdot)$  是指示函数。

边际函数衡量的是正确分类在  $(\mathbf{X}, Y)$  的平均投票数超过任何其他分类的平均投票数的程度。差距越大, 对分类的信心就越大。一般泛化误差由下式给出:

$$\text{PE}^* = \Pr_{\mathbf{X}, Y}(\text{mg}(\mathbf{X}, Y) < 0)$$

随机森林的边际函数为

$$\text{mg}(\mathbf{X}, Y) = \Pr_{\Theta}(h(\mathbf{X}, \Theta) = Y) - \max_{j \neq Y} \Pr_{\Theta}(h(\mathbf{X}, \Theta) = j)$$

分类器集  $h(\mathbf{X}, \Theta)$  的强度定义为：

$$s = E_{\mathbf{X}, Y} \text{mg}(\mathbf{X}, Y).$$

不妨设  $s \geq 0$ , 由切比雪夫不等式可得下式成立：

$$\text{PE}^* \leq \text{var}(\text{mg}(\mathbf{X}, Y)) / s^2 \quad (6.3.5)$$

$\text{mg}(\mathbf{X}, Y)$  的方差的一个显式的表达为：

$$\widehat{j}(\mathbf{X}, Y) = \arg \max_{j \neq Y} \Pr_{\Theta}(h(\mathbf{X}, \Theta) = j)$$

故

$$\begin{aligned} \text{mg}(\mathbf{X}, Y) &= \Pr_{\Theta}(h(\mathbf{X}, \Theta) = Y) - \Pr_{\Theta}(h(\mathbf{X}, \Theta) = \widehat{j}(\mathbf{X}, Y)) \\ &= E_{\Theta}[I(h(\mathbf{X}, \Theta) = Y) - I(h(\mathbf{X}, \Theta) = \widehat{j}(\mathbf{X}, Y))] \end{aligned}$$

原始边际函数定义为：

$$\text{rmg}(\Theta, \mathbf{X}, Y) = I(h(\mathbf{X}, \Theta) = Y) - I(h(\mathbf{X}, \Theta) = \widehat{j}(\mathbf{X}, Y))$$

因此,  $\text{mg}(\mathbf{X}, Y)$  是  $\text{rmg}(\Theta, \mathbf{X}, Y)$  关于  $\Theta$  的期望。对于任意函数  $f$ , 当  $\Theta, \Theta'$  独立同分布时, 有下式成立:

$$[E_{\Theta} f(\Theta)]^2 = E_{\Theta, \Theta'} f(\Theta) f(\Theta')$$

即有

$$\text{mg}(\mathbf{X}, Y)^2 = E_{\Theta, \Theta'} \text{rmg}(\Theta, \mathbf{X}, Y) \text{rmg}(\Theta', \mathbf{X}, Y) \quad (6.3.6)$$

再由式 (6.3.2) 可得

$$\begin{aligned} \text{var}(\text{mg}(\mathbf{X}, Y)) &= E_{\Theta, \Theta'} (\text{cov}_{\mathbf{X}, Y} (\text{rmg}(\Theta, \mathbf{X}, Y) \text{rmg}(\Theta', \mathbf{X}, Y))) \\ &= E_{\Theta, \Theta'} (\rho(\Theta, \Theta') \text{sd}(\Theta) \text{sd}(\Theta')) \end{aligned}$$

其中,  $\rho(\Theta, \Theta')$  是  $\text{rmg}(\Theta, \mathbf{X}, Y)$  和  $\text{rmg}(\Theta')$  间的相关系数,  $\text{sd}(\Theta)$  是  $\text{rmg}(\Theta, \mathbf{X}, Y)$  的标准差。那么,

$$\begin{aligned} \text{var}(\text{mg}) &= \rho(E_{\Theta} \text{sd}(\Theta))^2 \\ &\leq \rho E_{\Theta} \text{var}(\Theta) \end{aligned} \quad (6.3.7)$$

其中,  $\rho$  是相关系数的平均值, 即

$$\rho = E_{\Theta, \Theta'} (\rho(\Theta, \Theta') \text{sd}(\Theta) \text{sd}(\Theta')) / E_{\Theta, \Theta'} (\text{sd}(\Theta) \text{sd}(\Theta'))$$

$$\begin{aligned} \text{E}_\Theta \text{var}(\Theta) &\leq \text{E}_\Theta (\text{E}_{\mathbf{X}, Y} \text{rmg}(\Theta, \mathbf{X}, Y))^2 - s^2 \\ &\leq 1 - s^2 \end{aligned} \quad (6.3.8)$$

由式 (6.3.7) 及 (6.3.8) 可得到随机森林分类误差的上界。

**定理 6.4** 随机森林泛化误差的上界由下式给出

$$\text{PE}^* \leq \rho(1 - s^2)/s^2$$

由定理6.4 可知，分类错误的上界受随机森林中单个分类器强弱及分类间的相关性影响。

## 6.4 随机森林实践

### 6.4.1 R 语言实践

本文将应用随机森林来预测网络文章信息传播热度和文章热度的分类问题。

#### 数据来源及预处理

数据来源于 UCI (Machine Learning Repository) 中 Online News Popularity 数据集，该数据提取来自于 Mashable 网站 2013 年 1 月 7 日至 2015 年 1 月 7 日两年内所有发表文章的内容，共 39797 条文章记录，每篇文章特征变量有 60 项，主要包括词语、链接、数字媒介、时间、关键词及语言特征等方面。考虑到文章传播热度需要足够时间，舍弃部分数据集，获得文章出版时间间隔小于 21 天的文章记录，该类型数据记录占总数据集的 2%，经筛选确定最终样本量为 39000 条有效文章记录。

表 6.1 特征变量及其数据类型

特征变量	类型	特征变量	类型
词语		链接	
1. 内容字数 2. 平均词语长度 3. 连续词语的概率	数字 数字 数字	21. 链接数 22. 与其他文章的链接数 23. 相关文章索引次数的最小值 24. 相关文章的索引次数平均值	数字 数字 数字 数字
关键词		时间	
5. 元数据中关键词的数量 6. 最弱关键词的最小索引次数 7. 最强关键词的最小索引次数 8. 最强关键词的最大索引次数 9. 最强关键词的平均索引次数 10. 平均关键词的最小索引次数 11. 平均关键词的最大索引次数 12. 平均关键词的平均索引次数	数字 数字 数字 数字 数字 数字 数字 数字	25. 时间间隔 26. 文章是否发表于周一 27. 文章是否发表于周二 28. 文章是否发表于周三 29. 文章是否发表于周五 30. 文章是否发表于周六 31. 文章是否发表于周末	数字 类型 类型 类型 类型 类型 类型 类型
内容主题		语言特征	
13. 数据通道是否归属于“娱乐” 14. 数据通道是否归属于“商业” 15. 数据通道是否归属于“社交” 16. 数据通道是否归属于“科技” 17. 与热门话题 0 的接近程度 18. 与热门话题 1 的接近程度 19. 与热门话题 2 的接近程度 20. 与热门话题 4 的接近程度	类型 类型 类型 类型 数字 数字 数字 数字	32. 标题客观性 33. 文章客观性水平 34. 绝对主观性水平 35. 文中积极词汇概率 36. 非中性词汇中消极词汇的概率 37. 积极词汇倾向性平均值 38. 积极词汇倾向性最小值 39. 消极词汇倾向性最大值	比率 比率 比率 比率 比率 比率 比率 比率
数字媒介		目标变量	
4. 图片数	数字	40. 文章分享次数	数量

为提高数据挖掘效率需要对高维数据进行变量选择，寻找对目标变量具有最强解释力的特征变量集合。本文采用 LASSO (Least Absolute Shrinkage and Selection Operator) [78] 算法实现特征变量的精简，其主要思想是：通过最小化  $L_1$  惩罚函数，来进行参数估计和变量选择。LASSO 估计能产生某些严格等于 0 的回归系数，达到变量选择的目的。本数据集通过实现该算法最终选择 40 个特征变量，结果见表6.1。

## 样本生成

为了考虑问题的方便，我们将按下列方式来生成两类样本及训练和检验样本：

(1) 目标变量：文章分享次数，是数量型变量。鉴于本文的目的是要建立对文章热度的二分类预测判别模型，为保证数据结构的平衡性和稳定性，本文根据目标变量的中位数（1400 次）将样本分为两类：文章分享次数大于 1400 次的样本对象定义为“热门文章”（popular），文章分享次数小于或等于 1400 次的样本对象定义为“非热门文章”（unpopular），各类别样本数量结果见表6.2。

(2) 数据集划分：将数据集划分为训练集和测试集，采用随机抽样法从原始数据中抽取 70% 的样本量作为训练集，30% 的样本量为测试集，各数据集结构见表6.3。

表 6.2 各类别样本数

类别	样本数
热门文章	19244
非热门文章	19756

表 6.3 数据集结构

样本类别	样本量	变量数目
训练样本集	27300	40
测试样本集	11700	40

## 结果展示

### 1、不同分类方法的效果比较：

下面选用三种典型的机器学习分类方法：SVM、KNN 和神经网络，与基于随机森林的网络文章热度分类方法进行性能比较，所采用的数据均来自于同一数据集。

对于 KNN 方法，可构成非线性判别边界，简单易于实现，但对异常点和不平衡数据较为敏感，且结果的可解释性较差。本文基于参数近邻个数  $K$  的取值选择最优模型，近邻数  $K$  与误判率的关系见图6.1。

对于 SVM 方法，本文基于向量分类机类型和核函数的选择优化模型，各模型选择的预测错误个数见表6.4。由表6.4可知，C-classification 分类模型无论采用何种核函数结合，其模型判别错误均相对较小，其中，C-classification 和 radial 核函数结合建立的模型判别错误最少。进一步利用交叉验证对 radial 核函数里的参数（gamma、cost）加以调校，得到 gamma=0.5, cost=10 的最优预测模型。

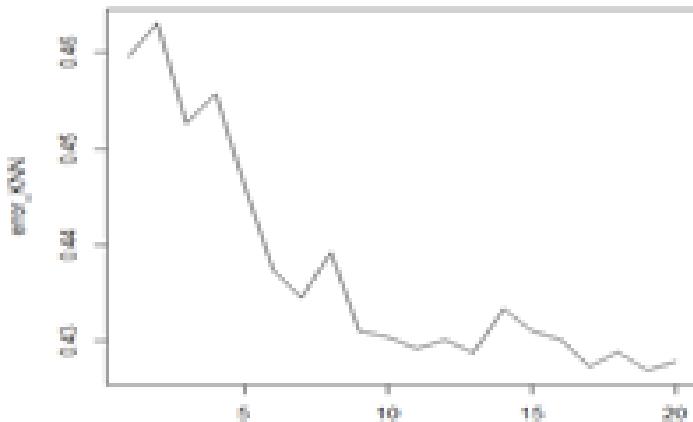
图 6.1 近邻数  $K$  与误判率关系图

表 6.4 各模型选择的预测错误个数

	linear	polynomial	radial	sigmoid
C-classification	3512	1360	987	4457
nu-classification	4077	2662	2800	4660
one-classification	8860	5050	5001	5003

对于神经网络方法，是一种应用类似于大脑神经突触链接的结构进行信息处理的数学运算模型，通过统计学的标准数学方法构建函数表达的局部结构空间。因而容易陷入局部极值，导致过度拟合。本文基于隐藏节点数目和迭代次数选择最优模型，各参数选择与模型误判率的关系见图6.2和图6.3，得到参数隐藏节点数 size=7 和训练周期 maxit=200 的最优模型。

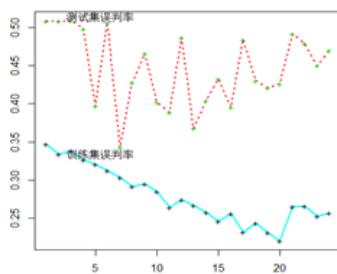
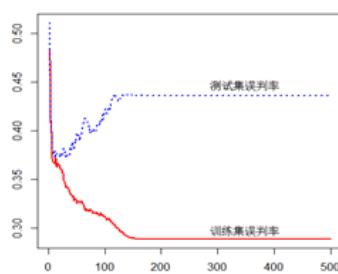


图 6.2 各隐藏节点个数下神经网络模型的误判率  
图 6.3 各训练周期下神经网络模型的误判率



对于随机森林方法，在实例运用过程中，随机森林对高维数据处理的准确性高且速度快，能够对大样本并行化处理，此算法优点与网络信息数据高维、大量的特点相契合。基于 R 语言循环语句设计对模型进行参数调校，并实现结果可视化（图6.4和图6.5）。由决策树节点分支所选择变量个数和决策树数量与模型误判率的关系比较可知，最优参数分别为  $k=5$ 、 $M=400$ ，基于此建立最优随机森林模型，并对测试集数据进行预测。

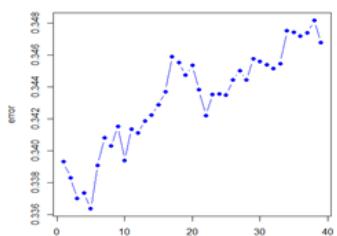
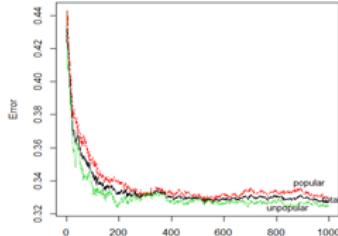


图 6.4 模型误差与决策树节点所选变量个数关系  
图 6.5 各训练周期下神经网络模型的误判率



进一步，各方法经交叉验证后的预测精度及 AUC 值 (Area Under ROC Curve) 见表6.5。

表 6.5 四种方法的预测判别效果

方法	预测准确率	AUC 值
随机森林	67.29%	0.66
SVM	60.83%	0.61
KNN	57.34%	0.58
神经网络	55.28%	0.57

由表6.5比较分析可知，四种分类方法所建立的模型中，随机森林的预测准确率最高，AUC 值最大，分别高达 67.29% 和 0.66，说明随机森林的预测分类性能最好；其次是 SVM，预测准确率比随机森林低 6.46%，AUC 值为 0.61，但实例应用中算法运行效率较低；最后，KNN 和神经网络的预测准确率和 AUC 值相近，二者所建模型的预测分类效果相对较弱，KNN 的算法效率高但因易受邻近个数  $K$  的选择影响从而导致预测效果不稳定，神经网络因输出结果的可解释性较弱而影响其实际应用性。总体而言，在大样本下对于网络文章热度的预测，四种分类方法的预测性能效果存在明显差异，依次为：随机森林 > SVM > KNN > 神经网络，其中，“A>B” 表示 A 的表现优于 B。

## 2、特征变量的重要性评价：

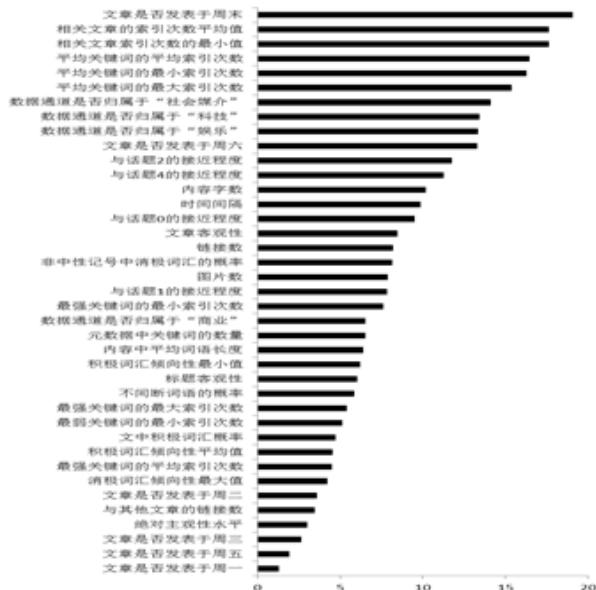


图 6.6 网络文章各特征变量重要性排序

由随机森林输出的基于平均精度减少量的变量重要性程度，可以筛选出对文章热度分类预测影响显著的相关变量。由图6.6可知，文章是否发表于周末、相关文章的索引次数、平均关键词的索引次数以及数据通道来源等相关变量的重要程度较大；其次，内容字数、词语长度以及与热门话题接近程度等相关变量的对文章热度的影响也较为重要；而文章是否发表于非周末时间段和语言极性情感特征等相关变量的整体重要性较小。

### 6.4.2 Python 语言实践

本文将介绍如何使用 Python 已有的 sklearn 机器学习库实现随机森林实例，并在之后引入 Python 编程应用随机森林方法预测病人心脏健康状况的实例。

#### Python 随机森林包

在 Python 的编辑器中，我们可以通过导入 sklearn 模块中的 RandomForestClassifier 和 RandomForestRegressor 来创建随机森林实例，其中，RandomForestClassifier 用于分类任务，RandomForestRegressor 用于回归任务。导入代码如下。

```
from sklearn.ensemble import RandomForestClassifier, RandomForestRegressor
```

接下来我们将说明上述两个随机森林实例的参数及其可选范围。

- **n\_estimators**: int 型，默认值为 100

随机森林中的树数量；

- **criterion**: string 型，{gini, entropy}，默认值为 gini

树分支的分裂准则，支持的准则有 gini——基尼不纯度和 entropy——信息增益；

- **max\_features**: int, float, string 或 None 等等多种类型，默认值为 auto  
寻找最佳分割时需要考虑的特征数目：

如果是 int，就要考虑每一次分割处的 max\_features 个特征；

如果是 float，那么 max\_features 就是一个百分比数值， $(\text{max\_features} * \text{n\_features})$  个特征是每个分割处需要考虑的；

如果是 auto，那么  $\text{max\_features} = \sqrt{\text{n\_features}}$ ，即 n\_features 的平方根值；

如果是 sqrt，同 auto；

如果是 log2，那么  $\text{max\_features} = \log_2(\text{n\_features})$ ；

如果是 None，那么  $\text{max\_features} = \text{n\_features}$ ；

- **max\_depth**: int, None, 默认值为 None

树的最大深度：

如果值为 None，那么会扩展节点，直到所有的叶子是纯净的，或者直到所有叶子包含少于 min\_sample\_split 的样本；

- **min\_samples\_split**: int, float, 默认值为 2

分割内部节点所需要的最小样本数量：

如果是 int，那么 min\_samples\_split 作为每个分割的最小样本数量；

如果是 float, 那么 min\_samples\_split 是一个百分比，并把  $\text{ceil}(\text{min\_samples\_split} * n_{\text{samples}})$  作为每个分割的最小样本数量；

- **min\_samples\_leaf**: int, float, 默认值为 1

每个叶结点上所需的最小样本数量：

如果是 int，那么 min\_samples\_leaf 作为每个叶节点所需的最小样本数量；

如果是 float, 那么 min\_samples\_leaf 是一个百分比，并把  $\text{ceil}(\text{min\_samples\_leaf} * n_{\text{samples}})$  作为每个叶节点所需的最小样本数量；

- **min\_weight\_fraction\_leaf**: float, 默认值为 0

一个叶节点所需的权重总和（所有的输入样本）的最小加权分数，当取默认值时，样本具有相同权重；

- **max\_leaf\_nodes**: int, None, 默认值为 None

以最优方法来生成最大叶节点数为 max\_leaf\_nodes 的树，如果是 None，那么将不限制叶节点的数量；

- **min\_impurity\_decrease**: float, 默认值为 0

如果节点的分裂导致的不纯度的下降程度大于或等于这个节点的值，那么这个节点将会被分裂；

- **bootstrap**: boolean, 默认值为 True

生成树时，是否使用有放回抽样；

- **oob\_score**: boolean, 默认值为 False

是否使用袋外样本来估计泛化精度；

- **n\_jobs**: int, 默认值为 1

用于拟合和预测的并行运行的作业数量。如果值为 -1，那么作业数量将被设置为计算机核的数量；

- **random\_state**: int, RandomState 实例, None, 默认值为 None

如果是 int，那么 random\_state 是随机数生成器使用的种子；

如果是 RandomState 实例，那么 random\_state 是随机数生成器；

如果是 None，那么随机数生成器是 *np.random* 使用的 RandomState 实例；

- **verse**: int, 默认值为 0  
控制树建立过程的冗余度;

- **warn\_start**: boolean, 默认值为 False

当设置为 True 时, 使用之前存在的树来进行拟合和添加更多的树; 反之则拟合一个全新的森林;

我们可以看出 Python 的随机森林实例包含了许多可以调整的超参数, 读者可以通过调整实例中与任务相关的参数, 从而训练出一个较优的随机森林模型。读者也可以访问 sklearn 模块的官方文档 (<https://scikit-learn.org/stable/>) 查看关于随机森林实例更详尽的描述以及其他机器学习实例的相关信息。接下来我们将使用 Python 的随机森林实例训练一个用于分类任务的随机森林模型。

## 数据来源及预处理

数据来源于 kaggle 中 Heart Attack Analysis 数据集, 该数据集收集了 303 位病人的信息, 包含 11 个与患心脏病相关的特征及一个分类标签变量 (即病人心脏病发作的可能性是否更大)。特征集包含病人的年龄、性别、静息血压、最大心率等相关信息, 通过数据预处理, 观察到数据不存在缺失值、异常值、特征集维数较低且都与自变量有较高的相关度。因此我们将不进行特征选择, 保留原始所有数据进行后续的模型训练及评估。数据集的详细描述见表 6.6。

表 6.6 数据集描述

变量名	类型	变量含义
年龄	数值	病人年龄
性别	类别	0: 男性;1: 女性
心绞痛类型	类别	0: 典型;1: 非典型;2: 其他;3: 无症状
静息血压	数值	病人静息血压值 (mm/Hg)
通过 BMI 传感器获取的胆固醇	数值	通过 BMI 传感器获取的胆固醇数值 (mg/dl)
空腹血糖超过 120mg/dl	类别	0: 否;1: 是
静息心电图结果	类别	0: 正常;1:ST-T 波正态性;2: 左心室肥厚
最大心率	数值	病人的最大心率值 (bpm)
运动性心绞痛	类别	0: 否;1: 是
上一个高峰	数值	病人静息心电图上一个高峰值
主血管数量	数值	病人主血管的数量
发作可能性是否更大	类别	0: 否;1: 是

## 数据集划分

将数据集划分为训练集和测试集，考虑到样本数量较少，我们采用随机抽样法从原始数据中抽取 80% 的样本量作为训练集，20% 的样本量作为测试集，各数据集结构见表6.7、6.8。

表 6.7 数据集标签变量分布

标签类别	样本数
更大可能心脏病发作 (1)	165
更小可能心脏病发作 (0)	138

表 6.8 训练集、测试集样本量

训练样本集	242
测试样本集	61

## 结果展示

我们将使用 RandomForestClassifier 包进行分类，将参数设置为其默认参数，在训练集上训练得到一个随机森林分类器，并将训练到的分类器用于测试集预测，计算其精确率、召回率、f1 分数等评价指标，见表6.9

表 6.9 评价指标

指标类别	数值
accuracy_score	0.8852
precision_score	0.8788
recall_score	0.9063
f1_score	0.8923

接下来，我们将使用基尼系数和信息增益两个不同的树节点分裂准则分别去训练随机森林分类器，并将基分类树数量设置在 1-400 区间内，对比其模型效果，见图6.7。由分裂准则和基分类树数量与模型误差关系图可见，最优分裂准则为基尼系数，最优基分类树数量位于 10-40 区间。接下来，我们再将树分裂节点所选特征数、树最大深度两个超参数分别设定在 1-11、1-20 进行超参数全局搜索，得到最优超参数组合，用最优随机森林分类器预测测试集，并画出模型 ROC 曲线，结果见图6.8。

计算各项模型评价指标、混淆矩阵，结果见表6.10、6.11。

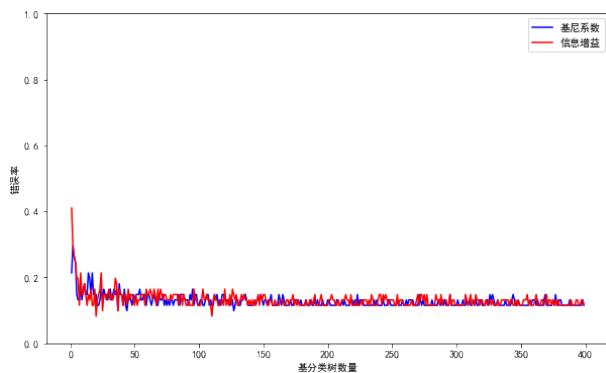


图 6.7 模型误差与基分类树数量、分裂准则关系图

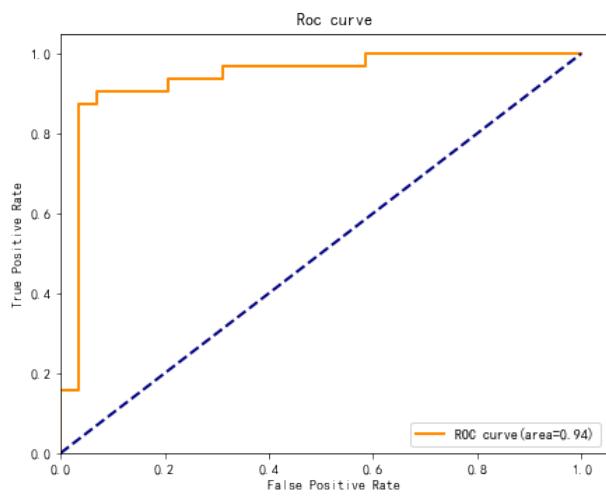


图 6.8 ROC 曲线

表 6.10 评价指标

指标类别	数值
accuracy_score	0.8689
precision_score	0.8529
recall_score	0.9063
f1_score	0.8788
auc	0.9356

表 6.11 混淆矩阵

真实值情况	预测值: 1	预测值: 0
1	24	5
0	3	29

## 6.5 小结

为了简便，对于回归树本节仅介绍了在每一个叶子节点用一个常数来预测所有的因变量，常数预测有一定的缺陷，比如当因变量出现极端值的情况，预测会很不好。为了克服这一问题 Quinlan [79] 提出了在每一个叶子节点建立线性模型来进行预测。关于利用其他的模型来进行预测可以参考 [76] 和 [80]。同理，对于分类树，在每个叶子节点，本节仅介绍了类似 K-近邻的分类方法，我们也可以采用其他的分类方法如：Logistic 回归，线性判别分析，二次判别分析，支持向量机等。

这里介绍的随机森林是专门基于树的结构提出的集成学习方法，类似于集成方法 Bagging (Bootsrap AGGRegatING) [82] 和 Boosting，随机森林的思想也可以推广到其他的统计模型或弱分类器中。Bagging 是利用 bootstrap 抽样来建立单棵树，但是每次都是利用所有  $p$  个预测来建立单棵树。因此，与 Bagging 相比，随机森林具有两个明显的优势，第一，随机森林只用  $k$  个特征，可以节省计算量和计算机内存；第二，当预测中存在高度相关的自变量时，在减少树之间的相关性方面，随机森林具有明显的优势。关于集成学习 Bagging 和 Boosting 更多的介绍，有兴趣的读者可以参考 [71] 和 [70] 两本书。

单棵树的 R 软件包主要有 2 个：rpart 和 party。R 包 rpart 是基于 CART 方法和 rpart 函数；R 包 party 主要是基于条件树函数 ctree。具体的计算代码为：

```
library(rpart)
repartTree<-rpart(y~, data=trainData)
```

```
library(party)
cTree<-ctree(y~., data=trainData)
```

epart 函数中的控制参数可以通过 repart.control 来实施，其中有两个重要的控制参数是修枝时用的复杂度参数 cp 和最大节点数 maxdepth。party 软件包可以通过 ctree\_control 来实施。两个经常用的参数是 mincriterion 和 maxdepth，为了连续分裂 mincriterion 定义的统计标准必须满足，maxdepth 给出树最大深度。以上是在每一个叶子节点利用常数来预测得回归树模型，基于模型得回归树可以利用 RWeka R 软件包来实现。具体的代码为：

```
library(RWeka)
repartTree<-M5P(y~., data=trainData)
```

随机森林的 R 软件包为 randomForest。具体的计算代码为：

```
library(randomForest)
rf<-randomForest(TrainX,TrainY)
或者
Rf<-randomforest(y~., data=trainData)
```

随机森林中有两个重要的参数需要设置分别是：mtry 是建立单棵树的自变量数（本书用  $k$  表示），默认值为  $p/3$ ；ntree 是树的数目，默认值为 500。如果需要输出变量重要性评估图，R 实现语言为：

```
Rf<-randomforest(y~., data=trainData, importance=TRUE).
```

其他的集成学习的 R 软件包括：实现 Bagging 的 R 软件包 ipred；基于树的 Boosting 方法可用 R 软件包 gbm 来实现。

## 6.6 习题

- 1、证明定理6.1。
- 2、集成学习是否可以避免过拟合？
- 3、随机森林的不同基决策树模型差异体现在那些地方？
- 4、随机森林和 Bagging 类方法有什么异同？
- 5、随机森林的投票方式是怎么样的？
- 6、编写程序实现表6.5的内容，并画出各方法的 ROC 曲线。
- 7、用 Logistic 回归代替随机森林中树的分类方法，并与本章所介绍方法进行比较，并讨论各自方法的优缺点。
- 8、模拟或者找一个实际数据  $Y$  是连续随机变量， $\mathbf{X}$  是随机向量，分别用 R 软件包 rpart，party 和 RWeka 等运行这一数据并讨论各算法的优缺点。

9、模拟或者找一个实际数据  $Y$  是连续随机变量,  $\mathbf{X}$  是随机向量, 对这一数据比较 Bagging、Boosting 和随机森林等集成方法的表现, 并总结各自方法的优缺点。

# 第七章 Boosting 方法

## 7.1 简介

如何根据历史数据，建立精确的预测模型，是机器学习方法研究的主要目的之一。例如，我们拟构建一个可以区分垃圾邮件和正常邮件的电子邮件过滤器。通常机器学习方法解决此问题的思路如下：首先收集尽可能多的垃圾电子邮件和非垃圾电子邮件的案例。然后，应用已有的机器学习算法对收集到的邮件和其对应的标签进行训练，建立一个分类准则，基于已知数据建立的准则可以对新的、未加标签的电子邮件进行自动的分类。一个自然的期望是建立的分类（预测）准则能够对新的电子邮件做出精确的预测。

通常，建立一个高度准确的预测或分类准则往往是困难的。但是，建立系列相对准确、比随机猜测稍好的经验法则就容易实现的多。对上例的电子邮件问题，可以建立这样一个简单规则：如果电子邮件中出现“立即购买”的短语，那么就预测它是垃圾邮件。一方面，此规则并不能涵盖所有垃圾邮件；另一方面，这一规则显然要好于随机猜测。对于任何一个遭受过垃圾邮件骚扰的人来说，一些识别垃圾邮件的规则会迅速在我们脑海中闪现。例如，如果邮件里有“购买”、“Viagra”这种词，那么该邮件很可能就是垃圾邮件。但是，根据个人经验总结的规则对于垃圾邮件的正确区分是远远不够的。一方面，例如把出现“购买”的邮件都划分为垃圾邮件，其他邮件都视作正常邮件，那么也会出错。但是另一方面，这些规则也并非毫无用处，至少给我们提供了有价值的信息。虽然其准确率较低，但至少比随机猜测的效果要好。另外，发现这些“弱”的准则也相对容易。

Boosting 方法正是基于这样一种考虑：找到许多粗糙的、比随机猜测稍好的经验准则要比找到单一的、高度精确的预测规则容易得多。这些粗糙的、比随机猜测稍好的经验准则或算法，通常被称为[弱学习算法](#)或者[基学习算法](#)。

[Boosting 算法](#)反复调用这些弱学习算法，每次让其学习训练样本的不同子集。更准确地说，是训练数据的不同分布或权重。每次调用时，基学习算法都会生成一个新的弱学习算法（学习器）。重复建立多个弱学习算法后，Boosting 算法最终将这些弱学习算法（学习器）有效组合成最后的强学习算法（强学习器），强学习器有望比任何一个弱学习器都准确得多。

因此，Boosting 的核心思想就是：对于一个复杂的任务而言，有效的综合多个专家的预测进而所得出的新的预测，要优于其中任何一个专家的单独预测。即通常

所说的“三个臭皮匠顶个诸葛亮”的道理。

Boosting 集成学习方法基本流程如图 7.1 所示。

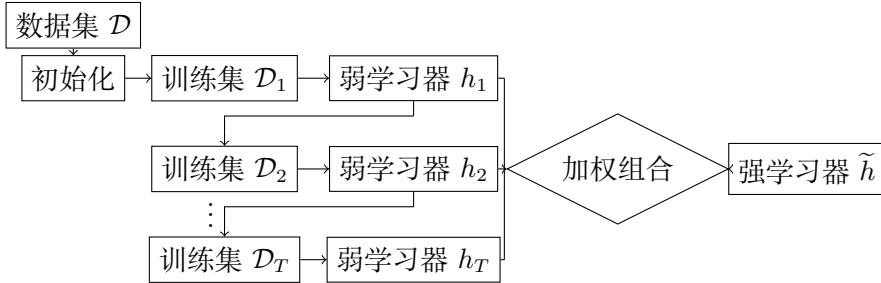


图 7.1 Boosting 集成学习方法基本流程

当然，要使上述 Boosting 方法切实可行，还有以下两个基本问题需要解决：

- 1、在每一轮训练中，如何改变训练数据的权值或概率分布；
- 2、最终，如何将所得到的弱分类器组合成为一个强学习器。

关于第 1 个问题，Boosting 算法家族中的代表性算法 AdaBoost 的处理方式是：提高那些被前一轮弱学习器错误分类样本的权值，而降低那些被正确分类样本的权值。这样一来，那些没有得到正确分类的数据，由于其权值的加大而受到后一轮弱学习器的更大关注。于是得到的系列弱学习器之间具有“互补”的特点。至于第 2 个问题，AdaBoost 算法通常采取加权投票的方法。具体地，增加分类误差率小的弱学习器的权值，使其在表决中起较大作用；减少分类误差率大的弱学习器的权值，使其在表决中起较小的作用。

### 7.1.1 Boosting 方法起源

Boosting 方法起源于一个纯理论性的公开问题，即所谓的 **Boosting 公开问题**。

1984 年，Valiant [83] 在研究 PAC (probably approximately correct) 学习框架时，给出了强可学习(或称可学习)和弱可学习的概念。1988 年，Kearns 和 Valiant [84] 在研究 PAC 学习模型时，针对以上弱可学习与强可学习提出如下公开问题：是否可以将性能仅略好于随机猜测的弱学习算法“Boosting”提升为具有任意精确度的强学习算法？即 Boosting 公开问题。此问题非常重要，因为获得一个弱学习器要比获得一个强学习器要容易的多。如果该问题的答案是肯定的，那么任何弱学习器都有可能被提升为强学习器。这一思想对今后机器学习算法尤其是集成学习的发展产生了深远的影响。

1989 年，Schapire 给出了肯定的答复，在 PAC 框架下，强可学习与弱可学习是等价的，即：一个概念是强可学习的充要条件是这个概念是弱可学习的。同时，他在文章中给出的构造性证明也成为最早的 Boosting 算法 [85]。随后，在 1990 年，Freund

开发了一个效率更高的且具有某种最优性的 Boosting 算法 [86]。Drucker 等人利用这些早期的 Boosting 方法在 OCR (Optical Character Recognition) 任务上进行了首次实验验证 [87]。然而, 由于上述算法需要提前知晓弱学习器的错误率上界, 这通常在实际应用中是未知的。因此上述 Boosting 算法并不具备实际应用性。

Freund 与 Schapire 在随后的研究中发现, “Online” 学习与 Boosting 问题之间存在着极大的共性。他们将其与加权投票的相关研究成果进行融合, 并在 Boosting 问题中进行相应推广, 得到了著名的 AdaBoost 算法。特别地, 此算法不需要提前预知弱学习器的分类精度等相关的任何先验知识, 在实践中获得了极大成功。凭借此工作, Freund 和 Schapire 获得了 2003 年度理论计算机的最高奖——哥德尔奖 (Godel Prize)。AdaBoost 一举成为最具影响力的集成算法之一, 被评为数据挖掘十大算法之一 [88]。

### 7.1.2 AdaBoost 算法

#### AdaBoost 通用算法

AdaBoost 算法可以做出非常精准的预测, 其过程却非常简单。现举例如下:

以二元问题为例, 设  $\mathcal{X}$  为自变量组成的样本空间, 其中的样本都是从分布  $\mathcal{D}$  中随机抽取, 且满足独立同分布性; 其学习目标函数记为  $Y$ 。设  $\mathcal{X}$  由  $\mathcal{X}_1$ 、 $\mathcal{X}_2$  和  $\mathcal{X}_3$  三部分组成, 每个部分占  $1/3$ , 通过随机猜测工作的学习器在这个问题上有  $0.5$  的分类误差。我们想在这个问题上得到一个精确的 (例如零误差) 学习器。可借助的只有一个弱学习器  $h_1$ , 它在样本空间  $\mathcal{X}_1$  和  $\mathcal{X}_2$  中有正确的分类, 在  $\mathcal{X}_3$  有错误的分类。如何将此弱学习器  $h_1$  “Boosting” 为强学习器呢?

一个自然的想法就是纠正  $h_1$  所犯的错误。首先, 通过  $\mathcal{D}$  派生出新的分布  $\mathcal{D}_1$ 。例如通过提高那些被  $h_1$  错误学习样本的权值, 降低那些被  $h_1$  正确预测的样本的权值诱导出新分布  $\mathcal{D}_1$ 。显然在  $\mathcal{D}_1$  上,  $h_1$  的错误被彰显。然后用  $\mathcal{D}_1$  训练得到学习器  $h_2$ 。此时得到的学习器  $h_2$  极有可能也是一个弱学习器。假设它在  $\mathcal{X}_1$  和  $\mathcal{X}_3$  中有正确的预测, 在  $\mathcal{X}_2$  有错误的预测。通过以某种适当的方式组合  $h_1$  和  $h_2$ , 组合的学习器将在  $\mathcal{X}_1$  中具有正确的预测, 并且可能在  $\mathcal{X}_2$  和  $\mathcal{X}_3$  中仍有错误。类似的, 为了使组合学习器的错误彰显, 我们再次派生出一个新的分布  $\mathcal{D}_2$ , 并从  $\mathcal{D}_2$  训练出新的学习器  $h_3$ , 使  $h_3$  在  $\mathcal{X}_2$  和  $\mathcal{X}_3$  有正确的预测。最后, 通过组合  $h_1$ 、 $h_2$  和  $h_3$ , 就得到一个强学习器, 因为在  $\mathcal{X}_1$ 、 $\mathcal{X}_2$  和  $\mathcal{X}_3$  的每个空间中, 至少有两个学习器是正确的。

简而言之, Boosting 方法就是顺序训练一族弱学习器, 并将它们组合以形成强学习器来进行预测。训练过程中, 让后建立的学习器更多地关注前序学习器的错误预测样本。通用 Boosting 算法如下:

---

### 通用 Boosting 算法

---

**输入:** 样本权值分布  $\mathcal{D}$ ;

基学习算法  $\mathcal{L}$ ;

学习轮数  $T$ .

**过程:**

1.  $\mathcal{D}_1 = \mathcal{D}$ ; % 初始化分布
2. **for**  $t = 1, \dots, T$ :
3.  $h_t = \mathcal{L}(\mathcal{D}_t)$ ; % 依分布训练弱学习器
4.  $\text{err}_{\mathcal{D}_t} = \Pr_{\mathcal{D}_t}(h_t(\mathbf{X}) \neq Y)$ ; % 度量误差
5.  $\mathcal{D}_{t+1} = \text{Adjust\_Distribution}(\mathcal{D}_t, \text{err}_{\mathcal{D}_t})$ ; % 更新分布
6. **end**

**输出:**  $\mathcal{H}(\mathbf{X}) = \text{Combine\_Outputs}(\{h_1(\mathbf{X}), \dots, h_T(\mathbf{X})\})$

---

### AdaBoost 算法

---

**输入:** 训练数据集  $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$ ;

基学习算法  $\mathcal{L}$ ;

学习轮数  $T$ .

**过程:**

1.  $\mathcal{D}_1(\mathbf{X}_i) = 1/n$ ; % 初始化权重分布
2. **for**  $t = 1, \dots, T$ :
3.  $h_t(\mathbf{X}) = \mathcal{L}(\mathcal{D}, \mathcal{D}_t)$ , 且  $h_t(\mathbf{X}) : \mathcal{X} \rightarrow \{-1, 1\}$ ; % 依分布训练输出弱学习器
4.  $\text{err}_{\mathcal{D}_t} = \sum_{i=1}^n \Pr_{\mathbf{X}_i \sim \mathcal{D}_t}(h_t(\mathbf{X}_i) \neq Y_i)$ ; % 度量误差
5. **if**  $\text{err}_{\mathcal{D}_t} \geq 0.5$  **then break**
6.  $\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \text{err}_{\mathcal{D}_t}}{\text{err}_{\mathcal{D}_t}} \right)$ ; % 计算弱学习器  $h_t$  的权重
7.  $\mathcal{D}_{t+1}(\mathbf{X}_i) = \frac{\mathcal{D}_t(\mathbf{X}_i)}{Z_t} \exp(-\alpha_t Y_i h_t(\mathbf{X}_i))$ , 其中  
 $Z_t = \sum_{i=1}^n \mathcal{D}_t(\mathbf{X}_i) \exp(-\alpha_t Y_i h_t(\mathbf{X}_i))$ ; % 更新分布,  $Z_t$  为规范化因子
8. **end**

**输出:**  $\mathcal{H}(\mathbf{X}) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(\mathbf{X}))$

---

### AdaBoost 具体算法

对上述 Boosting 通用算法, 将  $\text{Adjust\_Distribution}(\cdot, \cdot)$  和  $\text{Combine\_Outputs}(\cdot, \dots, \cdot)$  取成不同的形式, 则会衍生出各种类型的 Boosting 算法, 如 AdaBoost.M1、

AdaBoost.MR、FilterBoost、GentleBoost、GradientBoost、LogitBoost 等，进而形成庞大的 Boosting 算法族。我们下面重点介绍 AdaBoost 算法。

假设  $\mathcal{X} \subset \mathbf{R}^{n \times p}$  是自变量生成的样本空间， $\mathcal{Y}$  是标签集，通常假设  $\mathcal{Y} = \{-1, 1\}$ 。给定一个二元训练数据集  $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$  每个训练数据由自变量样本与标签组成，其中  $\mathbf{X}_i \in \mathcal{X}$  为自变量样本空间， $Y_i \in \mathcal{Y}$  为标签， $i = 1, 2, \dots, n$ 。AdaBoost 算法如上。

**注 7.1** 上述算法通常被称为“离散型 AdaBoost”，因为弱学习器  $h_t(\mathbf{X})$  返回离散的标签。如果弱学习器返回的预测值为连续实数，则可以对 AdaBoost 做恰当的修改，参见文献 [89]。此外，上述算法中的关键量  $h_t$  和  $\alpha_t$  的选取原理将在下一章给出具体的解释。

**注 7.2** 对上述 AdaBoost 算法做一些简单说明：

- A. 为保证第 1 步能够输出基本学习器  $h_1(\mathbf{X})$ ，我们需要假设原始训练数据具有已知的权值分布。为简单起见，通常假定训练数据具有均匀分布。
- B. 根据已有的符号表示，可以将训练数据权值的分布写成如下形式：

$$\mathcal{D}_t = (w_1^{(t)}, w_2^{(t)}, \dots, w_i^{(t)}, \dots, w_n^{(t)}), \quad t = 1, 2, \dots, T, \quad (7.1.1)$$

$$\text{其中, } w_i^{(1)} = \frac{1}{n}, i = 1, 2, \dots, n.$$

- C. 计算基本学习器  $h_t(\mathbf{X})$  在分布为  $\mathcal{D}_t$  的训练数据集上的误差率：

$$\begin{aligned} \text{err}_{\mathcal{D}_t} &= \sum_{i=1}^n \Pr_{\mathbf{X}_i \sim \mathcal{D}_t} (h_t(\mathbf{X}_i) \neq Y_i) \\ &= \sum_{h_t(\mathbf{X}_i) \neq Y_i} w_i^{(t)} \end{aligned} \quad (7.1.2)$$

此式表明弱学习器  $h_t(\mathbf{X})$  在加权训练数据集上的分类误差等于被  $h_t(\mathbf{X})$  误分的样本权值之和。

- D. 上述算法给出弱学习器的系数表达式为

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \text{err}_{\mathcal{D}_t}}{\text{err}_{\mathcal{D}_t}} \right) \quad (7.1.3)$$

显然当  $\text{err}_{\mathcal{D}_t} \leq 0.5$  时，其系数  $\alpha_t \geq 0$ ，且  $\alpha_t$  为  $\text{err}_{\mathcal{D}_t}$  的减函数。所以误差率  $\text{err}_{\mathcal{D}_t}$  越小的学习器其权重  $\alpha_t$  就越大。

- E. 根据算法中的第 7 个式子以及式 (7.1.1) 可知，更新权值分布可以写成：

$$w_i^{(t+1)} = \frac{w_i^{(t)}}{Z_t} \exp(-\alpha_t Y_i h_t(\mathbf{X}_i))$$

$$= \begin{cases} \frac{w_i^{(t)}}{Z_t^{(t)}} e^{-\alpha_t}, & h_t(\mathbf{X}_i) = Y_i \\ \frac{w_i^{(t)}}{Z_t^{(t)}} e^{\alpha_t}, & h_t(\mathbf{X}_i) \neq Y_i \end{cases} \quad (7.1.4)$$

其中

$$\begin{aligned} Z_t &= \sum_{i=1}^n \mathcal{D}_t(\mathbf{X}_i) \exp(-\alpha_t Y_i h_t(\mathbf{X}_i)) \\ &= \sum_{i=1}^n w_i^{(t)} \exp(-\alpha_t Y_i h_t(\mathbf{X}_i)) \\ &= \sum_{Y_i=h_t(\mathbf{X}_i)} w_i^{(t)} e^{-\alpha_t} + \sum_{Y_i \neq h_t(\mathbf{X}_i)} w_i^{(t)} e^{\alpha_t} \end{aligned} \quad (7.1.5)$$

式 (7.1.4) 说明, 如果样本  $\mathbf{X}_i$  被正确学习, 即  $h_t(\mathbf{X}_i) = Y_i$ , 则其权重会减小; 如果  $\mathbf{X}_i$  被错误学习, 即  $h_t(\mathbf{X}_i) \neq Y_i$ , 则其权重会增大。这样, 误分学习样本在下一轮学习中就会受到基学习器的更大关注而起到更大作用。

#### F. 记

$$\tilde{h}_T(\mathbf{X}) = \sum_{t=1}^T \alpha_t h_t(\mathbf{X}) \quad (7.1.6)$$

则 AdaBoost 的最后输出结果为

$$\mathcal{H}(\mathbf{X}) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(\mathbf{X})) \quad (7.1.7)$$

此式表明对于任何数据的分类采用加权投票表决的方法, 其中  $\mathcal{H}(\mathbf{X})$  的符号决定数据  $\mathbf{X}$  的类。需要注意的是, 基本学习器  $h_t(\mathbf{X})$  的系数  $\alpha_t$  未必具有归一性。

#### G. AdBoost 具备以下两个特点:

- (a) AdBoost 不更新训练数据, 只是不断更新训练数据的权值分布, 使得训练数据在基本分类器的学习中起不同的作用。
- (b) AdBoost 最终分类器  $\tilde{h}_T(\mathbf{X})$  的构成是利用基本分类器  $h_t$ ,  $t = 1, 2, \dots, T$  的线性组合构建, 因而可视其为加法模型。

### 7.1.3 AdaBoost 实例

**例 7.1** 给定如表 7.1 所示训练数据。假设弱学习器 (或称阈值函数) 由  $X < v$  或  $X > v$  产生, 其阈值  $v$  使该学习器在训练数据集上分类误差率最低。试用 AdBoost 算法学习一个强学习器。

表 7.1 训练数据表

序号	1	2	3	4	5	6	7	8	9	10
$X$	1	2	3	4	5	6	7	8	9	10
$Y$	1	1	1	-1	1	-1	-1	1	-1	-1

解：初始化数据权值分布

$$\mathcal{D}_1 = (w_1^{(1)}, w_2^{(1)}, \dots, w_n^{(1)})$$

$$w_i^{(1)} = 0.1, \quad i = 1, 2, \dots, 10$$

1、对  $t = 1$ :

(1) 在权值分布为  $\mathcal{D}_1$  的训练数据上，阈值  $v$  取 3.5 时分类误差率最低，故基本分类器为

$$h_1(X) = \begin{cases} 1, & X < 3.5 \\ -1, & X \geq 3.5 \end{cases}$$

(2)  $h_1(X)$  在训练数据集上的误差率:  $\text{err}_{\mathcal{D}_1} = \sum_{i=1}^n \Pr_{X_i \sim \mathcal{D}_1} (h_1(X_i) \neq Y_i) = 0.2$

(3) 计算  $h_1(X)$  的系数:  $\alpha_1 = \frac{1}{2} \ln \frac{1 - \text{err}_{\mathcal{D}_1}}{\text{err}_{\mathcal{D}_1}} = 0.6931$

(4) 更新训练数据的权值分布:

$$\mathcal{D}_2 = (w_1^{(2)}, w_2^{(2)}, \dots, w_n^{(2)})$$

$$w_i^{(2)} = \frac{w_i^{(1)}}{Z_1} \exp(-\alpha_1 Y_i h_1(X_i)), \quad i = 1, 2, \dots, 10$$

$$\mathcal{D}_2 = (0.0625, 0.0625, 0.0625, 0.0625, 0.2500, 0.0625, \\ 0.0625, 0.2500, 0.0625, 0.0625)$$

$$\tilde{h}_1(X) = 0.6931 h_1(X)$$

分类器  $\text{sign}(\tilde{h}_1(X))$  在训练数据集上有 2 个误分类点。

2、对  $t = 2$ :

(1) 在权值分布为  $\mathcal{D}_2$  的训练数据上，阈值  $v$  是 8.5 时分类误差率最低，基本分类器为

$$h_2(X) = \begin{cases} 1, & X < 8.5 \\ -1, & X \geq 8.5 \end{cases}$$

- (2)  $h_2(X)$  在训练数据集上的误差率  $\text{err}_{\mathcal{D}_2} = 0.1875$
- (3) 计算  $\alpha_2 = 0.7332$
- (4) 更新训练数据权值分布:

$$\begin{aligned}\mathcal{D}_3 = & (0.0385, 0.0385, 0.0385, 0.1667, 0.1539, 0.1667, \\ & 0.16667, 0.1539, 0.0385, 0.0385)\end{aligned}$$

$$\tilde{h}_2(X) = 0.6931h_1(X) + 0.7332h_2(X)$$

分类器  $\text{sign}(\tilde{h}_2(X))$  在训练数据集上有 3 个误分类点。

3、对  $t = 3$ :

- (1) 在权值分布为  $\mathcal{D}_3$  的训练数据上, 阈值  $v$  是 5.5 时分类误差率最低, 基本分类器为

$$h_3(X) = \begin{cases} 1, & X \geq 5.5 \\ -1, & X < 5.5 \end{cases}$$

- (2)  $h_3(X)$  在训练数据集上的误差率  $\text{err}_{\mathcal{D}_3} = 0.1820$
- (3) 计算  $\alpha_3 = 0.7514$
- (4) 更新训练数据权值分布:

$$\begin{aligned}\mathcal{D}_4 = & (0.125, 0.125, 0.125, 0.102, 0.102, 0.102, \\ & 0.065, 0.065, 0.065, 0.125)\end{aligned}$$

$$\tilde{h}_3(X) = 0.4236h_1(X) + 0.6496h_2(X) + 0.7514h_3(X)$$

分类器  $\text{sign}(\tilde{h}_3(X))$  在训练数据集上有 0 个误分类点。

于是最终分类器为:

$$\mathcal{H}(X) = \text{sign}(\tilde{h}_3(X)) = \text{sign}\{0.4236h_1(X) + 0.6496h_2(X) + 0.7514h_3(X)\}$$

## 7.2 AdaBoost 算法的误差分析

### 7.2.1 AdaBoost 算法的训练误差

AdaBoost 最基本的理论性质涉及其减少训练误差的能力, 即训练集上的误差率。Freund 和 Schapire 给出了误差界 [90]:

**定理 7.1** 若 AdaBoost 在每一轮迭代中生成弱分类器  $h_t$  的误差率是  $\text{err}_{\mathcal{D}_t}$ ,  $t = 1, 2, \dots, T$ . 则最终分类器  $\mathcal{H}(\mathbf{X})$  的训练误差率  $\text{err}_{\mathcal{D}} = \sum_{i=1}^n \Pr_{\mathbf{X}_i \sim \mathcal{D}} (\mathcal{H}(\mathbf{X}_i) \neq Y_i)$  满足

$$\text{err}_{\mathcal{D}} \leq 2^T \prod_{i=1}^T \sqrt{\text{err}_{\mathcal{D}_t}(1 - \text{err}_{\mathcal{D}_t})}$$

此定理的证明具有一定的技巧性与特殊性, 因而在后续研究中极少被采用。随后 Schapire 和 Singer 进行了更深入研究 [91], 得到如下结论:

**定理 7.2** AdaBoost 算法最终分类器的训练误差满足

$$\frac{1}{n} \sum_{i=1}^n I(\mathcal{H}(\mathbf{X}_i) \neq Y_i) \leq \frac{1}{n} \sum_{i=1}^n \exp(-Y_i \tilde{h}_T(\mathbf{X}_i)) = \prod_{t=1}^T Z_t \quad (7.2.1)$$

其中  $I()$  是示性函数,  $Z_t$ ,  $\tilde{h}_T(\mathbf{X})$  和  $\mathcal{H}(\mathbf{X})$  分别见式 (7.1.5), (7.1.6) 和 (7.1.7)。

**证明** 当  $\tilde{h}_T(\mathbf{X}_i) \neq Y_i$  时,  $Y_i \tilde{h}_T(\mathbf{X}_i) < 0$ , 因而  $\exp(-Y_i \tilde{h}_T(\mathbf{X}_i)) \geq 1$ 。由此可知 (7.2.1) 的不等式成立。

下面证明 (7.2.1) 中等式是成立的。首先, 由式 (7.1.4) 可知:

$$w_i^{(t)} \exp(-\alpha_t Y_i \tilde{h}_t(\mathbf{X}_i)) = Z_t w_i^{(t+1)} \quad i = 1, 2, \dots, n.$$

根据上式得到:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n \exp(-Y_i \tilde{h}_T(\mathbf{X}_i)) &= \frac{1}{n} \sum_{i=1}^n \exp \left( - \sum_{t=1}^T \alpha_t Y_i h_t(\mathbf{X}_i) \right) \\ &= \sum_{i=1}^n w_i^{(1)} \prod_{t=1}^T \exp(-\alpha_t Y_i h_t(\mathbf{X}_i)) \\ &= Z_1 \sum_{i=1}^n w_i^{(2)} \prod_{t=2}^T \exp(-\alpha_t Y_i h_t(\mathbf{X}_i)) \\ &= Z_1 Z_2 \sum_{i=1}^n w_i^{(3)} \prod_{t=3}^T \exp(-\alpha_t Y_i h_t(\mathbf{X}_i)) \\ &= \dots \\ &= Z_1 Z_2 \cdots Z_{T-1} \sum_{i=1}^n w_i^{(T)} \exp(-\alpha_T Y_i h_T(\mathbf{X}_i)) \end{aligned} \quad (7.2.2)$$

再由式 (7.1.5) 中  $Z_t$  的定义可知:

$$Z_T = \sum_{i=1}^n w_i^{(T)} \exp(-\alpha_T Y_i h_T(\mathbf{X}_i)) \quad (7.2.3)$$

证毕。 □

**注 7.3** 1、定理7.2的左边是训练误差，右边是所有归一化因子  $Z_t$  ( $t = 1, 2, \dots, T$ ) 的乘积。通过上式可以知道，要想得到对训练样本拟合精度高的强学习器，需要选择弱学习器  $h_t(\mathbf{X})$  以及其权值  $\alpha_t$ ，使  $\prod_{t=1}^T Z_t$  最小化。虽然上式给出的是相对松弛的训练误差界，但因其更好的解释性和可操作性，可得到更广泛的应用。

2、要实现  $\prod_{t=1}^T Z_t$  的最小化，每加入一个新的弱学习器，都可能要修改已有弱学习器的集成方式，其复杂度太高。AdaBoost 的思想是不改变已有预测准则的形式，以线性加和的方式加入新的弱学习器，只最小化当前迭代的归一化因子  $Z_t$ 。

### 定理 7.3

$$\prod_{t=1}^T Z_t = 2^T \prod_{t=1}^T \sqrt{\text{err}_{\mathcal{D}_t}(1 - \text{err}_{\mathcal{D}_t})} = \prod_{t=1}^T \sqrt{1 - 4\gamma_t^2} \leq \exp(-2 \sum_{t=1}^T \gamma_t^2) \quad (7.2.4)$$

其中  $\gamma_t = \frac{1}{2} - \text{err}_{\mathcal{D}_t}$ ，称为  $h_t$  的 **边界** (edge)。

本定理证明过程略，感兴趣的读者可参考 [92]。

**推论 7.1** 如果存在  $\gamma > 0$ ，对所有  $t$  有  $\gamma_t \geq \gamma$ ，则

$$\frac{1}{n} \sum_{i=1}^n I(\mathcal{H}(\mathbf{X}_i) \neq Y_i) \leq \exp(-2T\gamma^2) \quad (7.2.5)$$

因为  $\text{err}_{\mathcal{D}_t} = \frac{1}{2} - \gamma_t$ ，说明边界  $\gamma_t$  度量第  $t$  个弱学习器  $h_t$  的误差率比随机猜测的错误率好。

此推论告诉我们这样一个事实：只要弱学习器的学习能力比随机猜测稍好 ( $\gamma_t = \frac{1}{2} - \text{err}_{\mathcal{D}_t} \geq \gamma > 0$ )，那么通过 AdaBoost 输出的强学习器  $\mathcal{H}$  的训练误差就是随着轮数  $T$  呈指数下降的。例如令  $\gamma = 0.1$ ，即所有弱学习器的误差率不超过 0.4，则上式表明强预测准则  $\mathcal{H}$  的训练误差率至多是

$$\left( \sqrt{1 - 4 \times 0.1^2} \right)^T \approx (0.9798)^T$$

但其实 AdaBoost 算法及其分析并不需要知道这个下界  $\gamma$ ，而且能适应弱学习器各自的训练误差率，由此获得了自适应 Boosting 的名称，即 AdaBoost (Ada 是 Adaptive 的简写)。如果某些  $\gamma_t$  很大，那么训练误差界的减少将会更大。

此外，通过此推论可知：若想使训练误差率  $\text{err}_{\mathcal{D}} \leq \epsilon$ ，则可让训练轮数  $T$  为

$$\left( \frac{1}{2\gamma^2} \ln \frac{1}{\epsilon} \right)$$

上述式子说明可以通过有限个学习器就可以使得训练误差足够的小。

### 7.2.2 AdaBoost 算法的泛化误差

定理 7.1, 7.2 和 7.3 讨论了 AdaBoost 算法训练误差的下界。然而，在机器学习中我们真正关心的是它在测试数据上的泛化能力。事实上，任一种能够降低训练误差的算法不一定有资格作为 Boosting 算法。Boosting 算法是一种可以使泛化误差 (generalization error) 任意接近零的算法。直观来说，它是一种对测试数据具有接近完美的预测能力的学习算法。当然，以上结论的成立需要对算法提供合理规模的训练样本，所使用的弱学习算法也一直有比随机猜测好的弱学习器。

许多实验都表明 AdaBoost 算法在迭代次数很高时似乎不容易出现过拟合。特别地，在训练误差率降到 0 以后，继续增大学习轮数  $T$ ，AdaBoost 的测试误差率在某种程度上仍在降低。例如，Schapire 绘制了效果曲线，如图 7.2 (a) 所示。对于 AdaBoost 算法，在最初几轮训练误差率已降至 0 后的很长一段时间内，测试误差率仍持续降低。从表面上看，随着更多子分类器的加入，集成分类器形式趋于复杂，分类精度却仍在提高，似乎违背了科学研究中的“奥卡姆剃刀”原理。因此，如何解释 AdaBoost 为什么不容易过拟合成为了 AdaBoost 算法中最迷人的理论问题，吸引了大量关注。

为了解释这一现象，分析 AdaBoost 算法的强泛化能力，Schapire 将统计学习中分类间隔（例如支持向量机中定义的间隔）的相关分析理论引入 AdaBoost 算法 [93]。由此，Schapire et al. 指出：AdaBoost 之所以没有过拟合，是因为在训练误差达到 0 后，强分类器随着学习轮数增加，其间隔还在增大，因而泛化误差仍在减小。这也成为了目前最流行的分析方法。

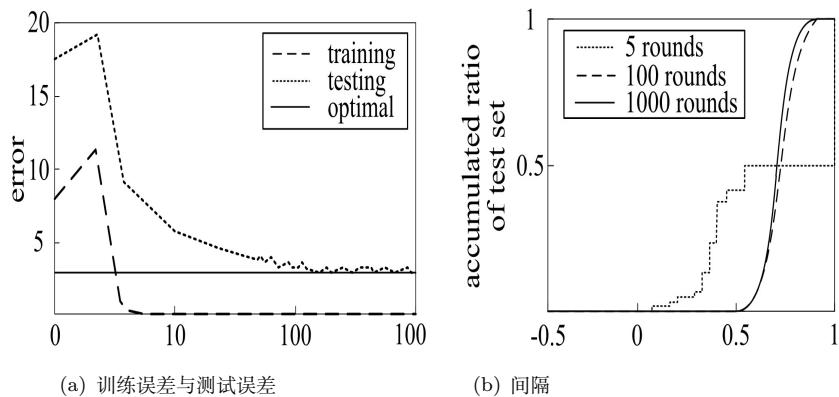


图 7.2 AdaBoost 的间隔解释

根据泛化误差依赖于最小间隔  $\theta$  的值。我们可以最大化最小间隔来得到更紧的泛化误差。这也正是 Breiman 设计 Arc-gv 算法的主要思想。在每轮迭代中，Arc-gv

根据上述思想更新  $\alpha_t$  形式如下：

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 + \gamma_t}{1 - \gamma_t} \right) - \frac{1}{2} \ln \left( \frac{1 + \theta_t}{1 - \theta_t} \right)$$

其中  $\gamma_t = \frac{1}{2} - \text{err}_{\mathcal{D}_t}$ ,  $\theta_t$  是截止当前学习轮次的组合分类器的最小间隔。

与 AdaBoost 相比, Arc-gv 能够得到更大的最小分类间隔以及更好的分类间隔分布。而且, 基于间隔理论, Breiman 给出了更紧的泛化误差界 [94]。

按照理论分析, Arc-gv 具有更强的泛化能力, Arc-gv 应当比 AdaBoost 性能要好。然而 Breiman 在实验中发现: 尽管 Arc-gv 能够得到比 AdaBoost 更大的最小间隔, 但 Arc-gv 的测试误差率却总是高于 AdaBoost。因此, Breiman 对 Schapire 的间隔分析理论提出了质疑, 使得 AdaBoost 的间隔分析方法受到了极大的挑战。

随后的几年, 虽然有诸多学者给出了 AdaBoost 更精致的泛化误差界, 并指出其实是分类间隔分布影响着 AdaBoost 的泛化误差界。但是关于 Breiman 的质疑, 都没有给出正面回答。七年后, Reyzin 和 Schapire 发现了一个有趣的现象。考虑到泛化误差界与间隔  $\theta$ 、学习轮数  $T$  以及子学习器的复杂度相关, 因此为了研究间隔的影响, 需要固定另外的两个因素。Breiman 在比较 Arc-gv 和 AdaBoost 时, 在实验中以 CART 为弱分类器, 通过固定叶节点数目来控制弱分类器复杂度。然而, Reyzin 和 Schapire 却进一步发现, 当叶节点数目相同时, Arc-gv 与 AdaBoost 生成的决策树有着很大的不同 [95]。Arc-gv 生成的树通常拥有较大的深度, 而 AdaBoost 生成的树则宽度较大。一般情况下, 更深的决策树由于进行了更多分裂, 可能具有更大的模型复杂度。在这种情况下比较 Arc-gv 与 AdaBoost 有失公平。Reyzin 和 Schapire 重新进行了 Breiman 的实验, 但使用复杂度相同的决策树作为基分类器。此时, AdaBoost 比 Arc-gv 的间隔分布要优。虽然 Reyzin 和 Schapire 指出分类间隔分布是影响算法泛化能力的关键, 但并没有给出比 Breiman 更精致地泛化误差界。

为彻底解决此类问题, 需要通过分类间隔分布给出 AdaBoost 算法更精致的泛化误差界。而且, 要想正面回答 Breiman 的问题, 这个泛化误差界应该比 Breiman 给出的基于最小分类间隔的泛化误差界更精致。也就是说, 是分类间隔分布, 而不是最小分类间隔决定了算法的泛化能力。2008 年, 王立威等构造了一种与分类间隔分布有关, 与最小分类间隔几乎无关的平衡分类间隔 (Equilibrium margin, Emargin), 给出了基于平衡分类间隔更精致的 Emargin Bounds [96]。通过实验证明: 与 Arc-gv 相比, AdaBoost 有着更大的平衡分类间隔以及更低的测试误差率, 实现了实验测试与理论分析结果的一致。但是证明过程中考虑了比 Breiman 和 Schapire 模型更多的信息, 因而无法直接说明平衡分类间隔比最小间隔更本质。

为更进一步解决这个问题, 2013 年, 周志华进行了更深入的研究 [97]。首先, 他们引入了第  $k$  间隔界 (the  $k$ th margin bound), 并且研究其与最小间隔界与 Emargin 界的关系。然后, 通过改进 Bernstein 界最终给出了更好的泛化误差界。总之, 周志华他们的结果不仅正面回答了 Breiman 的疑问, 合理解释了 AdaBoost 不会过拟合的原因, 而且进一步巩固了间隔分析理论知识结构。

## 7.3 AdaBoost 算法原理探析

AdaBoost 算法最初的方法融合了“Online”学习与加权投票的思想。但随着研究的深入，不同学者从不同的视域提出了各种新的理论模型，越来越多的将 Boosting 算法设计与优化理论联系起来，从不同的视角解释 AdaBoost 算法的原理和有效性。这些不同的视域分析也为新算法的设计提供了更广阔的思路。本章将分别从几种主流的视域对 AdaBoost 算法进行分析，并对其算法的原理进行讨论。

### 7.3.1 损失函数最小化视域

多数统计与机器学习问题都可视为对目标函数或者损失函数的优化。例如，最简单的一元线性回归问题。给定样本  $\{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ 。最小二乘估计的目标是找到参数  $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ ，使其最小化残差平方和

$$\text{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^n (Y_i - \beta_1 - \mathbf{X}_i^T \boldsymbol{\beta}_2)^2.$$

其中， $\text{RSS}(\boldsymbol{\beta})$  就是损失函数。许多其他方法，如神经网络、最大似然估计、支持向量机、Logistic 回归等，都可视为某种程度下的目标函数优化问题。

现在的问题是：AdaBoost 是否也可视为是最优化目标函数的过程？后面的研究表明，AdaBoost 确实可视为优化某种损失函数的一种算法。

#### 指数损失函数

那么 AdaBoost 关联的损失函数如何定义呢？我们前面得到分类器的误差率为

$$\frac{1}{n} \sum_{i=1}^n I\{\mathcal{H}(\mathbf{X}_i) \neq Y_i\} \quad (7.3.1)$$

那么 AdaBoost 算法就是最小化 (7.3.1) 所示的目标函数吗？其实并不是，而是最小化式 (7.3.1) 的一个上界。因为  $\mathcal{H}(X) = \text{sign}(\tilde{h}_T(\mathbf{X}))$ ，则

$$\frac{1}{n} \sum_{i=1}^n I\{\mathcal{H}(\mathbf{X}_i) \neq Y_i\} = \frac{1}{n} \sum_{i=1}^n I\{\text{sign}(\tilde{h}_T(\mathbf{X}_i)) \neq Y_i\} = \frac{1}{n} \sum_{i=1}^n I\{\tilde{h}_T(\mathbf{X}_i) Y_i \leq 0\}. \quad (7.3.2)$$

利用  $I\{x \leq 0\} \leq e^{-x}$ ，则知误差率上界可取为

$$\mathcal{L}_{\text{exp}}(\tilde{h} | \mathcal{D}) = \mathbb{E}_{\mathbf{X} \sim \mathcal{D}}[e^{-Y \tilde{h}_T(\mathbf{X})}] \quad (7.3.3)$$

其中,  $Y\tilde{h}_T(\mathbf{X})$  称为假定问题的分类间隔。

要使上述损失函数达到最小, 关键要解决以下两个问题:

1、如何确定一系列的弱学习器  $h_t$ ?

2、如何确定合适的权重  $\alpha_t$ ?

利用式 (7.1.6) 可知, 只需求

$$\mathcal{L}_{\text{exp}}(\alpha_t h_t | \mathcal{D}_t) = \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_t} [e^{-Y\alpha_t h_t}]. \quad (7.3.4)$$

在分布  $\mathcal{D}_t$  下关于权重  $\alpha_t$  和学习器  $h_t$  的最小值。

$$\begin{aligned} & \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_t} [e^{-Y\alpha_t h_t}] \\ &= \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_t} [e^{-\alpha_t} I(Y = h_t(\mathbf{X})) + e^{\alpha_t} I(Y \neq h_t(\mathbf{X}))] \\ &= e^{-\alpha_t} \Pr_{\mathbf{X} \sim \mathcal{D}_t} (Y = h_t(\mathbf{X})) + e^{\alpha_t} \Pr_{\mathbf{X} \sim \mathcal{D}_t} (Y \neq h_t(\mathbf{X})) \\ &= e^{-\alpha_t} (1 - \text{err}_{\mathcal{D}_t}) + e^{\alpha_t} \text{err}_{\mathcal{D}_t} \end{aligned} \quad (7.3.5)$$

其中  $\text{err}_{\mathcal{D}_t} = \Pr_{\mathbf{X} \sim \mathcal{D}_t} (Y \neq h_t(\mathbf{X}))$ 。为得到最优  $\alpha_t$ , 将指数损失函数求导令其为零, 则

$$\frac{\partial \mathcal{L}_{\text{exp}}(\alpha_t h_t | \mathcal{D}_t)}{\partial \alpha_t} = -e^{-\alpha_t} (1 - \text{err}_{\mathcal{D}_t}) + e^{\alpha_t} \text{err}_{\mathcal{D}_t} = 0 \quad (7.3.6)$$

解得

$$\alpha_t = \frac{1}{2} \ln \left( \frac{1 - \text{err}_{\mathcal{D}_t}}{\text{err}_{\mathcal{D}_t}} \right) \quad (7.3.7)$$

这正是 AdaBoost 算法中计算  $\alpha_t$  的方法。下面讨论  $h_t$  的选取。

假设已获得了一系列弱学习器并知其权重, 则组合形成  $\tilde{h}_{t-1}$ 。关于  $h_t$  的理想选择就是纠正  $\tilde{h}_{t-1}$  的错误, 最小化指数损失函数

$$\begin{aligned} & \mathcal{L}_{\text{exp}}(\tilde{h}_{t-1} + h_t | \mathcal{D}) \\ &= \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} [e^{-Y(\tilde{h}_{t-1}(\mathbf{X}) + h_t(\mathbf{X}))}] \\ &= \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} [e^{-Y\tilde{h}_{t-1}(\mathbf{X})} \cdot e^{-Yh_t(\mathbf{X})}] \end{aligned} \quad (7.3.8)$$

对  $e^{-Yh_t(\mathbf{X})}$  利用 Taylor 展式, 则

$$\begin{aligned} & \mathcal{L}_{\text{exp}}(\tilde{h}_{t-1} + h_t | \mathcal{D}) \\ &\approx \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} \left[ e^{-Y\tilde{h}_{t-1}(\mathbf{X})} \left( 1 - Yh_t(\mathbf{X}) + \frac{Y^2 h_t(\mathbf{X})^2}{2} \right) \right] \\ &= \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} \left[ e^{-Y\tilde{h}_{t-1}(\mathbf{X})} \left( 1 - Yh_t(\mathbf{X}) + \frac{1}{2} \right) \right] \end{aligned} \quad (7.3.9)$$

因此,  $h_t$  的理想选择为

$$\begin{aligned}
 h_t(\mathbf{X}) &= \operatorname{argmin}_h \mathcal{L}_{\text{exp}}(\tilde{h}_{t-1} + h_t | \mathcal{D}) \\
 &= \operatorname{argmin}_h \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} \left[ e^{-Y\tilde{h}_{t-1}(\mathbf{X})} \left( 1 - Yh_t(\mathbf{X}) + \frac{1}{2} \right) \right] \\
 &= \operatorname{argmax}_h \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} \left[ e^{-Y\tilde{h}_{t-1}(\mathbf{X})} Yh_t(\mathbf{X}) \right] \\
 &= \operatorname{argmax}_h \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} \left[ \frac{e^{-Y\tilde{h}_{t-1}(\mathbf{X})}}{\mathbb{E}_{\mathbf{X} \sim \mathcal{D}}[e^{-Y\tilde{h}_{t-1}(\mathbf{X})}]} Yh_t(\mathbf{X}) \right]
 \end{aligned} \tag{7.3.10}$$

记

$$\mathcal{D}_t(\mathbf{X}) = \frac{\mathcal{D}(\mathbf{X}) e^{-Y\tilde{h}_{t-1}(\mathbf{X})}}{\mathbb{E}_{\mathbf{X} \sim \mathcal{D}}[e^{-Y\tilde{h}_{t-1}(\mathbf{X})}]} \tag{7.3.11}$$

则利用期望的定义, 可得

$$\begin{aligned}
 h_t(\mathbf{X}) &= \operatorname{argmax}_h \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} \left[ \frac{e^{-Y\tilde{h}_{t-1}(\mathbf{X})}}{\mathbb{E}_{\mathbf{X} \sim \mathcal{D}}[e^{-Y\tilde{h}_{t-1}(\mathbf{X})}]} Yh_t(\mathbf{X}) \right] \\
 &= \operatorname{argmax}_h \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_t}[Yh_t(\mathbf{X})]
 \end{aligned} \tag{7.3.12}$$

又由于  $Yh_t(\mathbf{X}) = 1 - 2I(Y \neq h_t(\mathbf{X}))$ , 因而  $h_t(\mathbf{X})$  满足

$$h_t(\mathbf{X}) = \operatorname{argmin}_h \mathbb{E}_{\mathbf{X} \sim \mathcal{D}_t}[I(Y \neq h_t(\mathbf{X}))] \tag{7.3.13}$$

通过式 (7.3.11), 可得

$$\begin{aligned}
 \mathcal{D}_{t+1}(\mathbf{X}) &= \frac{\mathcal{D}(\mathbf{X}) e^{-Y\tilde{h}_t(\mathbf{X})}}{\mathbb{E}_{\mathbf{X} \sim \mathcal{D}}[e^{-Y\tilde{h}_t(\mathbf{X})}]} \\
 &= \frac{\mathcal{D}(\mathbf{X}) e^{-Y\tilde{h}_{t-1}(\mathbf{X})} \cdot e^{-Y\alpha_t h_t(\mathbf{X})}}{\mathbb{E}_{\mathbf{X} \sim \mathcal{D}}[e^{-Y\tilde{h}_t(\mathbf{X})}]} \\
 &= \mathcal{D}_t(\mathbf{X}) \cdot e^{-Y\alpha_t h_t(\mathbf{X})} \frac{\mathbb{E}_{\mathbf{X} \sim \mathcal{D}}[e^{-Y\tilde{h}_{t-1}(\mathbf{X})}]}{\mathbb{E}_{\mathbf{X} \sim \mathcal{D}}[e^{-Y\tilde{h}_t(\mathbf{X})}]}
 \end{aligned} \tag{7.3.14}$$

这恰好是 AdaBoost 算法更新分布的方法。

**注 7.4** 从优化损失函数的角度分析 AdaBoost 算法, 有以下优点:

1、帮助我们明朗化 AdaBoost 算法的学习目标, 有助于理解 AdaBoost 的理论原理及相关性质。

2、实现学习目标 (损失函数最小化) 与实现目标 (数值方法) 之间的“解耦”(decoupling)。一方面可以修改目标函数以适应新的学习模型, 另一方面可以引入快速实用的数值方法。从而可以建立其他类型的 AdaBoost 算法 [98]。如优化任意可导函数的 AnyBoost; 优化基于分类间隔损失函数的 MarginBoost 等 [99]。

### Logistic 损失函数

对于 AdaBoost 算法, 如注 7.4 所言, 可以根据需要使用不同的损失函数。

对于指数损失函数  $e^{-Y\tilde{h}_T(\mathbf{X})}$ , 其优点是性质好且容易处理。但如果预测错误, 则  $Y\tilde{h}_T(\mathbf{X})$  为负, 从而  $e^{-Y\tilde{h}_T(\mathbf{X})}$  为指数增长。这意味着此时对扰动很敏感, 即稳健性较差。

为此, 在当代 Boosting 方法中, 常用 Logistic 损失函数:

$$\mathcal{L}_{\text{log}}(\tilde{h}|\mathcal{D}) = \mathbb{E}_{\mathbf{X} \sim \mathcal{D}} \left[ \ln \left( 1 + e^{-Y\tilde{h}_T(\mathbf{X})} \right) \right]. \quad (7.3.15)$$

显然根据指数函数与对数函数的性质知式 (7.3.3) 与式 (7.3.15) 会被相同的函数最小化。事实上, 可以通过极大似然的思想来解释  $\ln \left( 1 + e^{-Y\tilde{h}_T(\mathbf{X})} \right)$  的来源。已知  $Y_i \in \{-1, 1\}$ , 且服从 Logit 模型, 则有:

$$\Pr(Y_i|\mathbf{X}_i) = \begin{cases} \frac{1}{1 + e^{-\tilde{h}_T(\mathbf{X}_i)}}, & \text{if } Y_i = 1 \\ \frac{1}{1 + e^{\tilde{h}_T(\mathbf{X}_i)}}, & \text{if } Y_i = -1 \end{cases}$$

考虑到  $Y_i$  的取值为  $-1$  或者  $1$ , 将上式合并成

$$\Pr(Y_i|\mathbf{X}_i) = \begin{cases} \frac{1}{1 + e^{-Y_i\tilde{h}_T(\mathbf{X}_i)}}, & \text{if } Y_i = 1 \\ \frac{1}{1 + e^{-Y_i\tilde{h}_T(\mathbf{X}_i)}}, & \text{if } Y_i = -1 \end{cases}$$

即条件概率的表达式统一为

$$\Pr(Y_i|\mathbf{X}_i) = \frac{1}{1 + e^{-Y_i\tilde{h}_T(\mathbf{X}_i)}}$$

从而极大似然函数为

$$\prod_{i=1}^n \Pr(Y_i|\mathbf{X}_i) = \prod_{i=1}^n \frac{1}{1 + e^{-Y_i\tilde{h}_T(\mathbf{X}_i)}} = \prod_{i=1}^n \left( 1 + e^{-Y_i\tilde{h}_T(\mathbf{X}_i)} \right)^{-1} \quad (7.3.16)$$

上式 (7.3.16) 的最大值等价成取负对数的最小值, 即式 (7.3.15).

**注 7.5** 指数损失函数与 Logistic 损失函数都可视作对 0-1 损失函数的光滑近似。相比较而言, Logistic 的近似程度要更好。见图 7.3。

Friedman et al. 认为, 与其用 AdaBoost 中的拟牛顿更新策略, 不如通过梯度下降的方式优化 Logistic 损失函数, 并据此给出了 LogitBoost 算法 [89]。

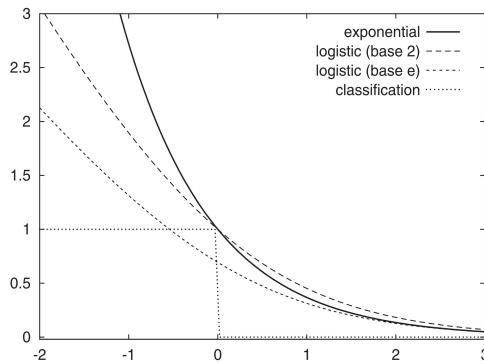


图 7.3 指数损失、Logistic 损失、分类损失

### LogitBoost 算法

**输入:** 训练数据集  $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$ ;

基学习算法  $\mathcal{L}$ ;

学习轮数  $T$ .

**过程:**

1.  $Y_0(\mathbf{X}_i) = Y_i$ ; % 初始化目标
2.  $\tilde{h}_0(\mathbf{X}) = 0$ ; % 初始化函数
3. **for**  $t = 1, \dots, T$ :
4.  $p_t(\mathbf{X}) = \frac{1}{1 + e^{-2\tilde{h}_{t-1}(\mathbf{X})}}$ ; % 计算概率
5.  $Y_t(\mathbf{X}) = \frac{Y_{t-1}(\mathbf{X}) - p_t(\mathbf{X})}{p_t(\mathbf{X})(1 - p_t(\mathbf{X}))}$ ; % 更新目标
6.  $\mathcal{D}_t(\mathbf{X}) = p_t(\mathbf{X})(1 - p_t(\mathbf{X}))$ ; % 更新权重
7.  $h_t = \mathcal{L}(\mathcal{D}, Y_t, \mathcal{D}_t)$ ; % 训练分类器  $h_t$  在分布  $\mathcal{D}_t$  下拟合样本集中的  $Y_t()$
8.  $\tilde{h}_t = \tilde{h}_{t-1} + \frac{1}{2}h_t$ ; % 更新组合分类器
9. **end**

**输出:**  $\mathcal{H}(\mathbf{X}) = \text{sign}\left(\sum_{t=1}^T h_t(\mathbf{X})\right)$

### 7.3.2 向前逐段可加视域

#### 向前逐段算法

对于二元分类问题的 AdaBoost 算法，其最终表达式为

$$\mathcal{H}(\mathbf{X}) = \text{sign}(\tilde{h}_T(\mathbf{X})) = \text{sign}\left(\sum_{t=1}^T \alpha_t h_t(\mathbf{X})\right)$$

可将  $h_t(\mathbf{X}) \in \{-1, 1\}$  视为基函数，类似于泰勒展式的函数项。不失一般性，将希望学到的函数  $f(x)$  做基函数展开，可以得到加法模型：

$$f(x) = \sum_{t=1}^T \alpha_t h(x; \gamma_t) \quad (7.3.17)$$

其中， $\alpha_t$  为展开系数， $h(x; \gamma_t)$  为基函数， $\gamma_t$  为参数。

为估计展开式系数  $\alpha_t$  与参数  $\gamma_t$ ，可以最小化以下目标函数：

$$\min_{\alpha'_t s, \gamma'_t s} \sum_{i=1}^n L \left( Y_i, \sum_{t=1}^T \alpha_t h(\mathbf{X}_i; \gamma_t) \right) \quad (7.3.18)$$

其中， $L(Y_i, f(\mathbf{X}_i))$  为损失函数。例如  $L(Y_i, f(\mathbf{X}_i)) = (Y_i - f(\mathbf{X}_i))^2$ ， $L(Y_i, f(\mathbf{X}_i)) = |Y_i - f(\mathbf{X}_i)|$  或者 0-1 损失函数  $I(Y_i \neq f(\mathbf{X}_i))$ 。

显然，这是一个复杂的优化问题。**向前分段算法** (forward stagewise algorithm) 求解这一优化问题的想法是：利用加性模型的性质，考虑从向前的分步计算，每一步只学习一个基函数及其系数，逐步逼近优化目标函数式 (7.3.18)。具体地，每步只需优化如下损失函数：

$$\min_{\alpha, \gamma} \sum_{i=1}^n L(Y_i, \alpha h(\mathbf{X}_i; \gamma)) \quad (7.3.19)$$

### 向前分段算法与 AdaBoost 算法

由向前分段算法可以推导出 AdaBoost，用如下定理叙述这一关系。

**定理 7.4** 二元分类问题的 AdaBoost 算法是向前分段加法算法的特例。其中模型是由基本分类器组成的加法模型，损失函数为  $L(Y, f(\mathbf{X})) = e^{-Y f(\mathbf{X})}$ 。

该定理的证明关键是找出  $h_t(\mathbf{X})$  和  $\alpha_t$ ，其证明过程类似于前面的推导过程，可参见 [19]。

## 7.4 Boosting 算法的演化

基于 AdaBoost 的理论剖析，人们对 AdaBoost 算法进行了各种改进与推广，并进行了广泛应用。

### 7.4.1 回归问题的 Boosting 算法

最初的 AdaBoost 算法是针对于分类问题而设计的，但其算法思想极其具有一般性，因而可将其应用于其他问题，下而重点介绍回归问题的 LSBoost 方法。

对于回归问题，一般采用均方误差为损失函数：

$$\mathcal{L}_{\text{mse}} = \left( Y - \tilde{h}_T(\mathbf{X}) \right)^2 \quad (7.4.1)$$

将 7.3.2 节中的向前逐段加法模型代入该损失函数如下，并且最小化该损失函数可得  $\alpha_t, \gamma_t$ ：

$$\begin{aligned} & \min_{\alpha, \gamma} \sum_{i=1}^n \mathcal{L}_{\text{mse}} \left( Y_i, \tilde{h}_{t-1}(\mathbf{X}_i) + \alpha \cdot \tilde{h}(\mathbf{X}_i; \gamma) \right) \\ &= \min_{\alpha, \gamma} \sum_{i=1}^n \left( Y_i - \tilde{h}_{t-1}(\mathbf{X}_i) - \alpha \cdot \tilde{h}(\mathbf{X}_i; \gamma) \right)^2 \\ &= \min_{\alpha, \gamma} \sum_{i=1}^n \left( r_{ti} - \alpha \cdot \tilde{h}(\mathbf{X}_i; \gamma) \right)^2. \end{aligned} \quad (7.4.2)$$

其中， $r_{ti}$  为当前阶段模型的残差。因而算法的更新思想是以当前残差  $r_{ti}$  为因变量，对自变量  $\tilde{h}(\mathbf{X}_i; \gamma)$  进行回归即可。进行回归时，可以采用普通的线性回归方法，也可以使用如下所述的回归树。

#### 树模型

**分类与回归树模型** (classification and regression tree, CART) 主要由 Quinlan、Breiman 等人创立 [67] [79]，是一种基本的分类与回归方法。用于分类问题时，称为**分类树** (classification tree)；用于回归问题，称为**回归树** (regression tree)。这些是前面介绍过的内容。

对于分类问题，我们回忆例 7.1 的做法。基本分类器可以看作是由一个根节点连接两个叶节点的简单决策树，常称为**决策树桩** (decision stump)。最终的强分类器形式为  $\tilde{h}_T(\mathbf{X}) = \sum_{t=1}^T h_t(\mathbf{X})$ 。抽象成一般形式应用于回归问题，则回归 Boosting 树模型可以表示以决策树为基函数的加法模型：

$$\tilde{h}_T(\mathbf{X}) = \sum_{t=1}^T G(\mathbf{X}; \Theta_t) \quad (7.4.3)$$

其中， $G(\mathbf{X}; \Theta_t)$  表示决策树， $\Theta_t$  为决策树的参数， $T$  为树的数量。

#### 回归问题的 Boosting 树算法

假设  $\mathcal{X}$  为输入空间， $\mathcal{Y}$  为输出空间， $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$  为训练数据集。

一棵回归树由输入空间的一个划分以及在划分单元上的取值两部分决定。假设已将输入空间  $\mathcal{X}$  划分为  $K$  个单元:  $R_1, \dots, R_K$ , 并且在每个单元  $R_k$  上的取值为  $c_k$ , 于是回归树具有如下形式:

$$G(\mathbf{X}; \Theta) = \sum_{k=1}^K c_k I(\mathbf{X} \in R_k) \quad (7.4.4)$$

其中, 参数  $\Theta = \{(R_1, c_1), (R_2, c_2), \dots, (R_K, c_K)\}$  表示树的区域划分和各区域上的常数。 $K$  是回归树的复杂度, 即叶节点个数。

回归问题的 Boosting 树算法采用向前分段算法。首先确定初始 Boosting 树  $\tilde{h}_0(\mathbf{X}) = 0$ , 设  $\tilde{h}_{t-1}(\mathbf{X})$  为当前模型, 则第  $t$  步的模型是

$$\tilde{h}_t(\mathbf{X}) = \tilde{h}_{t-1}(\mathbf{X}) + G(\mathbf{X}; \Theta_t)$$

其中  $\Theta_t$  的选取遵循损失最小化原则, 即

$$\Theta_t = \operatorname{argmin}_{\Theta} \sum_{i=1}^n \mathcal{L}(Y_i, \tilde{h}_{t-1}(\mathbf{X}_i) + G(\mathbf{X}_i; \Theta))$$

由于树的线性组合可以很好的拟合训练数据, 所以回归问题的 Boosting 树是一个很高效的学习方法。此算法代码如下。

### 回归问题的 Boosting 树算法

**输入:** 训练数据集  $\{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$ ;

损失函数  $\mathcal{L}_{\text{mse}}(Y, \tilde{h}_T(\mathbf{X}))$ , 基函数集  $\{G(\mathbf{X}; \Theta_t)\}$ ;

学习轮数  $T$ .

**过程:**

1.  $\tilde{h}_0(\mathbf{X}) \equiv 0$ ; % 初始化基函数

2. **for**  $t = 1, \dots, T$ :

3.  $\Theta_t = \operatorname{argmin}_{\Theta} \sum_{i=1}^n \mathcal{L}_{\text{mse}}(Y_i, \tilde{h}_{t-1}(\mathbf{X}_i) + G(\mathbf{X}_i; \Theta))$ ; % 拟合残差得到回归树

4.  $\tilde{h}_t(\mathbf{X}) = \tilde{h}_{t-1}(\mathbf{X}) + G(\mathbf{X}; \Theta_t)$ ; % 更新加法模型

5. **end**

**输出:**  $\tilde{h}_T(\mathbf{X}) = \sum_{t=1}^T G(\mathbf{X}; \Theta_t)$ .

### 回归问题其他损失函数

对于回归问题的损失函数, 可以根据需要取成不同的形式。

1、Laplace 损失函数：

$$\mathcal{L}_l(Y, \tilde{h}_T(\mathbf{X})) = |Y - \tilde{h}_T(\mathbf{X})|$$

相比较于平方损失函数  $\mathcal{L}_{\text{mse}}$ ，Laplace 损失函数不易受到异常值的影响，因而更稳健。

2、Huber 损失函数：

$$\mathcal{L}_{hl}(Y, \tilde{h}_T(\mathbf{X})) = \begin{cases} \frac{1}{2} [Y - \tilde{h}_T(\mathbf{X})]^2, & \text{if } |Y - \tilde{h}_T(\mathbf{X})| \leq \delta \\ \delta \left[ |Y - \tilde{h}_T(\mathbf{X})| - \frac{\delta}{2} \right], & \text{if } |Y - \tilde{h}_T(\mathbf{X})| > \delta \end{cases}$$

其中， $\delta$  为松弛参数。通过其表达式可以看出，如果误差绝对值较小时，使用平方损失函数；如果误差绝对值较大，则使用绝对值损失函数。因此，Huber 损失函数结合了平方损失函数与绝对值损失函数的优点。

#### 7.4.2 梯度 Boosting 方法

##### 梯度 Boosting 算法

根据前面的理论剖析，AdaBoost 可视为加性模型中利用坐标下降法优化指数型损失函数的过程。在此基础上，Friedman 提出了梯度 Boosting 算法（gradient boosting），简称为 GBM（gradient boosting machine）算法 [100]。GBM 思想：以非参数方法估计基函数，并在“函数空间”使用梯度下降法求解。

假设训练数据  $\mathcal{D} = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ ，若以函数  $\tilde{h}_T(\mathbf{X})$  来预测  $Y$ ，则在总体中的平均损失（expected loss function）为

$$\mathbb{E}_{Y, \mathbf{X}} \mathcal{L}[Y, \tilde{h}_T(\mathbf{X})]$$

其中  $\mathcal{L}[\cdot, \cdot]$  为损失函数。目标变成找出能够最小化损失的函数  $\tilde{h}^*(\mathbf{X})$ ，即

$$\tilde{h}^*(\mathbf{X}) = \operatorname{argmin}_{\tilde{h}} \mathbb{E}_{Y, \mathbf{X}} \mathcal{L}[Y, \tilde{h}_T(\mathbf{X})]$$

利用期望的性质将上式写成

$$\begin{aligned} \tilde{h}^*(\mathbf{X}) &= \operatorname{argmin}_{\tilde{h}} \mathbb{E}_{Y, \mathbf{X}} \mathcal{L}(Y, \tilde{h}_T(\mathbf{X})) \\ &= \operatorname{argmin}_{\tilde{h}} \mathbb{E}_{\mathbf{X}} (\mathbb{E}_Y (\mathcal{L}(Y, \tilde{h}_T(\mathbf{X}))) | \mathbf{X})) \end{aligned}$$

因而最小化问题等价为

$$\min_{\tilde{h}} \mathbb{E}_Y [\mathcal{L}(Y, \tilde{h}_T(\mathbf{X})) | \mathbf{X}] \quad (7.4.5)$$

使用非参数思想，将  $\tilde{h}_T(\mathbf{X})$  在每个  $\mathbf{X}$  的取值均看作参数。由此， $\tilde{h}_T(\mathbf{X})$  可视为无穷维向量，故有无穷多参数。在函数空间上，进行“泛函梯度下降”方法找最优解。

假设模型为加法模型：

$$\tilde{h}^*(\mathbf{X}) = \sum_{t=1}^T f_t(\mathbf{X})$$

利用梯度下降算法，则

$$f_t(\mathbf{X}) = -\rho_t g_t(\mathbf{X}) \quad (7.4.6)$$

其中  $\rho_t$  为步长，也称为学习率（learning rate），且

$$g_t(\mathbf{X}) = \left[ \frac{\partial \mathbb{E}_Y(\mathcal{L}(Y, \tilde{h}_T(\mathbf{X})) | \mathbf{X})}{\partial \tilde{h}_T(\mathbf{X})} \right]_{\tilde{h}(\mathbf{X}_i) = \tilde{h}_{t-1}(\mathbf{X}_i)}, \quad \tilde{h}_{t-1}(\mathbf{X}) = \sum_{j=1}^{t-1} f_j(\mathbf{X}) \quad (7.4.7)$$

假设泛函正则性足够好，能够交换微积分次序，则

$$g_t(\mathbf{X}) = \mathbb{E}_Y \left[ \frac{\partial \mathcal{L}(Y, \tilde{h}_T(\mathbf{X}))}{\partial \tilde{h}_T(\mathbf{X})} | \mathbf{X} \right]_{\tilde{h}(\mathbf{X}_i) = \tilde{h}_{t-1}(\mathbf{X}_i)}. \quad (7.4.8)$$

$\rho_t$  通过下式给出

$$\rho_t = \operatorname{argmin}_\rho \mathbb{E}_{Y, \mathbf{X}} \mathcal{L}(Y, \tilde{h}_{t-1}(\mathbf{X}) - \rho g_t(\mathbf{X})) \quad (7.4.9)$$

因为训练数据的有限性，这种非参数统计的方法在实际应用中是行不通的。为解决此问题，首先假设

$$\tilde{h}_T(\mathbf{X}; \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{t=1}^T \alpha_t h(\mathbf{X}; \gamma_t) \quad (7.4.10)$$

其中  $h(\mathbf{X}; \gamma_t)$  由基学习器决定。根据训练数据，则其梯度方向为

$$g_t(\mathbf{X}_i) = \left[ \frac{\partial \mathcal{L}(Y_i, \tilde{h}_T(\mathbf{X}_i))}{\partial \tilde{h}_T(\mathbf{X}_i)} \right]_{\tilde{h}(\mathbf{X}_i) = \tilde{h}_{t-1}(\mathbf{X}_i)}, \quad i = 1, 2, \dots, n \quad (7.4.11)$$

因而，选择与负梯度方向最为接近的弱学习器，即

$$\gamma_t = \operatorname{argmin}_\gamma \sum_{i=1}^n [-g_t(\mathbf{X}_i) - h(\mathbf{X}_i; \gamma)]^2. \quad (7.4.12)$$

求得最优  $h(\mathbf{X}_i; \gamma_t)$  之后，则函数更新为

$$\tilde{h}_t = \tilde{h}_{t-1} + \alpha_t h(\mathbf{X}; \gamma_t) \quad (7.4.13)$$

其中，步长  $\rho_t$  利用直线搜索（line search）取为

$$\alpha_t = \operatorname{argmin}_\alpha \sum_{i=1}^n \mathcal{L}(Y_i, \tilde{h}_{t-1}(\mathbf{X}_i) + \alpha \cdot h(\mathbf{X}_i; \gamma_t)) \quad (7.4.14)$$

## GBDT 方法

以 GBM 算法为基础，结合树的思想给出 梯度提升决策树 (gradient boosting decision tree, GBDT) 方法。GBDT 以决策树为弱学习器的学习模型，利用梯度 Boosting 的方法，完成机器学习任务。直接给出算法如下

---

### GBDT 算法

---

**输入：**训练数据集  $D = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$ ;

损失函数  $\mathcal{L}(\cdot, \cdot)$ ;

学习轮数  $T$ .

**过程：**

$$1. \tilde{h}_0(\mathbf{X}) = \operatorname{argmin}_c \sum_{i=1}^n \mathcal{L}(Y_i, c); \quad \% \text{ 构造初始学习器}$$

2. **for**  $t = 1, \dots, T$ :

$$3. r_{ti} = - \left[ \frac{\partial \mathcal{L}(Y_i, \tilde{h}_T(\mathbf{X}_i))}{\partial \tilde{h}_T(\mathbf{X}_i)} \right]_{\tilde{h}(\mathbf{X}_i)=\tilde{h}_{t-1}(\mathbf{X}_i)}; \quad \% \text{ 计算样本负梯度}$$

4. 利用  $D_t = \{(\mathbf{X}_1, r_{t1}), \dots, (\mathbf{X}_n, r_{tn})\}$  学习得到新的叶节点  $R_{tj}$ ;  $\%$

$$5. c_{tj} = \operatorname{argmin}_c \sum_{i \in R_{tj}} \mathcal{L}(Y_i, \tilde{h}_{t-1}(\mathbf{X}_i) + c),$$

$R_{tj}$  表示第  $t$  棵树的第  $j$  个叶节点区域,  $j = 1, 2, \dots, J$ .  $\%$  生成决策树

$$6. \tilde{h}_t(\mathbf{X}) = \tilde{h}_{t-1}(\mathbf{X}) + \sum_{j=1}^J c_{tj} I(\mathbf{X} \in R_{tj}); \quad \% \text{ 更新决策树}$$

7. **end**

$$\text{输出: } \tilde{h}_T(\mathbf{X}) = \sum_{t=1}^T \sum_{j=1}^J c_{tj} I(\mathbf{X} \in R_{tj}).$$


---

**注 7.6** 对上述 GBDT 算法做一些说明:

- (1) 算法第 1 步构造初始学习器，即只有一个根节点的初始决策树。
- (2) 对于决策树模型，损失函数的常用方式有均方误差损失函数、Laplace 损失函数、Huber 损失函数等。
- (3) 算法第 3 步计算出  $r_{ti}$ ，将其作为残差估计，称其为**拟残差** (pseudo residuals)。
- (4) 算法第 4 步建立新的训练样本集  $D_t$ :

$$D_t = \{(\mathbf{X}_1, r_{t1}), \dots, (\mathbf{X}_n, r_{tn})\} \quad (7.4.15)$$

并用  $D_t$  作为训练样本集构造一棵决策树，取此决策树为第  $t+1$  个基学习器。

- (5) 算法第 5 步表示基学习器  $\tilde{h}_T(\mathbf{X})$  中每个叶节点的输出均使得上一轮迭代所取得模型的预测误差达到最小。

**注 7.7** 将 GBDT 应用于回归问题。假设损失函数取为均方损失函数，则

$$\mathcal{L}(Y, \tilde{h}_T(\mathbf{X})) = \frac{1}{2}[Y - \tilde{h}_T(\mathbf{X})]^2$$

则此时步骤 3 中拟残差的表达式变为

$$\begin{aligned} r_{ti} &= - \left[ \frac{\partial \mathcal{L}[Y_i, \tilde{h}_T(\mathbf{X}_i)]}{\partial \tilde{h}_T(\mathbf{X}_i)} \right]_{\tilde{h}(\mathbf{X}_i)=\tilde{h}_{t-1}(\mathbf{X}_i)} \\ &= - \left[ \frac{\partial \frac{1}{2}[Y_i - \tilde{h}_T(\mathbf{X}_i)]^2}{\partial \tilde{h}_T(\mathbf{X}_i)} \right]_{\tilde{h}(\mathbf{X}_i)=\tilde{h}_{t-1}(\mathbf{X}_i)} \\ &= [Y_i - \tilde{h}_T(\mathbf{X}_i)]_{\tilde{h}(\mathbf{X}_i)=\tilde{h}_{t-1}(\mathbf{X}_i)} \\ &= Y_i - \tilde{h}_{t-1}(\mathbf{X}_i) \end{aligned} \quad (7.4.16)$$

此时对于回归问题，拟残差  $r_{ti}$  就是真正的残差  $Y_i - \tilde{h}_{t-1}(\mathbf{X}_i)$ 。进一步计算最优步长

$$\begin{aligned} \alpha_t &= \operatorname{argmin}_\alpha \sum_{i=1}^n (Y_i - \tilde{h}_{t-1}(\mathbf{X}_i) - \alpha \cdot h(\mathbf{X}_i; \gamma_t))^2 \\ &= \operatorname{argmin}_\alpha \sum_{i=1}^n (r_{ti} - \alpha \cdot h(\mathbf{X}_i; \gamma_t))^2 \end{aligned} \quad (7.4.17)$$

因此知当基学习器为决策树时，GBDT 恰好是前面的回归 Boosting 方法。

## 7.5 AdaBoost 算法实践

AdaBoost 算法作为集成算法的杰出代表，不仅能够显著改善子分类器预测精度，而且本身有深厚的理论支撑，是理论与应用的完美结合。因而一经提出就受到不同领域研究人员的关注，在解决各行业应用问题中都获得了极大的成功。其应用领域举例如下：

- 手写字体识别
- 人脸检测
- 文本分类、文本过滤与信息检索

除此之外，AdaBoost 及其变种算法在机器视觉领域中被应用于目标检测、车辆识别、视频文字定位等；在计算机安全领域用于垃圾邮件分类、钓鱼网站检测等；在计算生物信息学中用于 DNA 序列分析；在高炉炼铁中应用于温度控制。以下两应用举例均来自参考文献《机器学习及 R 应用》，稍有改动 [101]。

### 7.5.1 R 语言实践

利用 AdaBoost 算法研究波士顿房价问题，其 R 语言代码如下：

```
install.packages("gbm")
library(gbm)
library(MASS)
set.seed(1)
train_index<-sample(506,354)
train<-Boston[train_index,]#设定训练集
test<-Boston[-train_index,]#设定测试集
set.seed(123)
#调用梯度提升算法：假设输出变量服从高斯分布；迭代次数为5000；基学习器深度为4
fit<-gbm(medv~.,data=train,distribution="gaussian",n.trees=5000, cv.fold=5,
interaction.depth=4,shrinkage=0.01,bag.fraction=0.5,n.minobsinnode=5)
summary(fit)
#分别画出偏依赖图
plot(fit,i.var="rm",main="Partial Importance Plot",ylab="medv")
plot(fit,i.var="lstat",main="Partial Importance Plot",ylab="medv")
#使用 predict() 函数，在测试集中进行预测，并计算均方误差
pred<-predict(fit,newdata=test,n.trees=5000)
y.test<-test[, "medv"]
mean((pred-y.test)^2)
```

从图 7.4 中看出，影响最大的变量依次是 rm 与 lstat。下面使用 gbm.perf() 考察决策树数目  $M$  对于训练误差与交叉验证误差的影响，结果如图 7.7。黑实线表示训练误差，绿色实线表示交叉验证误差，而蓝色虚线表示使交叉验证最小的决策树数目。上图可知，超出最优决策树数目 4960 后，随着迭代次数都加，训练误差趋于 0，而交叉验证误差则变大。

```
gbm.perf(fit,method="cv")
abline(h=0,lty=2)
legend("top",legend=c("Training Error","CV Error"),lty=1, col=c("black","green"))
#训练误差与交叉验证误差
min(fit$train.error)
which.min(fit$train.error)
min(fit$cv.error)
which.min(fit$cv.error)
```

考察决策树数目 [4500,5000] 区间的交叉验证误差，见图 7.8。

```
plot(4500:5000,fit$cv.error[4500:5000],type="l",xlab="Trees",
```

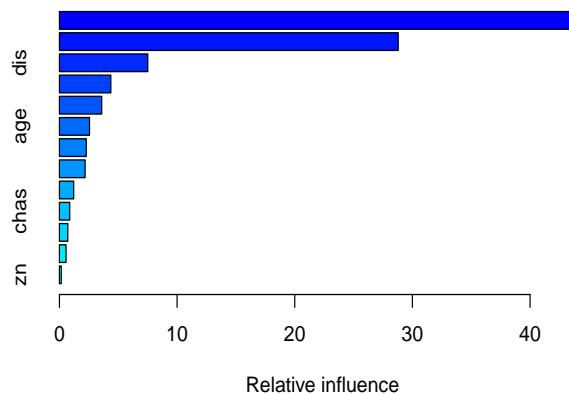


图 7.4 变量重要性

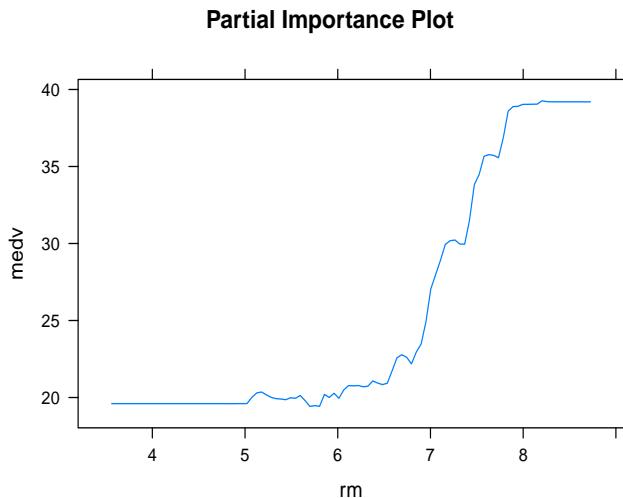


图 7.5 变量 rm 偏依赖图

```

ylab="Squared error loss",main="CV Error")
abline(v=which(fit$cv.error),lty=2)
#选择参数的最优值，使用下面的 expand.grid() 函数
grid <- expand.grid(shrinkage = c(.01,.1),interaction.depth = c(2, 4),
n.minobsinnode = c(5, 10),bag.fraction = c(.5, 1),optimal_trees = 0,min_MSE = 0)
grid
#进行网格搜索，找出最优值。将结果的前 6 行进行显示如下
for(i in 1:nrow(grid))
{
  set.seed(123)
  gbm.tune <- gbm(medv~,data=train,distribution = "gaussian",
}

```

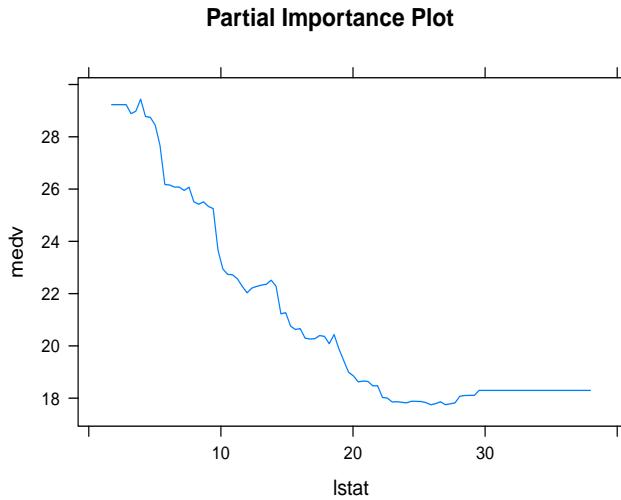


图 7.6 变量 lstat 偏依赖图

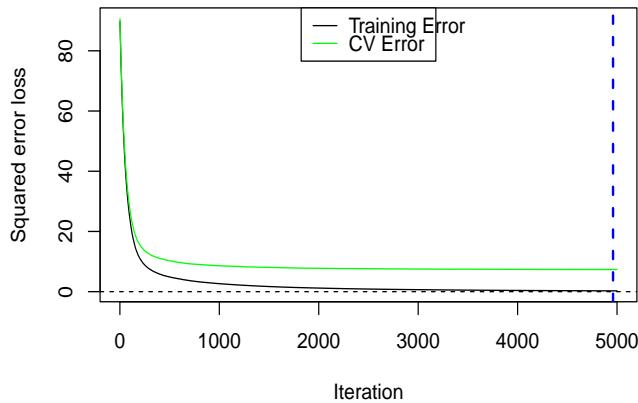


图 7.7 决策树数目对于训练误差与交叉验证误差的影响

```

n.trees = 5000, interaction.depth = grid$interaction.depth[i],
shrinkage = grid$shrinkage[i], n.minobsinnode = grid$n.minobsinnode[i],
bag.fraction = grid$bag.fraction[i], cv.folds = 5)
grid$optimal_trees[i] <- which.min(gbm.tune$cv.error)
grid$min_MSE[i] <- min(gbm.tune$cv.error)
}
head(grid[order(grid$min_MSE),])

```

根据网格搜索找出的最优值，最优参数分别为：收缩值  $\eta = 0.01$ ，交互深度  $d = 4$ ，终节点最小规模为 5，子抽样比例为 0.5。

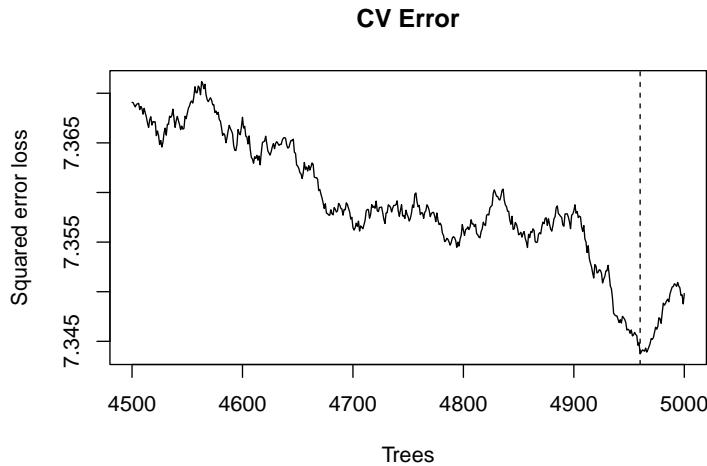


图 7.8 决策树数目对于交叉验证误差的影响

### 7.5.2 Python 语言实践

本小节将基于 Python 语言 sklearn.ensemble 函数库中的 AdaBoostClassifier 函数实现 AdaBoost 算法。具体实现过程如下：

1、导入需要的函数库，并生成数据集。

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import AdaBoostClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.datasets import make_gaussian_quantiles
X1, y1 = make_gaussian_quantiles(cov=2.0, n_samples=500, n_features=2, n_classes=2,
random_state=1)
X2, y2 = make_gaussian_quantiles(mean=(3, 3), cov=1.5, n_samples=400, n_features=2,
n_classes=2, random_state=1)
#生成两组二维正态分布数据
X = np.concatenate((X1, X2))#合成一组数据
y = np.concatenate((y1, -y2 + 1))
plt.scatter(X[:, 0], X[:, 1], marker='o', c=y) #数据集可视化
```

我们通过可视化观察我们的分类数据。如图7.9所示，它有两个特征，两个输出类别，用颜色区别，可以看到数据有些混杂。

2、我们现在用基于决策树的 AdaBoost 来做分类拟合，这里我们选择了 SAMME 算法，最多 200 个弱分类器，步长 0.8，在实际运用中可能需要通过交叉验证调参而选择最好的参数。拟合完成后，绘制拟合区域图像并查看拟合分数。具体实现代码如下：

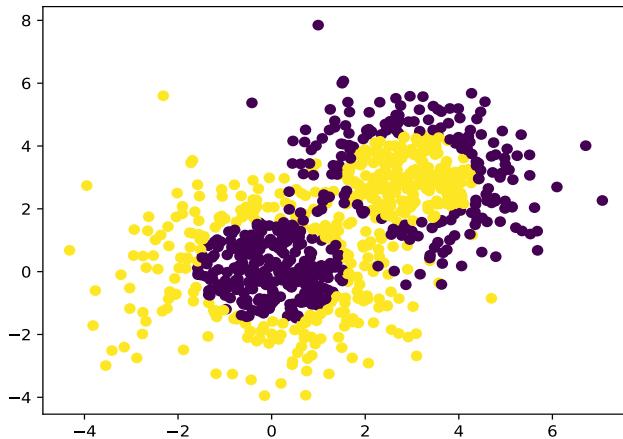


图 7.9 可视化数据集

```
bdt = AdaBoostClassifier(DecisionTreeClassifier(max_depth=2, min_samples_split=20,
min_samples_leaf=5), algorithm="SAMME", n_estimators=200, learning_rate=0.8)
bdt.fit(X, y)#进行分类拟合
x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, 0.02), np.arange(y_min, y_max, 0.02))
Z = bdt.predict(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
cs = plt.contourf(xx, yy, Z, cmap=plt.cm.Paired)
plt.scatter(X[:, 0], X[:, 1], marker='o', c=y)
plt.show()#查看拟合区域
print("Score:", bdt.score(X,y))#输出拟合分数
```

拟合的区域如图7.10所示。AdaBoost 的拟合效果较为不错。拟合分数为 0.91 也同样不错，但需要注意的是拟合分数不一定越高越好，因为可能出现过拟合的情况。

3、现在我们将最大弱分离器个数从 200 增加到 300。再来看看拟合分数，具体实现代码如下：

```
bdt = AdaBoostClassifier(DecisionTreeClassifier(max_depth=2, min_samples_split=20,
min_samples_leaf=5), algorithm="SAMME", n_estimators=300, learning_rate=0.8)
bdt.fit(X, y)
print("Score:", bdt.score(X,y))
#输出结果:
Score: 0.962222222222
```

这说明弱分离器个数越多，则拟合程度越好，当然也越容易过拟合。

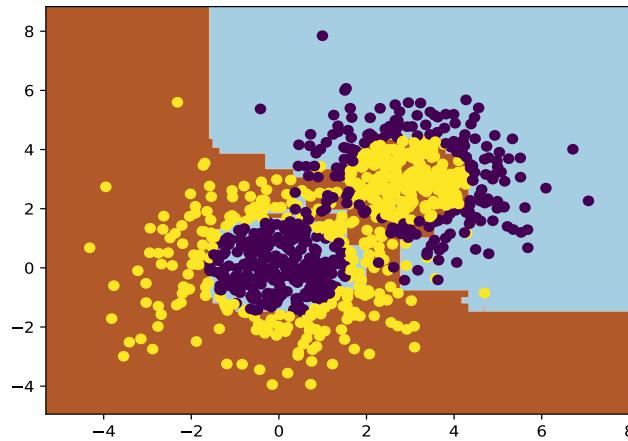


图 7.10 拟合区域图

## 7.6 总结

本章首先介绍 Boosting 方法及其代表性算法 AdaBoost；然后进行 AdaBoost 算法的误差分析，进一步解释 AdaBoost 算法提高学习精度的原因并叙述 Boosting 方法更具体的实例——Boosting tree（提升树）；最后，讨论 AdaBoost 及其变种算法和它们在实际问题中的应用。

接着讲述 Python 程序实现相关问题：Python AdaBoost 包在 Python 的编辑器中，可以通过导入 sklearn.ensemble 模块中的 AdaBoostClassifier 和 AdaBoostRegressor 来创建 AdaBoost 实例。其中，AdaBoostClassifier 用于分类任务，AdaBoostRegressor 用于分类任务。导入代码如图所示。

下来说明上述 AdaBoostClassifier、AdaBoostRegressor 实例的参数及其可选范围。

1、AdaBoostClassifier：

(1)  $base\_estimator$ : object 型，默认值为 None 基分类器，默认是决策树，理论上可以是任意一个分类器，但是如果使用其他分类器时需要指明样本权重。

(2)  $n\_estimator$ : int 型，默认值为 50 集成的基分类器数量，取值必须大于等于 1。

(3)  $Learning\_rate$ : float 型，默认值为 1.0 每个基分类器的集成权重。

algorithm 默认值为“SAMME.R”，基分类器权重的度量。支持“SAMME”和“SAMME.R”两种方式，前者是使用对样本集分类效果作为基分类器的权重，后者使用对样本集的预测概率大小作为基分类器的权重，如果使用的是‘SAMME.R’，则基分类器参数  $base\_estimator$  必须限制使用支持概率预测的分类器。

$random\_state$ : int 型，默认值为 None，随机种子设置

3、AdaBoostRegressor:

(1)  $base\_estimator$ : 同上

(2)  $n_{estimator}$ : 同上

(3)  $learning\_rate$ : 同上

loss 默认值为 “linear” 作为损失函数的形式。支持的形式有 “linear” 线性损失、 “square” 平方损失、 “exponential” 指数损失。

## 7.7 习题

1、给定如表 7.2 所示训练数据。假设弱分类器为阈值函数，选择弱分类器的标准是其阈值  $s$  使该分类器在训练数据集上分类误差率最低。试用 AdBoost 算法学习一个强分类器，并尝试给出算法代码。

表 7.2 训练数据表

序号	1	2	3	4	5	6	7	8	9	10
$X_i$	0	1	2	3	4	5	6	7	8	9
$Y_i$	-1	-1	-1	1	1	1	-1	-1	-1	1

2、某公司招聘职员考察身体、业务能力、发展潜力这 3 项。身体分为合格 1、不合格 0 两级；业务能力和发展潜力分为上 1、中 2、下 3 三级；分类标签分为合格 1、不合格 -1 两类。已知 10 人数据，如表 7.3 所示。假设弱分类器为决策树桩，试用 AdaBoost 算法学习一个强分类器，并尝试给出算法代码。

表 7.3 应聘人员情况数据表

序号	1	2	3	4	5	6	7	8	9	10
身体	0	0	1	1	1	0	1	1	1	0
业务能力	1	3	2	1	2	1	1	1	3	2
发展潜力	3	1	2	3	3	2	2	1	1	1
分类	-1	-1	-1	-1	-1	-1	1	1	-1	-1

3、已知如表 7.4 所示的训练数据， $X_i$  的取值范围为区间  $(0.5, 10.5)$ ， $Y_i$  的取值范围为区间  $(5.0, 10.0)$ ，请使用回归问题的 Boosting 树算法进行建模分析，考虑只用树桩作为基函数。

4、现有某工厂四位工人的考评信息及月薪数据，如表 7.5 所示。试根据该数据集和 GBDT 学习算法构造包含两个个体学习器的集成模型，并使用该集成模型预测工龄为 25 年，绩效得分为 65 分的工人的月薪。

表 7.4 训练数据表

$X_i$	1	2	3	4	5	6	7	8	9	10
$Y_i$	5.56	5.70	5.91	6.40	6.80	7.05	8.90	8.70	9.00	9.05

表 7.5 员工信息表

编号	工龄 (年)	绩效得分	月薪 (万元)
1	5	20	1.1
2	7	30	1.3
3	21	70	1.7
4	30	60	1.8

5、第 4 题的损失函数为均方损失函数，尝试分别利用 Laplace 损失函数、Huber 损失函数，仍然根据表 7.5，利用 GBDT 算法构造包含 3 个个体的学习器。

# 第八章 模型选择

## 8.1 简介

当我们考虑用模型来拟合数据，刻画数据的结构预测变量的时候，需要考虑两个基本问题：一是采用什么样的模型；二是采用哪些变量，即判断哪些变量是重要的，哪些变量是不重要的。由于统计推断和预测是在选定模型之后进行的，模型选择的好坏将直接影响到统计推断和预测的效果，因此模型选择问题得到了极大的重视和广泛的研究。

回归模型包含的预测变量并不是越多越好。首先，收集更多变量的数据会浪费时间和精力；其次，使用全部的预测变量会出现多重共线性的问题；最后，有些预测变量对响应变量并无影响，增加不必要的变量会带来估计噪声。因此我们要选择简单的且“足够好”的模型。

我们先介绍一个概念：[方差—偏差平衡](#)。当模型越来越复杂的时候，模型的预测能力也会越好。但问题是，模型只是在训练集上表现不错，而在测试集上表现比较差。具体来说：模型的偏差虽然小了，但是方差很大，即模型对“新”数据表现出“不稳定性”。这就是[过拟合](#)（over-fitting）了。模型过于复杂，连“噪声”也不放过，导致[训练误差](#)（training error）非常小，[测试误差](#)（test error）非常大。

如果一个模型对未知数据（通常是测试集的数据）的预测能力很好，我们就称它的[泛化能力](#)（generalization ability）很好，显然过拟合意味着模型的泛化能力很差。

如图8.1所示，当[模型复杂度](#)（degree of complexity）太低，训练误差和测试误差都很高，这个时候模型是[欠拟合](#)（under-fitting）的。随着模型复杂度的提升，测试误差先减小后增大，但是训练误差不断地减小，最后过拟合。因此，控制模型复杂度可以实现方差—偏差的平衡。而控制模型复杂度可以转为选择最优模型。[模型选择](#)（model selection）就是从多个候选模型中，基于某评价准则和训练数据选出最优的模型。它的主要思想是通过训练数据来估计期望的测试误差，用数学语言描述模型选择就是：给定数据集  $D$ ，依据某个模型评价准则，从候选模型集合  $S_m$ 's 中选出最优模型  $S^*$ ，即

$$S^* = \operatorname{argmin}_m (\operatorname{crit}(S_m; D))$$

其中  $\operatorname{crit}$  表示模型评价准则。不同的评价准则对应不同的模型选择方法。

接下来介绍两种常用的模型选择方法：子集选择法和压缩估计法。在子集选择法中我们将介绍基于检验的模型选择方法和基于准则的模型选择方法，在压缩估计

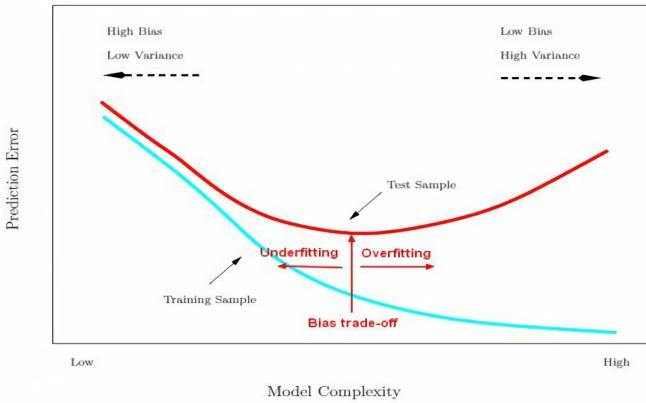


图 8.1 模型复杂度图

法中介绍 Lasso、非凸惩罚函数、群组变量选择方法和双层变量选择方法。

## 8.2 基于检验的方法

在基于检验的方法中介绍两种最常见的方法：最优子集法和逐步选择法。

### 8.2.1 最优子集法

对于含有  $p$  个预测变量组成的集合来说，最优子集选择即对  $p$  个预测变量的所有可能组合分别使用最小二乘回归进行拟合，即对含有一个预测变量的模型，拟合  $p$  个模型；对含有两个预测变量的模型，拟合  $p(p-1)/2$  个模型，依此类推最后我们一共拟合了  $C(p,0)+C(p,1)+\cdots+C(p,p-1)+C(p,p)$  即  $2^p$  个模型，再根据指标从所有可能模型中选取一个最优的模型。**最优子集法**具体过程可以概括为：

- 1、记不包含预测变量的零模型为  $M_0$ 。
- 2、对  $k = 1, \dots, p$ 
  - (1) 从  $p$  个预测变量中任意选择  $k$  个，拟合  $C(p,k)$  个模型。
  - (2) 在  $C(p,k)$  个模型中选择最优的一个（RSS 最小或  $\bar{R}^2$  最大），记为  $M_k$ 。
- 3、根据交叉验证预测误差、 $C_p$ 、AIC、BIC 或者修正的决定系数  $\bar{R}^2$  等指标，从  $M_0, \dots, M_p$  个模型中选择一个最优模型。

步骤 2 先在相同数量预测变量的模型中选择一个最优模型（内循环）；步骤 3 在不同数量预测变量的模型中进行模型选择（外循环）。通过外循环选择最优模型，需要改变准则。因为随着模型中预测变量数目增加，这  $p+1$  个模型的 RSS 会下降。

如果我们仅依据 RSS 进行模型选择的话，最后选出来的最优模型将包含所有变量。因此对于变量个数相同的模型，我们可通过 RSS 来选择最优模型；对于变量个数不同的模型，我们使用交叉验证预测误差、 $C_p$ 、AIC、BIC 或者  $\bar{R}^2$  等指标进行模型选择。例如，根据修正决定系数  $\bar{R}^2$  的定义：

$$\bar{R}^2 = 1 - \frac{\text{SSE}/(n - k - 1)}{\text{SST}/(n - 1)}. \quad (8.2.1)$$

我们知道， $\bar{R}^2$  越大，模型拟合程度越好。这里引入了预测变量数量  $k$  从而对预测变量增加进行了惩罚，即模型中包含的预测变量  $k$  增加时会抑制  $\bar{R}^2$  的增加，使选出来的模型具有更小的测试误差。通俗的来讲，修正决定系数并不会“偏爱”预测变量较多的模型。

显然，最优子集法运用了穷举的思想。其优势很明显：简单直接，遍历所有可能的情况，最后的选择一定是最优的；劣势也很明显：当  $p$  越大时，可选模型数量在迅速增加，计算量明显增大，降低计算效率。因此，最优子集法只适用于  $p$  较小的情况。一般来说，当  $p < 10$  时，我们可以考虑最优子集法。

### 8.2.2 逐步选择法

最优子集法在  $p$  很大时运算效率低，并且随着搜索空间的增大找到的模型泛化能力差。因此，本章引入逐步选择法避免上述缺点。

在 [逐步选择法](#) 中，模型会在原有变量的基础上一次增加一个显著的预测变量或剔除一个不显著的预测变量，直到既没有显著的预测变量选入回归方程，或也没有不显著的预测变量从回归方程中剔除为止。最优子集法与逐步选择法的区别在于，最优子集法可选择任意  $k$  个变量进行建模，而逐步选择法只能在之前所选的  $k$  个变量的基础上建模。常用的逐步选择法包括向前逐步选择、向后逐步选择、双向选择。

#### 向前逐步选择

[向前逐步回归](#)的思想是由少到多，逐个引入预测变量。我们首先考虑一个不包含任何预测变量的零模型，从零模型开始，依次引入一个预测变量，直至没有可引入的变量为止。所谓没有可引入的变量是指要添加的任何新变量都不会使模型有所改进。

具体来说就是从零号模型  $M_0$  开始，这个模型只有截距项而没有任何预测变量。然后，将  $p$  个预测变量依次加入模型中（每次只加入一个预测变量，考虑  $p$  次，得到  $p + 1$  个模型），保留 RSS 最小或  $\bar{R}^2$  最大的那个预测变量，此时模型含有一个预测变量，记为  $M_1$ 。然后在此基础上，将剩余的  $p - 1$  个预测变量依次分别加入，仍然保留 RSS 最小或  $\bar{R}^2$  最大的那个预测变量，此时得到的模型含有 2 个预测变量记为  $M_2$ 。这样重复操作，直至包含  $p$  个预测变量的模型  $M_p$ 。最后根据交叉验证预测

误差、 $C_p$ 、AIC、BIC 或者修正的  $\bar{R}^2$  等指标，从  $p+1$  个最优模型中选择一个最优模型。

向前逐步回归的算法过程如下：

1、记不包含任何预测变量的零模型为  $M_0$ 。

2、对  $k=0,1,2,\dots,p-1$ ：

(1) 基于上一步选取的最优模型  $M_k$  的  $k$  个预测变量，将剩余的  $p-k$  个预测变量分别加入，这样就得到  $p-k$  个模型；

(2) 在上述  $p-k$  个模型中选择 RSS 或  $\bar{R}^2$  最高的模型作为最优模型，记为  $M_{k+1}$ 。

3、进一步使用交叉验证误差、 $C_p$ 、AIC、BIC 或者修正的决定系数  $\bar{R}^2$  等指标，从  $p+1$  个最优模型中选择一个最优模型。

与最优子集选择法要对  $2^p$  个模型进行拟合不同，向前回归只需要对零模型以及第  $k$  次迭代所包含的  $p-k$  个模型进行拟合，其中  $k = 0, 1, 2, \dots, p-1$ 。相当于拟合  $\sum_{k=0}^{p-1} (p-k) = 1 + p(p+1)/2$  个模型，特别的，当  $p = 30$  时，最优子集需要拟合 1073741824 个模型，而向前逐步回归只需要拟合 466 个模型。

但是我们需要注意的是向前逐步回归无法保证最后得到的模型是  $2^p$  个模型中最优的。例如在给定的包含三个变量  $X_1, X_2, X_3$  的数据集中，最优的单变量模型只包含  $X_2$ ，最优的双变量模型只包含  $X_1, X_3$ ，且通过一些指标得出最优双变量模型在  $2^3 = 8$  个模型中是最优的。则通过向前逐步回归无法得到该最优模型，因为  $M_1$  包含变量  $X_2$ ，而  $M_2$  只能包含  $X_2, X_3$  或者  $X_2, X_1$ ，而无法得到包含  $X_1, X_3$  的双变量模型。

下面我们用一个例子说明向前回归的具体过程。

假设  $\{X_1, X_2, \dots, X_p\}$  为待选择的预测变量的集合，首先可以将每一个单一变量看作是一个子集，构成  $p$  个模型，在这  $p$  个模型中选择 RSS 最小或  $\bar{R}^2$  最大的为最优模型，在此不妨假设  $\{X_1\}$  是最优的单一变量模型；接下来我们在上一步的基础上添加一个变量，得到  $p-1$  个包含两个预测变量的模型，再基于上述准则选出最优模型，不妨假设  $\{X_1, X_2\}$  为最优，且优于  $\{X_1\}$ ，则第二步选择结束；依次类推，直到第  $k+1$  轮选择的最优模型拟合效果不如第  $k$  轮选出最优模型，则整个子集选择过程结束，并将第  $k$  轮选出的模型作为最后的最优模型。

## 向后逐步选择

与向前逐步回归法的选择方向相反，[向后逐步回归法](#)以一个包含全部预测变量的全模型为起点，逐次迭代，每次剔除一个对模型效果最不利的冗余变量，直到继续剔除变量却不能使模型质量有所改进为止。具体来说就是从包含全部  $p$  个预测变量的模型  $M_p$  开始，然后一个个地移除  $p$  个预测变量，保留 RSS 最小或  $\bar{R}^2$  最大的那个模型，保留下来的模型是包含  $p-1$  个预测变量的模型，记为  $M_{p-1}$ ，基于该模

型继续移除剩余的  $p - 1$  个预测变量。这样重复操作，直至包含 0 个预测变量的模型  $M_0$ 。然后根据交叉验证预测误差、 $C_p$ 、AIC、BIC 或者修正的  $\bar{R}^2$  等指标，从  $p + 1$  个最优模型中选择一个最优模型。

向后逐步回归的算法过程如下：

- 1、记不包所有预测变量的全模型为  $M_p$ 。
- 2、对  $k=p, p-1, \dots, 1$ ：
  - (1) 基于上一步选取的最优模型的  $k$  个预测变量，依次逐个剔除，这样就可以构建  $k$  个模型；
  - (2) 在上述  $k$  个模型中选择最优模型，记为  $M_{k-1}$ 。
- 3、进一步使用交叉验证预测误差、 $C_p$ 、AIC、BIC 或者修正的决定系数  $\bar{R}^2$  等指标，从  $p + 1$  个最优模型中选择一个最优模型，为最后得到的最优模型。

同样，向后逐步回归共需对  $1 + p(p+1)/2$  个模型进行搜索，比最优子集选择更高效。但是，向后逐步回归无法保证得到的模型是包含  $p$  个预测变量子集的最优模型。这是因为向后逐步回归同样都依赖于上一步的迭代结果。此处需要注意，因为向后逐步回归法是从全模型开始的，如果使用最小二乘法进行拟合，需要满足  $n > p$  的条件。而向前逐步回归法则不需要满足这一条件。当预测变量向量的维度  $p$  比较高时，应优先选择向前逐步回归法。

向前逐步回归和向后逐步回归都属于贪心算法，能够局部达到最优，但是从全局来看不一定是最优的。向前逐步回归虽然每一次都能选取最显著的一个预测变量，但在实际情况下，很可能有的预测变量在开始时是显著的，但是在其余预测变量添加进去之后，它就变得不显著了，但向前逐步回归此时不会剔除该变量。而向后逐步回归则很有可能会遗漏一些很重要的变量，比如刚开始变量  $X_1$  不显著，但是在剔除  $m$  个变量后其变得显著，此时向后逐步回归并不会重新加入这个变量。总的来说这些问题是由向前逐步回归不会剔除已经加入进来的变量，向后逐步回归不会加入已经剔除的变量导致的，下面介绍的双向选择会解决这些问题。

## 双向选择

**双向选择**是向前和向后逐步选择法的结合。每次引入一个变量，但与此同时也会剔除对模型没有贡献的变量。引入一个变量后，我们首先通过 F 检验，验证新变量是否会使得模型发生显著性变化。若原有变量由于新变量的加入变得不再显著时，剔除此变量。若模型依旧显著，再对所有变量进行  $t$  检验，剔除不显著变量。直到既没有新的显著的预测变量加入回归方程，也没有原有的不显著的预测变量从回归方程中剔除为止，最终我们得到一个最优模型。在这个过程中，某一特定预测变量可能会被反复引入和删除。该方法在试图达到最优子集选择效果的同时也保留了向前和向后逐步回归在计算效率上的优势。

## 8.3 基于准则的方法

基于准则的方法将介绍  $C_p$  准则、AIC 准则、BIC 准则与交叉验证。

### 8.3.1 各种准则

上节已经讨论过，包含所有预测变量的模型总是有最小的 RSS 和最大的  $\bar{R}^2$ ，因为这些统计量是衡量训练数据的误差，通常训练集的均方误差比测试集的均方误差要低，而我们在选择模型时是希望得到一个具有最小测试误差的模型。并且训练误差随着更多变量加入模型中将会逐渐降低。因此，RSS 并不适用于对包含不同预测变量的模型进行模型选择。

为了基于测试误差选择最优模型，那么就需要去估计测试误差，通常有两种方法估计测试误差：

- 1、间接估计测试误差：根据过拟合导致的偏差对训练误差进行调整。
- 2、直接估计测试误差：通过交叉验证方法直接估计测试误差。

接下来就介绍几种适用于对包含不同预测变量的模型进行模型选择的方法，是间接估计测试误差的方法，包括  $C_p$  准则、AIC 准则、BIC 准则。

#### $C_p$ 统计量

$C_p$  统计量的思想是通过在训练集 RSS 的基础上增加惩罚项，来调整训练误差倾向于低估测试误差这一现象。具体来说就是训练集 RSS 随着模型中预测变量的增加而降低，但是此时测试集的 RSS 要比训练集的 RSS 要高，所以就在训练集 RSS 的基础上增加一项，使得该增加项的大小随着模型中预测变量个数的增加而增加。采用最小二乘法拟合一个包括  $d$  个预测变量的模型， $C_p$  值计算如下：

$$C_p = \frac{1}{n} (RSS + 2k\hat{\sigma}^2)$$

其中， $\hat{\sigma}$  是响应变量观测误差的方差的估计值， $k$  是模型中变量的个数。 $C_p$  统计量在训练集 RSS 的基础上增加惩罚项  $2d\hat{\sigma}^2$ ，这可以看作是在考虑拟合残差同时，依变量个数施加的惩罚。显然，惩罚项随模型中变量个数的增加而增大，用于调整由于变量个数增加而不断降低的训练集 RSS。如果  $\hat{\sigma}$  是观测误差方差的无偏估计，则  $C_p$  是测试均方误差的无偏估计。 $C_p$  值越小，表示模型的准确性越高。

#### AIC 准则

AIC 准则是衡量统计模型拟合优良性的一种准则，由日本统计学家赤池弘次 [103] 在 1974 年提出，因而又叫赤池信息准则。它是权衡估计模型复杂度和拟合数据优良性的准则。AIC 准则的定义公式为：

$$AIC = 2k - 2\ln L$$

其中  $K$  是模型参数个数,  $L$  是极大似然函数。由于似然函数的值越大得到的估计量就越好, 因此  $-2\ln(L)$  越小越好, 该项刻画了模型的精度或拟合程度, 又考虑到模型包含的预测变量的个数不能太多, 故加入  $2K$  对模型参数个数进行惩罚。即一方面要使似然函数值较大, 另一方面通过惩罚部分限制模型的复杂度。因此, AIC 是寻找可以最好地解释数据但包含最少自由参数的模型。

假设模型的随机误差服从独立正态分布。此时  $-2\ln(L) = n\ln\left(\frac{RSS}{n}\right) + \text{常数}$ , 那么 AIC 变为:

$$AIC = 2k + n\ln\left(\frac{RSS}{n}\right)$$

其中  $n$  为样本量,  $k$  为参数个数, RSS 为残差平方和。

### BIC 准则

Schwarz [104] 为了能够克服在大样本数据情形下 AIC 准则容易失效的不足, 提出了贝叶斯信息准则。其定义公式为:

$$BIC = k\ln(n) - 2\ln L$$

假设条件是模型的误差服从独立正态分布, BIC 准则的公式为:

$$BIC = k\ln(n) + n\ln\left(\frac{RSS}{n}\right)$$

其中,  $n$  为样本容量。当样本容量很大时, 在 AIC 准则中的残差平方和的所在项会受到样本容量影响而放大, 而参数个数的惩罚因子却未随样本容量的变化而变化。因此当样本容量很大时, 使用 AIC 准则选择的模型不收敛到真实的最优模型。事实上, 它通常比真实模型所含的未知参数个数要多。为引入对预测变量过多的惩罚, BIC 准则把 AIC 中的 2 换成了  $\ln(n)$ 。所以, 当  $n > e^2$  时, BIC 统计量相比于 AIC 统计量给包含多个变量的模型施以较重的惩罚, 所以与 AIC 统计量相比, BIC 得到的模型规模更小。如果真实模型是有限维参数, BIC 准则会表现得更好。

### AIC 和 BIC 比较

AIC 和 BIC 的原理是不同的, AIC 是从预测角度, 选择一个好的模型用来预测, BIC 是从拟合角度, 选择一个对现有数据拟合最好的模型, 从贝叶斯因子的解释来讲, 就是边际似然最大的那个模型。

**相同点:** 构造这些统计量所遵循的统计思想是一致的, 就是在考虑拟合残差的同时, 依预测变量个数施加“惩罚”。

**不同点：**BIC 的惩罚项比 AIC 大，考虑了样本个数，可以防止模型精度过高造成的模型复杂度过高。AIC 和 BIC 前半部分是惩罚项，当  $n \geq 8$  时， $k \ln(n) \geq 2k$ ，所以，BIC 相比 AIC 在大数据量时对模型参数惩罚得更多，导致 BIC 更倾向于选择参数少的简单模型。

### 8.3.2 交叉验证

**交叉验证** (cross-validation) 它是一种没有任何前提假定直接估计泛化误差的模型选择方法。由于没有任何假定，它可以应用于各种模型选择中。它的基本思想是将数据分为两部分，一部分数据用来进行模型的训练，通常我们叫做**训练集**，另一部分数据用来测试训练生成模型的误差，我们叫做**测试集**。由于两部分数据的不同，泛化误差的估计是在新的数据上进行，这样的泛化误差的估计可以更接近真实的泛化误差。在数据足够的情况下，我们可以很好估计出真实的泛化误差，但是在实际应用中，往往只有有限的数据可用，我们必须对数据进行重用，即对数据进行多次切分来得到好的估计，自从交叉验证提出以后，人们提出了不同的数据切分方式，因此产生了多种形式的交叉验证方法，下面我们介绍一些主要的交叉验证方法。

#### 验证集方法

用给定的观测集估计使用某种模型拟合所产生的测试误差，一种最简单、直接的方法是**验证集方法**。首先，我们将给定的样本观测集数据随机地分为不重复的两部分，然后用训练集来训练模型，在测试集上验证模型及参数。接着，我们再把样本打乱，重新选择训练集和测试集，继续训练数据和检验模型。最后我们选择损失函数评估最优的模型和参数。这就是验证集方法的基本原理。下面以  $5 \times 2$  交叉验证为例介绍训练集和测试集的选择方法。

**$5 \times 2$  交叉验证法**的主要思想是将数据集  $V$  平均分为两部分  $V_1^{(1)}$  和  $V_1^{(2)}$ ，首先用  $V_1^{(1)}$  作为训练集， $V_1^{(2)}$  作为验证集，然后互换角色，用  $V_1^{(2)}$  作为训练集， $V_1^{(1)}$  作为验证集，这样就得到了第一折，即第一次对折。为了得到第二折，我们将数据集重新打乱并划分为新的两个等份  $V_2^{(1)}$  和  $V_2^{(2)}$ 。将  $V_2^{(1)}$  作为训练集， $V_2^{(2)}$  作为验证集，然后对调两者的角色，得到第二折，重复以上作法五次，会得到十个训练集和验证集。当然我们可以进行超过五次的对折得到更多的训练集和验证集，但在五次对折之后，各个集合共享了许多样本，使得计算出来的泛化误差估计变得相互依赖无法增加新的信息。

可以看出验证集方法的优点：处理简单，只需随机把原始数据分为两组即可；并且测试集和训练集是分开的，避免了过拟合的现象。但是其缺点也有两个部分：

1、由于是随机的将原始数据分组，所以最后验证集的分类准确率与原始数据的划分有很大的关系，最终模型的确定将强烈地依赖于训练集和验证集的划分方式，测

试误差会根据划分方式的不同而变化。

2、由于只使用部分数据进行训练，得到的结果并不具有说服性。在实际应用中，当用于模型训练的观测数据越多时，模型的效果就会越理想。而验证集方法无法充分利用所有的观测数据，对模型的效果会产生影响。

接下来介绍交叉验证法，针对验证集方法的上述问题进行了改进。

### 留一交叉验证法

**留一法** (Leave-One-Out Cross-Validation, LOOCV) 的基本思想是每次从个数为  $n$  样本集中取出一个样本作为验证集，剩下的  $n - 1$  个样本作为训练集，重复进行多次，依次取遍所有  $n$  个数据作为验证集。也就是说每次只留下一个样本做测试集，其它样本做训练集。操作步骤如下：

1. 如果设原始数据有  $n$  个样本，即  $\{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_{i-1}, Y_{i-1}), (\mathbf{X}_{i+1}, Y_{i+1}), \dots, (\mathbf{X}_n, Y_n)\}$ ，那么以每个样本单独作为验证集，其余的  $n - 1$  个样本作为训练集进行训练，我们会得到  $n$  个模型。

2、对于每个在训练集上拟合的模型而言，我们考虑其相应的验证集进而得到该模型的均方误差。

$$(1) \text{ 回归问题: } \text{MSE}_i = (Y_i - \hat{Y}_i)^2 \quad (1.1.8)$$

$$(2) \text{ 分类问题: } \text{Err}_i = I(Y_i \neq \hat{Y}_i) \quad (1.1.9)$$

3、将  $n$  个模型的  $n$  个均方误差取均值后，我们得到测试均方误差的 LOOCV 估计：

$$(1) \text{ 回归问题: } \text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^n \text{MSE}_i \quad (1.1.10)$$

$$(2) \text{ 分类问题: } \text{CV}_{(n)} = \frac{1}{n} \sum_{i=1}^n \text{Err}_i \quad (1.1.11)$$

LOOCV 是交叉验证方法中最常见的一种方法，相比于验证集方法，它有如下特点：

**优点：**每一回合中几乎所有的样本都用于训练模型，这样评估所得的结果比较可靠；实验中没有随机因素，整个过程是可重复的。

**缺点：**计算成本高，当  $n$  非常大时，计算耗时。

于是就有了一种折中的方法—— $K$  折交叉验证法 ( $K$ -fold CV)。

### $K$ 折交叉验证法

**$K$  折交叉验证法** ( $K$ -fold Cross Validation, K-CV) 的原理：

1、将原始数据分成  $K$  组（一般是均分），或者说折 (fold)；然后，让每个子集数据分别做一次验证集，其对应的剩余  $K - 1$  组子集数据作为训练集，这样会得到  $K$  个模型。

2、同 LOOCV 方法一样，基于 K-CV 的实验建立的  $K$  个模型，分别在训练集上拟合模型，再将拟合的模型用于保留的验证集上计算均方误差。

$$(1) \text{ 回归问题: } \text{MSE}_i = (Y_i - \hat{Y}_i)^2 \quad (1.1.12)$$

$$(2) \text{ 分类问题: } \text{Err}_i = I(Y_i \neq \hat{Y}_i) \quad (1.1.13)$$

3、用这  $K$  个模型最终的验证集的均方误差的平均数作为此 K-CV 的估计。

$$(1) \text{ 回归问题: } \text{CV}_{(K)} = \frac{1}{K} \sum_{k=1}^K \text{MSE}_k \quad (1.1.14)$$

$$(2) \text{ 分类问题: } \text{CV}_{(K)} = \frac{1}{K} \sum_{k=1}^K \text{Err}_k \quad (1.1.15)$$

在实作上， $K$  要够大才能使各回合中的训练样本数够多。根据经验一般选取  $K = 5$  或  $K = 10$ ，这两个取值会使测试误差的估计不会有过大的偏差或方差， $K = 10$  是相当足够了。

$K$  折交叉验证法有如下特点：

**与 LOOCV 方法的关系：**LOOCV 方法就是  $K$  折交叉验证法当  $K = n$  时的一个特例。由于  $K < n$ ， $K$  折交叉验证法拟合模型次数  $K$  便小于 LOOCV 方法，这样就降低了计算成本。

**优点：**K-CV 使得每一个样本数据都既被用作训练数据，也被用作测试数据，可以有效的避免过拟合以及欠拟合的发生，最后得到的结果也比较具有说服性。

**缺点：** $K$  折交叉验证法的缺点在于  $K$  的确定。一方面， $K$  越大——样本划分的组多——训练集包含的观测数据多——拟合的偏差越小。特例是当  $K = n$  时也就是 LOOCV，我们能够得到一个近似无偏的测试误差估计。另一方面， $K$  越大就意味着每次用于拟合模型的训练集的观测数据重合度高，模型也就更相似。还是考虑特殊情况 LOOCV，每一次训练的观测数据几乎是相同的，拟合的结果之间是高度（正）相关的，由于许多高度相关的量的均值要比相关性相对较小的量的均值具有更高的波动性——方差也将更大。

与  $C_p$  准则、AIC 准则、BIC 准则、调整  $\bar{R}^2$  相比，这些方法的优势在于直接给出了测试误差的直接估计，符合我们最初选择测试误差最小的模型的目的。

## 8.4 正则化方法

子集选择方法通过保留预测变量的一个子集，并舍弃其它变量，产生一个简单的具有一定解释性的模型。然而子集选择法也有缺点，主要体现在两个方面：首先，在进行最优变量子集筛选的过程时离散，通常具有很高的变异性，不够稳定；其次，忽略了变量选择过程中的随机误差。

模型选择的另一个典型方法是正则化，**正则化**就是通过对模型参数进行调整（数量和大小），降低模型的复杂度，以达到可以避免过拟合的效果。具体的解决方法是

在模型的损失函数中加入正则项。如果不加入正则项，我们的目标是最小化损失函数；加入正则项以后，变成最小化损失同时复杂度不要太高，这一方法称为**结构风险最小化**。通过惩罚函数约束模型的回归系数，同步实现变量选择和系数估计，模型估计是一个连续的过程，更正了子集选择法的缺陷。

在详细介绍这一方法之前，我们先介绍目标函数和范数的概念。

### 目标函数

求解优化问题的关键就是最小化目标函数。假设样本矩阵为  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$  是  $n \times p$  矩阵，因变量记为  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ 。我们考虑每组样本中的预测变量为  $\mathbf{X}_i$ ，响应变量为  $Y_i$ ， $p$  维参数为  $\beta$ 。目标函数的一般形式为：

$$\min_{\beta} L(\beta) = \min_{\beta} \left\{ Q(\beta | \mathbf{Y}, \mathbf{X}) + \sum_{j=1}^p p_{\lambda}(\|\beta_j\|) \right\}$$

其中  $Q(\beta | \mathbf{Y}, \mathbf{X})$  是**损失函数**，不同模型的损失函数形式不同，线性模型的损失函数为最小二乘函数、Logistic 回归的损失函数为极大似然函数的负向变换。

$p_{\lambda}(\|\beta_j\|)$  为**正则化项**，所谓正则化项，也叫惩罚项，描述的其实是模型的复杂度，模型的复杂度越高，过拟合的风险也就越大。惩罚项的惩罚力度越大时，模型的复杂度越低。例如前面章节讲到的岭回归，是在极小化损失函数的基础上通过对系数添加  $L_2$  范数惩罚来进行系数的连续收缩的，它是 Hoerl 和 Kennard [105] 于 1970 年提出的，是统计学著名的系数收缩方法之一。

岭回归的目标函数是：

$$\min_{\beta} L(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$

其中第一项是残差平方和 RSS（线性模型的损失函数），第二项是  $L_2$  正则化项， $\lambda$  为**调节参数**， $\lambda \geq 0$ ，为非负数，其作用是控制损失函数和正则化项对回归系数估计的相对影响程度， $\lambda = 0$  时，岭回归估计值与最小二乘估计值相同， $\lambda$  越大，则为了使目标函数最小，回归系数  $\beta$  的估计值就越接近于 0。

### 2、范数

我们都知道，函数与几何图形往往是有对应的关系，这个很好想象，特别是在三维以下的空间内，函数是几何图像的数学概括，而几何图像是函数的高度形象化。但当函数与几何超出三维空间时，就难以获得较好的想象，于是就有了映射的概念，进而引入范数的概念。

当我们有了范数的概念后，我们就可以引出两个向量的距离的定义，这个向量可以是任意维数的。通过距离的定义，进而我们可以讨论逼近程度，从而讨论收敛性、求极限。

**$L_0$ -范数：**即向量中非零元素的个数。其公式为：

$$\|\mathbf{X}\|_0 = \sum_i \mathbb{I}(X_j \neq 0)$$

**L<sub>1</sub>-范数**: 即向量各元素绝对值之和。其公式为:

$$\|\mathbf{X}\|_1 = |x_1| + |x_2| + \dots + |x_n|$$

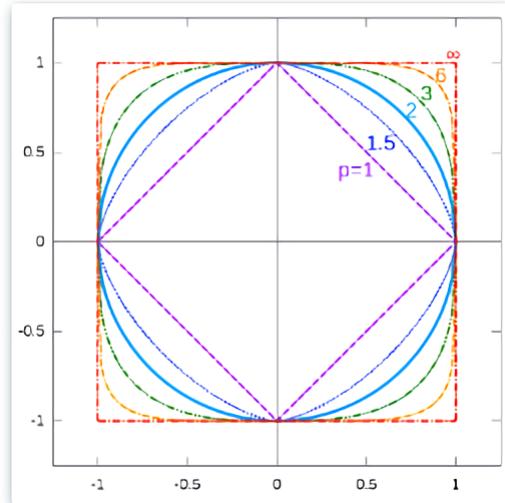
**L<sub>2</sub>-范数**: 欧几里得范数, 常用计算向量长度, 即向量元素绝对值的平方和再开方:

$$\|\mathbf{X}\|_2 = (\|x_1\|^2 + \|x_2\|^2 + \dots + \|x_n\|^2)^{1/2}$$

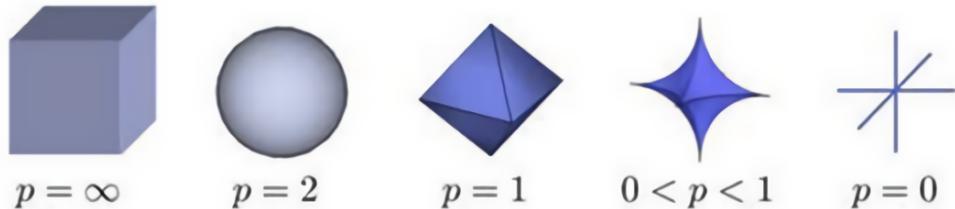
**L<sub>p</sub>-范数**: 即向量元素绝对值的  $p$  次方和的  $1/p$  次幂:

$$\|\mathbf{X}\|_p = (\|x_1\|^p + \|x_2\|^p + \dots + \|x_n\|^p)^{1/p}$$

下图展示了直角坐标系中  $p$  取不同的值时各个范数下单位向量终点的轨迹:



下图表示了  $p$  从无穷到 0 变化时, 三维空间中到原点的距离(范数)为 1 的点构成的图形的变化情况, 以常见比如  $p = 2$  时为例, 此时的范数也即欧氏距离, 空间中到原点的欧氏距离为 1 的点构成了一个球面:



其中, L<sub>1</sub> 范数正则化、L<sub>2</sub> 范数正则化都有助于降低过拟合风险, 我们在此主要运用 L<sub>1</sub> 和 L<sub>2</sub> 范数正则化, 或者二者相结合的方法。对于线性回归模型, 使用 L<sub>1</sub> 正

则化的模型叫做 Lasso 回归，使用  $L_2$  正则化的模型叫做 岭回归 (Ridge 回归)，两者相结合的方法叫做 弹性网。

### 8.4.1 Lasso 回归

与岭回归类似，Lasso 回归是在极小化残差平方和的基础上通过对系数添加  $L_1$  范数惩罚来收缩系数的，1996 年由 Robert Tibshirani [78] 首次提出，全称为：Least absolute shrinkage and selection operator。该方法是一种压缩估计，当  $\lambda$  充分大时，可以把某些待估系数精确的收缩到零，因而特别适用于参数数目的缩减与参数的选择。

由于该方法以岭回归为基础，增加了稳定性。Lasso 的核心是稀疏性，稀疏性的优势在于它可以解释拟合模型，并且计算简单。

Lasso 回归的目标函数是：

$$\min_{\beta} L(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda\|\beta\|_1$$

#### 岭回归和 Lasso 的比较

岭回归和 Lasso 回归的主要区别就是在正则化项，岭回归用的是  $L_2$  正则化，而 Lasso 回归用的是  $L_1$  正则化。 $L_2$  正则能够有效的防止模型过拟合，解决非满秩下求逆困难的问题； $L_1$  正则化最大的特点是能稀疏矩阵，进行庞大预测变量数量下的预测变量选择。

如果把二者目标函数最小化看成是有约束的极值问题，那么岭回归可以表达为：

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

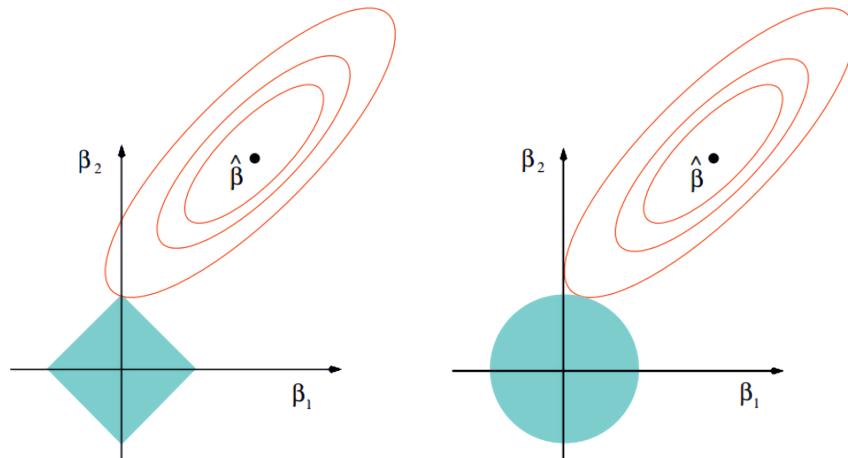
$$s.t. \quad \|\beta\|_2^2 \leq t$$

Lasso 回归可以表达为：

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

$$s.t. \quad \|\beta\|_1 \leq t$$

因此岭回归和 Lasso 回归可以看作是两个同样目标函数但在不同的约束区域下求解的问题。



如图，二维空间为例，不添加正则项的目标函数可以用一圈圈等值线表示，约束区域对应图中的蓝色区域。等值线和约束域的切点就是目标函数的最优解。岭方法对应的约束域是圆，其切点只会存在于圆周上，不会与坐标轴相切，则在任一维度上的取值都不为0，因此没有稀疏；对于 Lasso 方法，其约束域是正方形，会存在与坐标轴的切点，使得部分回归系数变为零，因此很容易产生稀疏的结果。

所以，Lasso 方法可以达到变量选择的效果，将不显著的变量系数压缩至0，在估计参数的同时实现了变量选择；而岭方法虽然也对原本的系数进行了一定程度的压缩，但是任一系数都不会压缩至0，只有压缩功能，没有选择功能，最终模型保留了所有的变量。Lasso 回归的计算量将远远小于岭回归，但无论对于岭回归还是 Lasso 回归，它们的本质都是通过调节  $\lambda$  来实现模型偏差和方差的平衡调整。

### 8.4.2 非凸惩罚函数回归——SCAD 和 MCP

#### SCAD

Lasso 总是有偏估计，这是因为 Lasso 的调节参数  $\lambda$  对所有参数一视同仁，导致大的系数被过分压缩，带来较大的估计偏差，而且 Lasso 的结果具有不稳定性。由 Fan 和 Li [106] 在 2001 年提出了 SCAD (Smoothly Clipped Absolute Deviation Penalty) 较好的弥补了 Lasso 的不足，更进一步地，该方法具有 Oracle 性质，使预测效果与真实模型别无二致。

Fan 和 Li 指出，一个好的惩罚函数应该使得最后得到的估计量满足三个性质：  
1. 无偏性。即当未知参数的真实值很大时，估计值应当近乎无偏。  
2. 稀疏性。即可以自动地将很小的估计系数压缩为0，降低模型的复杂度。  
3. 连续性。即估计的系数应当是连续的，保证模型预测的稳定性。

该方法也是在损失函数的基础上施加惩罚项，例如基于二次损失的惩罚最小二

乘目标函数可以表示为：

$$L(\boldsymbol{\beta}) = \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p p_\lambda(|\beta_j|) \quad (8.4.1)$$

SCAD 的惩罚函数为：

$$p_\lambda(|\beta_j|; \alpha) = \begin{cases} \lambda |\beta_j| & \text{若 } 0 \leq |\beta_j| < \lambda \\ -\frac{|\beta_j|^2 - 2\alpha\lambda|\beta_j| + \lambda^2}{2(\alpha - 1)} & \text{若 } \lambda \leq |\beta_j| < \alpha\lambda \\ (\alpha + 1)\lambda^2/2 & \text{其它} \end{cases}$$

其中， $\alpha$  和  $\lambda$  是两个调节参数。

考虑  $\beta_j \in [0, +\infty)$  的情况，SCAD 惩罚函数的导数为：

$$p'_\lambda(|\beta_j|; \alpha) = \begin{cases} \lambda, & \beta_j < \lambda \\ \frac{\alpha\lambda - \beta_j}{\alpha - 1}, & \lambda < \beta_j < \alpha\lambda \\ 0 & \beta_j > \alpha\lambda. \end{cases}$$

可以看出 SCAD 惩罚函数的导数在原点附近与 Lasso 相同，离原点越远导数值越小，直至为零，也就是说，SCAD 方法对较大的参数施加较少的惩罚，当参数大于  $\alpha\lambda$ ，不施加惩罚。因此显著的变量更容易被选入模型。

### MCP

2010 年，Zhang [107] 提出的 MCP 方法保留了 SCAD 的渐进无偏优点，其惩罚函数同样是分段函数：

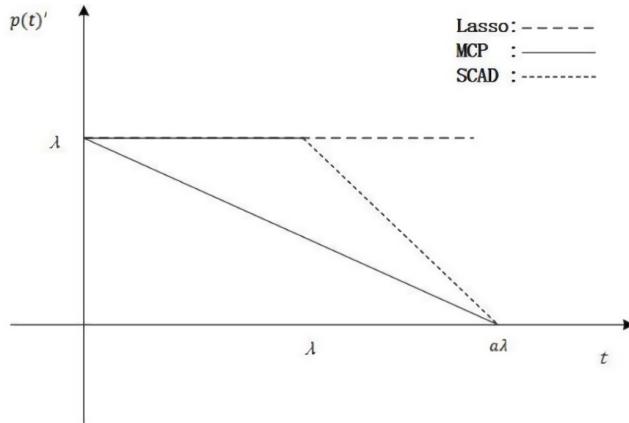
$$p_\lambda(|\beta_j|; \lambda, \alpha) = \begin{cases} \lambda |\beta_j| - \frac{|\beta_j|^2}{2\alpha}, & \text{若 } |\beta_j| \leq \alpha\lambda \\ \frac{\alpha\lambda^2}{2}, & \text{若 } |\beta_j| > \alpha\lambda \end{cases}$$

其惩罚函数的导数为：

$$p'_\lambda(|\beta_j|; \lambda, \alpha) = \begin{cases} \lambda - \frac{|\beta_j|}{\alpha} & \text{若 } |\beta_j| \leq \alpha\lambda \\ 0 & \text{若 } |\beta_j| > \alpha\lambda \end{cases}$$

可以看出，当  $\beta_j$  近似于零时，MCP 和 Lasso 的惩罚力度一致，而随  $\beta_j$  从零开始增大，MCP 的惩罚力度逐渐缩减为零。当  $|\beta| > \alpha\lambda$  时，MCP 对大系数不施加惩罚，因此 MCP 同 SCAD 类似，也实现了回归系数的有差别惩罚，实现了更精确的估计。

下图展示了回归系数变化时 Lasso、SCAD 和 MCP 的惩罚力度的变化情况：



### 8.4.3 群组变量选择方法

随着科技发展，数据的种类与维数的不断增加，导致了回归建模时，预测变量的群组结构成为了一种普遍现象。比如，变量间的相关性无法忽略时，建模时应当处理共线性问题；或者在医学上研究基因对疾病的影响时，一般把同一基因下的变量作为一组变量来进行处理；或者在处理多分类变量时，经常使用的虚拟变量也是一类组变量，在进行组变量选择时它们要当成一个整体去对待。在这种情况下，相关的选择问题就变成选择组而不仅仅是单个变量，群组变量选择模型及其算法成为当下高维数据建模的主要研究方向，在此我们介绍两种代表性方法。

#### 弹性网 (Elastic Net)

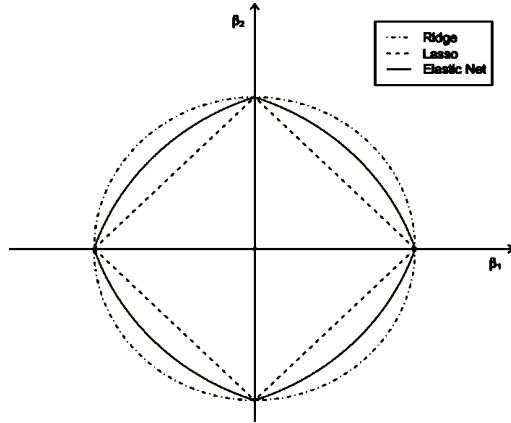
岭回归结果表明，岭回归虽然一定程度上可以拟合模型，但容易导致回归结果失真；Lasso 回归虽然能刻画模型代表的现实情况，但是模型过于简单，不符合实际，当多个变量和另一个变量高度相关的时候，Lasso 倾向于随机选择其中一个。但如果这几个变量都是对模型重要的变量，那么 Lasso 只能从多个相关强预测变量中选一个。

Zou 和 Hastie [108] 在 2005 年提出的弹性网结合了岭回归和 Lasso 算法，既能做到岭回归不能做的预测变量提取，又能实现 Lasso 不能做的预测变量分组，是最早解决共线性问题的变量选择方法，能够将显著的强相关变量同时选入模型。其目标函数为：

$$\min_{\beta} (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) + \lambda [\alpha \|\beta\|_1 + (1 - \alpha) \|\beta\|_2^2]$$

显然，这种方法是利用  $L1 + L2$  范数的凸组合来实现约束的，其中  $L1$  正则化用于变量选择， $L2$  正则化用于解决共线性问题。除了加入调节参数  $\lambda$ ，我们还引入参数  $\alpha$  来调节  $L1$  范数和  $L2$  范数的平衡。当  $\alpha = 0$  时，弹性网回归即为岭回归；当

$\alpha = 1$  时，弹性网回归即为 Lasso 回归。当  $\alpha = 0.5$  的时候，弹性网的几何结构如下图所示：



由图形可知，弹性网回归在坐标轴上有尖角，在各个象限内部呈弧形，因此既具备 Lasso 的变量选择功能，又具备 Ridge 的系数收缩功能。

### Group Lasso

Yuan 和 Lin [109] 在 2006 年将 Lasso 方法推广到组群结构上面，诞生了 Group Lasso。我们可以将所有变量分组，然后在目标函数中加总每个组系数的 L2 范数，这样达到的效果就是可以将一整组的系数同时消成零，即抹掉一整组的变量，这种方法叫做 Group Lasso 算法。其目标函数为：

$$\min_{\beta} \left( \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \sum_{g=1}^G \|\sqrt{q_l} \beta_{I_g}\|_2 \right)$$

其中， $I_g$  是  $g$  组的预测变量下标， $g = 1, 2, \dots, G$ ， $G$  代表组的个数， $\lambda$  是调整参数， $\sqrt{q_l}$  是每一组的加权，可按需调节。其惩罚函数中，在组内，是只具有压缩功能而没有选择功能的岭惩罚；在组间是具有变量选择功能的 Bridge 惩罚，因而整组变量会同时被选入或删除。我们在此提到的 Bridge 惩罚，即  $0 < p < 1$  的  $L_p$  罚函数，从罚函数形式来看，Lasso 对系数  $\beta_j$  的惩罚强度是固定的，而 Bridge 对  $\beta_j$  的惩罚强度随  $\beta_j$  的增大而减小。当  $\beta_j$  趋于 0 时，Bridge 的惩罚强度趋于无穷大，使得其稀疏性较 Lasso 更加显著，而当  $\beta_j$  很大时，Bridge 的惩罚强度变得很小，使得其参数估计近似无偏。

由于 Group Lasso 是一种通过对惩罚项的修改将 Lasso 方法在群组结构下的拓展，因此与 Lasso 具备相同的缺点，惩罚率不随组系数的大小而变化，因此对较大的系数过分压缩，导致估计偏差。

#### 8.4.4 双层变量选择方法

Group Lasso 方法由于只进行组变量的选择，不能选择组内重要的单变量，使得同组变量全被选择或全被删除，因此有些显著单变量在进行变量选择时会随着组变量被淘汰掉，从而导致模型误差过大，精确度降低。为了同时既考虑组的选择又考虑组内重要变量选择，我们引入双层变量选择方法。

##### Group Bridge

Huang 等 [110] 提出的 Group Brige 是一种可以满足组间、组内选择的双层变量选择方法，是最早实现双层变量选择的方法。其惩罚函数的形式为：

$$P(\beta; \lambda, \gamma) = \lambda \sum_{j=1}^W p_j^\gamma \|\beta^{(j)}\|_1^\gamma$$

其中  $0 < \gamma < 1$ ,  $p_j$  是第  $j$  组变量所包含个数， $\beta^{(j)} = (\beta_1^{(j)}, \dots, \beta_{p_j}^{(j)})$  是第  $j$  组预测变量的回归系数，其中回归系数共分为  $W$  个组。其惩罚函数是由组内惩罚和组间惩罚函数复合而成，其中，组内选择的是 Lasso 惩罚函数，组间选择的是 Bridge 惩罚函数，从而实现了组间变量与组内变量的选择。

##### Composite MCP

Composite MCP 是 Huang 等人 2009 年提出的另一种复合惩罚类的变量选择方法 [111]，其惩罚函数形式为：

$$P(\beta; \lambda, a, b) = \sum_{j=1}^W P_{\text{MCP}} \left( \sum_{m=1}^{p_j} P_{\text{MCP}}(|\beta_j^{(m)}|; \lambda, a); \lambda, b \right)$$

其组内和组间均使用 MCP 惩罚，均具备单变量选择的功能。

##### SGL (Sparse Group Penalty)

在双层变量选择方法中也存在另一种形式，它的惩罚函数不是由两个具有单变量选择的功能惩罚函数复合而成的，而是由单个变量惩罚和仅选择组变量惩罚的线性组合构成的函数。这种惩罚方法叫做稀疏组惩罚。

Simon 等人 2013 年提出的 SGL [112] 就属于此类方法，其惩罚函数形式为：

$$P_{\text{SGL}}(\beta; \lambda_1, \lambda_2) = \lambda_1 \sum_{j=1}^W \|\beta^{(j)}\| + \lambda_2 \|\beta\|_1$$

其中第一项是选择重要组变量，选择的是 Group Lasso 方法的惩罚函数；第二个惩罚是选择单个变量，用的是 Lasso 方法的惩罚函数，其中  $\lambda_1$ 、 $\lambda_2$  分别为作用在单个变量和组变量上的调整参数。

##### Adaptive SGL

SGL 相对复合函数型方法计算简单，但它对于所有系数及组系数的惩罚力度相同，过度压缩大系数，引起估计偏差。基于此 Fang 等 2014 年提出了更一般化

的Adaptive SGL [113]，其惩罚函数形式为：

$$P_{\text{adSGL}}(\boldsymbol{\beta}; \lambda_1, \lambda_2) = \lambda_1 \sum_{j=1}^W w_j \|\boldsymbol{\beta}^{(j)}\|_2 + \lambda_2 \sum_{j=1}^W \xi^{(j)T} |\boldsymbol{\beta}^{(j)}|$$

它通过引入权重  $\xi = (\xi^1, \dots, \xi^W)$  和  $w = (w_1, \dots, w_W)$  分别对单个系数和组系数施加不同程度的惩罚，权重依据样本数据而定，系数的真实值越大，给予权重越小，惩罚力度越小，增进了预测精度。

## 8.5 模型选择实践

### 8.5.1 R 语言实现

下面介绍怎么用 R 语言实现上面提到的模型选择方法。

#### 子集选择法

本节使用 ISLR 包中的 College 数据集，该数据集包含了自 1995 年以来美国大学的 777 条数据，每条数据含有 18 个变量。其中将变量“Apps”作为因变量，其余 17 个变量作为预测变量进行下面的分析。具体实现过程如下：

##### 1、最优子集法

```
library(ISLR)
#R 语言的包leaps中的regsubsets()函数可用于筛选最优因变量子集
library(leaps)
data(College)
names(College)
dim(College)

#Apps为因变量、College为数据、nvmax表示所输出结果最优模型所包含的变量个数，默认值为8
subset.full<-regsubsets(Apps ~ .,College,nvmax = 3)
full.summary<-summary(subset.full)

#输出结果如下：
Selection Algorithm: exhaustive
      PrivateYes Accept Enroll Top10perc Top25perc F.Undergrad
1 (1)   " "     "*"   " "    " "    " "    " "
2 (1)   " "     "*"   " "    "*"   " "    " "
3 (1)   " "     "*"   " "    "*"   " "    " "
          P.Undergrad Outstate Room.Board Books Personal PhD Terminal
1 (1)   " "        " "      " "    " "    " "    " "
2 (1)   " "        " "      " "    " "    " "    " "
3 (1)   " "        " "      " "    " "    " "    " "
          S.F.Ratio perc.alumni Expend Grad.Rate
1 (1)   " "        " "      " "    " "
```

```

2 (1)   " "
3 (1)   " "
#summary()函数的结果
names(full.summary)
full.summary$rsq
full.summary$cp

```

结果中的星号“\*”表示该列对应的变量出现在该行对应的最优模型中，最前面的阿拉伯数字表示该模型中含有的变量数，如第三行结果表示该最优模型含有3个变量分别为Accept、Expend、Top10perc。除此之外summary()函数还包含 $\bar{R}^2$ 、Rss、调整 $\bar{R}^2$ 、 $C_p$ 、BIC等结果。

从结果中可以看出随着模型中包含变量个数的增加， $\bar{R}^2$ 不断增大， $C_p$ 值先不断地减小再增加，在变量数为12时达到最小。最优子集法选择模型时，根据不同的准则选择，例如在一定条件下选择 $\bar{R}^2$ 最大、 $C_p$ 、AIC、BIC值最小的模型。可以用图形表示出模型中变量个数与相应指标值的关系，从而更加方便明确的选择最优模型：

```

par(mfrow = c(1,3))
which.min(full.summary$cp) # CP
plot(full.summary$cp,xlab="Number of Variables",ylab="CP",type="b")
points(12,full.summary$cp[12],col="red",cex=2,pch=20)
which.min(full.summary$bic) # BIC
plot(full.summary$bic,xlab="Number of Variables",ylab="BIC",type="b")
points(10,full.summary$bic[10],col="red",cex=2,pch=20)
which.max(full.summary$adjr2) # Adjust Rsq
plot(full.summary$adjr2,xlab="Number of Variables",ylab="Adjusted Rsq",type="b")
points(13,full.summary$adjr2[13],col="red",cex=2,pch=20)
#提取含有10个变量的最优模型
coef(subset.full,10)

```

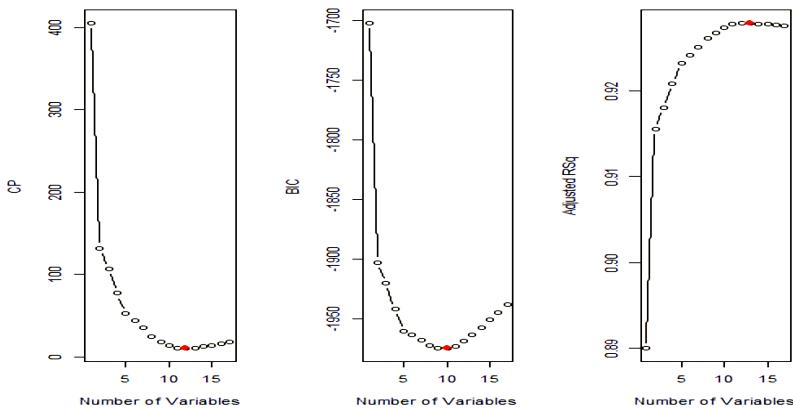


图 8.2 模型选取的变量个数与不同统计指标的关系

从图8.2中可以看出随着模型中变量个数的增加 $C_p$ 、BIC的值是先下降后上升，

变量数分别为 12 和 10 时相应指标值达到最小，调整  $\bar{R}^2$  的值是先上升后下降，在变量数为 13 时达到最大。

## 2、向前逐步选择法与向后逐步选择法

向前逐步选择法与向后逐步选择法仍是使用 `regsubsets()` 函数，分别将参数 `method` 设置为 `forward` 和 `backward`，并分别利用 `coef()` 函数提取含有 8 个变量的最优模型

```
#向前逐步选择法
subset.fwd<-regsubsets(Apps ~ ., College, nvmax = 17, method = "forward")
summary(subset.fwd)

#向后逐步选择法
subset.bwd<-regsubsets(Apps ~ ., College, nvmax = 17, method = "backward")
summary(subset.bwd)
coef(subset.full,8)
coef(subset.fwd,8)
coef(subset.bwd,8)
```

## 惩罚方法

下面说明这三种方法分别在线性回归和 logistic 回归下的变量选择。

### 1、线性回归

首先产生样本量为 100 的样本，每个样本变量个数  $p=20$ ，各预测变量独立同分布于标准正态分布，由向量 `beta` 给出回归系数的真实值，误差项满足标准正态分布。在实现 LASSO 惩罚、MCP 惩罚以及 SCAND 惩罚这三种方法的变量选择时，使用 R 语言中的 `glmnet` 包和 `ncvreg` 包，其中 `glmnet` 包的 `cv.glmnet()` 函数默认使用 LASSO 方法选择变量，`ncvreg` 包中的 `cv.ncvreg()` 函数默认使用 MCP 方法选择变量。

#### (1) LASSO 惩罚

```
x<-matrix(rnorm(100*20), 100, 20)
#设定第7、8、9、10个变量对应的回归系数为零
beta<-c(seq(1, 2, length.out = 6), 0, 0, 0, 0, rep(1, 10))
y<-x %*% beta + rnorm(100)
library(glmnet)
fit1<-cv.glmnet(x, y, family = "gaussian")
beta.fit1<-coef(fit1)#提取参数的估计值
beta.fit1
resid1 <-(x %*% beta.fit1 [ -1 ] + beta.fit1 [ 1 ] - y)
MSE1<-sum(resid1 ^ 2)#计算残差平方和
MSE1
```

#### (2) MCP 惩罚和 SCAD 惩罚

```
library(ncvreg)
fit2<-cv.ncvreg(x, y, family = "gaussian")#MCP
fit3<-cv.ncvreg(x, y, family = "gaussian", penalty = "SCAD")
fit.mcp<-fit2$fit
```

```

fit.scad<-fit3$fit
beta.fit2<-fit.mcp$beta [, fit2$min ]#提取参数的估计值
beta.fit3<-fit.scad$beta [, fit3$min ]#提取参数的估计值
round(beta.fit2, 3)#保留三位小数
round(beta.fit3, 3)
resid2 <-(x %*% beta.fit2 [-1] + beta.fit2 [1] - y)
resid3 <-(x %*% beta.fit3 [-1] + beta.fit3 [1] - y)
MSE2<-sum(resid2 ^ 2)#计算残差平方和
MSE3<-sum(resid3 ^ 2)

```

## 2、Logistic 回归

在进行 Logistic 回归时使用的数据是 ncvreg 包中自带的 heart 数据集，该数据共 462 个样本，每个样本含有 10 个变量，以变量“chd”为因变量，其余 9 个变量作为预测变量。

### (1) LASSO 惩罚

```

library(ncvreg)
data(heart)
x<-as.matrix(heart[,1:9])
y<-heart$chd
library(glmnet)
fit1<-cv.glmnet(x, y, family = "binomial")
beta.fit1<-coef(fit1)#提取参数的估计值
beta.fit1
round(beta.fit1, 3)

```

### (2) MCP 和 SCAD 惩罚

```

fit2<-cv.ncvreg(x, y, family = "binomial")#MCP
fit3<-cv.ncvreg(x, y, family = "binomial", penalty="SCAD")#SCAD
fit.mcp<-fit2$fit
fit.scad<-fit3$fit
beta.fit2<-fit.mcp$beta [, fit2$min ]
beta.fit3<-fit.scad$beta [, fit3$min ]
round(beta.fit2, 3)
round(beta.fit3, 3)

```

## 组模型选择

### 1、线性回归

使用 R 语言中的 grpreg 包来处理 Group LASSO 问题，使用 glmnet 包处理弹性网问题，下面以 grpreg 包中的 birthwt.grpreg 数据为例进行分析。

### (1) Group LASSO

```

library(grpreg)
data(birthwt.grpreg)
X<-as.matrix(birthwt.grpreg [,-1 : -2])
y<-birthwt.grpreg$bwt
colnames(X)#变量组结构，数字相同的为同一类变量

```

```
group<-c(1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 8)
cvfit<-cv.grpreg(X, y, group, penalty = "grLasso")
beta.cvfit<-coef(cvfit)#系数估计值
round(beta.cvfit,3)
```

## (2) 弹性网

```
library(glmnet)
library(grpreg)
data(birthwt.grpreg)
X<-as.matrix(birthwt.grpreg [, -1:-2])
y<-birthwt.grpreg$bwt
fit.enet<-cv.glmnet(X, y, family = "gaussian", alpha = 0.2)
beta.enet<-coef(fit.enet)
round(beta.enet,3)
```

## 2、Logistic 回归

进行 Logistic 回归时仍使用 birthwt.grpreg 数据集，由于 cv.grpreg() 函数可以自动识别因变量的数据类型从而进行对应的回归，所以在使用 cv.grpreg() 时没有加 family 参数。

### (1) Group LASSO

```
library(grpreg)
data(birthwt.grpreg)
X<-as.matrix(birthwt.grpreg [, -1 : -2])
y<-birthwt.grpreg$low
group<-c(1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 8)#变量的分组结构
cvfit<-cv.grpreg(X, y, group, penalty = "grLasso")
beta.cvfit<-coef(cvfit)
summary(cvfit)
plot(cvfit)
```

大家可以输出一下结果，可知有 6 个变量没有选入模型，且从 summary() 的结果可以看出，最优的  $\lambda$  值为 0.0379，在该最优  $\lambda$  值下模型的交叉验证误差达到最小为 0.2。图8.3展示了连续  $\lambda$  对数值下交叉验证误差的变化：

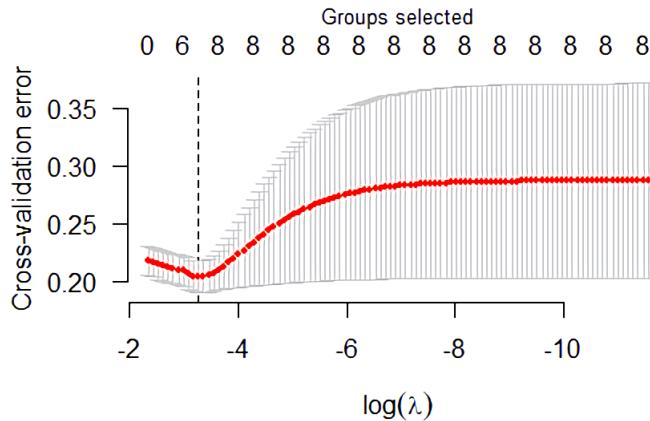


图 8.3 birthwt.grpreg 使用 Group LASSO 方法的交叉验证图

## (2) 弹性网

```
y<-birthwt.grpreg$low
fit.enet<-cv.glmnet(X, y, family = "binomial", alpha = 0.1)
beta.enet<-coef(fit.enet)
round(beta.enet,3)
```

## 双层模型选择

通过 R 语言中的 grpreg 包实现 Group Bridge 和 Composite MCP，通过 SGL 包实现 SGL 惩罚，下面仍然以 Birthwt.grpreg 为例说明这两种方法分别在线性回归和 Logistic 回归中的变量选择。

### 1、线性回归

#### (1) 复合函数双层选择：Group Bridge 和 Composite MCP

```
#Group Bridge
library(grpreg)
data(birthwt.grpreg)
X<-as.matrix(birthwt.grpreg [, -1 : -2])
y<-birthwt.grpreg$bwt
group<-c(1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 8)#变量的分组结构
cvfit.b<-gBridge(X, y, group)#L1 group bridge
beta.cvfit.b<-select(cvfit.b)$beta
round(beta.cvfit.b,3)
cvfit.m<-cv.grpreg(X, y, group, penalty = "cMCP", gama = 2.5)#Composite MCP
coef(cvfit.m)
```

#### (2) 稀疏组惩罚：SGL

```
library(SGL)
```

```

library(grpreg)
data(birthwt.grpreg)
X<-as.matrix(birthwt.grpreg [ , -1 : -2 ])
y<-birthwt.grpreg$bwt
group<-c(1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 8)#变量的分组结构
data<-list(x = X, y = y)
cvFit<-cvSGL(data, group, type = "linear")#SGL
lambda.min<-which.min(cvFit$lldiff)
beta.cvFit<-cvFit$fit$beta [ , lambda.min ]
round(beta.cvFit,3)

```

## 2、Logistic 回归

### (1) 复合函数双层选择: Group Bridge 和 Composite MCP

```

# Group Bridge
library(grpreg)
data(birthwt.grpreg)
X<-as.matrix(birthwt.grpreg [ , -1 : -2 ])
y<-birthwt.grpreg$low
cvfit.b<-gBridge(X, y, group, family = "binomial")
beta.cvfit.b<-select(cvfit.b)$beta
round(beta.cvfit.b,3)
#Composite MCP
cvfit.m<-cv.grpreg(X, y, group, penalty = "cMCP", family = "binomial")
beta.cvfit.m<-coef(cvfit.m)
round(beta.cvfit.m,3)

```

### (2) 稀疏组惩罚: SGL

```

library(SGL)
library(grpreg)
data(birthwt.grpreg)
X<-as.matrix(birthwt.grpreg [ , -1 : -2 ])
y<-birthwt.grpreg$low
group<-c(1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 5, 6, 7, 8, 8, 8)#变量的分组结构
data<-list(x = X, y = y)
cvFit<-cvSGL(data, group, type = "logit")
lambda.min<-which.min(cvFit$lldiff)#最优值
beta.cvFit<-cvFit$fit$beta [ , lambda.min ]#最优值时回归系数的估计结果
round(beta.cvFit,3)
cvFit$fit$intercepts [ lambda.min ]

```

## 8.5.2 Python 语言实现

本节使用 python 中 sklearn 模块中自带的数据集。sklearn 中含有很多类型的数据集，例如，“20newsgroups”，“20newsgroups\_vectorized”用于文本分析；“california\_housing”进行回归；“wine”进行分类。本书选择“california\_housing”即加州房屋数据集作为代码部分的支撑。

下面先对该数据集进行简单的介绍：该数据集是基于 1990 年加州普查的数据，数据集中的每一个数据都代表着一块区域内房屋和人口的基本信息，总共包括 9 项：该地区中心的纬度 (latitude)、该地区中心的经度 (longitude)、区域内所有房屋屋龄的中位数 (housingMedianAge)、区域内总房间数 (totalRooms)、区域内总卧室数 (totalBedrooms)、区域内总人口数 (population)、区域内总家庭数 (households)、区域内人均收入中位数 (medianIncome)、该区域房价的中位数 (medianHouseValue)。这里的地域单位是美国做人口普查的最小地域单位，平均一个地域单位中有 1400 多人。

本书主要以房屋价值中位数为标签，其余属性为预测变量，通过使用 python 编程，利用 python 相关库，如 numpy、pandas、statsmodels 等库，对加州地区的数据进行预测变量选择。

## 逐步回归

首先导入相应模块，并定义用于逐步回归的 stepwise\_select 函数

```
import numpy as np
import pandas as pd
import statsmodels.api as sm
from statsmodels.formula.api import ols #加载ols模型
#####逐步回归
def stepwise_select(data, label, cols_all, method='forward'):
    ...
    data: 数据源(数据框) label: 标签(字符串) cols_all: 逐步回归的全部字段
    method: 方法(forward:向前, backward:向后, both:双向)
    函数返回值: select_col: 最终保留的字段列表(列表) summary: 模型参数
    AIC: 赤池信息量
    ...
#1. 前向回归
#从一个变量都没有开始，将变量逐个加入到模型中，直至没有可以再加入的变量结束
if method == 'forward':
    add_col = []
    AIC_None_value = np.inf
    while cols_all:
        # 单个变量加入，计算aic
        AIC = {}
        for col in cols_all:
            print(col)
            X_col = add_col.copy()
            X_col.append(col)
            X = sm.add_constant(house_data[X_col])
            y = house_data[label]
            # 将预测变量名连接起来
            formula = "{}~{}".format(label, "+".join(add_col + [col]))
            AIC[col] = ols(formula=formula, data=house_data).fit().aic
        AIC_min_value = min(AIC.values())
        AIC_min_key = min(AIC, key=AIC.get)
        # 如果最小的aic小于不加该变量时的aic，则加入变量，否则停止
        if AIC_min_value < AIC_None_value:
            add_col.append(AIC_min_key)
            cols_all.remove(AIC_min_key)
            AIC_None_value = AIC_min_value
```

```

        if AIC_min_value < AIC_None_value:
            cols_all.remove(AIC_min_key)
            add_col.append(AIC_min_key)
            AIC_None_value = AIC_min_value
        else:
            break
    select_col = add_col
#2. 后向回归
# 从全部变量都在模型中开始，一个变量一个变量的删除，直至没有可以再删除的变量结束
elif method == 'backward':
    # 全部变量，一个都不剔除，计算初始aic
    X_col = cols_all.copy()
    X = sm.add_constant(house_data[X_col])
    y = house_data[label]
    formula = "{}-{}".format(label, "+".join(cols_all)) # 将预测变量名连接起来
    LR = ols(formula=formula, data=house_data).fit()
    AIC_None_value = LR.aic
    while p:
        # 删除一个字段提取aic最小的字段
        AIC = {}
        for col in cols_all:
            print(col)
            X_col = [i for i in cols_all if i != col]
            X = sm.add_constant(house_data[X_col])
            cols_all_remove = cols_all.copy()
            cols_all_remove.remove(col)
            # 将预测变量名连接起来
            formula = "{}-{}".format(label, "+".join(cols_all_remove))
            AIC[col] = ols(formula=formula, data=house_data).fit().aic
        AIC_min_value = min(AIC.values())
        AIC_min_key = min(AIC, key=AIC.get)
        # 如果最小的aic小于不删除该变量时的aic，则删除该变量，否则停止
        if AIC_min_value < AIC_None_value:
            cols_all.remove(AIC_min_key)
            AIC_None_value = AIC_min_value
            p = True
        else:
            break
    select_col = cols_all
    ##### 3. 双向选择
elif method == 'both':
    p = True
    add_col = []
    # 全部变量，一个都不剔除，计算初始aic
    X_col = cols_all.copy()
    X = sm.add_constant(house_data[X_col])
    y = house_data[label]
    formula = "{}-{}".format(label, "+".join(cols_all)) # 将预测变量名连接起来
    LR = ols(formula=formula, data=house_data).fit()
    AIC_None_value = LR.aic
    while p:
        # 删除一个字段提取aic最小的字段

```

```
AIC = {}
for col in cols_all:
    print(col)
X_col = [i for i in cols_all if i != col]
X = sm.add_constant(house_data[X_col])
# 将预测变量名连接起来
formula = "{}~{}".format(label, "+".join(add_col + [col]))
LR = ols(formula=formula, data=house_data).fit()
AIC[col] = LR.aic
AIC_min_value = min(AIC.values())
AIC_min_key = min(AIC, key=AIC.get)
if len(add_col) == 0: # 第一次只有删除操作, 不循环加入变量
    if AIC_min_value < AIC_None_value:
        cols_all.remove(AIC_min_key)
        add_col.append(AIC_min_key)
        AIC_None_value = AIC_min_value
        p = True
    else:
        break
else:
    # 单个变量加入, 计算aic
    for col in add_col:
        print(col)
        X_col = cols_all.copy()
        X_col.append(col)
        X = sm.add_constant(house_data[X_col])
        formula = "{}~{}".format(label, "+".join(add_col + [col])) #
将预测变量名连接
起来
        LR = ols(formula=formula, data=house_data).fit()
        AIC[col] = LR.aic
        AIC_min_value = min(AIC.values())
        AIC_min_key = min(AIC, key=AIC.get)
        if AIC_min_value < AIC_None_value:
            # 如果aic最小的字段在添加变量阶段产生, 则加入该变量, 如果aic
            最小的字段在删除阶段产生, 则删除该变量
            if AIC_min_key in add_col:
                cols_all.append(AIC_min_key)
                add_col = list(set(add_col) - set(AIC_min_key))
                p = True
            else:
                cols_all.remove(AIC_min_key)
                add_col.append(AIC_min_key)
                p = True
                AIC_None_value = AIC_min_value
        else:
            break
    select_col = cols_all
#模型
X = sm.add_constant(house_data[select_col])
# 将预测变量名连接起来
formula = "{}~{}".format(label, "+".join(select_col))
```

```

LR = ols(formula=formula, data=house_data).fit()
summary = LR.summary()
AIC = LR.aic
return select_col, summary, AIC

```

再导入加利福尼亚房屋价值数据，调用上文定义的 stepwise\_select 函数进行模型选择

```

#加载加利福尼亚房屋价值数据
from sklearn.datasets import fetch_california_housing as fch
data=fch() #导入数据
house_data=pd.DataFrame(data.data) #将预测变量转换成dataframe格式，便于查看
house_data.columns=data.feature_names #命名预测变量
house_data.loc[:, "value"] = data.target #合并预测变量，因变量数据
house_data.shape #查看数据量
house_data.head(10) #查看前10行数据
cols_all=set(house_data.columns) #将字段名转换成字典类型
cols_all.remove('value')
label='value'
print(stepwise_select(house_data, label, cols_all, method='both'))

```

## 岭回归与 Lasso 回归

```

from sklearn.linear_model import Lasso
from sklearn.preprocessing import StandardScaler
import pandas as pd
from sklearn.datasets import fetch_california_housing as fch
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import Lasso
from sklearn.linear_model import Ridge
iris = fch()
data = pd.DataFrame(iris.data)
print(data.corr())#看一下各变量之间的相关性
scaler = StandardScaler()
x = scaler.fit_transform(iris.data)
y = iris.target
lr = LinearRegression()
lr.fit(x,y)
print('lr.coef')
print(lr.coef_)
lasso = Lasso(alpha=0.1)
lasso.fit(x,y)
print('Lasso.coef')
print(lasso.coef_)
ridge = Ridge(alpha=10)
ridge.fit(x,y)
print('Ridge.coef')
print(ridge.coef_)

```

## 群组变量选择方法

首先是弹性网，代码如下

```
import numpy as np
from sklearn.linear_model import ElasticNet
from sklearn.linear_model import SGDRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import fetch_california_housing as fch
import pandas as pd
iris = fch()
data = pd.DataFrame(iris.data)
print(data.corr())#看一下各变量之间的相关性
scaler = StandardScaler()
x = scaler.fit_transform(iris.data)
y = iris.target
elastic_net = ElasticNet(alpha=0.0001, l1_ratio=0.15)
print(elastic_net.fit(x, y).coef_)
```

然后是 GroupLasso

```
from groupLasso import GroupLassoRegressor
import numpy as np
np.random.seed(0)
#create sample dataset:
X = np.random.randn(10, 3)
y = X[:, 0] + np.random.randn(10) * 0.1
#Set group_ids, which specify group membership:
# 0th feature and 1st feature are the same group.
group_ids = np.array([0, 0, 1])
model = GroupLassoRegressor(group_ids=group_ids, random_state=42, verbose=False,
alpha=1e-1)
model.fit(X, y)
print(model.coef_)
```

## 8.6 练习

1、试证明  $\hat{\beta} = (\hat{\beta}_1^T, \mathbf{0}^T)^T \in R^p$  是基于惩罚最小二乘目标函数 (8.4.1) 的局部极小值的充分条件是：

$$n^{-1}\mathbf{X}_1^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) - p'_\lambda(|\hat{\beta}_1|) \operatorname{sgn}(\hat{\beta}_1) = \mathbf{0} \quad (8.6.1)$$

$$\left\| n^{-1}\mathbf{X}_2^T(\mathbf{Y} - \mathbf{X}\hat{\beta}) \right\|_\infty \leq p'_\lambda(0+) \quad (8.6.2)$$

$$\lambda_{\min}(n^{-1}\mathbf{X}_1^T\mathbf{X}_1) \geq \kappa(p_\lambda; \hat{\beta}_1) \quad (8.6.3)$$

其中设计矩阵  $\mathbf{X}_1$  对应的变量与  $\hat{\beta}_1$  对应的变量一致,  $\mathbf{X}_2$  是零系数估计量对应的变量。 $\lambda_{\min}(\mathbf{A})$  代表矩阵  $\mathbf{A}$  的最小预测变量值,  $\|\mathbf{a}\|_\infty = \max_j |a_j|$ 。 $\kappa(p_\lambda; \mathbf{v})$  描述了  $p_\lambda(\cdot)$  在  $\mathbf{v} = (v_1, \dots, v_q)^T$  的局部凹性质, 且定义为

$$\kappa(p_\lambda; \mathbf{v}) = \lim_{\epsilon \rightarrow 0+} \max_{1 \leq j \leq q, t_1 < t_2 \in ([v_j] - \epsilon, [v_j] + \epsilon)} -\frac{p'_\lambda(t_2) - p'_\lambda(t_1)}{t_2 - t_1}$$

根据凹函数性质, 在  $[0, \infty)$ ,  $\kappa(p_\lambda; \mathbf{v}) \geq 0$ 。Lasso 惩罚的  $\kappa(p_\lambda; \mathbf{v}) = 0$ ; 对于 SCAD 惩罚, 当  $|\mathbf{v}|$  取值于  $[\lambda, a\lambda]$ ,  $\kappa(p_\lambda; \mathbf{v}) = (a-1)^{-1}\lambda^{-1}$ , 其余地方  $\kappa(p_\lambda; \mathbf{v}) = 0$ 。

### 2、基于练习 1 的结论, 证明

(1) Lasso 惩罚函数自动满足第三个 KKT 条件 (8.6.3), 且 (8.6.1) - (8.6.2) 转变为

$$\begin{aligned} n^{-1}\mathbf{X}_1^T (\mathbf{Y} - \mathbf{X}_1\hat{\beta}_1) - \lambda \operatorname{sgn}(\hat{\beta}_1) &= \mathbf{0} \\ \left\| (n\lambda)^{-1}\mathbf{X}_2^T (\mathbf{Y} - \mathbf{X}_1\hat{\beta}_1) \right\|_\infty &\leq 1 \end{aligned}$$

$$(2) \left\| n^{-1}\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\hat{\beta}) \right\|_\infty \leq \lambda$$

(3) 当  $\lambda > \left\| n^{-1}\mathbf{X}^T \mathbf{Y} \right\|_\infty$ ,  $\hat{\beta} = 0$ 。因此调谐参数  $\lambda$  必须在区间  $[0, \left\| n^{-1}\mathbf{X}^T \mathbf{Y} \right\|_\infty]$  上取值。

### 3、考虑如下简单的惩罚最小二乘形式:

$$\hat{\theta}(z) = \operatorname{argmin}_\theta \left\{ \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|) \right\}.$$

(1) 如果  $p_\lambda(|\theta|) = \frac{\lambda^2}{2}I(|\theta| \neq 0)$ , 即  $L_0$  惩罚, 证明  $\hat{\theta}_H(z | \lambda) = zI(|z| \geq \lambda)$ 。

(2) 如果  $p_\lambda(\theta) = \frac{1}{2}\lambda^2 - \frac{1}{2}(\lambda - \theta)_+^2$ , 我们得到基于惩罚最小二乘硬阈估计量

$\hat{\theta}_H(z) = zI(|z| > \lambda)$ 。并且解释硬阈惩罚函数在哪些方面优于  $L_0$  惩罚。

(3) 若  $p_\lambda(\theta)$  是 Lasso 函数, 我们得到软阈估计量:

$$\hat{\theta}_{\text{soft}}(z) = \operatorname{sgn}(z)(|z| - \lambda)_+$$

(4) 若  $p_\lambda(\theta)$  是 SCAD 函数, 估计量是:

$$\hat{\theta}_{\text{SCAD}}(z) = \begin{cases} \operatorname{sgn}(z)(|z| - \lambda)_+, & \text{when } |z| \leq 2\lambda \\ \operatorname{sgn}(z)[(\alpha-1)|z| - \alpha\lambda]/(\alpha-2), & \text{when } 2\lambda < |z| \leq \alpha\lambda \\ z, & \text{when } |z| \geq \alpha\lambda \end{cases}$$

并且当  $\alpha = \infty$  时, SCAD 估计量变成软阈估计量。

(5) 若  $p_\lambda(\theta)$  是 MCP 函数 ( $\alpha \geq 1$ ), 估计量是:

$$\hat{\theta}_{\text{MCP}}(z) = \begin{cases} \text{sgn}(z)(|z| - \lambda)_+ / (1 - 1/\alpha), & \text{when } |z| < \alpha\lambda \\ z, & \text{when } |z| \geq \alpha\lambda \end{cases}$$

(6) 计算基于 Elastic Net 惩罚函数  $p_\lambda(\theta) = \lambda \{(1 - \alpha)\theta^2 + \alpha|\theta|\}$  的估计量  $\hat{\theta}(z | \lambda)$ 。

4、根据 Lasso 惩罚最小二乘的第一个 KKT 条件, 证明 Lasso 非零回归系数估计量是有偏的, 并计算偏差。

# 第九章 模型平均

## 9.1 简介

### 9.1.1 模型不确定性

在使用统计方法进行数据分析时，研究者通常根据分析的问题假定统计模型，并假定其为真实模型，即生成实际数据的模型。这种所谓的真实模型一般是根据研究者的经验或研究者使用数据里面的一些先验信息来建立的。但这种假设往往是不正确的，因为这个事先给定的模型只是真实模型的一个近似，并且有可能是错误的，并且在很多实际问题的数据分析研究中真实模型都是未知的，这就是模型不确定性带来的不良影响。从建模角度看，模型不确定性一方面体现在采用的模型形式不确定，例如函数形式、分布假定、模型结构或使用不同的预测变量等，另一方面是由于模型选择导致的不确定性，例如不同的模型选择方法对应不同的模型选择结果或同一模型选择方法的选择结果会不稳定。

### 9.1.2 模型选择与模型平均

考虑到模型不确定性时，传统的一种做法是利用模型选择方法从众多的候选模型中选出最优模型，并将其看作真实模型进行后续的统计推断。前面章节已经详细介绍了模型选择，解决了过度复杂或简单的模型均可能使估计或者预测的方差偏大的问题。研究者们常采用的模型选择方法和准则包括逐步回归、AIC、BIC、 $C_p$ 、交叉验证、Lasso 回归、SCAD 或 MCP 方法等，这些方法通过选择最优模型的过程看似在一定程度上解决了模型不确定性问题。但是模型选择方法导致了人们忽视模型选择过程所带来的不确定性，即模型选择本身就是不确定性的。

模型平均方法 (Model Averaging) 起源于 20 世纪 60 年代，早期最有影响的是 Bates 教授和诺贝尔经济学奖获得者 Granger 所做的工作，他们通过组合两个无偏的预测来说明组合预测的优越性 [114]。他们把来自不同模型的估计或者预测通过一定的权重平均起来，有时也称为模型组合，一般包括组合估计和组合预测。模型平均方法没有追求最优模型，而以特定的权重对所有候补模型的统计结果进行平均，在避免了遗失有用信息的同时也充分考虑了模型不确定性，并且使得估计更加稳健。

因此，模型平均解决了模型选择带来的一系列问题：

(1) . 模型平均使用连续的权重去组合来自不同模型的估计，在表示形式上，模型选择可以看作模型平均的特例，但它的权重只取 0 或者 1，因而模型平均估计一般更加稳健。

(2) . 模型选择通过一系列的模型选择方法选出一个最优模型，这样就可能导致遗失其他模型的信息或者其他变量所特有的影响因素信息。但是，模型平均方法不会把某个选定的模型当作真实模型，因为模型平均不轻易地排除任何模型，给每一个模型赋予一个权重去充分利用每一个模型的信息，因而减少信息的遗失。

(3) . 模型选择也会导致模型不确定性，因为可能选到一个与真实模型相差甚远的模型，统计推断就可能存在很高的风险。模型平均提供了一种保障机制，规避了这种风险，也可以说是避免了把鸡蛋放在同一个篮子里。

总的来说，模型选择是统计推断的基础，模型选择旨在选定一个最优模型，基于该模型进行统计推断。但是，模型选择的不确定性会影响到统计推断，从而使得分析结果会出现问题。模型平均方法从估计和预测角度来看是模型选择的推广，相比传统的模型选择方法来确定唯一的最优模型，模型平均方法通过组合不同的模型进行估计和预测，解决模型选择带来的模型不确定性问题，常常能够减小估计风险，得到更加有效的结论。模型平均本质上是一类集成学习方法。所谓集成学习方法，就是使用多种学习器进行学习，并使用某种规则把各个学习结果进行集成汇总，从而获得比单个学习器更好的学习效果的一种机器学习方法，它可以用于分类问题集成、回归问题集成等。

在模型平均的理论研究过程中，如何确定组合权重是最重要的问题。下面通过组合预测的角度，介绍模型平均的两类方法：贝叶斯模型平均 (BMA: Bayesian Model Averaging) 和频率模型平均 (FMA: Frequentist Model Averaging)。

## 9.2 贝叶斯模型平均

**贝叶斯模型平均**是一种基于贝叶斯理论并且将模型本身的不确定性考虑在内的方法。基本步骤是：设定待组合模型的先验概率和各个模型中参数的先验分布，然后用经典的贝叶斯方法进行统计分析。

设  $Y$  为组合预测随机变量， $D$  为已获得的数据。因为并不知道哪一个模型是最优模型，即模型本身存在着不确定性，我们设定  $\mathcal{M} = \{M_1, M_2, \dots, M_K\}$  代表所有可能模型组成的模型空间，

根据贝叶斯模型平均方法，组合预测随机变量  $Y$  的后验分布为：

$$\Pr(y|D) = \sum_{k=1}^K \Pr(y, M_k|D) \quad (\text{全概率公式})$$

$$= \sum_{k=1}^K \Pr(M_k|D) \Pr(y|M_k, D) \quad (\text{条件概率公式}) \quad (9.2.1)$$

其中  $\Pr(M_k|D)$  为给定数据  $D$  的条件下模型  $M_k$  的后验分布, 反映了研究人员对于真实模型的不确定性。根据贝叶斯公式, 其形式为

$$\begin{aligned} \Pr(M_k|D) &= \frac{\Pr(D|M_k)\Pr(M_k)}{\Pr(D)} \quad (\text{贝叶斯公式}) \\ &= \frac{\Pr(D|M_k)\Pr(M_k)}{\sum_{k=1}^K \Pr(D|M_k)\Pr(M_k)} \quad (\text{全概率公式}) \end{aligned} \quad (9.2.2)$$

其中  $\Pr(M_k)$  为候补模型  $M_k$  的先验概率, 在没有特别先验信息的条件下可取均匀分布, 即  $\Pr(M_k) = 1/K$ ;  $\Pr(D|M_k)$  为模型  $M_k$  的似然函数。在参数模型假设下, 假定  $\boldsymbol{\theta}_k$  为模型  $M_k$  的参数向量, 例如线性回归模型  $\boldsymbol{\theta}_k$  包含了回归系数和误差的方差,  $\Pr(\boldsymbol{\theta}_k|M_k)$  是给定模型  $M_k$  下  $\boldsymbol{\theta}_k$  的先验概率函数,  $\Pr(D|\boldsymbol{\theta}_k, M_k)$  是给定模型  $M_k$  下参数化的似然函数。可以推出

$$\Pr(D|M_k) = \int \Pr(D|\boldsymbol{\theta}_k, M_k)\Pr(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k \quad (9.2.3)$$

从组合预测随机变量的后验分布可以发现, 贝叶斯模型平均方法实际上是以模型的后验分布为权重, 对所有模型的预测后验分布进行加权。使用贝叶斯模型平均的关键在于确定组合的模型以及各单项模型的后验概率, 即权重。另外也可以看出, 贝叶斯模型平均有两个困难: 第一是  $\Pr(D|M_k)$  计算中涉及到积分运算, 如果模型复杂, 积分也会变得困难; 第二是候补模型的个数也需要科学方法去确定。

记  $\mu_k$  和  $\sigma_k^2$  为给定数据  $D$  和候补模型  $M_k$ , 则组合预测随机变量  $Y$  的条件期望和条件方差表示为  $E(Y | D, M_k) = \mu_k$  和  $\text{var}(Y | D, M_k) = \sigma_k^2$ 。令候补模型  $M_k$  的后验概率为  $\Pr(M_k | D) = w_k$ , 则组合预测随机变量  $Y$  条件期望分解形式:

$$\begin{aligned} \mu(\mathbf{w}) &= E(Y | D) = E_M[E(Y | D, M)] \\ &= \sum_{k=1}^K \Pr(M_k | D) E(Y | D, M_k) \\ &= \sum_{k=1}^K w_k \mu_k \end{aligned}$$

$\mathbf{w} = (w_1, \dots, w_K)^T$ 。在参数模型假设下, 我们可以使用贝叶斯估计框架, 使用方程 (9.2.3) 得到似然函数  $\Pr(D|M_k)$  的估计。接下来使用公式 (9.2.2) 和给定的候补模型  $M_k$  的先验概率  $\Pr(M_k)$ , 估计出所有候补模型的权重  $\hat{w}_k$ ’s。另外, 我们根据基于模型  $M_k$  的后验概率分布  $\Pr(y|M_k, D)$  估计出  $\hat{\mu}_k$ , 则 [贝叶斯模型平均](#) 估计的预测值是

$$\hat{\mu}(\hat{\mathbf{w}}) = \sum_{k=1}^K \hat{w}_k \hat{\mu}_k$$

另外，条件方差  $\text{var}(Y | D)$  可以分解为：

$$\begin{aligned}\text{var}(Y | D) &= \text{E}_{\mathcal{M}}[\text{var}(Y | D, \mathcal{M})] + \text{var}_{\mathcal{M}}[\text{E}(Y | D, \mathcal{M})] \\ &= \sum_{k=1}^K w_k \sigma_k^2 + \sum_{k=1}^K w_k (\mu_k - \mu)^2\end{aligned}$$

研究者认为上述条件方差被作为衡量模型不确定性的基础。 $\text{E}_{\mathcal{M}}[\text{var}(Y | D, \mathcal{M})]$  被认为是模型结构内方差， $\text{var}_{\mathcal{M}}[\text{E}(Y | D, \mathcal{M})]$  则是结构间方差，是由模型结构的不确定性引起的，原因是当能够确定真实模型时， $\text{var}_{\mathcal{M}}[\text{E}(Y | D, \mathcal{M})]$  为 0。

### 9.3 频率模型平均

与贝叶斯模型平均相比较，频率模型平均方法不需要考虑如何设置候补模型（Candidate Model）的先验概率，模型估计和权重估计完全由数据确定。频率模型平均过程为：假设有  $K$  个模型（或者  $K$  种估计方法），每个模型（或方法）估计出一个预测值，我们不妨将其写为

$$\hat{u}_k, \quad k = 1, \dots, K,$$

那么频率模型平均方法得到的最终估计的预测值为：

$$\hat{u} = w_1 \hat{u}_1 + w_2 \hat{u}_2 + \dots + w_K \hat{u}_K$$

其中  $\mathbf{w} = (w_1, \dots, w_K)$  为权重向量，通常满足  $0 \leq w_k \leq 1$ ， $\sum_{k=1}^K w_k = 1$ 。

若我们定义权重系数为示性函数，即令

$$w_k = I\{\text{第 } k \text{ 个模型被选到}\}$$

则模型平均变为模型选择，从这里可以看出模型选择是模型平均的一个特例，模型平均是模型选择的推广。和贝叶斯模型平均方法一样，频率模型平均法也需要确定权重值，那么怎么求取候补模型对应的权重呢？下面介绍几种权重选择方法。

### 9.4 权重选择方法

### 9.4.1 基于信息准则

所谓信息准则 (IC) 的权重选择方法即对于每一个候补模型给出一个基于信息准则的得分，最后依据得分多少来描述候补模型的重要性。上节在介绍模型选择时已给出了 AIC 和 BIC 的公式，我们可以先计算每一个候补模型的 AIC 和 BIC 值，然后通过如下公式计算各个组合权重：

$$w_k = \frac{\exp(-\frac{IC_k}{2})}{\sum_{k=1}^K \exp(-\frac{IC_k}{2})} \quad (9.4.1)$$

其中  $K$  表示候补模型集合中模型的个数， $w_k$  表示第  $k$  个候补模型的权重； $IC_k$  为第  $k$  个候补模型的 AIC 或 BIC 值。由上式可看出基于信息准则的组合权重计算方法比较简单，只需计算每个候补模型下的 AIC 或 BIC 值即可，对应的模型平均方法称为 AIC 模型平均和 BIC 模型平均。由于计算方便，基于信息准则的思路是比较常用的权重选择方法。

### 9.4.2 基于 Mallows 准则

在模型选择准则的介绍中介绍了  $C_p$  准则， $C_p$  准则的全称为 Mallows's  $C_p$  准则。我们接下来要介绍的 Mallows 准则是对 Mallows's  $C_p$  准则的推广，基于该准则的权重选择方法是最小化 Mallows 准则来得到组合预测的权重。下面以线性模型为例介绍具体的过程。

记响应变量的观测样本  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ ，估计的预测值记为  $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_n)$ ，协变量样本矩阵记为  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$ ，其中  $\mathbf{X}_i = (1, X_{i1}, \dots, X_{ip})^T$ ， $p$  是变量的个数，误差项  $\boldsymbol{\epsilon} = (e_1, \dots, e_n)^T$ 。协变量与响应变量之间的线性关系表示为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

其中  $\boldsymbol{\beta}$  为协变量对应的系数向量，形式为  $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ 。

假设研究者采用  $K$  个候补模型去获得  $\hat{\boldsymbol{\mu}}$ ，其中第  $k$  个采用的候补模型为

$$M_k : \mathbf{Y} = \mathbf{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}$$

其中  $\mathbf{X}_k = (\mathbf{X}_{k1}, \dots, \mathbf{X}_{kn})^T$  是  $n \times p_k$  的第一列为 1 的矩阵，即它是由  $\mathbf{X}$  除去第一列外的任意  $p_k - 1$  列组成的， $\boldsymbol{\beta}_k$  是其相应的  $p_k$  维的系数向量。一般情况下  $\mathbf{X}_k^T \mathbf{X}_k$  是可逆的，故  $\boldsymbol{\beta}_k$  基于第  $k$  个候补模型的最小二乘估计为  $\hat{\boldsymbol{\beta}}_k = (\mathbf{X}_k^T \mathbf{X}_k)^{-1} \mathbf{X}_k^T \mathbf{Y}$ 。相应地， $\boldsymbol{\mu}_k$  的估计为  $\hat{\boldsymbol{\mu}}_k = \mathbf{X}_k (\mathbf{X}_k^T \mathbf{X}_k)^{-1} \mathbf{X}_k^T \mathbf{Y} = H_k \mathbf{Y}$ 。其中  $H_k = \mathbf{X}_k (\mathbf{X}_k^T \mathbf{X}_k)^{-1} \mathbf{X}_k^T$  是帽子矩阵。记权重向量  $\mathbf{w} = (w_1, \dots, w_K)^T$ ，且满足：

$$H_n = \{\mathbf{w} \in [0, 1]^K : \sum_{k=1}^K w_k = 1\}$$

那么  $\mu$  的模型平均估计为

$$\hat{\mu}(\mathbf{w}) = \sum_{k=1}^K w_k \hat{\mu}_k \quad (9.4.2)$$

接下来，采用以下 Mallows 准则估计权重：

$$C_n(\mathbf{w}) = \mathbf{w}^T \hat{\mathbf{E}}^T \hat{\mathbf{E}} \mathbf{w} + 2\hat{\sigma}^2 \mathbf{w}^T \phi$$

其中  $\hat{\mathbf{E}} = (\hat{e}_1, \dots, \hat{e}_K)$ ,  $\hat{e}_k = \mathbf{Y} - \hat{\mu}_k$  为基于第  $k$  个候补模型估计的残差向量,  $\phi = (p_1, \dots, p_K)^T$ ,  $\hat{\sigma}^2 = \frac{\hat{e}_{\tilde{k}}^T \hat{e}_{\tilde{k}}}{n - \phi_{\tilde{k}}}$ ,  $\tilde{k}$  满足  $\phi_{\tilde{k}} = \max\{p_1, \dots, p_K\}$ 。通过极小化 Mallows 准则所得到的权重为：

$$\hat{\mathbf{w}}_M = \operatorname{argmin}_{\mathbf{w} \in H_n} C_n(\mathbf{w}) \quad (9.4.3)$$

该权重对应的模型平均估计成为 [Mallows 模型平均](#) (MMA: Mallows Model Average) 估计, 将其带入式 (9.4.2) 可得到 MMA 估计的观测值  $\hat{\mu}(\hat{\mathbf{w}})$ 。

### 9.4.3 基于 Jackknife 准则

基于 Jackknife 准则的模型平均方法是通过最小化 Jackknife 准则得到组合预测的权重。该方法适用于随机误差项异方差的情形, 即当  $\operatorname{cov}(e|\mathbf{X}) = \operatorname{diag}(\sigma_1^2, \dots, \sigma_n^2)$  的时候可用基于 Jackknife 准则的模型平均方法去选择权重。

假设研究者依然以线性模型为例, 并且采用  $K$  个候补模型, 其中第  $k$  个采用的候补模型使用的协变量观测矩阵是  $\mathbf{X}_k = (\mathbf{X}_{k1}, \dots, \mathbf{X}_{kn})^T$ 。使用 Jackknife 估计方法对第  $k$  个候补模型估计预测值的具体流程是：对于  $i = 1, \dots, n$ ,

(1) . 删除第  $i$  个样本, 使用最小二乘得到回归参数的估计

$$\hat{\beta}_{(-i)} = (\mathbf{X}_{k(-i)}^T \mathbf{X}_{k(-i)})^{-1} \mathbf{X}_{k(-i)}^T \mathbf{Y}_{(-i)}$$

其中  $\mathbf{X}_{k(-i)}$  和  $\mathbf{Y}_{(-i)}$  分别是去掉第  $i$  个样本之后的  $(n-1) \times p_k$  协变量观测矩阵和  $(n-1)$  维的响应变量观测向量。

(2) . 估计出第  $i$  个样本的预测值为  $\tilde{\mu}_{ki} = \mathbf{X}_{ki}^T \hat{\beta}_{(-i)}$ 。

最后得到基于第  $k$  个候补模型的对应  $n$  个样本的预测向量  $\tilde{\mu}_k = (\tilde{\mu}_{k1}, \dots, \tilde{\mu}_{kn})^T$ 。注意到  $\mathbf{X}_{k(-i)}^T \mathbf{X}_{k(-i)} = \mathbf{X}_k^T \mathbf{X}_k - \mathbf{X}_{ki} \mathbf{X}_{ki}^T$ , 根据 Sherman–Morrison 公式, 我们可以推导出

$$\tilde{\mu}_{ki} = \sum_{j \neq i} \frac{H_{k,ij}}{1 - H_{k,ii}} Y_j$$

其中  $H_{k,ij}$  代表第  $k$  个模型的帽子矩阵第  $i$  行第  $j$  列的元素。上式表明  $\tilde{\mu}_{ki}$  不依赖于  $Y_i$ 。可以用矩阵表示第  $k$  个候补模型的预测向量为

$$\tilde{\mu}_k = (D_k(H_k - I_n) + I_n) \mathbf{Y}$$

其中  $D_k$  是对角阵，它的第  $i$  个对角元素为  $(1 - H_{k,ii})^{-1}$ ，并且  $D_{k,ii}(H_{k,ii} - 1) = \frac{H_{k,ii} - 1}{1 - H_{k,ii}} = -1$ ， $I_n$  是单位阵。

令  $\tilde{\mathbf{E}} = (\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2, \dots, \tilde{\mathbf{e}}_K)$ ， $\tilde{\mathbf{e}}_k = \mathbf{Y} - \tilde{\boldsymbol{\mu}}_k$  为基于第  $k$  个模型估计的残差向量，而且  $\tilde{\mathbf{e}}_k$  是弃一的交叉验证的误差向量，在统计学中弃一的交叉验证又称为 Jackknife。Jackknife 准则定义为：

$$J_n(\mathbf{w}) = \mathbf{w}^T \tilde{\mathbf{E}}^T \tilde{\mathbf{E}} \mathbf{w} \quad (9.4.4)$$

并且通过极小化 Jackknife 准则所得到的权重为：

$$\hat{\mathbf{w}}_J = \arg \min_{\mathbf{w} \in H_n} J_n(\mathbf{w})$$

并且基于 Jackknife 准则得出的组合模型权重对应的模型平均方法为 [Jackknife 模型平均](#) (JMA: Jackknife Model Average)。

## 9.5 Python 语言实现

下面介绍怎么用 Python 语言实现上面提到的模型平均方法。

### 贝叶斯模型平均-线性模型

```
#载入需要用到的包
import numpy as np
import pandas as pd
from statsmodels.regression.linear_model import OLS
from statsmodels.tools import add_constant
import statsmodels.api as sm
from itertools import combinations
#读取数据
df = pd.read_csv('D:/Guber1999data.csv') #此处为数据存放的路径
df.head(5) #查看数据的前5行
```

查看输出的“Guber1999data”数据的前 5 行：

	State	Spend	StuTeaRat	Salary	PrcntTake	SATV	SATM	SATT
0	Alabama	4.405	17.2	31.144	8	491	538	1029
1	Alaska	8.963	17.6	47.951	47	445	489	934
2	Arizona	4.778	19.3	32.175	27	448	496	944
3	Arkansas	4.459	17.1	28.934	6	482	523	1005
4	California	4.992	24.0	41.078	45	417	485	902

接下来建立贝叶斯模型平均线性回归模型。首先定义 BMA 类

```
#贝叶斯模型平均
class BMA:
    def __init__(self, y, X, **kwargs):
        # 设置基本的变量
        self.y = y
        self.X = X
        self.names = list(X.columns)
        self.nRows, self.nCols = np.shape(X)
        self.likelihoods = np.zeros(self.nCols)
        self.coefficients = np.zeros(self.nCols)
        self.probabilities = np.zeros(self.nCols)
        self.names = list(X.columns)
        #检查模型的最大尺寸 (模型中使用的预测变量的最大数量)
        if 'MaxVars' in kwargs.keys():
            self.MaxVars = kwargs['MaxVars']
        else:
            self.MaxVars = self.nCols
        #如果提供了先验参数, 则准备好这些参数
        if 'Priors' in kwargs.keys():
            if np.size(kwargs['Priors']) == self.nCols:
                self.Priors = kwargs['Priors']
            else:
                print("WARNING: Provided priors error. Using equal priors instead.")
                print("The priors should be a numpy array of length equal to the
                      number of regressor variables.")
                self.Priors = np.ones(self.nCols)
        else:
            self.Priors = np.ones(self.nCols)

    def fit(self):
        #将所有模型的可能性之和初始化为0
        likelihood_sum = 0
        #为了方便在所有可能的模型中进行迭代, 我们从遍历模型中元素的数量开始
        for num_elements in range(1,self.MaxVars+1):
            #列出所有这个尺寸的模型的索引集
            model_index_sets = list(combinations(list(range(self.nCols)),
                                                num_elements))
            #遍历给定大小的所有可能的模型
            for model_index_set in model_index_sets:
                #计算线性回归模型
                model_X = self.X.iloc[:,list(model_index_set)]
                model_regr = OLS(self.y, model_X).fit()
                #计算模型的似然 (乘以先验)
                model_likelihood = np.exp(-model_regr.bic/2)*np.prod(
                    self.Priors[list(model_index_set)])
                print("Model Variables:",model_index_set,
                      "likelihood=", model_likelihood)
                likelihood_sum = likelihood_sum + model_likelihood
            #将此可能性(乘以先验)加到模型中每个变量的似然值
            for idx, i in zip(model_index_set, range(num_elements)):
                self.likelihoods[idx] = self.likelihoods[idx] + model_likelihood
```

```

        self.coefficients[idx] = self.coefficients[idx] +
        model_regr.params[i]*model_likelihood

    #用贝叶斯定理除以分母来标准化概率
    self.probabilities = self.likelihoods / likelihood_sum
    self.coefficients = self.coefficients / likelihood_sum

    #返回新的BMA对象
    return self

#将BMA的结果作为一个数据框返回，以便于查看。
def summary(self):
    df = pd.DataFrame([self.names, list(self.probabilities),
    list(self.coefficients)], ["Variable Name", "Probability",
    "Avg. Coefficient"]).T
    return df

```

将数据代入，运用定义的 BMA 类进行分析

```

X = df[['Spend', 'StuTeaRat', 'Salary', 'PrcntTake']]
y = df['SATT']
result = BMA(y, add_constant(X)).fit()
result.summary()

```

查看输出结果：

```

Model Variables: (0,) likelihood= 7.193780355030914e-126
Model Variables: (1,) likelihood= 1.8226074496883656e-152
...
Model Variables: (1, 2, 3, 4) likelihood= 3.4556604391428603e-134
Model Variables: (0, 1, 2, 3, 4) likelihood= 2.26958690901297e-110

      Variable Name  Probability  Avg. Coefficient
0            const       1.0      1015.358718
1          Spend      0.701118       8.388865
2      StuTeaRat      0.296262      -1.032458
3        Salary      0.300254       0.513476
4    PrcntTake       1.0      -2.822242

```

首先，我们观察到 PrcntTake 变量的概率是 1。这意味着不包括这个变量的模型的可能性很小。第二，关于支出和 SAT 成绩之间的关系的问题，我们看到，在其他条件相同的情况下，支出的增加对应于 SAT 成绩的增加。

## 贝叶斯模型平均-logistic 模型

```

import numpy as np
import pandas as pd
import statsmodels.api as sm
from statsmodels.tools import add_constant
from itertools import combinations
from sklearn.model_selection import train_test_split#加载需要用到的包
df = pd.read_csv('D:/CHDdata.csv')#读取数据
df['famhist'] = (df['famhist'] == 'Present')*1#将famhit转换为0(没有hist)和1(有hist)
df.head() #查看数据的前5行

```

查看“CHDdata”数据前5行的输出结果：

	sbp	tobacco	ldl	adiposity	famhist	typea	obesity	alcohol	age	chd
0	160	12.00	5.73	23.11	1	49	25.30	97.20	52	1
1	144	0.01	4.41	28.61	0	55	28.87	2.06	63	1
2	118	0.08	3.48	32.28	1	52	29.14	3.81	46	0
3	170	7.50	6.41	38.03	1	51	31.99	24.26	58	1
4	134	13.60	3.50	27.78	1	60	25.99	57.34	49	1

接下来先建立 logistic 回归模型，并拟合数据。

```
# 建立logistic回归并拟合数据
X = df.drop(["chd"], axis=1)
y = df["chd"]
log_reg = sm.Logit(y, add_constant(X)).fit()
log_reg.summary()
```

再定义 BMA 类，建立贝叶斯模型平均 logistic 回归模型。

```
from mpmath import mp
mp.dps = 50
class BMA:
    def __init__(self, y, X, **kwargs):
        # 设置基本的变量
        self.y = y
        self.X = X
        self.names = list(X.columns)
        self.nRows, self.nCols = np.shape(X)
        self.likelihoods = mp.zeros(self.nCols, 1)
        self.likelihoods_all = {}
        self.coefficients_mp = mp.zeros(self.nCols, 1)
        self.coefficients = np.zeros(self.nCols)
        self.probabilities = np.zeros(self.nCols)
        # 检查模型的最大尺寸（模型中使用的预测变量的最大数量），这可以减少运行时间
        if 'MaxVars' in kwargs.keys():
            self.MaxVars = kwargs['MaxVars']
        else:
            self.MaxVars = self.nCols
        if 'Priors' in kwargs.keys():
            if np.size(kwargs['Priors']) == self.nCols:
                self.Priors = kwargs['Priors']
            else:
                print("WARNING: Provided priors error. Using equal priors instead.")
                print("The priors should be a numpy array of length equal to the number of regressor variables.")
                self.Priors = np.ones(self.nCols)
        else:
            self.Priors = np.ones(self.nCols)
        if 'Verbose' in kwargs.keys():
            self.Verbose = kwargs['Verbose']
        else:
            self.Verbose = False
        if 'RegType' in kwargs.keys():
            self.RegType = kwargs['RegType']
        else:
            self.RegType = "BMA"
```

```

        self.RegType = kwargs['RegType']
    else:
        self.RegType = 'LS'
def fit(self):
    # 进行贝叶斯模型平均化
    # 将所有模型的似然之和初始化为零。
    # 这将是贝叶斯定理中的“归一化”分母。
    likelihood_sum = 0
    # 为了方便在所有可能的模型中进行迭代，我们从迭代模型中的元素数量开始
    max_likelihood = 0
    for num_elements in range(1, self.MaxVars+1):
        if self.Verbose == True:
            print("Computing BMA for models of size: ", num_elements)
        # 列出这个尺寸的模型的所有索引集的列表
        Models_next = list(combinations(list(range(self.nCols)), num_elements))
        # Occam's window - 计算下一次迭代要使用的候补模型
        # Models_previous: 前一次迭代中满足以下条件的模型集合
        (likelihood > max_likelihood/20)
        # Models_next: 下一次迭代的候补模型集
        # Models_current: 来自Models_next的一系列的模型集，该模型集可以通过在模型Models_previous上一个模型中添加一个新变量
        if num_elements == 1:
            Models_current = Models_next
            Models_previous = []
        else:
            idx_keep = np.zeros(len(Models_next))
            for M_new, idx in zip(Models_next, range(len(Models_next))):
                for M_good in Models_previous:
                    if(all(x in M_new for x in M_good)):
                        idx_keep[idx] = 1
                        break
                else:
                    pass
            Models_current = np.asarray(Models_next)
            [np.where(idx_keep==1)].tolist()
            Models_previous = []
        # 遍历给定尺寸的所有可能的模型
        for model_index_set in Models_current:
            # 计算此给定模型的线性回归
            model_X = self.X.iloc[:,list(model_index_set)]
            if self.RegType == 'Logit':
                model_regr = sm.Logit(self.y, model_X).fit(disp=0)
            else:
                model_regr = OLS(self.y, model_X).fit()
            # 计算模型的似然值（乘以先验）
            model_likelihood = mp.exp(-model_regr.bic / 2) * np.prod(
                self.Priors[list(model_index_set)])
            if (model_likelihood > max_likelihood/20):
                if self.Verbose == True:
                    print("Model Variables:", model_index_set, "likelihood=", model_likelihood)
                self.likelihoods_all[str(model_index_set)] = model_likelihood

```

```

likelihood_sum = mp.fadd(likelihood_sum, model_likelihood)
for idx, i in zip(model_index_set, range(num_elements)):
    self.likelihoods[idx] = mp.fadd(
        self.likelihoods[idx],
        model_likelihood, prec=1000)
    self.coefficients_mp[idx] = mp.fadd(self.coefficients_mp[idx]
                                         ,model_regr.params[i]*model_likelihood, prec=1000)
    # add this model to the list of good models
    Models_previous.append(model_index_set)
    # get the new max likelihood if it is this model
    max_likelihood = np.max([max_likelihood,model_likelihood])
else:
    if self.Verbose == True:
        print("Model Variables:",model_index_set,
              "rejected by Occam's window")
# 除以贝叶斯定理中的分母，使概率之和为1
self.likelihood_sum = likelihood_sum
for idx in range(self.nCols):
    self.probabilities[idx] = mp.fdiv(self.likelihoods[idx],likelihood_sum,
                                       prec=1000)
    self.coefficients[idx] = mp.fdiv(self.coefficients_mp[idx],
                                     likelihood_sum, prec=1000)
# 返回新的BMA并输出
return self
def predict(self, data):
    data = np.asarray(data)
    if self.RegType == 'Logit':
        try:
            result = 1 / (1 + np.exp(-1 * np.dot(self.coefficients, data)))
        except:
            result = 1 / (1 + np.exp(-1 * np.dot(self.coefficients, data.T)))
    else:
        try:
            result = np.dot(self.coefficients,data)
        except:
            result = np.dot(self.coefficients,data.T)
    return result
def summary(self):
    # 将BMA结果作为一个数据框架返回，便于查看
    df = pd.DataFrame([self.names, list(self.probabilities),
                       list(self.coefficients)], ["Variable Name", "Probability",
                       "Avg. Coefficient"]).T
    return df

```

利用定义好的 BMA 类，代入数据进行模型拟合

```

result = BMA(y,add_constant(X), RegType = 'Logit', Verbose=True).fit()
result.summary() #查看结果的概要

```

## 基于 AIC 与 BIC 信息准则

我们基于 AIC 和 BIC 信息准则对“Credit”数据集进行建模。

```
#载入需要用到的包
import itertools
import time
import numpy as np
import pandas as pd
import seaborn as sns
import statsmodels.api as sm
from statsmodels.regression.linear_model import OLS
from statsmodels.tools import add_constant
from itertools import combinations
import matplotlib.pyplot as plt
from sklearn import linear_model
from sklearn.metrics import mean_squared_error
%matplotlib inline
plt.style.use('ggplot')
#读取数据
credit = pd.read_csv('D:/Credit.csv', usecols=list(range(1,12)))
credit.head(3) #查看数据的前3行
```

查看“Credit”数据前 3 行的输出结果：

	Income	Limit	Rating	Cards	...	Student	Married	Ethnicity	Balance
0	14.891	3606	283	2	...	No	Yes	Caucasian	333
1	106.025	6645	483	3	...	Yes	Yes	Asian	903
2	104.593	7075	514	4	...	No	No	Asian	580

[3 rows x 11 columns]

```
#将定性数据进行编码
credit = pd.get_dummies(credit, columns=['Gender', 'Student','Married','Ethnicity'],
drop_first = True)
credit.head(3)
```

我们对定性数据进行编码，将女性编码为 1，男性编码为 0；将学生编码为 1，非学生编码为 0；将已婚编码为 1，未婚编码为 0；将亚洲人编码为 1，非亚洲人编码为 0。查看“Credit”数据前 3 行的输出结果：

	Income	Limit	Rating	...	Married_Yes	Ethnicity_Asian	Ethnicity_Caucasian
0	14.891	3606	283	...	1	0	1
1	106.025	6645	483	...	1	1	0
2	104.593	7075	514	...	0	1	0

[3 rows x 12 columns]

建立模型，并进行接下来的分析。

```
#拟合线性回归，返回RSS和R方
def fit_linear_reg(X,Y):
    model_k = linear_model.LinearRegression(fit_intercept = True)
```

```

model_k.fit(X,Y)
RSS = mean_squared_error(Y,model_k.predict(X)) * len(Y)
R_squared = model_k.score(X,Y)
return RSS, R_squared

##选择subset
from tqdm import tnrange, tqdm_notebook
#变量初始化
Y = credit.Balance
X = credit.drop(columns = 'Balance', axis = 1)
k = 11
RSS_list, R_squared_list, feature_list = [],[],[]
numb_features = []
#在X中循环k = 1到k = 11个特征
for k in tnrange(1,len(X.columns) + 1, desc = 'Loop...'):
    ##循环所有可能的组合:从11个中选择k个
    for combo in itertools.combinations(X.columns,k):
        tmp_result = fit_linear_reg(X[list(combo)], Y)
        RSS_list.append(tmp_result[0])
        R_squared_list.append(tmp_result[1])
        feature_list.append(combo)
        numb_features.append(len(combo))
#存储在数据框中
df = pd.DataFrame({'numb_features': numb_features,'RSS': RSS_list, 'R_squared': R_squared_list, 'features':feature_list})

```

接下来为不同数量的特征找到最佳子集。

```

#为不同数量的特征找到最佳子集
df_min = df[df.groupby('numb_features')['RSS'].transform(min) == df['RSS']]
df_max = df[df.groupby('numb_features')['R_squared'].transform(max) == df['R_squared']]
display(df_min.head(3))
display(df_max.head(3))

```

进一步，对上面输出的数据框的格式进行调整。

```

#调整上面输出的数据框
df['min_RSS'] = df.groupby('numb_features')['RSS'].transform(min)
df['max_R_squared'] = df.groupby('numb_features')['R_squared'].transform(max)
df.head()

```

查看将数据框调整后的输出结果：

	numb_features	RSS	...	min_RSS	max_R_squared
0	1	6.620874e+07	...	2.143512e+07	0.745848
1	1	2.171566e+07	...	2.143512e+07	0.745848
2	1	2.143512e+07	...	2.143512e+07	0.745848
3	1	8.370950e+07	...	2.143512e+07	0.745848
4	1	8.433963e+07	...	2.143512e+07	0.745848

[5 rows x 6 columns]

再分别计算各模型的 RSS

```

#初始化变量
Y = credit.Balance
X = credit.drop(columns = 'Balance', axis = 1)
k = 11
remaining_features = list(X.columns.values)
features = []
RSS_list, R_squared_list = [np.inf], [np.inf]
features_list = dict()
for i in range(1,k+1):
    best_RSS = np.inf
    for combo in itertools.combinations(remaining_features,1):
        #存储临时结果
        RSS = fit_linear_reg(X[list(combo) + features],Y)
        if RSS[0] < best_RSS:
            best_RSS = RSS[0]
            best_R_squared = RSS[1]
            best_feature = combo[0]
    #为下个循环更新变量
    features.append(best_feature)
    remaining_features.remove(best_feature)
    #将值储存起来
    RSS_list.append(best_RSS)
    R_squared_list.append(best_R_squared)
    features_list[i] = features.copy()
##结果展示
print('Number of features |', 'Features |', 'RSS')
display([(i,features_list[i], round(RSS_list[i])) for i in range(1,5)])

```

接下来为最优模型的 AIC、BIC 以及调整后  $R^2$  计算过程。

```

#定义用到的新数据框
df1 = pd.concat([pd.DataFrame({'features':features_list}),pd.DataFrame(
{'RSS':RSS_list, 'R_squared': R_squared_list})], axis=1, join='inner')
df1['numb_features'] = df1.index
#初始化用到的变量
m = len(Y)
p = 11
hat_sigma_squared = (1 / (m - p - 1)) * min(df1['RSS'])
#计算 AIC 和 BIC
df1['AIC'] = (1/(m*hat_sigma_squared)) * (df1['RSS'] + 2 * df1['numb_features'] *
hat_sigma_squared )
df1['BIC'] = (1/(m*hat_sigma_squared)) * (df1['RSS'] + np.log(m) *
df1['numb_features'] * hat_sigma_squared )
df1['R_squared_adj'] = 1 - ((1 - df1['R_squared']) * (m - 1) /
(m - df1['numb_features'] - 1))
df1 #输出数据框df1

```

## 基于 Mallows 与 Jackknife 准则

基于上面的代码以及“Credit”数据，进行这一部分的建模分析。

```
#计算
df1['C_p'] = (1/m) * (df1['RSS'] + 2 * df1['numb_features'] * hat_sigma_squared)
df1
```

首先进行数据的预处理

```
#设置X,Y变量，添加常数项
Y = credit.Balance
X = credit[['Ethnicity_Asian','Ethnicity_Caucasian']]
X = sm.add_constant(X)
X.head()
```

再根据上面提到的准则建模，代码如下。

```
#拟合模型并输出结果
model_1 = sm.OLS(Y, X)
result_1 = model_1.fit()
result_1.summary()
```

## 9.6 练习

- 假设观察到随机样本  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ ,  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$  是  $n \times p$  的设计矩阵, 且满足线性模型

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

其中  $\varepsilon \sim N(\mathbf{0}, \sigma^2 I_n)$ 。对于参数  $\beta$ ,  $\sigma^2$ , 我们考虑标准正态伽马共轭先验, 即

$$\beta \sim N(\mu, \sigma^2 V),$$

$$\frac{\nu\lambda}{\sigma^2} \sim \chi_\nu^2,$$

其中  $\nu$ ,  $\lambda$ ,  $\mu$ ,  $V$  是超参数且已知。若采用的候补模型是全模型, 试证明  $p(D|M) = f(\mathbf{Y}|\mathbf{X})$  服从非中心的  $t$  分布, 自由度是  $\nu$ , 期望是  $\mathbf{X}\mu$ , 方差是  $[\nu/(\nu-2)]\lambda(I_n + \mathbf{X}V\mathbf{X}^T)$ 。

- 请推导 Jackknife 准则中的结论:  $\tilde{\mu}_k = (D_k(H_k - I_n) + I_n)\mathbf{Y}$ ,  $D_k$  是对角阵, 它的第  $i$  个对角元素为  $(1 - H_{k,ii})^{-1}$ , 而  $H_{k,ii}$  是  $H_k$  的第  $i$  个对角元素。

(提示: 使用 Sherman–Morrison 公式:  $(A + \mathbf{U}\mathbf{V}^T)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{U}\mathbf{V}^TA^{-1}}{1 + \mathbf{V}^TA^{-1}\mathbf{U}}$ , 其中  $A$  是  $r \times r$  可逆方阵,  $\mathbf{U}, \mathbf{V}$  分别是  $r \times 1$  列向量,  $1 + \mathbf{V}^TA^{-1}\mathbf{U} \neq 0$ )

- 本数据来自 1991 年的收入和项目参与的调查 (SIPP), 由 9915 个观测样本和 44 个变量组成。我们对个人总财富 (total wealth) 变量感兴趣, 因此选取为响应变量。基于 R 或 Python 语言,

(1). 建立线性模型, 并使用最小二乘方法估计参数, 并计算 MSE:  $\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ 。

- (2) . 根据 (1) 中回归系数估计量的绝对值的大小对变量的重要性进行排序，并根据变量排序，构造 4 个候补模型，每个候补模型 11 个变量。分别使用基于 AIC、BIC、Mallows 准则和 Jackknife 准则的模型平均方法获得预测值  $\hat{Y}_i$ ，并计算 MSE。
- (3) . 对比 (1) 和 (2) 预测结果并解释。

# 第十章 特征筛选

## 10.1 简介

在前面章节中，我们介绍了模型选择以及变量选择方法。尽管变量选择方法可以用来识别重要变量，但是在维度  $p$  非常高的情况下，即维度  $p$  随着样本数量  $n$  的增加呈现指数速率增长， $\log(p) = O(n^\delta)$ ,  $0 < \delta < 1$ ，例如基因数据，此时用于优化的算法仍然非常昂贵。在实践中，我们可以自然地考虑一个两阶段的方法：先特征筛选，然后变量选择。首先强调一下，在机器学习领域普遍使用特征表示统计学里的变量，两个概念本质上是等价的。具体地说，我们使用特征筛选方法将超高维  $p$  降低至中等尺度  $d \leq n$ ，一般情况下，维度  $d$  随着样本数量  $n$  的增加呈现幂增长， $d = O(n^\zeta)$ ,  $0 < \zeta < 1$ 。然后再用变量选择方法从剩下的变量中选择真实模型。如果在每一个降维阶段都保留了所有重要变量，那么这个两阶段方法要经济得多。接下来我们介绍诸多特征筛选方法，其目标是尽可能多地丢弃噪声特征，同时保留所有的重要特征。

在本章中，我们采用以下符号。设  $Y$  为响应变量， $\mathbf{X} = (X_1, \dots, X_p)^T$  由  $p$  维预测变量组成，由此得到  $n$  个独立的随机样本  $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ 。 $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ ， $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^T$  是  $n \times p$  的设计矩阵。设  $\mathbf{X}_j$  为  $\mathbf{X}$  的第  $j$  列，则  $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ 。我们稍微滥用了符号  $\mathbf{X}$  和  $\mathbf{X}_j$ ，但是在上下文中它们的含义是清楚的。令  $\varepsilon$  是一个一般随机误差， $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ 。设  $\mathcal{M}_*$  代表一个尺寸为  $s = |\mathcal{M}_*|$  的真实模型， $\widehat{\mathcal{M}}$  为尺寸为  $d = |\widehat{\mathcal{M}}|$  的选择模型。对于不同的模型和背景， $\mathcal{M}_*$  和  $\widehat{\mathcal{M}}$  的定义可能有所不同。

## 10.2 基于边际模型的特征筛选

### 10.2.1 边际最小二乘

对于线性回归模型，其矩阵形式为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (10.2.1)$$

当  $p \gg n$  时,  $\mathbf{X}^T \mathbf{X}$  是奇异的, 因此  $\boldsymbol{\beta}$  的最小二乘估计没有很好的定义。在这种情况下, 岭回归特别有用。模型 (10.2.1) 的岭回归估计量由下式给出

$$\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^T \mathbf{Y}$$

其中  $\lambda$  是一个岭参数。回归分析章节中介绍过岭回归解决了多重共线性的问题, 从正则项看, 岭回归估计量是线性模型的带  $L_2$  惩罚的惩罚最小二乘的解。当  $\lambda \rightarrow 0$  且  $\mathbf{X}$  为满秩时,  $\hat{\boldsymbol{\beta}}_\lambda$  趋于最小二乘估计量, 而当  $\lambda \rightarrow \infty$  时,  $\lambda \hat{\boldsymbol{\beta}}_\lambda$  趋于  $\mathbf{X}^T \mathbf{Y}$ 。这意味着当  $\lambda \rightarrow \infty$  时,  $\hat{\boldsymbol{\beta}}_\lambda \propto \mathbf{X}^T \mathbf{Y}$ 。

假设所有的协变量和响应变量标准化, 它们的样本均值和方差分别为 0 和 1。因此  $\frac{1}{n} \mathbf{X}^T \mathbf{Y}$  成为由响应变量和所有协变量之间的 Pearson 相关系数的样本形式组成的向量。这促使人们使用 Pearson 相关系数作为特征筛选的边际统计量。具体来说, 首先我们标准化  $\mathbf{X}_j$  和  $\mathbf{Y}$ , 然后计算

$$\omega_j = \frac{1}{n} \mathbf{X}_j^T \mathbf{Y}, \quad j = 1, 2, \dots, p \quad (10.2.2)$$

即第  $j$  个预测变量和响应变量之间的样本相关系数。

直观来说,  $X_j$  和  $Y$  之间的相关性越高, 说明  $X_j$  越重要。Fan 和 Lv [115] 提出根据  $|\omega_j|$  对预测变量  $X_j$  的重要性进行排序 (SIS, Sure Independence Screening), 并开发了一种基于 Pearson 相关系数的特征筛选过程, 也成为确定性筛选, 如下所示。对于预先指定的比例  $\gamma \in (0, 1)$ , 选择排名靠前的  $\lceil \gamma n \rceil$  个预测变量来获得子模型

$$\widehat{\mathcal{M}}_\gamma = \{1 \leq j \leq p : |\omega_j| \text{ 是排序在前 } \lceil \gamma n \rceil \text{ 中大的}\}$$

其中  $\lceil \gamma n \rceil$  表示  $\gamma n$  的整数部分。它将超高维度降至一个相对适中的尺度  $\lceil \gamma n \rceil$ , 即  $\widehat{\mathcal{M}}_\gamma$  的大小, 然后对子模型  $\widehat{\mathcal{M}}_\gamma$  再采用变量选择方法。我们需要设置  $\gamma$  值来进行筛选, 一般地,  $\lceil \gamma n \rceil$  的值可以取为  $\lceil n / \log(n) \rceil$ 。

### 10.2.2 边际极大似然

假设  $\{\mathbf{X}_i, Y_i\}_{i=1}^n$  是前面介绍的广义线性模型的随机样本。基于第  $i$  个样本  $\{Y_i, \mathbf{X}_i\}$ , 用下式

$$\ell(Y_i, \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}) = Y_i (\beta_0 + \mathbf{X}_i^T \boldsymbol{\beta}) - b(\beta_0 + \mathbf{X}_i^T \boldsymbol{\beta})$$

表示使用正则连接函数  $b$  的似然函数的负对数 (不失一般性, 离散参数取  $\phi = 1$ )。

注意, 当  $p \gg n$  时, 负对数似然  $\sum_{i=1}^n \ell(Y_i, \beta_0 + \mathbf{X}_i^T \boldsymbol{\beta})$  的最小值无法给出很好地定义。

接下来我们考虑边际极大似然方法筛选出重要预测变量。与线性回归模型的边际最小二乘估计相似，假设每个预测变量进行了标准化，均值为 0，标准差为 1，并定义第  $j$  个预测变量  $X_j$  的边际极大似然估计量（MMLE: Marginal Maximum Likelihood Estimator） $\hat{\beta}_j^M$  为

$$\hat{\beta}_j^M = \left( \hat{\beta}_{j0}^M, \hat{\beta}_{j1}^M \right) = \arg \min_{\beta_{j0}, \beta_{j1}} \sum_{i=1}^n \ell(Y_i, \beta_{j0} + \beta_{j1} X_{ij})$$

可以将  $\hat{\beta}_{j1}^M$  的大小作为边际筛选统计量，对  $X_j$  的重要性排序，并通过给定的阈值  $\kappa_n$  选择子模型，即

$$\widehat{\mathcal{M}}_{\kappa_n} = \left\{ 1 \leq j \leq p : \left| \hat{\beta}_{j1}^M \right| \geq \kappa_n \right\}$$

阈值  $\kappa_n$  的作用与  $[\gamma n]$  中的  $\gamma$  是一样的，都是选择一个合适的值确定子模型。

### 10.2.3 边际非参估计

假设我们有一个随机样本集  $\{(X_i, Y_i)\}_{i=1}^n$ ，来自可加模型：

$$Y = \sum_{j=1}^p f_j(X_j) + \varepsilon \quad (10.2.3)$$

其中  $\mathbf{X} = (X_1, \dots, X_p)^T$ ,  $\varepsilon$  是条件均值为 0 的随机误差。为了快速识别 (10.2.3) 式中的重要变量，我们考虑以下  $p$  个边际非参数回归问题：

$$\min_{f_j} E [Y - f_j(X_j)]^2 \quad (10.2.4)$$

(10.2.4) 式中的最小值是  $f_j(X_j) = E(Y | X_j)$ ，即  $Y$  在  $X_j$  上的投影。我们根据  $E f_j^2(X_j)$  对 (10.2.3) 式中协变量的统计量进行排序，并通过阈值选择一小组协变量。

为了获得边际非参数回归的样本形式，我们使用 B 样条基。令  $\mathcal{S}_n$  为维度  $l \geq 1$  的多项式样条的空间， $\{\Psi_{jk}, k = 1, \dots, d\}$  表示具有  $\|\Psi_{jk}\|_\infty \leq 1$  的 B 样条基，其中  $\|\cdot\|_\infty$  是 sup 范数。在某些平滑条件下，非参数投影  $\{f_j\}_{j=1}^p$  可以被  $\mathcal{S}_n$  中的函数  $f_{nj}$  很好地近似计算，即  $f_j \approx f_{nj}$ 。并且对于任意的  $f_{nj} \in \mathcal{S}_n$ ，存在某些系数  $\{\beta_{jk}\}_{k=1}^{q_n}$ ，我们有

$$f_{nj}(x) = \sum_{k=1}^{q_n} \beta_{jk} \Psi_{jk}(x), \quad 1 \leq j \leq p$$

此时，边际回归问题的样本形式可以表示为

$$\min_{f_{nj} \in \mathcal{S}_n} E (Y - f_{nj}(X_j))^2 = \min_{\beta_j \in \mathcal{R}^{q_n}} E (Y - \boldsymbol{\Psi}_j^T \boldsymbol{\beta}_j)^2 \quad (10.2.5)$$

其中  $\Psi_j \equiv \Psi_j(X_j) = (\Psi_1(X_j), \dots, \Psi_{q_n}(X_j))^T$  表示  $q_n$  维基函数。基于上述最小化的目标函数，我们首先得到回归系数  $\beta_j$  的总体形式

$$\beta_j = (E\Psi_j\Psi_j^T)^{-1} E(\Psi_j Y)$$

其中  $E$  表示期望。进而可以  $f_{nj}(X_j)$  的最小二乘解的总体形式表示出来，如下所示：

$$f_{nj}(X_j) = \Psi_j^T (E\Psi_j\Psi_j^T)^{-1} E(\Psi_j Y), \quad j = 1, \dots, p$$

基于观测的随机样本集  $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ ，利用矩估计方法，即  $E\Psi_j\Psi_j^T$  和  $E(\Psi_j Y)$  用样本矩代替，我们就可以得到  $\beta_j$  的估计了。

$$\widehat{\beta}_j = \left( \frac{1}{n} \sum_{i=1}^n \Psi_j(X_{ij}) \Psi_j^T(X_{ij}) \right)^{-1} \frac{1}{n} \sum_{i=1}^n \Psi_j(X_{ij}) Y_i$$

接下来，计算目标  $E f_j^2(X_j)$  的统计量的表示形式。首先， $E f_j^2(X_j)$  用  $E f_{nj}^2(X_j)$  去代替， $E f_{nj}^2(X_j)$  用  $E(\Psi_j^T(X_j) \widehat{\beta}_j)^2$  去近似。最后  $E\Psi_j^T(X_j) \widehat{\beta}_j$  所对应的估计量为  $\|\widehat{f}_{nj}\|_n^2$ ，具有如下形式：

$$\|\widehat{f}_{nj}\|_n^2 = \frac{1}{n} \sum_{i=1}^n \widehat{f}_{nj}(X_{ij})^2 = \frac{1}{n} \sum_{i=1}^n \left\{ \Psi_j^T(X_{ij}) \widehat{\beta}_j \right\}^2$$

接下来，引入阈值  $\kappa_n$ ，根据边际非参数估计量  $\|\widehat{f}_{nj}\|_n^2$  对重要变量进行排序 (NIS, Nonparametric Independence Screening)。

$$\widehat{\mathcal{M}}_{\kappa_n} = \left\{ 1 \leq j \leq p : \left\| \widehat{f}_{nj} \right\|_n^2 \geq \kappa_n \right\}$$

## 10.3 基于边际相关系数的特征筛选

### 10.3.1 广义和秩相关系数

上一节中的 SIS 方法对于具有超高维预测变量的线性回归模型表现良好。众所周知，Pearson 相关系数是用来衡量线性相关性的。然而对于超高维数据要确定一个回归结构非常困难。如果线性模型指定错误，SIS 会失败，因为 Pearson 相关系数只能捕获每个预测变量和响应变量之间的线性关系。因此，SIS 最有可能错过一些非线性的重要预测变量，NIS 方法也需要可加性的条件。如果存在更加复杂的非线性关系，例如协变量外面是更加复杂且未知的函数变换等，NIS 的特征筛选效果也会不

好。因此，这促使我们需要将转换后的协变量和响应变量之间的 Pearson 相关系数计算出来，视为边际筛选统计量进行特征筛选。

为了捕获这种类型的非线性，Hall 和 Miller [116] 定义第  $j$  个预测变量  $X_j$  和  $Y$  的广义相关系数为：

$$\rho_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\text{cov}\{h(X_j), Y\}}{\sqrt{\text{var}\{h(X_j)\} \text{var}(Y)}}, \quad j = 1, \dots, p$$

其中  $\mathcal{H}$  是包含所有线性函数的一类函数，例如，它是一类给定次数的多项式函数。请注意，如果  $\mathcal{H}$  是所有线性函数的类，则  $\rho_g(X_j, Y)$  是  $h(X_j)$  和  $Y$  之间 Pearson 相关系数。因此， $\rho_g(X, Y)$  被认为是传统 Pearson 相关系数的推广。假设  $\{(X_{ij}, Y_i), i = 1, 2, \dots, n\}$  是总体  $(X_j, Y)$  的随机样本。广义相关系数  $\rho_g(X_j, Y)$  可以通过下式来估计。

$$\hat{\rho}_g(X_j, Y) = \sup_{h \in \mathcal{H}} \frac{\sum_{i=1}^n \{h(X_{ij}) - \bar{h}_j\} (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n \{h(X_{ij}) - \bar{h}_j\}^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

其中， $\bar{h}_j = n^{-1} \sum_{i=1}^n h(X_{ij})$ ， $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ 。如果已知函数类  $\mathcal{H}$ ，我们可以使用  $\hat{\rho}_g(X_j, Y)$  筛选重要特征。

另外，我们也可以对响应变量进行函数转换，并定义转换后的响应变量和协变量之间的相关系数。一般地，转换回归模型被定义为

$$H(Y_i) = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i \quad (10.3.1)$$

Li 等 [117] 通过在模型 (10.3.1) 中假定  $H(\cdot)$  严格单调，提出使用秩相关系数衡量每个预测变量的重要性。他们没有使用之前定义的样本 Pearson 相关系数，而是提出了边际秩相关系数

$$\hat{\omega}_j = \frac{1}{n(n-1)} \sum_{i \neq l}^n I(X_{ij} < X_{lj}) I(Y_i < Y_l) - \frac{1}{4}$$

来衡量第  $j$  个预测变量  $X_j$  对响应变量  $Y$  的重要性，并将方法命名为 RCS (Rank Correlation Screening)。注意边际秩相关系数等于响应变量与第  $j$  个预测变量之间的 Kendall  $\tau$  相关系数的四分之一。因此，我们可以基于秩相关系数  $\hat{\omega}_j$  的绝对值筛选重要预测变量，即

$$\widehat{\mathcal{M}}_{\kappa_n} = \{1 \leq j \leq p : |\hat{\omega}_j| > \kappa_n\}$$

其中  $\kappa_n$  为预设定的阈值。

### 10.3.2 确定独立秩筛选

我们知道,  $X_j$  和  $Y$  之间的 Pearson 相关系数仅仅刻画了  $X_j$  和  $Y$  之间的线性相关性。Zhu 等人 [118] 提出使用  $X_j$  和  $I(Y < y)$  的 Pearson 相关系数来刻画  $X_j$  和  $Y$  之间的非线性相关性, 因为示性函数具有单调变换不变性, 即针对单调函数  $g(\cdot)$ , 有  $I(Y < y) = I(g(Y) < g(y))$ 。当  $j = 1, \dots, p$  时, 假设  $E(X_j) = 0$ ,  $\text{Var}(X_j) = 1$ , 我们得到随机变量  $X_j$  和  $I(Y < y)$  相关系数表示为

$$\Omega_j(y) = \text{cov}\{X_j, I(Y < y)\} = E[X_j E\{I(Y < y) | X_j\}]$$

直观地说, 如果  $X_j$  和  $Y$  是独立的, 对于任意的  $y$ , 都有  $\Omega_j(y) = 0$ 。另一方面, 如果  $X_j$  和  $Y$  是相关的, 则存在  $y$  使得  $\Omega_j(y) \neq 0$ 。他们提出使用

$$\omega_j = E\{\Omega_j^2(Y)\}, \quad j = 1, \dots, p$$

作为特征筛选的边际统计量。由于  $\omega_j$  为正, 可以使用  $\omega_j$  对所有预测变量进行排序。

假设  $\{(X_i, Y_i), i = 1, \dots, n\}$  是来自  $\{X, Y\}$  的随机样本。为了便于表示, 我们假设样本预测变量都是标准化的。对于任意给定的  $y$ ,  $\Omega_j(y)$  的样本矩估计量为

$$\widehat{\Omega}_j(y) = n^{-1} \sum_{i=1}^n X_{ij} I(Y_i < y)$$

因此,  $\omega_j$  的估计量为

$$\widehat{\omega}_j = \frac{1}{n} \sum_{j=1}^n \widehat{\Omega}_j^2(Y_j) = \frac{1}{n} \sum_{j=1}^n \left\{ \frac{1}{n} \sum_{i=1}^n X_{ij} I(Y_i < Y_j) \right\}^2, \quad j = 1, \dots, p$$

接下来, 使用  $\widehat{\omega}_j$  对所有候选预测变量  $X_j, j = 1, \dots, p$  的重要性进行排序, 然后选择最前面的几个作为重要预测变量, 此过程被命名为确定独立秩筛选, 简记为 SIRS (Sure Independent Ranking Screening)。

### 10.3.3 距离相关系数

这一章介绍另一种利用相关系数来有效地度量预测变量与响应变量之间的线性与非线性相关关系, 来进行超高维数据的特征筛选。Li, Zhong 和 Zhu [119] 提出了一种基于距离相关系数筛选方法 (DCS, Distance Correlation Screening)。前面介绍的 Pearson 相关系数、秩相关系数和广义相关系数仅仅定义了两个随机变量的相关关系, 而距离相关系数是定义了两个不同维度的随机向量的关系。

首先, 定义两个随机向量  $\mathbf{U} \in R^{q_1}$  和  $\mathbf{V} \in R^{q_2}$  之间的距离协方差为

$$\text{dcov}^2(\mathbf{U}, \mathbf{V}) = \int_{R^{q_1+q_2}} \|\phi_{\mathbf{U}, \mathbf{V}}(\mathbf{t}, \mathbf{s}) - \phi_{\mathbf{U}}(\mathbf{t})\phi_{\mathbf{V}}(\mathbf{s})\|^2 w(\mathbf{t}, \mathbf{s}) d\mathbf{t} d\mathbf{s} \quad (10.3.2)$$

其中  $\phi_U(\mathbf{t})$  和  $\phi_V(\mathbf{s})$  是  $\mathbf{U}$  和  $\mathbf{V}$  的边际特征函数,  $\phi_{\mathbf{U}, \mathbf{V}}(\mathbf{t}, \mathbf{s})$  是  $\mathbf{U}$  和  $\mathbf{V}$  的联合特征函数, 并且

$$w(\mathbf{t}, \mathbf{s}) = \left\{ c_{q_1} c_{q_2} \|\mathbf{t}\|_{q_1}^{1+q_1} \|\mathbf{s}\|_{q_2}^{1+q_2} \right\}^{-1}$$

$c_d = \pi^{(1+d)/2}/\Gamma\{(1+d)/2\}$  (这个选择是为了方便微分的计算)。这里  $\|\phi\|^2 = \phi\bar{\phi}$ ,  $\phi$  表示复值函数,  $\bar{\phi}$  是  $\phi$  的共轭。由定义 (10.3.2) 可知, 当且仅当  $\mathbf{U}$  和  $\mathbf{V}$  是独立时,  $\text{dcov}^2(\mathbf{U}, \mathbf{V}) = 0$ 。Székely, Rizzo 和 Bakirov [120] 证明了

$$\text{dcov}^2(\mathbf{U}, \mathbf{V}) = S_1 + S_2 - 2S_3$$

其中

$$\begin{aligned} S_1 &= E(\|\mathbf{U} - \tilde{\mathbf{U}}\| \|\mathbf{V} - \tilde{\mathbf{V}}\|), \quad S_2 = E(\|\mathbf{U} - \tilde{\mathbf{U}}\|)E(\|\mathbf{V} - \tilde{\mathbf{V}}\|) \\ S_3 &= E\{E(\|\mathbf{U} - \tilde{\mathbf{U}}\| \mid \mathbf{U})E(\|\mathbf{V} - \tilde{\mathbf{V}}\| \mid \mathbf{V})\} \end{aligned}$$

并且  $(\tilde{\mathbf{U}}, \tilde{\mathbf{V}})$  是  $(\mathbf{U}, \mathbf{V})$  的独立副本。因此,  $\mathbf{U}$  和  $\mathbf{V}$  之间的距离协方差可以通过代入对应的样本来估计。具体来说, 基于总体  $(\mathbf{U}, \mathbf{V})$  中的随机样本  $\{(\mathbf{U}_i, \mathbf{V}_i), i = 1, \dots, n\}$ , 我们有

$$\widehat{\text{dcov}}^2(\mathbf{U}, \mathbf{V}) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3$$

其中,

$$\begin{aligned} \widehat{S}_1 &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{U}_i - \mathbf{U}_j\| \|\mathbf{V}_i - \mathbf{V}_j\| \\ \widehat{S}_2 &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{U}_i - \mathbf{U}_j\| \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{V}_i - \mathbf{V}_j\| \\ \widehat{S}_3 &= \frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n \sum_{l=1}^n \|\mathbf{U}_i - \mathbf{U}_l\| \|\mathbf{V}_j - \mathbf{V}_l\| \end{aligned}$$

根据  $\mathbf{U}$  和  $\mathbf{V}$  的距离相关系数定义

$$\text{dcorr}(\mathbf{U}, \mathbf{V}) = \text{dcov}(\mathbf{U}, \mathbf{V}) / \sqrt{\text{dcov}(\mathbf{U}, \mathbf{V}) \text{dcov}(\mathbf{V}, \mathbf{V})}$$

根据以上距离协方差阵的估计过程, 可以得到距离相关系数估计量

$$\widehat{\text{dcov}}(\mathbf{U}, \mathbf{V}) / \sqrt{\widehat{\text{dcov}}(\mathbf{U}, \mathbf{U}) \widehat{\text{dcov}}(\mathbf{V}, \mathbf{V})}$$

假设响应变量  $\mathbf{Y}$  是多维度的, 如果从  $p$  个预测变量  $X_j$  中筛选出重要的预测变量, 我们提出使用如下边际距离相关系数估计量:

$$\widehat{\omega}_j = \widehat{\text{dcov}}(\mathbf{Y}, X_j) / \sqrt{\widehat{\text{dcov}}(\mathbf{Y}, \mathbf{Y}) \widehat{\text{dcov}}(X_j, X_j)}$$

并且使用  $\widehat{\omega}_j^2$  来对预测变量的重要性进行排序。

## 10.4 高维分类数据的特征筛选

### 10.4.1 Kolmogorov-Smirnov 统计量

Fan 和 Fan [121] 提出在高维二值分类中使用双样本  $t$  检验统计量作为特征筛选的边际统计量。尽管基于双样本  $t$  检验统计量 [121] 的特征筛选在高维分类问题中表现很好，但它可能会被重尾分布或者具有异常值的数据破坏，而且它是基于模型的，当数据结构不满足假定模型时，筛选结果也是失效的。

为了克服这些缺点，Mai 和 Zou [122] 提出了一种新的基 Kolmogorov-Smirnov 统计量的二值分类特征筛选方法。为了便于标记，重新标记  $Y = +1, -1$  为类标签。令  $F_{1j}(x) = \Pr(X_j \leq x | Y = 1)$  和  $F_{2j}(x) = \Pr(X_j \leq x | Y = -1)$  分别为给定  $Y = 1, -1$  时  $X_j$  的条件分布函数。因此，如果  $X_j$  和  $Y$  独立，则  $F_{1j}(x) \equiv F_{2j}(x)$ 。基于这一观察，上述条件分布函数的差可用于构建特征筛选的方法。

因此 Mai 和 Zou [122] 提出 Kolmogorov-Smirnov 边际效用：

$$\omega_j = \sup_{x \in \mathcal{R}} |F_{1j}(x) - F_{2j}(x)|$$

并且将

$$\widehat{\omega}_j = \sup_{x \in \mathcal{R}} |\widehat{F}_{1j}(x) - \widehat{F}_{2j}(x)|$$

作为特征筛选的边际统计量，Mai 和 Zou [122] 将这种特征筛选方法命名为 **Kolmogorov-Smirnov 统计量** (KS)，其中  $\widehat{F}_{1j}(x)$  和  $\widehat{F}_{2j}(x)$  为对应的经验条件分布函数，即

$$\widehat{F}_{1j}(x) = \frac{1}{n_1} \sum_{i:Y_i=1} I(X_{ij} \leq x), \quad \widehat{F}_{2j}(x) = \frac{1}{n_2} \sum_{i:Y_i=-1} I(X_{ij} \leq x)$$

其中  $n_1 = \sum_{i=1}^n I(Y_i = 1)$  和  $n_2 = \sum_{i=1}^n I(Y_i = -1)$ ，最后使用  $\widehat{\omega}_j$  来对预测变量的重要性进行排序。

### 10.4.2 均值-方差统计量

Cui, Li 和 Zhong [123] 提出了一种利用均值-方差指数进行超高维分类问题的确定独立筛选方法，形式如下：

$$\text{MV}(X_j, Y) = \mathbb{E}_{X_j} [\text{Var}_Y (F(X_j | Y))] \tag{10.4.1}$$

它不仅保留了 Kolmogorov 滤波器的优点，而且允许分类响应变量具有  $O(n^\kappa)$  个发散的类别，其中  $\kappa \geq 0$ 。假设分类响应变量  $Y$  有  $K$  个类别  $\{y_1, \dots, y_K\}$ 。令  $F_j(x) = \Pr(X_j \leq x)$  表示第  $j$  个特征  $X_j$  的边际分布函数， $F_{jk}(x) = \Pr(X_j \leq x | Y = y_k)$  表示给定  $Y = y_k$  时  $X_j$  的条件分布函数。由于  $Y$  是离散响应变量，Cui, Li 和 Zhong [123] 推导出

$$\text{MV}(X_j, Y) = \sum_{k=1}^K p_k \int [F_{jk}(x) - F_j(x)]^2 dF_j(x)$$

如果  $X_j$  和  $Y$  在统计上是独立的，那么对于任意的  $k$  和  $x$ ，理论上都有  $F_{jk}(x) = F_j(x)$ 。通过上式推导，均值-方差这是给定  $Y = y_k$  时  $X_j$  的条件分布函数与  $X_j$  的无条件分布函数之间的 Cramér-von Mises 距离的加权平均，其中  $p_k = \Pr(Y = y_k)$ 。他们进一步表明当且仅当  $X_j$  和  $Y$  在统计上独立时， $\text{MV}(X_j | Y) = 0$ 。

设  $\{(X_{ij}, Y_i) : 1 \leq i \leq n\}$  是总体  $(X_j, Y)$  中大小为  $n$  的随机样本，则均值方差估计量是

$$\widehat{\text{MV}}(X_j, Y) = \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^n \widehat{p}_k \left[ \widehat{F}_{jk}(X_{ij}) - \widehat{F}_j(X_{ij}) \right]^2$$

其中  $\widehat{p}_k = n^{-1} \sum_{i=1}^n I\{Y_i = y_k\}$ ,  $\widehat{F}_{jk}(x) = n^{-1} \sum_{i=1}^n I\{X_{ij} \leq x, Y_i = y_k\} / \widehat{p}_k$ ,  $\widehat{F}_j(x) = n^{-1} \sum_{i=1}^n I\{X_{ij} \leq x\}$ 。因此我们可以使用  $\widehat{\text{MV}}(X_j, Y)$  筛选出重要的预测变量。这个过程被称为基于 MV 的确定独立筛选方法。

#### 10.4.3 类别自适应筛选统计量

假设观察到具有  $K$  ( $K > 2$ ) 类的分类响应变量  $Y$ ，即  $\{y_1, y_2, \dots, y_K\}$ ，对于所有的  $k = 1, \dots, K$  都有  $p_k = \Pr(Y = y_k) > 0$ 。在大数据时代，不同分类变量响应变量的样本的来源可能不同，例如在不同的时间段或者在不同的实验方法下收集到。换句话说，具有分类响应变量的高维数据通常是异构的。因此，我们考虑

- (i) (异构性) 一组重要的预测变量  $\mathcal{A}_k \equiv \{1 \leq j \leq p : \Pr(W_k \leq w_k | \mathbf{X})\}$  依赖于  $X_j\}$ ，其中对于不同的  $k = 1, \dots, K$ ,  $W_k \equiv I(Y = y_k)$  重要变量集合  $\mathcal{A}_k$  可能不同；
- (ii) (稀疏性) 对于某个常数  $\alpha > 0$  的维度  $p = o\{\exp(n^\alpha)\}$ ， $|\mathcal{A}_k| = s_k = o(n)$ ，其中  $|\mathcal{A}_k|$  是  $\mathcal{A}_k$  的基数， $n$  是样本大小。

为了寻找不同类别下的重要变量，Xie 等 [124] 考虑了给定  $I(Y = y_k)$  时  $X_j$  的条件分布函数，即  $F_{jk}(x) = \Pr(X_j \leq x | Y = y_k) = \frac{\Pr(X_j \leq x, Y = y_k)}{\Pr(Y = y_k)}$ ，并提出如下边际筛选指数

$$\tau_{jk} = E_{X_j}\{F_{jk}(X_j)\} - \frac{1}{2}$$

很明显, 如果  $I(Y = y_k)$  与  $X_j$  独立, 则  $\tau_{jk} = 0$ 。令  $\{(\mathbf{X}_i, Y_i), i = 1, \dots, n\}$  是独立同分布随机样本。定义  $\hat{p}_k = \frac{1}{n} \sum_{i=1}^n I(Y_i = y_k)$ ,  $k = 1, \dots, K$ 。我们得到  $\tau_{jk}, j = 1, \dots, p$  的样本估计为

$$\hat{\tau}_{jk} = \frac{1}{n+1} \sum_{l=1}^n \left\{ \frac{1}{n} \sum_{i=1}^n \frac{I(X_{ij} \leq X_{lj}, Y_i = y_k)}{\hat{p}_k} \right\} - \frac{1}{2} \quad (10.4.2)$$

接下来, 我们根据  $|\hat{\tau}_{jk}|$  值的大小, 对所有候选预测变量  $X_j, j = 1, \dots, p$  的重要性进行排序。 $|\hat{\tau}_{jk}|$  也成为类别自适应筛选统计量, 并简记为 CAS (Category Adaptive Screening) 方法。

如果研究者对影响一个类集  $\mathcal{G} \subset \{1, 2, \dots, K\}$  的重要变量感兴趣, 可以采用如下筛选统计量

$$\hat{\tau}_{j,\mathcal{G}} \equiv \sup_{k \in \mathcal{G}} |\hat{\tau}_{jk}|$$

特别地, 如果筛选出影响到所有类别的重要变量, 我们采用

$$\hat{\tau}_j \equiv \sup_{k \in \{1, 2, \dots, K\}} |\hat{\tau}_{jk}|$$

因此, 类别自适应筛选统计量方法非常灵活, 既可以筛选出影响某一类的重要变量, 也可以用于筛选影响某一类集以及所有类的重要变量。

## 10.5 R 语言实现

下面介绍怎么用 R 语言实现上面提到的特征筛选方法。

### 10.5.1 基于边际模型拟合的特征筛选

首先介绍如何使用 R 语言实现基于边际模型拟合的特征筛选方法。首先我们需要定义生成数据的函数。

```
####SIS 和 MMLE 生成数据的函数 ####
gdat_linear<- function(n,p,beta,rho){
  M1 <- matrix(rep(1:p,p),ncol=p,byrow=F)
  corM <- rho^(abs(M1-t(M1)))
  X<- mvrnorm(n,rep(0,p),corM)#生成多元正态数
  epsilon <- rnorm(n)#生成n个随机正态分布的数
  Y <- beta[1]*X[,1]+beta[2]*X[,2]+beta[3]*X[,3] + epsilon
  dat <- list(X = X,Y = Y)
  return(dat)
}
```

接下来我们定义确定独立筛选 (SIS) [115] 的函数，即边际最小二乘估计。

```
SIS<-function(X,Y){
  p=dim(X)[2]
  n=length(Y)
  s<-matrix(0,nrow=p,ncol=1)
  Y<-Y-mean(Y)#centered y
  X<-apply(X,2, function(x) (x-mean(x))/sd(x))#X每列标准化
  for(j in 1:p){
    x0<-X[,j]
    s[j,] <-abs(mean(Y*X[,j]))
  }
  list(rank=(p+1)-rank(s))
}
```

下面是求取边际极大似然估计量 (MMLE) [125] 的函数。

```
MMLE<-function(X,Y){
  p=dim(X)[2]
  n=length(Y)
  s<-c()
  Y<-Y-mean(Y)#centered y
  X<-apply(X,2, function(x) (x-mean(x))/sd(x))
  for(j in 1:p){
    s[j] <-glm(Y ~ X[,j], family = gaussian())$coefficients[2] #广义线性模型
  }
  list(rank=(p+1)-rank(s))
}
```

定义了生成数据的函数、SIS 和 MMLE 函数，接下来的程序是调用函数。

```
library(MASS)
n = 50#样本量
p = 500#维数
rho<-0.5
beta <- c(1,1,0.8,rep(0,p-3))#参数真实值
d = floor(n/log(n))#前d个最重要的变量
####生成数据#####
zdata <- gdat_linear(n,p,beta,rho)
Y <- zdata$Y
X <- zdata$X
SIS(X,Y)$rank[1:d]
MMLE(X,Y)$rank[1:d]
```

经过运行函数，我们筛选出前  $\lceil n/\log(n) \rceil$  个重要特征：

```
###SIS###
[1] 3 1 2 7 9 10 4 87 245 459 69 452
###MMLE###
[1] 3 1 2 6 8 9 4 38 115 219 462 216
```

接下来再次定义生成数据的函数，以下程序是为非参数独立筛选 (NIS) 方法生成数据的函数。

```
gdat_nonlinear<- function(n,p,beta){
  X <- array(0,dim = c(n,p))
  for(s in 1:p){
    X[,s] = runif(n)#生成从0到1区间范围内的服从正态分布的随机数
  }
  epsilon <- rnorm(n)
  Y <- beta[1]*X[,1]+beta[2]*tan(X[,2])+beta[3]*X[,3]^3 + epsilon
  dat <- list(X = X,Y = Y)
  return(dat)
}
```

下面程序是非参数独立筛选 (NIS) [126] 的函数。

```
NIS<-function(X,Y){
  p=dim(X)[2]
  n=length(Y)
  s<-matrix(0,nrow=p,ncol=1)
  bfit<-matrix(0,nrow=n,ncol=p)
  Y<-Y-mean(Y)#centered y
  #X<-apply(X,2,function(x) (x-min(x))/(max(x)-min(x)))#X每列标准化
  for(j in 1:p){
    x0<-X[,j]
    knots=quantile(x0,c(1/4,2/4,3/4))#结点
    a=bs(x0, knots=knots,degree=1)#B样条
    b=lm(Y~a)#线性拟合
    s[j,] <-sum((b$fitted)^2)/n
    bfit[,j]<-b$fitted
  }
  list(rank=(p+1)-rank(s),fit=bfit,wk=s)
}
```

我们已经定义了生成数据的函数和非参数独立筛选 (NIS) 函数，下面程序是演示如何使用 NIS 筛选特征的过程。

```
####生成数据、运行定义筛选方法的函数#####
library(splines)
library(MASS)
n = 200#样本量
p = 2000#维数
beta <- c(4,2,5,rep(0,p-3))#参数真实值
d = floor(n/(4*log(n)))#前d个最重要的变量
####生成数据#####
zdata <- gdat_nonlinear(n,p,beta)
Y <- zdata$Y
X <- zdata$X
NIS(X,Y)$rank[1:d]
```

经过运行以上程序，我们筛选出前  $\lceil n/(4 * \log(n)) \rceil$  个特征：

```
####NIS#####
[1] 2 3 1 1046 1306 871 1905 1181 1167
```

### 10.5.2 基于边际相关系数的特征筛选

下面介绍如何 R 语言实现基于边际相关系数的特征筛选方法，我们需要定义生成数据的函数。以下程序为生成数据的函数。

```
##### 生成数据的函数 #####
gdat<- function(n,p,beta,rho){
  M1 <- matrix(rep(1:p,p),ncol=p,byrow=F)
  corM <- rho^(abs(M1-t(M1)))
  X<- mvrnorm(n,rep(0,p),corM)
  epsilon <- rnorm(n)
  Y <- beta[1]*exp(X[,1])+beta[2]*tan(X[,2])+beta[3]*X[,3]^3 + epsilon
  dat <- list(X = X,Y = Y)
  return(dat)
}
```

接下来我们定义秩相关筛选 (RCS) [117] 的函数。

```
RCS<-function(X,Y){
  p=dim(X)[2]
  n=length(Y)
  w<-c()
  for(j in 1:p){
    a<-0
    for(i in 1:n){
      a<-a+sum((X[,j]<X[i,j])*(Y<Y[i]))/n#找到 X[:,j] 中小于 X[i,j] 和 Y 中小于 Y[i] 的做计算
    }
    w[j]<-abs(a/(n-1)-1/4)
  }
  list(rank=(p+1)-rank(w))
}
```

下面是确定独立秩筛选 (SIRS) [118] 方法的函数。

```
SIRS<-function(X,Y){
  p=dim(X)[2]
  n=length(Y)
  w<-matrix(0, nrow=p, ncol=1)
  for(k in 1:p){
    w.k.j<-NULL
    for(j in 1:n){
      s<-(t(X[,k])%*%(1*(Y<Y[j]))/n)^2#将 X[:,k] 的转置与 (Y 中小于 Y[j] 做计算
      w.k.j<-c(w.k.j,s)
    }
    w[k,]<-(n^3*(mean(w.k.j)))/(n*(n-1)*(n-2))
  }
  list(rank=(p+1)-rank(w))
}
```

下面程序定义了距离相关系数筛选 (DCS) [119] 方法的函数。

```
DCS<-function(X,Y){
  p=dim(X)[2]
  n=length(Y)
  s<-matrix(0, nrow=p, ncol=1)
  for(j in 1:p){
    s[j]<-bcdcor(X[,j],Y)#距离相关系数
  }
  list(rank=(p+1)-rank(s))
}
```

我们已经定义了生成数据的函数、秩相关筛选 (RCS)、确定独立秩筛选 (SIRS)、距离相关系数筛选 (DCS) 方法的函数。下面程序演示了使用这三种方法进行特征筛选的过程。

```
####生成数据、运行定义筛选方法的函数#####
library(energy)
library(MASS)
n = 200#样本量
p = 1000#维数
rho<-0.5
beta <- c(4,2,5,rep(0,p-3))#参数真实值
d = floor(n/(4*log(n)))#前d个最重要的变量
zdata <- gdat(n,p,beta,rho)
Y <- zdata$Y
X <- zdata$X
RCS(X,Y)$rank[1:d]
SIRS(X,Y)$rank[1:d]
DCS(X,Y)$rank[1:d]
```

经过运行以上程序，我们分别用三种方法筛选出了  $\lceil n/(4 \log(n)) \rceil$  个特征：

```
###RCS###
[1] 2   3   1   4   36  25  493 402 131
###SIRS###
[1] 2   3   1   33 194 483 128 260 202
###DCS###
[1] 2   3   1   4   71  28 346 584 172
```

### 10.5.3 高维分类数据的特征筛选

下面介绍如何使用 R 语言实现高维分类数据的特征筛选方法。首先我们要定义生成数据的函数。

```
###生成数据的函数#####
gdat=function(n,p,R){
  epsilon <- array(rnorm(n*p),dim = c(n,p))
  P <- c(50/100,50/100)
```

```
#Y <- rbinom(n, 1, P) + 1
Y <- sample(1:R, n, replace = T, prob=P) #在1:R中以P的概率随机重复抽样n个数
X <- array(0, dim=c(n, p))
u <- array(0, dim=c(R, p))
diag(u) <- 1.5 #对角线为1.5
X <- u[Y, ] + epsilon
dat <- list(X=X, Y=Y)
return(dat)
}
```

下面程序定义的是通过 Kolmogorov-Smirnov 统计量 [122] 进行特征筛选的方法。

```
KS<-function(X,Y){
  n<-nrow(X)
  p<-ncol(X)
  data_ks = t(cbind(Y,X))
  ks_index = order(data_ks[,1])#按从小到大的顺序排，返回索引
  data_m0 = data_ks[,ks_index]
  n_1 = length(which(Y==1))
  n_2 = n - n_1
  tau_ks<-c()
  for(j in 1:p){
    group1 = data_m0[(j+1),1:n_1]
    group2 = data_m0[(j+1),(n_1+1):n]
    tau_ks[j] = ks.test(group1,group2)$statistic#做Kolmogorov-Smirnov检验
  }
  list(rank=(p+1)-rank(tau_ks))
}
```

接下来我们定义均值方差统计量筛选 [123] 函数。

```
MV <- function(X,Y){
  n = nrow(X)
  p = ncol(X)
  pr = rep(0,R)
  n_1 = length(which(Y==1))
  n_2 = n - n_1
  pr[1] = n_1/n
  pr[2] = n_2/n
  mv<-c()
  for(j in 1:p){
    F1 = apply(outer(X[which(Y==1),j],t(X[,j]),"<"),3,t)*1
    F2 = apply(outer(X[which(Y==2),j],t(X[,j]),"<"),3,t)*1
    FF = apply(outer(X[,j],t(X[,j]),"<"),3,t)*1
    mv[j] = sum(pr[1]*(apply(F1,2,sum)/n/pr[1] - apply(FF,2,sum)/n)^2 +
      pr[2]*(apply(F2,2,sum)/n/pr[2] - apply(FF,2,sum)/n)^2)/n
  }
  list(rank=(p+1)-rank(mv))
}
```

下面程序定义的是类别自适应特征筛选 [124] 函数。

```
CAS<- function(X,Y){
  n = nrow(X)
  p = ncol(X)
  tau1 = 0
  tau2 = 0
  n_1 = length(which(Y==1))
  n_2 = n - n_1
  n = n_1 + n_2
  for(i in 1:n_1){
    for(j in 1:n){
      tau1 = tau1 + (X[i,] <= X[j,])
    }
  }
  for(i in 1:n_2){
    for(j in 1:n){
      tau2 = tau2 + (X[i+n_1,] <= X[j,])
    }
  }
  Tau1 = tau1/n/n_1 - 1/2
  Tau2 = tau2/n/n_2 - 1/2
  Tau = cbind(abs(Tau1), abs(Tau2))
  tau_sp_max = apply(Tau,1,max)#按行找出最大值
  list(rank=(p+1)-rank(tau_sp_max))
}
```

在前面我们分别定义了通过 Kolmogorov-Smirnov 统计量进行特征筛选、均值方差特征筛选、类别自适应特征筛选的函数，以及生成数据的函数。接下来我们通过示例来展示如何使用这些方法进行特征筛选。

```
###生成数据、运行定义筛选方法的函数###
n = 50#样本量
p = 500#维数
R = 2#二值响应变量
d = floor(n/log(n))#前d个最重要的变量
zdata <- gdat(n,p,R)
Y <- zdata$Y
X <- zdata$X
data_full = cbind(Y,X)
data_full1 = data_full[which(Y == 1),]
data_full2 = data_full[which(Y == 2),]
outdata = rbind(data_full1,data_full2)
Y_full = outdata[,1]
X_full = outdata[,2:(p+1)]
MV(X_full,Y_full)$rank[1:d]
KS(X_full,Y_full)$rank[1:d]
CAS(X_full,Y_full)$rank[1:d]
```

经过运行以上程序，我们分别用三种方法筛选出了  $\lceil n/\log(n) \rceil$  个特征：

```
###MV###
[1] 1 2 416 343 198 357 76 474 166 396 115 175
```

```
###KS###
[1] 2.0 1.0 112.0 243.5 247.0 310.0 267.5 255.0 495.5 230.5 370.0 238.5
###CAS###
[1] 1.0 2.0 279.5 261.0 187.0 258.0 68.0 379.0 220.5 473.5 141.0 168.5
```

## 10.6 Python 语言实现

下面介绍怎么用 Python 语言实现上面提到的特征筛选方法。

### 10.6.1 基于边际模型拟合的特征筛选

首先介绍如何使用 Python 语言实现基于边际模型拟合的特征筛选方法。首先我们需要导入需要的库和函数。

```
import numpy as np
import math
import statsmodels.api as sm
from scipy.interpolate import BSpline
from sklearn import linear_model
```

接下来是定义生成数据的函数。

```
### SIS 和 MMLE 生成数据的函数 ####
def gdat_linear(n,p,beta,rho):
    a = np.arange(1, p + 1).reshape(p, 1)
    M1 = np.tile(np.mat(a), (1, p))#将a在列上重复p次
    corM = np.float_power(rho*np.ones_like(M1), abs(M1 - M1.T))
    X = np.random.multivariate_normal(np.arange(1, p + 1), corM, size=n)
    #生成多元正态分布矩阵
    epsilon = np.random.normal(size=(1, n))#随机产生服从正态分布的数
    Y = beta[0] * X[:, 0] + beta[1] * X[:, 1] + beta[2] * X[:, 2] + epsilon#构成Y
    return {"X":X, "Y":Y}
```

下面程序是确定独立特征筛选 (SIS) [115] 的函数，即边际最小二乘估计。

```
### Sure independence screening (SIS) by Fan and Lv (2008) ####
def SIS(X,Y):
    p = X.shape[1]#获取X矩阵的列数
    n = len(Y)#获取Y的长度
    s = []
    Y = Y - np.mean(Y) #centered y
    X = (X - np.mean(X, 0)) / np.std(X, 0)#标准化X
    for j in range(0,p):
        x0 = X[:, j]
        s.append(abs(np.mean(Y * x0)))#将Y与X中的每一列进行计算
```

```
rank=(p - 1) - np.argsort(s)#rank为s中的元素从大到小排列其对应的index(索引)
return rank
```

接下来是边际极大似然估计量 (MMLE) [125] 的函数。

```
### Marginal Maximum Likelihood Estimator (MMLE) by Fan and Song (2010)###
def MMLE(X, Y):
    p = X.shape[1] #列数
    n = len(Y)
    s = []
    Y = Y-np.mean(Y) #centered y
    f = lambda x:(x-np.mean(x))/np.std(x)#创建一个标准化函数
    X = np.apply_along_axis(f,1,X)#在X的每一列上应用这个标准化函数
    for j in range(p):
        model = sm.GLM(Y[0],X[:,j]).fit()#默认是gauss分布,广义线性模型
        s.append(model.params[0]) #找到模型的回归系数
    rank = (p - 1) - np.argsort(s)
    return rank
```

定义了一系列生成数据的函数、SIS 和 MMLE 函数，下面是运行函数。

```
### 运行函数 ####
n = 50### 样本量
p = 500### 维数
rho = 0.5
d = math.floor(n / math.log(n))###前d个最重要的变量
beta = np.hstack((np.array([1, 1, 0.8]), np.zeros(p + 1 - 3)))
zdata = gdat_linear(n, p, beta, rho)#运行函数gdat_linear
X = zdata["X"]
Y = zdata["Y"]
print(SIS(X, Y)[0:d])#选取SIS方法中前d个变量
print(MMLE(X, Y)[0:d])#选取MMLE方法中前d个变量
```

我们得到了以下  $\lceil n/\log(n) \rceil$  个特征。

```
###SIS###
[385 25 124 11 284 299 386 278 159 420 485 389]
###MMLE###
[250 246 256 257 240 255 243 239 232 244 271 233]
```

接下来是为非参数独立筛选 (NIS) 方法生成数据的函数。

```
### NIS生成数据的函数 ####
def gdat_nonlinear(n, p, beta):
    X = np.zeros((n, p))
    for s in range(p):
        X[:, s] = np.random.uniform(size=(1, n))#每一列生成n个从[0,1]中随机采样的数
    epsilon = np.random.normal(size=(1,n))#随机产生正态分布的数
    Y = beta[0]*X[:, 0] + beta[1]*np.tan(X[:, 1]) + beta[2]*X[:, 2]**3 + epsilon
    dat = {"X":X, "Y":Y}
    return dat
```

下面程序是非参数独立筛选 (NIS) [126] 的函数。

```
###Nonparametric Independent Screening (NIS) by Fan et al.(2011)###
def NIS(X,Y):
    from sklearn.linear_model import LinearRegression
    p = X.shape[1]
    n = len(Y)
    s = []
    bfit = []
    Y = Y - np.mean(Y) #centered y
    lm = linear_model.LinearRegression()#线性模型
    for j in range(p):
        x0 = X[:, j]
        knots = np.percentile(x0, (25, 50, 75, 100)) #结点
        spl = BSpline(t=knots, c=x0, k=1) #B样条函数
        b = np.array([spl(0), spl(25), spl(50), spl(75), spl(100)]).reshape(-1, 1)
        model = lm.fit(b.T, Y)#线性模型拟合
        s.append((sum(model.intercept_)**2) / n)#将线性模型的回归系数计算
        bfit = model.intercept_
    rank = (p-1) - np.argsort(s)
    fit = bfit
    wk = s
    return rank
```

我们已经定义了生成数据的函数和非参数独立筛选（NIS）函数，下面展示运行程序。

```
### 运行函数 ###
n = 200### 样本量
p = 2000### 维数
beta = np.hstack((np.array([4, 2, 5]), np.zeros(p+1-3)))###参数真实值
d = math.floor(n / (4 * math.log(n)))###前d个最重要的变量
###生成数据###
zdata = gdat_nonlinear(n,p,beta)#运行gdat_nonlinear函数
X = zdata["X"]
Y = zdata["Y"]
print(NIS(X,Y)[0:d])#选取NIS方法中前d个变量
```

经过运行以上函数，我们筛选出前  $\lceil n/(4 * \log(n)) \rceil$  个特征。

```
[1999 658 659 660 661 662 663 664 665]
```

## 10.6.2 基于边际相关系数的特征筛选

下面介绍如何 Python 语言实现基于边际相关系数的特征筛选方法，我们需要定义生成数据的函数。以下为本节程序需要用到的库。

```
import numpy as np
import math
import dcor
```

下面程序是定义生成数据的函数。

```
### 生成数据的函数 ####
def gdat(n,p,beta,rho):
    a = np.arange(1, p+1).reshape(p, 1)
    M1 = np.tile(np.mat(a), (1, p))
    corM = np.float_power(rho*np.ones_like(M1), abs(M1-M1.T))
    X = np.random.multivariate_normal(np.arange(1, p+1), corM, size=n)
    #生成多元正态分布矩阵
    epsilon = np.random.normal(size=(1,n))#随机产生服从正态分布的数
    Y = beta[0,0]*X[:,0] + beta[0,1]*np.tan(X[:,1]) + beta[0,2]*X[:,2]**3 + epsilon
    dat = {"X":X, "Y":Y}
    return dat
```

接下来是定义秩相关筛选 (RCS) [117] 的函数。

```
##Rank Correlation Screening (RCS) by Li et al. (2012) ####
def RCS(X, Y):
    p = X.shape[1]
    n = len(Y)
    w = []
    for j in range(p):
        a = 0
        for i in range(n):
            a = a + sum(((X[:,j] < X[i,j])*(Y < Y[i])))/n
            #找到X[:,j]中小于X[i,j]，和Y中小于Y[i]的值做计算
        w.append(abs(a/(n-1) - 1/4))
    rank = (p-1) - np.argsort(w)
    return rank
```

下面程序是确定独立秩筛选 (SIRS) [118] 方法的函数。

```
##Sure Independent Ranking and Screening (SIRS) by Zhu et al. (2011) ####
def SIRS(X, Y):
    p = X.shape[1]
    n = len(Y)
    w = []
    for k in range(p):
        w0 = []
        for j in range(n):
            s = np.dot(X[:, k].T, (1*(Y < Y[j]))/n)**2
            #将X[:,k]的转置与(Y中小于Y[j])做计算
            w0.append(s)
        w.append(n**3 * (np.mean(w0) / (n*(n-1)*(n-2))))
    rank = (p-1) - np.argsort(w)
    return rank
```

接下来定义的是距离相关系数筛选 (DCS) [119] 方法。

```
##Distance Correlation Screening (DCS) by Li et al. (2012) ####
def DCS(X, Y):
    p = X.shape[1]
    n = len(Y)
```

```

s = []
for j in range(p):
    s.append(dcor.distance_correlation(X[:,j], Y))#获取X[:,j]与Y的距离相关系数
rank = (p-1) - np.argsort(s)
return rank

```

在上面我们定义了生成数据的函数、秩相关筛选、确定独立秩筛选、距离相关系数筛选方法的函数。下面程序展示三种方法的运行过程。

```

##运行函数
n = 200### 样本量
p = 1000### 维数
rho = 0.5
beta = np.hstack((np.matrix((4, 2, 5)), np.zeros((1, p+1-3))))###参数真实值
d = math.floor(n/(4*math.log(n)))###前d个最重要的变量

zdata = gdat(n,p,beta,rho)#运行gdat函数
Y = zdata["Y"]
X = zdata["X"]
datafull = np.hstack((X, Y.T))#水平方向合并矩阵X与Y
Y = datafull[:,p]#最后的第p列为Y
X = datafull[:,0:p]#前p-1列为X
print(RCS(X, Y)[0:d])#选取RCS方法中前d个变量
print(SIRS(X, Y)[0:d])#选取SIRS方法中前d个变量
print(DCS(X, Y)[0:d])#选取DCS方法中前d个变量

```

经过以上过程，我们筛选出以下  $\lceil n/(4 * \log(n)) \rceil$  个特征。

```

####RCS####
[ 61 359 68 495 239 60 685 651 683]
####SIRS####
[999 998 997 996 995 994 993 992 991]
####DCS####
[662 405 969 962 165 680 464 280 283]

```

### 10.6.3 高维分类数据的特征筛选

下面介绍如何使用 Python 语言实现高维分类数据的特征筛选方法。首先我们要导入本节程序需要的库和函数。

```

import numpy as np
import itertools
from scipy.stats import ks_2samp
import math
import random

```

下面定义的是生成数据的函数。

```
### 生成数据的函数 ####
def gdat(n, p, R):
    epsilon = np.random.normal(size=(n, p))#随机产生服从正态分布的数据
    P = np.array((50/100, 50/100))
    Y = np.random.choice(np.array((1, R)), n, p=P)
    X = np.zeros((n, p))
    u = np.zeros((R, p))
    if R < p:#将矩阵的对角线元素为1.5
        for i in range(R):
            u[i, i] = 1.5
    else:
        for i in range(p):
            u[i, i] = 1.5

    X = u[Y-1, :] + epsilon #构成X
    dat = {"X":X, "Y":Y}
    return dat
```

接下来是通过 Kolmogorov-Smirnov 统计量 [122] 进行特征筛选的方法。

```
##Kolmogorov-Smirnov Statistics by Mai and Zhou (2013)##
def KS(X,Y):
    n = X.shape[0]#行数
    p = X.shape[1]#列数
    data_ks = np.hstack((np.mat(Y).T, X)).T#构成[Y.T,X]矩阵
    ks_index = np.lexsort(data_ks[0, :])#从小到大排序，返回数组
    data_m0 = data_ks[:, ks_index[0]]#按照索引ks_index构成data_m0
    n_1 = len(np.where(Y == 1)[0])#找到Y中等于1的数共有多少个
    n_2 = n - n_1
    tau_ks = []
    for j in range(p):
        group1 = data_m0[(j+1), 0:n_1]
        group2 = data_m0[(j+1), (n_1):n]
        group1 = list(itertools.chain(*group1.tolist()))
        group2 = list(itertools.chain(*group2.tolist()))
        tau_ks.append(ks_2samp(group1, group2)[0])#执行Kolmogorov-Smirnov检验
    rank = (p-1) - np.argsort(tau_ks)
    return rank
```

下面程序是均值方差统计量筛选 [123] 方法。

```
## Mean-Variance Screening by Cui et al., (2015)##
def MV(X,Y):
    n = X.shape[0]#行数
    p = X.shape[1]#列数
    pr = np.zeros(R)
    n_1 = len(np.where(Y == 1)[0])#找到Y中值为1的个数
    n_2 = n - n_1
    pr[0] = n_1 / n
    pr[1] = n_2 / n
    mv = []
    for j in range(p):
```

```

#找到Y中值为1的相对应的X[:,j]列的值，与X[:,j]中每个值做比较
F1 = [[1 if bb < aa else 0 for bb in X[np.where(Y == 1), j][0]]
for aa in X[:, j].T]
#找到Y中值为2的相对应的X[:,j]列的值，与X[:,j]中每个值做比较
F2 = [[1 if bb < aa else 0 for bb in X[np.where(Y == 2), j][0]]
for aa in X[:, j].T]
#将X[:,j]中每个值互相做比较
FF = [[1 if bb < aa else 0 for bb in X[:, j]] for aa in X[:,j].T]
#每列应用求和函数
F1_apply = np.apply_along_axis(lambda s:sum(s), 1, np.array(F1))/n/pr[0]
FF_apply = np.apply_along_axis(lambda s:sum(s), 1, np.array(FF))
F1_apply_sub_FF_apply = pr[0]*((F1_apply-FF_apply)**2)
F2_apply = np.apply_along_axis(lambda s:sum(s), 1, np.array(F2))/n/pr[1]
FF_apply = np.apply_along_axis(lambda s:sum(s), 1, np.array(FF))
F2_apply_sub_FF_apply = pr[1]*((F2_apply-FF_apply)**2)
mv.append(sum(F1_apply_sub_FF_apply+F2_apply_sub_FF_apply)/n)
return (p-1) - np.argsort(mv)

```

接下来是类别自适应特征筛选 [124] 方法。

```

### Category Adaptive Screening by Xie et al., (2020)###
def CAS(X, Y):
    n = X.shape[0] #行数
    p = X.shape[1] #列数
    tau1 = 0
    tau2 = 0
    n_1 = len(np.where(Y == 1)[0]) #找到Y中等于1的值共有几个
    n_2 = n - n_1
    n = n_1 + n_2
    for i in range(n_1):
        for j in range(n):
            tau1 = tau1 + (X[i, :] <= X[j, :])
            #找到X前n_1行中小于等于X[j,:J]中的个数累计做和
    for i in range(n_2):
        for j in range(n):
            tau2 = tau2 + (X[i+n_1, :] <= X[j, :])
            #找到X从n_1+1行到最后一行中小于等于X[j,:J]中的个数累计做和
    Tau1 = tau1/n/n_1 - 1/2
    Tau2 = tau2/n/n_2 - 1/2
    Tau = np.vstack((abs(Tau1), abs(Tau2)))
    tau_sp_max = Tau.max(axis=0) #选取每行的最大值
    rank = (p-1) - np.argsort(tau_sp_max)
    return rank

```

以上我们定义了通过 Kolmogorov-Smirnov 统计量进行特征筛选、均值方差特征筛选、类别自适应特征筛选的函数，以及为这些方法生成数据的函数。下面是运行函数的过程。

```

### 运行函数 ###
n = 50### 样本量
p = 500### 维数
R = 2### 二值响应变量

```

```

d = math.floor(n/math.log(n))###前d个最重要的变量
zdata = gdat(n,p,R)#运行函数gdat
Y = zdata["Y"]
X = zdata["X"]
data_full = np.hstack((np.mat(Y).T, X))#水平方向合并矩阵Y.T与X
data_full1 = [i.tolist() for i in data_full.A if i[0] == 1]
#找到data_full中Y值为1的行
data_full2 = [i.tolist() for i in data_full.A if i[0] == 2]
#找到data_full中Y值为2的行
outdata = np.array(data_full1+data_full2)#按照行合并data_full1和data_full2
Y_full = outdata[:, 0]#第0列为Y
X_full = outdata[:, 1:(p+1)]#1到p列为X
print(KS(X_full,Y_full)[0:d])#选取KS方法中前d个变量
print(CAS(X_full,Y_full)[0:d])#选取CAS方法中前d个变量
print(MV(X_full,Y_full)[0:d])#选取MV方法中前d个变量

```

运行以上程序，我们筛选出  $\lceil n/\log(n) \rceil$  个特征。

## 10.7 练习

1. 边际非参估计假设我们有一个随机样本  $\{(\mathbf{X}_i, Z_i, Y_i)\}_{i=1}^n$  来自于部分线性模型

$$Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + m(Z_i) + \varepsilon_i,$$

其中  $\mathbf{X}_i$  和  $Z_i$  分别是  $p$  维和 1 维的协变量，对响应变量分别产生线性和分线性影响。 $\varepsilon_i$  是条件均值为 0 的随机误差。为了快速识别  $\mathbf{X}_i$  中的重要变量，采用特征筛选方法。

- (1) 根据 10.2.1 和 10.2.3 节的内容，通过光滑样条逼近之后，考虑的  $p$  个边际回归模型对应的二次损失函数是什么？
- (2) 写出边际筛选统计量。
- (3) R 与 Python 语言程序实现。
2. 基于边际相关系数的特征筛选方法大多采用了传统统计的相关系数，基于此，
  - (1) 请写出 Spearman 相关系数，并解释其解决了何种统计模型的特征筛选问题。
  - (2) 请列举其他可以用来进行特征筛选的相关系数。
3. 在 10.4.3 节的高维分类数据的类别自适应筛选统计量，考虑了协变量是连续的情形。而当协变量是有序分类变量时，写出相应的自适应筛选统计量形式。
4. 请使用心肌病微阵列数据集确定对小鼠 G 蛋白偶联受体 (Ro1) 过度表达最具影响力的基因。此数据集的样本量  $n = 30$ ，变量个数  $p = 6319$ ，属于超维数据。基于 R 或 Python 语言，

(1) 使用边际最小二乘、广义和秩相关系数以及距离相关系数筛选出前  $\lceil n / \log(n) \rceil$  重要变量，并对比解释筛选结果。

(2) 请使用一步变量选择方法 Lasso, SCAD, 与“特征筛选 + 变量选择”两步模型拟合方法，即 SIS-Lasso, SIS-SCAD，对心肌病微阵列数据集进行线性回归分析。并使用样本内 MSE，即  $\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)$  对比模型预测结果。

5. 最近科研人员使用 mRNA 表达谱分析数据研究肺癌成因。本肺癌数据集包含 203 个快速冷冻肺肿瘤和正常肺的样本，mRNA 表达水平为 12600，即  $n = 203$ ,  $p = 12600$ 。这里，响应变量表示肿瘤类型：肺腺癌 (ADEN) 139 例，肺鳞状细胞癌 21 例癌 (SQUA)，小细胞肺癌 (SCLC) 6 例，20 例肺类癌 (COID)，其余 17 例正常肺样本 (正常)。基于 R 或 Python 语言，

(1) 使用均值方差统计量与类别自适应筛选统计量筛选出前  $\lceil n / \log(n) \rceil$  重要变量，并对比解释筛选结果。

(2) 请使用“特征筛选 + 变量选择”两步模型拟合方法，即 CAS-Lasso, CAS-SCAD，对心肌病微阵列数据集进行线性 Logistic 回归分析。并使用样本内错分率，即  $\frac{1}{n} \sum_{i=1}^n I[Y_i \neq \hat{Y}_i]$  对比预测结果。

# 第十一章 在线学习

## 11.1 简介

随着科技的快速发展，我们迎来了大数据时代。通过对海量数据的处理与分析，可以发现巨大的社会价值，服务于生产生活。日益增长的海量数据往往呈现“流”特征，亦称“数据流”，此类大数据对存储承载力与计算机的计算性能等方面提出了更高的要求，传统的基于离线优化的机器学习算法面对极大的挑战。因此，在线学习的出现解决了这一问题。[在线学习](#)算法可被理解为：在重复决策的过程中，算法基于之前的经验以及当前的数据做出预测，以实现实时决策并在不断地对模型进行改进，来提高预测的精度。在线学习算法已经被广泛应用于数据流的分析中。

在线学习是基于机器学习的方法，对海量数据流进行训练，基于之前的经验（即保存下来的估计），来不断更新最佳的预测，而非传统地以批量处理的方式运行。在数据流框架下，传统的批量学习方法具有时间和空间成本高、效率低、扩展性差的特点，在线学习方法仅仅基于当前数据和历史估计结果进行更新，算法的效率和可扩展性大大提高。

接下来介绍一个在线学习的应用实例，比如在云计算中使用并行处理时，需要实现一种分配资源和调度任务执行顺序的机制。由于资源和任务在过程中不断更新，后台研究人员无法一次性得到所有的信息进行离线训练。因此可以利用在线学习算法根据实际任务执行的更新信息动态调整资源分配，提高云的利用率。

在线学习适合大数据时代的分布式数据流的处理流程，目前，在线学习已经成为机器学习与人工智能领域的重要研究课题，其研究方向可能集中在对维度更敏感的高效在线图优化算法、理论方面上下限分析及最优解的寻找等方面。

下面以均值模型、线性模型、岭回归以及高维模型讲解在线学习算法，主要介绍3种方法：累积统计量，在线梯度下降以及正则化的在线梯度下降。

## 11.2 累积统计量在线学习

### 11.2.1 均值模型

平均值反映数据的位置，描述数据中心。对于一组  $p$  维的数据  $\mathbf{X}_1, \dots, \mathbf{X}_n$ ，我们可构建均值模型为

$$\mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\epsilon}_i, i = 1, \dots, n$$

模型通过误差  $\boldsymbol{\epsilon}_i$  来描述以  $\boldsymbol{\mu}$  为中心的样本值。对于参数  $\boldsymbol{\mu}$ ，用样本均值进行参数估计，即

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i$$

然而，对于实时获取的海量数据流来讲，我们会按照时间观测到数据  $\{\mathbf{X}_{ti} : i = 1, \dots, n_t, t = 1, \dots\}$ ， $\mathbf{X}_{ti}$  代表第  $t$  批次数据流第  $i$  个样本， $n_t$  是第  $t$  批次数据的样本量。数据的更新意味着需要对模型进行不断地重复计算。受限于数据存储承载力与计算机的计算性能，传统的离线算法不适用于数据流的均值估计。

因此，我们可以构建在线学习均值模型，即在保存的前面算法结果的基础上仅对新数据进行处理，从而不断更新总体期望的估计结果。截止到第  $t$  批次，均值估计过程如下，

$$\hat{\boldsymbol{\mu}}_t = \bar{\mathbf{X}}_t = \frac{1}{N_t} \sum_{j=1}^t \sum_{i=1}^{n_j} \mathbf{X}_{ji} = \frac{1}{N_t} \left\{ \sum_{j=1}^{t-1} \sum_{i=1}^{n_j} \mathbf{X}_{ji} + \sum_{i=1}^{n_t} \mathbf{X}_{ti} \right\}$$

其中  $N_t = n_1 + \dots + n_t$ 。通过在线学习均值模型的构建，在已知  $\sum_{j=1}^{t-1} \sum_{i=1}^{n_j} \mathbf{X}_{ji}$ （已保存，仅为  $p$  维向量）的基础上，只需计算  $\sum_{i=1}^{n_t} \mathbf{X}_{ti}$  以及更新  $N_t$  的值，便可得到新数据下参数  $\boldsymbol{\mu}$  的估计。这样一来，可从多输入源分布式地输入数据，算法的效率和可扩展性都大大提高。

在线更新样本均值算法流程可以表示如下：

表 11.1 均值模型在线学习算法

---

#### 算法 1：均值模型累积统计量在线学习算法

---

**输入：** 初始化数据  $\mathbf{X}_{11}, \dots, \mathbf{X}_{1n_1}$

**输出：** 实时的样本均值  $\bar{\mathbf{X}}_t$

for  $t = 2$  to  $\infty$  do

    计算新的样本总数  $N_t = N_{t-1} + n_t$

    计算并输出新的样本总和  $\sum_{j=1}^t \sum_{i=1}^{n_j} \mathbf{X}_{ji} = \sum_{j=1}^{t-1} \sum_{i=1}^{n_j} \mathbf{X}_{ji} + \sum_{i=1}^{n_t} \mathbf{X}_{ti}$

    计算新的样本均值  $\bar{\mathbf{X}}_t = \frac{1}{N_t} \sum_{j=1}^t \sum_{i=1}^{n_j} \mathbf{X}_{ji}$

---

当然，我们可以选择终止距离  $\varepsilon$ ，在某一时刻终止上述在线算法，得到收敛值，并作为最终的均值估计值。

### 11.2.2 线性模型

线性模型是一种描述因变量和自变量之间线性关系的模型，假设因变量  $Y$  和  $p$  个自变量  $X_1, \dots, X_p$  之间存在简单的线性关系，可构建如下形式：

$$Y = h_{\beta}(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

对于实时获取的海量回归数据流来讲，我们假设观测到数据  $\{(\mathbf{X}_{ti}^T, Y_{ti})^T : i = 1, \dots, n_t, t = 1, \dots\}$ ，令

$$\mathbf{Y}_t = \mathbf{X}_t \beta + \epsilon_t,$$

$$\mathbf{Y}_t = \begin{pmatrix} Y_{t1} \\ Y_{t2} \\ \vdots \\ Y_{tn_t} \end{pmatrix} \quad \mathbf{X}_t = \begin{pmatrix} 1 & X_{t1,1} & \cdots & X_{t1,p} \\ 1 & X_{t2,1} & \cdots & X_{t2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{tn_t,1} & \cdots & X_{tn_t,p} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \quad \epsilon_t = \begin{pmatrix} \epsilon_{t1} \\ \epsilon_{t2} \\ \vdots \\ \epsilon_{tn_t} \end{pmatrix}$$

截止到第  $t$  批次数据的到来，回归模型的参数估计过程如下，

$$\widehat{\beta}_t = (\sum_{j=1}^t \mathbf{X}_j^T \mathbf{X}_j)^{-1} (\sum_{j=1}^t \mathbf{X}_j^T \mathbf{Y}_j) = (\sum_{j=1}^{t-1} \mathbf{X}_j^T \mathbf{X}_j + \mathbf{X}_t^T \mathbf{X}_t)^{-1} (\sum_{j=1}^{t-1} \mathbf{X}_j^T \mathbf{Y}_j + \mathbf{X}_t^T \mathbf{Y}_t)$$

通过在线学习线性模型的构建，在已知  $\sum_{j=1}^{t-1} \mathbf{X}_j^T \mathbf{X}_j$ （已保存，仅为  $p \times p$  维矩阵）和  $\sum_{j=1}^{t-1} \mathbf{X}_j \mathbf{Y}_j$ （已保存，仅为  $p$  维向量）的基础上，只需计算  $\mathbf{X}_t^T \mathbf{X}_t$  以及  $\mathbf{X}_t^T \mathbf{Y}_t$  的值，便可得到新数据下参数  $\beta$  的估计。这样一来，可从多输入源分布式地输入数据，算法的效率和可扩展性都大大提高。

在线更新回归参数算法流程可以表示如下：

表 11.2 线性模型在线学习算法

<b>算法 2：线性模型累积统计量在线学习算法</b>	
<b>输入：</b> 初始化数据 $\mathbf{X}_1, \mathbf{Y}_1$	
<b>输出：</b> 实时的回归参数估计 $\hat{\boldsymbol{\beta}}_t$	
for $t = 2$ to $\infty$ do	
更新 $p \times p$ 矩阵 $\sum_{j=1}^t \mathbf{X}_j^T \mathbf{X}_j = \sum_{j=1}^{t-1} \mathbf{X}_j^T \mathbf{X}_j + \mathbf{X}_t^T \mathbf{X}_t$	
更新 $p$ 维向量 $\sum_{j=1}^t \mathbf{X}_j^T \mathbf{Y}_j = \sum_{j=1}^{t-1} \mathbf{X}_j^T \mathbf{Y}_j + \mathbf{X}_t^T \mathbf{Y}_t$	
估计回归参数 $\hat{\boldsymbol{\beta}}_t = (\sum_{j=1}^t \mathbf{X}_j^T \mathbf{X}_j)^{-1} (\sum_{j=1}^t \mathbf{X}_j^T \mathbf{Y}_j)$	

## 11.3 在线梯度下降

本章节介绍在线梯度下降算法 (OGD; online gradient descent)，并研究算法收敛性以及在线性模型和岭回归模型中的应用。

### 11.3.1 OGD 算法一般形式

假设第  $t$  次批量数据到来，在传统离线优化算法中，我们采用一般形式的损失函数去估计感兴趣参数  $\boldsymbol{\beta}$ :

$$L_t(\boldsymbol{\beta}) = \sum_{j=1}^t \ell_j(\boldsymbol{\beta})$$

其中  $\ell_j(\boldsymbol{\beta}) = \ell(h_{\boldsymbol{\beta}}(\mathbf{X}_j), \mathbf{Y}_j)$  代表第  $j$  批量数据下的损失函数，例如二次损失或负对数似然等，且满足 Lipschitz 连续性。然而，若采用上述形式损失函数更新参数，则需要使用数据流的所有样本进行估计，模型训练速度会随着时间的增加变得越来越慢。

接下来，我们采用在线梯度下降算法对其进行无约束优化，估计模型参数。梯度下降法是一种最优化算法，从几何意义上讲，梯度是函数变化增加最快的方向，沿着梯度向量相反的方向，梯度减少更快，更加容易找到函数的最小值。根据损失函数的选取、更新参数时使用样本数量的不同，常见的梯度下降法分为批量梯度下降法 BGD、随机梯度下降法 SGD、小批量梯度下降法 MBGD 等。

对于日益增加的样本量，若使用传统离线的优化方法，虽然模型拟合精度得到提高，但是计算复杂度不断提高，训练速度会越来越慢，并且历史数据也被重复

使用。因此, Zinkevich [127] 提出了在线梯度下降 (OGD; online gradient descent) 的在线学习算法。在线梯度下降算法思想是在第  $t + 1$  阶段, 仅使用第  $t$  阶段的数据对应的损失函数  $l_t(\beta)$  进行梯度下降。计算出直接下降结果后, 若可行域  $\mathcal{K}$  受限, 还需将梯度下降后的结果投影回可行空间。

在明确回归函数  $h_\beta(\mathbf{X})$  以及损失函数  $l_t(\beta)$  的具体形式后, OGD 算法流程的一般框架如下所示:

表 11.3 OGD 算法

---

**算法 3-1: OGD 算法**

---

**输入:** 初始参数  $\beta_0$ , 终止距离  $\varepsilon$ , 学习速率  $\alpha$ , 可行域  $\mathcal{K}$

**输出:** 参数  $\beta$  的收敛值

for  $t = 0$  to  $\infty$  do

    计算梯度下降点  $\theta_{t+1} = \beta_t - \alpha l'_t(\beta_t)$

    计算投影点  $\beta_{t+1} = \operatorname{argmin}_{\beta \in \mathcal{K}} \|\theta_{t+1} - \beta\|^2$

    计算当前损失  $l_t(\beta_{t+1})$  及其梯度  $l'_t(\beta_{t+1})$

    如果前后步距离  $\|\beta_{t+1} - \beta_t\| < \varepsilon$

end for

---

### 11.3.2 OGD 算法收敛性分析

定义 **遗憾函数** (Regret Function):

$$R_T = \sum_{t=1}^T [\ell_t(\beta_t) - \ell_t(\beta^*)] \quad (11.3.1)$$

其中  $\beta^* = \operatorname{argmin}_\beta \sum_{t=1}^T \ell_t(\beta)$ ,  $\beta_t$  表示第  $t$  轮迭代值。随着时间  $T$  的增加, 遗憾函数的值接近于一个常量, 则算法的收敛值  $\beta_T$  与最优的  $\beta^*$  是一致的, 即可证明此在线学习方法是有效的。

接下来, 我们分析 OGD 算法的收敛性。根据投影点的对应关系, 我们可以得到

$$\|\beta_{t+1} - \beta^*\|^2 \leq \|\theta_{t+1} - \beta^*\|^2 = \|\beta_t - \beta^*\|^2 - 2\alpha \langle \ell'_t(\beta_t), (\beta^* - \beta_t) \rangle + \alpha^2 \|\ell'_t(\beta_t)\|^2$$

由此式变形后, 可以得到

$$\langle \ell'_t(\beta_t), (\beta^* - \beta_t) \rangle \leq \frac{1}{2\alpha} \left( \|\beta_t - \beta^*\|^2 - \|\beta_{t+1} - \beta^*\|^2 \right) + \frac{\alpha}{2} \|\ell'_t(\beta_t)\|^2$$

另一方面, 根据损失函数的凸性可以得到

$$R_T = \sum_{t=1}^T [\ell_t(\beta_t) - \ell_t(\beta^*)] \leq \sum_{t=1}^T \langle \ell'_t(\beta_t), (\beta^* - \beta_t) \rangle$$

合并上述两式可得

$$R_T \leq \frac{1}{2\alpha} \|\beta_1 - \beta^*\|^2 + \frac{\alpha}{2} \sum_{t=1}^T \|\ell'_t(\beta_t)\|^2$$

再由损失函数的 Lipschitz 连续性 (假设常数为  $L$ ) 可知,  $\forall t, \|\ell'_t(\beta_t)\| \leq L$ , 因此得到  $R_T$  上界的最终形式如下

$$R_T \leq \frac{1}{2\alpha} \|\beta_1 - \beta^*\|^2 + \frac{\alpha L^2 T}{2}$$

当且仅当  $\frac{1}{2\alpha} \|\beta_1 - \beta^*\|^2 = \frac{\alpha L^2 T}{2}$  时, 即  $\alpha = \frac{\|\beta_1 - \beta^*\|}{L\sqrt{T}}$  时, 式右边取最小值, 即  $R_T \leq \|\beta_1 - \beta^*\| L\sqrt{T}$ 。因此, 我们推出 OGD 算法的性能最差情况是  $\mathcal{O}(\sqrt{T})$ 。

### 11.3.3 线性模型的 OGD 算法

注意到上述 OGD 算法是对于一般模型而言的, 实现了在线梯度下降的算法功能, 具有一定的包容性与可扩展性, 相比于传统离线优化算法大大减少了计算的复杂度, 因为损失函数  $l_t(\beta)$  仅使用第  $t$  阶段的数据。借助 OGD 的算法思想, 假设第  $t$  阶段的二次损失函数为

$$l_t(\beta) = \frac{1}{2} \|\mathbf{Y}_t - \mathbf{X}_t \beta\|_2^2$$

对应的梯度是

$$l'_t(\beta) = -\mathbf{X}_t^T (\mathbf{Y}_t - \mathbf{X}_t \beta)$$

在基于二次损失构造线性模型在线学习算法时, 我们不妨在第  $t$  阶段, 选取第  $t-1$  阶段求得的  $\beta_{t-1}$  计算梯度值, 这样不仅能通过新数据不断迭代更新参数估计, 还减少了第  $t$  阶段模型拟合带来的计算复杂度。因此第  $t$  阶段的梯度的计算公式如下:

$$l'_t(\beta_{t-1}) = -\mathbf{X}_t^T (\mathbf{Y}_t - \mathbf{X}_t \beta_{t-1})$$

另外, 考虑到线性模型的可行域为  $\mathcal{R}^{p+1}$ , 梯度下降后无需将中间迭代结果投影回可行域。因此基于二次损失的线性模型在线学习算法流程可以表示如下:

表 11.4 OGD 算法

**算法 3-2: OGD 算法-基于二次损失**


---

**输入:** 初始化参数  $\beta_0$ , 终止距离  $\varepsilon$ , 学习速率参数  $\lambda$

**输出:** 参数  $\beta$  的收敛值

for  $t = 1$  to  $\infty$  do

    计算第  $t$  阶段梯度  $l'_t(\beta_{t-1}) = -\mathbf{X}_t^T(\mathbf{Y}_t - \mathbf{X}_t\beta_{t-1})$

    更新学习速率  $\alpha = \lambda t^{-\frac{1}{2}}$

    更新参数  $\beta_t = \beta_{t-1} - \alpha l'_t(\beta_{t-1})$

    如果前后步距离  $\|\alpha l'_t(\beta_{t-1})\| < \varepsilon$

        end for

---

### 11.3.4 岭回归模型的 OGD 算法

岭回归方法是一种解决回归数据共线性问题的监督学习方法。其使用的目标函数是在二次损失的基础上，增加  $L_2$  正则项，详细介绍参考回归模型章节。下面介绍基于二次损失的岭回归在线学习算法。假设观测到第  $t$  次批量数据的到来，在传统岭回归模型离线优化算法中，我们采用如下二次损失函数去估计感兴趣参数  $\beta$ :

$$l_t(\beta) = \frac{1}{2n_t} \|\mathbf{Y}_t - \mathbf{X}_t\beta\|_2^2 + \frac{\lambda}{2} \beta^T \beta$$

对应的梯度是

$$l'_t(\beta) = -\frac{1}{n_t} \mathbf{X}_t^T (\mathbf{Y}_t - \mathbf{X}_t\beta) + \lambda\beta$$

考虑到模型的训练速度以及计算的复杂度，不妨借助上一节线性模型在线学习的思想，第  $t$  阶段对应的梯度函数，选取第  $t-1$  阶段求得的  $\beta_{t-1}$  进行下一步的迭代，这样不仅能够通过新数据不断迭代更新参数估计，还不用基于当前数据拟合模型，即此时的梯度是

$$l'_t(\beta_{t-1}) = -\frac{1}{n_t} \mathbf{X}_t^T (\mathbf{Y}_t - \mathbf{X}_t\beta_{t-1}) + \lambda\beta_{t-1}$$

以上岭回归模型的在线学习算法流程可以表示如下：

表 11.5 岭回归在线学习算法

**算法 4：岭回归在线学习算法-基于二次损失**

**输入：** 初始化参数  $\beta_0$ , 终止距离  $\varepsilon$ , 学习速率  $\alpha$ , 调谐参数  $\lambda$

**输出：** 参数  $\beta$  的收敛值

for  $t = 1$  to  $\infty$  do

$$\text{计算第 } t \text{ 阶段梯度 } l'_t(\beta_{t-1}) = -\frac{1}{n_t} \mathbf{X}_t^T (\mathbf{Y}_t - \mathbf{X}_t \beta_{t-1}) + \lambda \beta_{t-1}$$

$$\text{更新参数 } \beta_t = \beta_{t-1} - \alpha l'_t(\beta_{t-1})$$

$$\text{如果前后步距离 } \|\alpha l'_t(\beta_{t-1})\| < \varepsilon$$

end for

## 11.4 基于正则化的在线梯度下降

### 11.4.1 FTL 算法

FTL (Follow The Leader) 算法在损失函数强凸的情形下能有效解决在线优化问题, 其算法的思想是通过最小化累计损失来更新参数, 公式如下:

$$\beta_{t+1} = \operatorname{argmin}_{\beta} \sum_{j=1}^t l_j(\beta)$$

由归纳法得 FTL 算法的遗憾函数上限为:

$$R_T \leq \sum_{t=1}^T (l_t(\beta_t) - l_t(\beta_{t+1}))$$

可见, 在损失函数强凸的情形下,  $\beta_t$  会收敛到  $\beta^*$ 。但是, 在在线线性优化 (Online Linear Optimization) 问题中不一定成立。接下来, 我们考虑一维的兴趣参数  $\beta$ , 损失函数关于  $\beta$  是线性的, 假设为  $l_t(\beta) = G_t \beta$ ,  $\beta \in [-1, 1]$ 。此时  $G_t$  是损失函数的梯度, 其取值为:

$$G_t = \begin{cases} -0.5, & t=1 \\ 1, & t \text{ 为偶数} \\ -1, & t>1 \text{ 且 } t \text{ 为奇数} \end{cases}$$

由参数更新公式得:

$$\beta_{t+1} = \operatorname{argmin}_{\beta} \sum_{j=1}^t G_j \beta$$

使用 FTL 算法则会出现参数震荡现象：

$t$	$G_t$	$\beta_{t+1} = \operatorname{argmin}_{\beta} \sum_{j=1}^t G_j \beta$
1	-0.5	$\operatorname{argmin}_{\beta} \{-0.5\beta\} = 1$
2	1	$\operatorname{argmin}_{\beta} \{-0.5\beta + 1\beta\} = -1$
3	-1	$\operatorname{argmin}_{\beta} \{-0.5\beta + 1\beta - 1\beta\} = 1$
4	1	$\operatorname{argmin}_{\beta} \{-0.5\beta + 1\beta - 1\beta + 1\beta\} = -1$
...	...	...

经过计算， $\beta$  最终的“收敛值”在 1 和 -1 之间震荡，并不收敛，FTL 算法在这个的在线线性优化问题上失效。因此，FTL 算法在损失函数强凸的情形下虽然有效，但在一般凸函数情形下不满足遗憾函数次线性，导致参数不收敛，可见算法不稳定。

#### 11.4.2 FTRL 算法

为了解决上述问题，FTRL (Follow the Regularized Leader) 算法在 FTL 算法的基础上增加正则化项  $P_\alpha(\beta)$  来使算法更稳定。FTRL 算法的参数更新公式如下：

$$\beta_{t+1} = \operatorname{argmin}_{\beta} \left\{ \sum_{j=1}^t l_j(\beta) + P_\alpha(\beta) \right\}$$

容易证明 FTRL 算法的遗憾函数上限为：

$$R_T \leq P_\alpha(\beta^*) + \sum_{t=1}^T [l_t(\beta_t) - l_t(\beta_{t+1})]$$

针对上述的在线线性优化问题，不妨令  $P_\alpha(\beta) = \frac{1}{2\alpha} \|\beta\|_2^2$ ，求导得出参数更新公式为： $\beta_{t+1} = \beta_t - \alpha G_t$ ，注意此时我们考虑的是多维的兴趣参数  $\beta$ ， $G_t$  是梯度向量。

FTRL 算法的遗憾函数上限为：

$$R_T \leq \frac{1}{2\alpha} \|\beta^*\|_2^2 + \alpha \sum_{t=1}^T \|G_t\|_2^2$$

若参数  $\beta \in \{\beta : \|\beta\|_2 \leq B\}$ ，损失函数  $l$  满足  $L$ -Lipschitz 条件，即  $\frac{1}{T} \sum_{t=1}^T \|G_t\|_2^2 \leq L^2$ ，则

$$R_T \leq \frac{1}{2\alpha} B^2 + \alpha T L^2$$

当且仅当  $\frac{1}{2\alpha}B^2 = \alpha TL^2$ , 即  $\alpha = \frac{B}{L\sqrt{2T}}$  时间, 上式右边取得最大值。因此带入  $\alpha$ , 则遗憾函数满足

$$R_T \leq BL\sqrt{2T}$$

可以发现:

$$\lim_{T \rightarrow \infty} \frac{dR_T}{dT} = \lim_{T \rightarrow \infty} \frac{BL}{2\sqrt{2T}} = 0$$

因此,  $\beta_t$  收敛, 则 FTRL 为有效算法, 可见其相比于 FTL 算法更加稳定。

### 11.4.3 FTRL-Proximal 算法

FTRL-Proximal 算法是一种用于点击率预估的在线机器学习系统的核心算法, 其可以看做 RDA (Regularized Dual Averaging Algorithm) 算法与 FOBOS (Forward Backward Splitting) 算法的结合体, 在 FTL 算法的基础上通过增加  $L_1$ 、 $L_2$  正则项的方式来兼顾算法的稀疏性与精确度。

为了进一步提高算法的稀疏性, FTRL-Proximal 算法在 FTRL 算法的基础上添加  $L_1$  正则项。FTRL-Proximal 算法使用如下损失函数更新参数:

$$\beta_{t+1} = \operatorname{argmin} \left( \sum_{j=1}^t G_j^T \beta + \frac{1}{2} \sum_{j=1}^t \sigma_j \|\beta - \beta_j\|_2^2 + \lambda \|\beta\|_1 \right)$$

其中  $\sigma_j = \frac{1}{\alpha_j} - \frac{1}{\alpha_{j-1}}$ ,  $\alpha_t$  是  $t$  时刻的学习率,  $\lambda$  是  $L_1$  正则化调谐参数,  $\beta$  为感兴趣参数。值得注意的是,  $\sum_{j=1}^t G_j^T \beta$  保证向正确的方向更新, 而使用历史累计梯度保证不会过早地将重要特征的参数约束为 0。 $\frac{1}{2} \sum_{j=1}^t \sigma_j \|\beta - \beta_j\|_2^2$  要求新产生的参数不要偏离历史参数太远, 即参数更新不要太激进,  $\lambda \|\beta\|_1$  保证解的稀疏性。

在参数更新过程中, 我们将特征参数的各个维度拆解成独立的标量最小化问题, 从而简化算法, 具体计算过程如下:

令  $Z_t = \sum_{j=1}^t G_j - \sum_{j=1}^t \sigma_j \beta_j$ ,  $\sigma_t = \frac{1}{\alpha_t} - \frac{1}{\alpha_{t-1}}$ , 我们不妨构造函数  $F(\beta)$  为:

$$F(\beta) = Z_t^T \beta + \frac{1}{2\alpha_t} \|\beta\|_2^2 + \lambda \|\beta\|_1 + (\text{const})$$

即:  $\beta_{t+1} = \operatorname{argmin}_{\beta} F(\beta)$ 。下面求解使得  $F(\beta)$  最小的  $\beta$ 。由于  $\|\beta\|_1$  在零点处不可导, 我们考虑使用次梯度方法。

$$\partial_{\beta} F(\beta) = Z_t + \frac{\beta}{\alpha_t} + \lambda \partial_{\beta} \|\beta\|_1$$

令其次梯度为 0，并根据条件限制得到各维度参数更新公式如下：

$$\beta_{t+1,k} = \begin{cases} (\lambda - Z_{t,k})\alpha_t, & Z_{t,k} > \lambda \\ 0, & |Z_{t,k}| \leq \lambda \\ (-\lambda - Z_{t,k})\alpha_t, & Z_{t,k} < -\lambda \end{cases} \quad (11.4.1)$$

另外，FTRL-Proximal 算法考虑了数据在不同维度上的特征分布的不均匀性，建议每个维度采用的学习率  $\alpha_t$  是不一样的并改为如下形式的  $\alpha_{t,k}$ ：

$$\alpha_{t,k} = \frac{\omega}{\gamma + \sqrt{\sum_{j=1}^t G_{j,k}^2}}$$

其中， $\alpha_{t,k}$  为  $t$  时刻  $k$  维度的学习率， $\omega$ ， $\gamma$  为超参数， $G_{t,k}$  为  $t$  时刻  $k$  维度的损失函数的梯度。

FTRL-Proximal 算法流程如下：

表 11.6 FTRL-Proximal 算法

---

#### 算法 5：FTRL-Proximal 算法

---

**输入：**超参数  $\omega$ ， $\gamma$ ，正则化参数  $\lambda$ ，总时间  $T$ ，以及初始化参数  $Z_0$ ， $\alpha_0$ ， $\beta_1$

**输出：**  $\beta_T$

for  $t = 1$  to  $T$  do

    计算损失梯度向量  $\mathbf{G}_t = \mathbf{G}_t(\beta_t)$

    for  $k = 1$  to  $p$  do

        计算每一维度的学习速率  $\alpha_{t,k} = \frac{\omega}{\gamma + \sqrt{\sum_{j=1}^t G_{j,k}^2}}$

    end for

    计算对角阵  $\sigma_t = \text{diag}\left\{\frac{1}{\alpha_t}, \dots, \frac{1}{\alpha_{t-1}}\right\}$

    计算向量  $Z_t = Z_{t-1} + \mathbf{G}_t - \sigma_t \beta_t$

    使用 (11.4.1) 更新参数下一步迭代值  $\beta_{t+1}$

end for

---

## 11.5 在线学习实践

### 11.5.1 R 语言实践

#### 在线均值模型

下面介绍如何使用 R 语言实现均值模型的在线学习算法。首先我们要随机生成数组  $X$ ，并计算其初始样本均值。

```
set.seed(12345) # 设定随机数种子
N <- 100
X <- rnorm(N) # 随机生成生态分布后抽取 N 个值
X.mean <- mean(X) # 计算初始样本均值
```

如果查看初始的样本均值，可输入如下命令：

```
X.mean
[1] 0.2451972
```

其次，在循环下，我们可以构造得到新的数据，随之更新样本数量和样本总和，从而计算新的样本均值。

```
b <- 100
X.sum <- N*X.mean # 计算初始样本总和
for(j in 2:b){
  x <- rnorm(j*b) # 随机产生 j*b 个新样本
  N <- N + length(x) # 更新样本数量
  X.sum <- X.sum + sum(x) # 更新样本总和
  X.mean <- X.sum/N # 更新样本均值
}
```

如果查看最终的样本均值，可输入如下命令：

```
X.mean
[1] -6.21109e-05
```

可以从  $\bar{X}$  的结果对比中看出：经过数据的更新，样本均值从 0.2451972 大幅减少为-6.21109e-05。

#### 在线线性模型

下面介绍如何使用 R 语言实现线性模型在线学习的算法功能。首先我们要随机生成设计矩阵  $X$  和观测向量  $Y$ 。

```
N <- 2000
p <- 1
X <- matrix(rnorm(N*p), ncol=p) # 生成设计矩阵 X
beta <- c(-1, 1)
eps <- rnorm(N) # 随机生成扰动
y <- cbind(1, X) %*% beta + eps # 生成观测向量 Y
```

其次，初始化参数  $\beta$  和学习速率  $\alpha$ ， $\beta_{\text{final}}$  可在循环中与  $\beta_{\text{initial}}$  交换赋值实现迭代。

```
beta.initial <- rep(0, p+1) # 初始化 t 阶段参数 beta
alpha <- 0.5 # 初始化学习速率
beta.final <- matrix(0, nrow=N, ncol=p+1) # 初始化储存参数更新值矩阵
```

接着，利用循环不断根据新数据计算梯度、更新参数  $\beta$  的估计。另外，在优化的过程中还需对学习速率  $\alpha$  不断进行更新。

```
for(j in 1:N){
  grad <- -c(1, X[j,])* (y[j] - as.numeric(c(1, X[j,]) %*% beta.initial)) # 计算梯度
  alpha <- 0.5*j^(-1/2) # 更新学习速率
  beta.final[j,] <- beta.initial - alpha * grad # 更新参数
  beta.initial <- beta.final[j,] # 交换参数
}
```

最终，通过可视化方法观察参数迭代收敛过程。

```
# 可视化参数迭代过程
png("R-linear model.png", units="in", width=8, height=6, res=2000)
x <- c(1:N)
plot(x, beta.final[,1], type='p', col='blue', cex=0.1, pch=20, ylim=c(-1.5, 1.5),
      xlab = 'epoch', ylab='parameters')
abline(h=-1, lty=2, col='blue')
par(new = TRUE)
plot(x, beta.final[,2], type='p', col='red', cex=0.1, pch=20, axes = FALSE,
      xlab='', ylab='')
abline(h=1, lty=2, col='red')
dev.off()
```

图11.1表示迭代过程中参数值的变化。我们可以从图像中观察到参数逐渐收敛。

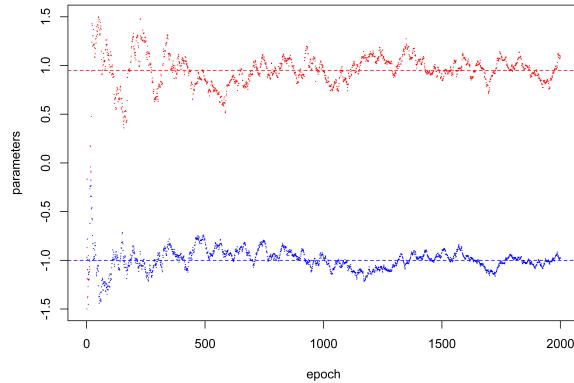


图 11.1 在在线性模型参数迭代值的变化

## 在线岭回归

下面介绍如何使用 R 语言实现岭回归模型的在线学习算法。

首先，我们定义一个读取数据的方程，需要分别提取文件中的自变量数据和因变量数据。另外，需要注意的是，我们在函数中将数据进行标准化处理，旨在提高运行速度。

```
## 用于读取数据的函数
read_data <- function(fname, sc){
  data <- read.csv(file = fname, head = TRUE, sep = ",")
  nr = dim(data)[1] # 提取数据行数
  nc = dim(data)[2] # 提取数据列数
  x = data[1:nr, 1:(nc-1)] # 提取自变量数据
  y = data[1:nr, nc] # 提取因变量数据
  if (isTRUE(sc)){
    x = scale(x) # 将数据进行标准化，提高运算速度
    y = scale(y)
  }
  return (list("x" = x, "y" = y))
}
```

其次，我们定义一个用于计算损失函数值与参数迭代矩阵的函数。

```
## 用于计算损失函数值与参数迭代矩阵的函数
sgd_train <- function(train_x, train_y, lambda, alpha, epsilon, max_epoch){
  Phi <- as.matrix(cbind('X0'= 1, train.data)) # 构造设计矩阵
  ## 计算 SGD 的最大迭代次数
  train_len = dim(train_x)[1]
  tau_max = max_epoch * train_len
  ## 初始化参数矩阵
  beta <- matrix(, nrow=tau_max, ncol=ncol(Phi)) # 初始化参数迭代矩阵
  obj_func_val <- matrix(, nrow=tau_max, ncol=1) # 初始化损失函数值的矩阵
  ## 第一次迭代
  tau = 1 # tau 为迭代次数
  set.seed(12345) # 设定种子
  ## 生成第一次迭代参数
  beta[1,] <- runif(ncol(Phi))
  ## 计算第一次迭代损失函数值
  obj_func_val[tau,1] = .5 * (train_y[1] - t(beta[tau,]) %*% Phi[1,])^2
  + .5 * lambda * t(beta[tau,]) %*% beta[tau,]
  ## 循环迭代开始
  while (tau <= tau_max){
    if (obj_func_val[tau,1] <= epsilon) {break} # 检查终止条件
    ## sample 函数用于重排数据
    train_index <- sample(1:train_len, train_len, replace = FALSE)
    ## 循环遍历每个数据点
    for (i in train_index) {
      tau <- tau + 1
      ## 检查终止条件
      if (tau > tau_max) {break}
      y_pred <- t(Phi[i,]) %*% beta[tau-1,] # 计算预测值 y_pred
      grad <- -(train_y[i] - y_pred) * Phi[i,] # 计算梯度
      beta[tau,] <- beta[tau-1,] * (1-alpha * lambda) - alpha * grad # 更新参数
      obj_func_val[tau,1] = .5 * (train_y[i] - t(beta[tau,]) %*% Phi[i,])^2
      + .5 * lambda * t(beta[tau,]) %*% beta[tau,] # 计算该阶段的损失函数值
    }
  }
}
```

```

    }
}

return(list('vals' = obj_func_val, 'beta' = beta))
}

```

接下来，我们将开始进行试验。首先我们需要加载 R 语言中相关库。

```

## 加载库
library(ggplot2)
library(reshape2)
library(mvtnorm)

```

其次，我们要添加上述定义函数的路径并用 source 函数将上述两个函数加载到主函数中。

```

## 添加路径
path="C:/Users/Administrator/Desktop"
setwd(path)
## source 函数执行 R 脚本
source('read_data.R')
source('sgd_train.R')

```

接着，我们应运用 read\_data 函数加载训练集数据及训练标签。training datasets 数据集由  $n \times p$  维自变量  $x$  与  $n \times 1$  维因变量  $y$  组成。

```

## 加载训练集数据和训练标签
train.data <- read_data('training_datasets.csv',TRUE)$x
train.label <- read_data('training_datasets.csv',TRUE)$y

```

此外，我们还需对最大循环次数、终止阈值、学习效率、正则化参数进行设定。

```

## 设置最大的循环次数
max_epoch = 15
## 设定相关参数
epsilon = .001#终止阈值
alpha = .04#学习效率
lambda= 0.3#正则化参数

```

接着，我们可以用 sgd\_train 函数计算迭代过程中损失函数值以及参数  $\beta$  的迭代矩阵。

```

## 函数返回损失函数值以及参数迭代矩阵
train_res = sgd_train(train.data, train.label, lambda, alpha, epsilon, max_epoch)

```

最后，我们画出迭代过程中参数迭代值的变化来评估模型的效果。

```

## 可可视化迭代过程中参数迭代值的变化
plot(train_res$beta[,1], type='p', col='blue',cex=0.1,pch=20, ylim=c(-2,2),
xlab ='epoch', ylab='parameters')
abline(h=1.5,lty=2, col='red')
par(new = TRUE)
plot(train_res$beta[,2], type='p', col='red',cex=0.1,pch=20,ylim=c(-2,2),
axes = FALSE,xlab='',ylab='')

```

```
abline(h=-1.5,lty=2, col='blue')
```

图11.2表示迭代过程中参数迭代值的变化。我们可以从图像中观察到参数收敛速度很快。

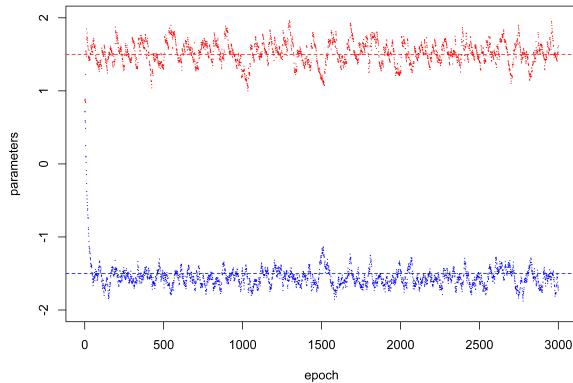


图 11.2 在线岭回归模型参数迭代值的变化

## 在线高维模型

下面介绍如何使用 R 语言实现高维模型的在线学习算法。首先，我们应生成符合逻辑回归的数据集，并进行数据读取。

```
## 生成逻辑回归数据集
set.seed(12345) #设定随机数种子
N <- 8000
p <- 1
beta <- c(-1,1)
eps <- rnorm(N)
X <- as.matrix(cbind('X0'= 1, matrix(rnorm(N*p), ncol=p)))#构造设计矩阵
y <- X %*% beta + eps#构造观测向量
logistic_y <- 1/(1+exp(-y))#逻辑回归
l <- length(logistic_y)
for(i in 1:l){
  if(logistic_y[i]<=0.5)#按0.5划分二分类
    logistic_y[i]=0
  else
    logistic_y[i]=1
}
## 读取数据
train.data <- X
train.label <- logistic_y
T <- length(train.label)#计算样本大小
N <- dim(train.data)[2]#计算参数向量的维度
```

其次，初始化超参数  $\alpha$ 、 $\beta$ ，正则化参数  $\lambda$ ，并初始化所需的各参数矩阵，用于存储迭代过程中生成的值。

```
## 初始化参数
alpha=0.1; beta=1#超参数
lambda=0.01#正则化参数
W <- matrix(0, nrow=T, ncol=N)#初始化参数迭代矩阵
Z <- matrix(0, nrow=T, ncol=N)#初始化中间矩阵
eta <- matrix(0, nrow=T, ncol=N)#初始化学习率矩阵
g <- matrix(0, nrow=T, ncol=N)#初始化损失梯度矩阵
sigma <- matrix(0, nrow=T, ncol=N)#初始化学习速率矩阵
```

最后，通过 for 循环对每个阶段的参数进行更新迭代。

```
for(t in 1:T){
  #利用for循环对每个维度更新参数W
  if(t != 1){
    for(i in 1:N){
      if(Z[t-1,i] > lambda){
        W[t,i] <- (lambda-Z[t-1,i])*eta[t-1,i]
      }else if(Z[t-1,i] < -lambda){
        W[t,i] <- (-lambda-Z[t-1,i])*eta[t-1,i]
      }else if(abs(Z[t-1,i]) <= lambda){
        W[t,i] = 0
      }
    }
  }
  #计算预测值pt
  tmp <- train.data[t,]*W[t,]
  pt <- 1/(1+exp(-tmp))#运用sigmoid函数计算
  #利用for循环对每个维度进行更新各参数值
  for(r in 1:N){
    g[t,r] <- (pt-train.label[t])*train.data[t,r]#计算损失梯度
    #计算g_{1:t}
    sum=0
    for(s in 1:t)
    {
      sum=sum+g[s,r]*g[s,r]
    }
    eta[t,r]=alpha/(beta+sqrt(sum))#计算学习率矩阵
    if(t != 1){
      sigma[t,r] <- 1/eta[t,r] - 1/eta[t-1,r]#计算学习速率
      Z[t,r]<- Z[t-1,r]+g[t,r]-sigma[t,r]*W[t,r]
    }
  }
}
```

最终，我们通过可视化方法观察参数迭代收敛过程。

```
png("FTRL.png",units="in", width=8, height=6,res=2000)
x <- c(1:T)
plot(x, W[,1], type='p', col='blue',cex=0.1,pch=20, ylim=c(-2,2),
xlab ='epoch', ylab='parameters')
```

```

abline(h=1.8,lty=2, col='red')
par(new = TRUE)
plot(x, W[,2], type='p', col='red',cex=0.1,pch=20,ylim=c(-2,2),
axes = FALSE,xlab='',ylab='')
abline(h=-1.8,lty=2, col='blue')
dev.off()

```

图11.3表示迭代过程中参数迭代值的变化。我们可以从图像中观察到，参数逐渐收敛。

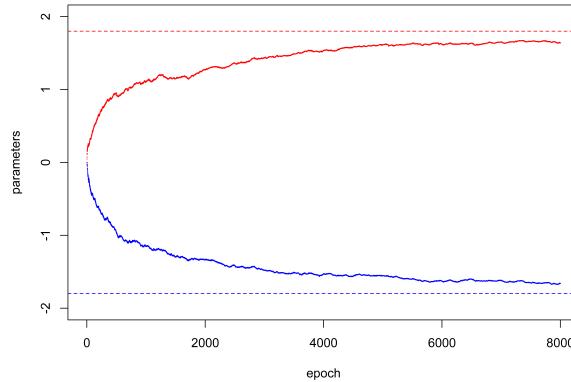


图 11.3 在线高维模型参数迭代值的变化

## 11.5.2 Python 语言实践

### 在线均值模型

下面介绍如何使用 Python 语言实现均值模型的在线学习算法。首先导入实验过程中需要的库并设定随机数种子。

```

#加载库
import numpy as np
import random
#设定随机数种子
random.seed(12345)
np.random.seed(12345)

```

其次，构造初始数据集并计算初始样本均值。

```

#构造初始数据集
X = np.random.randn(100) #生成多个服从标准正态分布的随机数组
X_length = len(X) #计算初始样本量
X_mean = sum(X) / len(X) #计算初始样本均值

```

```
X_sum = sum(X)          #计算初始样本总和
b = 100     #设置迭代次数
print("初始样本均值为: %10.10f" %X_mean)
```

其次，在循环下，我们可以构造得到新的数据，随之更新样本数量和样本总和，从而计算新的样本均值。

```
#利用for循环更新样本均值
for i in range(2,100):
    x = np.random.randn(i * b)      #随机生成新样本
    X_length = X_length + len(x)   #更新样本容量
    X_sum = X_sum + sum(x)         #更新样本总和
    X_mean = X_sum / X_length     #更新样本均值
print("更新后样本均值为: %10.10f" %X_mean)
```

可以从  $\bar{X}$  的结果对比中看出，经过数据的更新，样本均值从 0.0336143882 大幅减少为 0.0006315946。

## 在线线性模型

下面介绍如何使用 Python 语言实现线性模型的在线学习算法。首先导入实验过程中需要的库并设定随机数种子。

```
#加载库
import numpy as np
import random
import matplotlib.pyplot as plt
#设定随机数种子
random.seed(12345)
np.random.seed(12345)
```

其次，构造数据集，随机生成设计矩阵  $X$  和观测向量  $Y$ 。

```
#构造数据集
N = 2000    #设置样本容量
p = 2        #设置参数个数
X1 = np.random.randn(N * (p-1), p-1)
X = np.hstack((np.ones((N, 1)), X1))      #生成设计矩阵
beta = np.array([1, -1])      #设置参数
eps = np.random.randn(N)       #随机生成扰动
y = np.dot(X, beta) + eps    #生成观测向量
```

接着，初始化参数  $\beta$  和学习速率  $\alpha$ ， $\beta_{\text{final}}$  可在循环中与  $\beta_{\text{initial}}$  交换赋值实现迭代。

```
#初始化参数
beta_initial = np.zeros((p, 1))      #初始化参数beta
alpha = 0.5                          #初始化学习速率
beta_final = np.zeros((N-1, p))       #初始化储存参数更新矩阵
```

接着，利用循环不断根据新数据计算梯度、更新参数  $\beta$  的估计。另外，在优化的过程中还需对学习速率  $\alpha$  不断进行更新。

```
#利用for循环更新模型参数
for i in range(0,N-1):
    grad = -1 * X[i, :] * (y[i] - np.dot(X[i, :], beta_initial)) #计算梯度
    alpha = 0.5 * pow((i+1), (-1/2)) #更新学习速率
    beta_final[i, :] = beta_initial.T - alpha * grad #更新参数
    beta_initial = beta_final[i, :] #交换参数
```

最终，通过可视化方法观察参数迭代收敛过程。

```
#可视化参数迭代过程
fig = plt.figure()
plt.scatter(range(1, N), beta_final[:,0], color='r', s=0.7, marker='.')
plt.scatter(range(1, N), beta_final[:,1], color='b', s=0.7, marker='.')
plt.axhline(y=1, color='r', linestyle='--')
plt.axhline(y=-1, color='b', linestyle='--')
plt.xlabel('epoch')
plt.ylabel('parameters')
plt.show()
```

图11.4表示迭代过程中参数迭代值的变化。我们可以从图像中观察到，参数逐渐收敛。

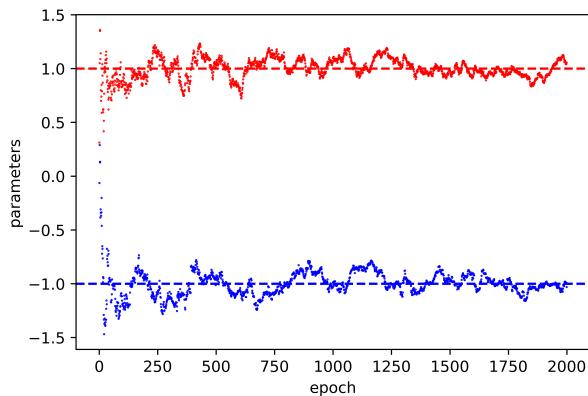


图 11.4 在在线性模型参数迭代值的变化

## 在线岭回归

下面介绍如何使用 Python 语言实现岭回归模型的在线学习算法。首先导入实验过程中需要的库并设定随机数种子。

```
#加载库
import numpy as np
import random
import matplotlib.pyplot as plt
```

```
from random import sample
#设定随机数种子
random.seed(12345)
np.random.seed(12345)
```

其次，我们定义一个用于计算损失函数值与参数迭代矩阵的函数。

```
# 岭回归标准方程法求解回归参数
def sgd_train(train_data, train_label, lambda1, eta, epsilon, max_epoch):
    #计算最大迭代次数
    train_len = train_data.shape[0]
    tau_max = max_epoch * train_len
    #初始化参数迭代矩阵
    beta = np.zeros((tau_max, train_data.shape[1]))
    value = np.zeros((tau_max, 1))
    #第一次迭代
    #迭代参数&损失函数值
    beta[0, :] = np.zeros((train_data.shape[1]))
    value[0] = 0.5 * pow((train_label[0] - np.dot(beta[0, :].T, train_data[0, :])), 2) + 0.5 * lambda1 * np.dot(beta[0, :].T, beta[0, :])
    tau = 1
    while tau <= tau_max:
        if(value[tau-1] <= epsilon):#检查终止条件
            break
        else:
            train_index = sample(range(0, train_len), train_len)
            #循环遍历每个数据点
            for i in train_index:
                tau = tau + 1
                if(tau > tau_max):#检查终止条件
                    break
                else:
                    t = tau - 1#矩阵从0开始计数
                    y_pred = np.dot(train_data[i, :].T, beta[t-1, :])#计算预测值
                    #计算梯度
                    grad = (-1) * np.dot((train_label[i] - y_pred), train_data[i, :])
                    beta[t, :] = beta[t-1, :] * (1-eta * lambda1) - eta * grad#更新参数
                    #计算损失函数值
                    value[t] = 0.5 * pow((train_label[i] - np.dot(beta[t, :].T, train_data[i, :])), 2) + 0.5 * lambda1 * np.dot(beta[t, :].T, beta[t, :])
    return value, beta
```

接着，我们构造初始数据集并对实验进行参数设置。

```
#构造数据集
N = 500#设置样本容量
p = 2#设置参数个数
X1 = np.random.randn(N * (p-1), p-1)
X = np.hstack((np.ones((N, 1)), X1))#生成设计矩阵
beta = np.array([1, -1])#设置参数
eps = np.random.randn(N)#随机生成扰动
y = np.dot(X, beta) + eps#生成观测向量
```

```
#参数设置
max_epoch = 30#最大循环次数
epsilon = 0.001#设置终止阈值
eta = 0.0005#学习效率
lambda1 = 0.05#正则化参数
```

最终，我们设置训练集数据及训练标签进行实验并对参数迭代矩阵进行可视化。

```
#设置训练集数据和训练标签
train_data = X
train_label = y
#计算参数更新过程中损失函数值以及参数迭代值
value, beta = sgd_train(train_data, train_label, lambda1, eta, epsilon, max_epoch)
#可视化迭代过程中损失函数值的变化与迭代矩阵的值
fig = plt.figure()
plt.scatter(range(1,len(beta)+1), beta[:, 0], color='r', s=0.3, marker='.')
plt.scatter(range(1,len(beta)+1), beta[:, 1], color='b', s=0.3, marker='.')
plt.axhline(y=1, color='r', linestyle='--')
plt.axhline(y=-1, color='b', linestyle='--')
plt.xlabel('epoch')
plt.ylabel('parameters')
plt.show()
```

图11.5表示迭代过程中参数迭代值的变化。我们可以从图像中观察到，参数逐渐收敛。

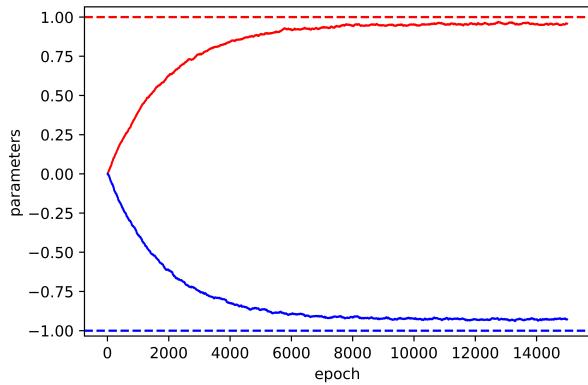


图 11.5 在线岭回归模型参数迭代值的变化

## 在线高维模型

下面介绍如何使用 Python 语言实现高维模型的在线学习算法。首先导入实验过程中需要的库并设定随机数种子

```
#加载库
import numpy as np
import random
import matplotlib.pyplot as plt
```

```
import math
#设定随机数种子
random.seed(12345)
np.random.seed(12345)
```

其次，我们生成符合逻辑回归的数据集，并进行数据读取。

```
#构造数据集
N = 10000#设置样本容量
p = 2#设置参数个数
X1 = np.random.randn(N * (p-1), p-1)
X = np.hstack((np.ones((N,1)), X1))#生成设计矩阵
beta = np.array([1, -1])#设置参数
eps = np.random.randn(N)#随机生成扰动
y = np.dot(X, beta) + eps#生成观测向量
#逻辑回归
logistic_y = 1/(1 + np.exp(-y))
l = len(logistic_y)
for i in range(0, l):
    if(logistic_y[i] <= 0.5):#以0.5进行二分类
        logistic_y[i] = 0
    else:
        logistic_y[i] = 1
#设置训练集数据和训练标签
train_data = X
train_label = logistic_y
```

接着，我们进行参数的设置。

```
#设置参数
omega = 0.1#超参数
gamma = 1#超参数
lambda1 = 0.01#正则化参数
T = len(train_label)#计算样本大小
N = train_data.shape[1]#计算参数向量的维度
beta_final = np.zeros((T, N))#初始化参数迭代矩阵
alpha = np.zeros((T, N))#初始化学习率矩阵
g = np.zeros((T, N))#初始化损失梯度矩阵
sigma = np.zeros((T, N))#初始化中间矩阵
Z = np.zeros((T, N))#初始化中间矩阵
```

再利用循环不断更新各维度参数值。

```
for t in range(0, T):
    #利用for循环对每个维度更新参数W
    if(t != 0):
        for i in range(0, N):
            if(Z[t-1,i] > lambda1):
                beta_final[t, i] = (lambda1 - Z[t-1, i]) * alpha[t-1, i]
            elif(Z[t-1,i] < -lambda1):
                beta_final[t, i] = (-lambda1 - Z[t-1, i]) * alpha[t-1, i]
            else:
                beta_final[t, i] = 0
```

```
#计算预测值
tmp = np.dot(train_data[t, :], beta_final[t, :].reshape(-1, 1))
pt = 1 / (1 + np.exp(-tmp))
#利用for循环对每个维度更新各参数的值
for r in range(0, N):
    g[t, r] = (pt - train_label[t]) * train_data[t, r]
    sum = 0
    for s in range(0, t):
        sum = sum + g[s, r] * g[s, r]
    alpha[t, r] = omega / (gamma + math.sqrt(sum))#更新学习率
    if(t != 0):
        sigma[t, r] = 1/alpha[t, r] - 1/alpha[t-1, r]
    Z[t, r] = Z[t-1, r] + g[t, r] - sigma[t, r] * beta_final[t, r]
```

最终，我们借助可视化的方法观察参数迭代过程。

```
#可视化参数迭代矩阵
fig = plt.figure()
plt.scatter(range(1, l+1), beta_final[:, 0], color='r', s=0.7, marker='.')
plt.scatter(range(1,l+1), beta_final[:,1], color='b', s=0.7, marker='.')
plt.axhline(y=1.8, color='r', linestyle='--')
plt.axhline(y=-1.8, color='b', linestyle='--')
plt.xlabel('epoch')
plt.ylabel('parameters')
plt.show()
```

图11.6表示迭代过程中参数迭代值的变化。我们可以从图像中观察到，参数逐渐收敛。

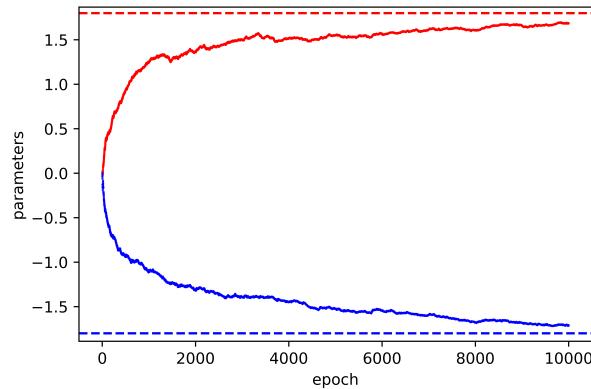


图 11.6 在线高维模型参数迭代值的变化

## 11.6 练习

1. 假设观测到一维数据  $\{X_{ti} : i = 1, \dots, n_t, t = 1, \dots\}$ ,  $X_{ti}$  代表第  $t$  批次数据流第  $i$  个样本,  $n_t$  是第  $t$  批次的样本量, 并且  $\text{var}(X_{ti}) = \sigma^2$ 。根据在线学习样本均值估计量的形式, 请写出在线方差估计量  $\hat{\sigma}^2$  形式以及具体更新算法。
2. 假设观测到回归数据  $\{(X_{ti}^T, Y_{ti}) : i = 1, \dots, n_t, t = 1, \dots\}$ , 满足模型

$$Y_{ti} = X_{ti}^T \beta + \epsilon_{ti},$$

根据在线学习样本均值估计量的形式 (不用 SGD), 请写出在线学习最小二乘估计量  $\hat{\beta}$  形式以及具体更新算法。

3. 根据 FTRL-Proximal, 写出基于 Lasso 惩罚函数的在线最小二乘估计过程以及程序实现。

4. 本练习中的数据是来自于摩洛哥北部泰图安市的 3 个区域的用电量以及影响到用电量的 6 个特征: 时间, 温度, 湿度, 风速, 一般漫反射流, 漫反射流。数据取自于监控和数据采集系统 (SCADA) 的 2017 年的全年数据, 数据每 10 分钟采集一次。为公用事业公司提供智能策略并帮助其改进重要任务, 采用统计模型对用电量进行分析。

- (1). 基于均值模型以及在线学习更新算法, 分别估计 3 个区域的用电量。
- (2). 建立线性模型分析区域 1 的用电量, 分别使用基于岭回归模型的在线学习算法和离线学习算法估计参数, 并比较估计结果。
- (3). 建立线性模型分析区域 1 的用电量, 分别使用基于 11.2.4 节高维在线学习算法和离线学习算法估计参数, 并比较估计结果。

# 第十二章 保形预测

## 12.1 简介

在进行预测的时候，也许我们会思考这样的问题：我们的预测有多好？如果我们对一个新对象的类别进行预测，我们有多少信心认为我们的预测结果  $\hat{Y}$  确实等于这个未知的真实的类别  $Y$  呢？如果类别是一个数字，我们得到的  $\hat{Y}$  与真实的类别  $Y$  有多接近？在机器学习中，这些问题通常以过去的经验相当粗略地回答。

因此，在使用机器学习进行预测时，需要注意预测值也是随机变量，存在不确定性。例如，使用一个股票自动交易系统中的机器学习方法预测股票价格。由于股票市场的高度不确定性，机器学习的点预测可能与实际值有很大不同。人工智能系统如果以高概率估计出覆盖目标真实值的范围，交易系统就可以计算出最好和最差的情况，并做出更明智的决定。所以，对实际值预测一个可信范围会更有意义。

**保形预测** (Conformal Prediction) 是目前机器学习中热门的且非常灵活的预测技术。即当我们在处理回归或分类问题时，给定输入，保形预测可以输出一个预测区间或者一个预测集。

关于保形预测的理解，跟传统的区间估计是类似的。给定一种回归问题的预测方法，保形预测产生一个 95% 的预测区间  $\Gamma^{0.05}$ ，即该区间包含  $Y$  的概率至少为 95%，通常  $\Gamma^{0.05}$  也包含预测值  $\hat{Y}$ 。我们称  $\hat{Y}$  为点预测，称  $\Gamma^{0.05}$  为区间预测。在分类的情况下，类别  $Y$  有有限个可能值， $\Gamma^{0.05}$  可能由这些值中的几个值组成，或者在理想的情况下  $\Gamma^{0.05}$  只是这些值中的一个值。

与传统区间估计所使用的方法不同的是，保形预测并没有研究预测值的分布或渐进分布，这也体现了保形预测的优势，可以适用于任何模型，因为研究渐进分布往往需要模型的假定。保形预测在处理连续因变量的回归预测问题时，对于单因变量的预测输出的是一个预测区间，对于多因变量的预测输出的是一个预测域。图12.1展示的是对单个连续因变量的区间预测。

保形预测在处理离散因变量的分类预测问题时，输出的是预测集。集合中的元素是由预测类别及对应的概率构成。图12.2 中我们展示了三种不同的输入值下，对松鼠类型预测的保形预测集。上述两图表明，预测区间和预测集都可以保证以较高的概率覆盖真实值。

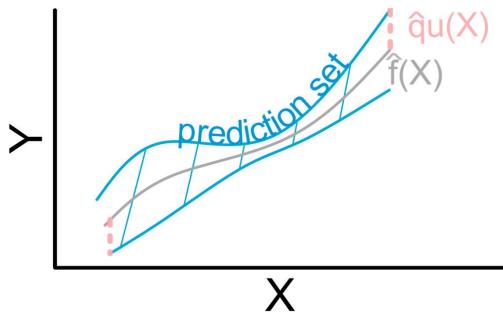


图 12.1 保形预测的预测区间



图 12.2 保形预测的预测集

### 12.1.1 简单运用

在介绍保形预测的流程之前，我们要先知道两个定义：一个是随机变量序列的可交换性，另一个是不符合度量。在后面章节中，保形预测的预测集既可代表回归问题的预测区间也可表示分类问题的预测集。

#### 可交换性 (Exchangeability)

一个有限的随机变量序列是可交换的 (exchangeable)，是指随机变量的联合概率分布对随机变量的排列不变。

$$\Pr(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n) = \Pr(\mathbf{X}_{\pi(1)}, \mathbf{X}_{\pi(2)}, \dots, \mathbf{X}_{\pi(n)})$$

这里  $\pi(1), \pi(2), \dots, \pi(n)$  代表自然数  $1, 2, \dots, n$  的任意一个排列。一个无限的随机变量序列是无限可交换 (infinitely exchangeable) 的，是指它的任意一个有限子序列都是可交换的。

如果一个无限随机变量序列  $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n, \dots$  是独立同分布的，那么它们是无限可交换的。反之不然。

## 不符合度量 (Nonconformity measure)

保形预测使用的不符合度量，度量了一个新例子与旧例子的不同程度，也可以说是度量预测值与真实值之间的不符合程度。因此，我们需要寻找一种方法来度量预测值与真实值之间的距离。

保形预测要求我们首先选择一个不符合度量，而度量预测值和真实值之间距离的方法与数据类型和预测器的选择相关：在回归问题中，最常用的不符合度量是残差的绝对值： $R = |Y - \hat{Y}|$ 。在分类问题中，最常用的不符合度量是： $H = 1 - \hat{Y}$ 。

给定一个不符合度量，保形算法对每一个误覆盖水平  $\alpha$  产生一个预测集  $\hat{C}_n$ 。 $\hat{C}_n$  为  $(1 - \alpha)$  预测集；它有至少  $(1 - \alpha)$  的概率包含真实值  $Y$ 。

不同  $\alpha$  得到的预测区间是嵌套的。因此，当  $\alpha_1 \geq \alpha_2$  时，置信度  $(1 - \alpha_1)$  低于  $(1 - \alpha_2)$ ，我们有  $\hat{C}_{n,\alpha_1} \subseteq \hat{C}_{n,\alpha_2}$ 。

在明确了以上两个概念之后，我们介绍运用保形预测来构造有效的预测集的几个基本步骤。假设我们有一个独立同分布的数据集：

$$\{(\mathbf{X}_1^T, Y_1), \dots, (\mathbf{X}_n^T, Y_n)\}$$

其中， $\mathbf{X}$  是输入的特征， $Y$  是类别， $n$  是数据点的数量。我们假设一个机器学习模型： $f: \mathbf{X} \rightarrow Y$  已经被训练。这个模型可以是一个经典的机器学习模型例如线性回归、支持向量机，或者深度学习技术例如全连接或卷积网络。目标是去估计模型输出的预测集。下面我们将描述保形预测的步骤：

### 保形预测的基本流程

#### a. 选择合适的不符合度量，计算不符合分数

选择一个适当的不符合度量  $S(\mathbf{X}, Y) \in \mathcal{R}$  来测量模型输出  $\hat{Y}$  与类别  $Y$  之间的差异，亦称评分函数，得到值成为不符合分数。这个不符合度量非常重要，因为它决定了我们能得到什么样的预测集。例如，在回归问题中，我们可以用  $|\hat{Y}_i - Y_i|$  作为评分函数。通过这种方式，所得到的保形预测集在预测值  $\hat{Y}_i$  周围的  $L_1$  范式球内；在分类问题中，我们可以用  $(1 - \hat{Y}_i)$  作为评分函数，其中  $\hat{Y}_i$  是真实类别对应的预测值。

#### b. 计算不符合分数的 $(1 - \alpha)$ 分位数

不符合分数的  $(1 - \alpha)$  分位数  $\hat{q}$  是通过计算  $n$  个不符合分数  $S_1 = S(\mathbf{X}_1, Y_1), \dots, S_n = S(\mathbf{X}_n, Y_n)$  的  $(1 - \alpha)$  分位数得到的。在完全保形预测方法中，我们需要训练  $m$  个模型来计算评分并构造预测集，其中  $m$  为可能的取值的个数，这无疑在计算上非常昂贵。为了降低计算复杂度，可以采用分裂保形预测的方法，分裂保形预测将整个训练集分割成合适的训练集和校准集 (calibration set)，然后，只对训练集进行训练，在校准集上计算不符合分数。

### c. 使用模型预测和不符合分数的 $(1 - \alpha)$ 分位数构造预测集

接下来，我们使用模型预测值和不符合分数的  $(1 - \alpha)$  分位数构造保形预测集。假设新的输入为  $\mathbf{X}_{n+1}$ ，针对分类问题，得到的保形预测集可以表示为： $\tau(\mathbf{X}_{n+1}) = \{Y : S(\mathbf{X}_{n+1}, Y) \leq \hat{q}\}$ 。例如，在使用保形预测进行回归分析时，不符合度量  $S(\mathbf{X}_{n+1}, Y) = |Y - \hat{f}(\mathbf{X}_{n+1})|$ ，其中  $\hat{f}$  是用训练数据集得到预测器。则此时得到的  $1 - \alpha$  预测区间可以表示为： $\hat{C}_n = [\hat{f}(\mathbf{X}_{n+1}) - \hat{q}, \hat{f}(\mathbf{X}_{n+1}) + \hat{q}]$ 。

#### 12.1.2 预测集的有效性

我们的保形预测总是有效的。Glenn Shafer 和 Vladimir Vovk 着眼于可交换性和独立性之间的关系，得出了可交换序列的大数定律，这为我们认为 95% 的预测区域在 95% 的时间内是正确的提供了信心基础。保形预测最重要的新颖之处在于其连续误差是概率独立的。这使得我们能够以一种不同寻常的直接方式来解释“95% 的正确率”。

对于保形预测的有效性的精确诊论实际上需要我们区分两种有效性：保守的和精确的。一般来说，保形预测是保守有效的：当它输出一个  $(1 - \alpha)$  集（也就是说，一组预测集置信水平为  $(1 - \alpha)$  时错误的概率不大于  $\alpha$ ，并且在预测连续的例子时，它所犯的错误之间几乎没有相关性。这意味着，根据大数定律，在置信水平  $(1 - \alpha)$  上的长期错误频率约为  $\alpha$  或更小。在实践中，保守性往往不是很大，尤其是当  $n$  很大时，经验结果显示，长期误差的频率与  $\alpha$  非常接近。然而，从理论的角度来看，为了获得准确的有效性，我们必须在预测过程中引入一个随机误差  $\alpha$ ，其中  $(1 - \alpha)$  集出现错误的概率正好是  $\alpha$ ，错误在不同的试验中是独立产生的，而长期错误的频率收敛到  $\alpha$ 。

保形预测可以提供数学上严格的保证。设  $Y_{n+1}$  为真值。 $Y$  可以是分类问题中的一个类别，也可以是回归问题中的一个实值。设  $\tau(\mathbf{X}_{n+1})$  是一个预测集（或区间）。如果  $Y_{n+1}$  在  $\tau(\mathbf{X}_{n+1})$  内，我们将其定义为  $\tau(\mathbf{X}_{n+1})$  覆盖  $Y_{n+1}$ ，即： $Y_{n+1} \in \tau(\mathbf{X}_{n+1})$ 。然后，给定一组同独立分布样本  $\{(\mathbf{X}_1^T, Y_1)^T, \dots, (\mathbf{X}_n^T, Y_n)^T\}$ ，保形预测集满足以下覆盖保证： $\Pr(Y_{n+1} \in \tau(\mathbf{X}_{n+1})) \geq 1 - \alpha$ 。

基于可交换性 (exchangeability) 假设的覆盖保证的证明可以在文章 A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification 的附录中找到。以上覆盖率保证在有限样本的情况下也是成立的，且与普通的统计推断通常都对变量的分布有更严格的假定（比如 Gaussianity）不同，该保证仅需要非常弱的条件（可交换性）。

### 12.1.3 效率

针对分类问题，保形预测得到的预测集可能包含一个类别，也可能包含多个类别。在处理回归问题时，预测集通常是包含预测值的一个区间。在保证一定置信度的情况下，保形预测集越小，则效率越高，即分类预测集的元素越少或者回归预测区间越短，则越好。这就是我们通常所说的高效率保形预测集。

## 12.2 保形回归

保形预测可以用于回归，也可以用于分类，接下来我们首先以回归为例，解释保形预测的思想和原理。本质上，保形预测理论背后的基本思想与样本分位数有关。

### 12.2.1 保形预测基本思想

假设有独立同分布的随机变量样本  $U_1, \dots, U_n$ （事实上，此处独立同分布的假设可以被更弱的假设—可交换性替代）。对于一个给定的误覆盖水平  $\alpha \in (0, 1)$  和另一个独立同分布的样本  $U_{n+1}$ ，当我们基于  $U_1, \dots, U_n$  定义样本分位数  $\hat{q}_{1-\alpha}$  为：

$$\hat{q}_{1-\alpha} = \begin{cases} U_{(\lceil(n+1)(1-\alpha)\rceil)} & \text{if } \lceil(n+1)(1-\alpha)\rceil \leq n \\ \infty & \text{otherwise} \end{cases},$$

可以得到：

$$\Pr(U_{n+1} \leq \hat{q}_{1-\alpha}) \geq 1 - \alpha,$$

其中， $U_{(1)} \leq \dots \leq U_{(n)}$  表示  $U_1, \dots, U_n$  的次序统计量。通过可交换性， $U_{n+1}$  在  $U_1, \dots, U_n, U_{n+1}$  中的排序是在集合  $\{1, \dots, n+1\}$  上均匀分布的，因此以上有限样本覆盖性质是很容易被验证的。

在回归问题中，我们观测到独立同分布的样本  $\mathbf{Z}_i = (\mathbf{X}_i^T, Y_i)^T \sim \Pr(\mathbf{z})$ ,  $i = 1, \dots, n$ ，我们可以考虑以下方法来构造  $Y_{n+1}$  在新特征值  $\mathbf{X}_{n+1}$  处的预测区间，其中  $(\mathbf{X}_{n+1}^T, Y_{n+1})^T$  是分布  $\Pr(\mathbf{z})$  中一个独立的随机变量。按照上述思想，我们可以构造以下预测区间：

$$\hat{C}_{\text{naive}}(\mathbf{X}_{n+1}) = \left[ \hat{f}(\mathbf{X}_{n+1}) - \hat{F}_n^{-1}(1 - \alpha), \hat{f}(\mathbf{X}_{n+1}) + \hat{F}_n^{-1}(1 - \alpha) \right],$$

其中， $\hat{f}$  是估计的回归函数预测器， $\hat{F}_n$  是拟合残差  $|Y_i - \hat{f}(\mathbf{X}_i)|$ ， $i = 1, \dots, n$  的经验分布， $\hat{F}_n^{-1}(1 - \alpha)$  是  $\hat{F}_n$  的  $(1 - \alpha)$  分位数。假设估计的回归函数  $\hat{f}$  是准确的，则该预测区间在大样本情况下是有效的（即估计的拟合残差分布的  $(1 - \alpha)$  分位数

$\widehat{F}_n^{-1}(1 - \alpha)$  足够接近总体残差  $|Y_i - f(\mathbf{X}_i)|$ ,  $i = 1, \dots, n$  的  $(1 - \alpha)$  分位数)。保证  $\widehat{f}$  的准确性通常需要数据分布  $\text{Pr}(\mathbf{z})$  和回归预测器  $\widehat{f}$  本身均满足一定的条件, 例如适当地选择的预测模型和调优参数。

### 12.2.2 完全保形预测

一般来说, 上述方法得到的预测区间可以粗略地覆盖真实值, 因为拟合残差分布往往是向下倾斜的。保形预测区间 [128] [129] [130] [131] 克服了上述原始方法预测区间的缺陷, 并且, 某种程度上值得注意的是, 保形预测可以保证提供适当的有限样本覆盖, 而无需对  $\text{Pr}(\mathbf{z})$  和回归预测器  $\widehat{f}$  进行任何假设 (除了  $f$  是数据点的对称函数)。

具体算法如下: 对于一个新的输入值  $\mathbf{X}_{n+1}$ , 其对应的  $Y_{n+1}$  是未知的。因此我们为  $Y_{n+1}$  考虑一个实验集合  $\mathbf{Y}_{\text{trial}}$ 。从实验集合中, 选取某一值  $Y \in \mathbf{Y}_{\text{trial}}$ , 我们基于增广数据集  $\mathbf{Z}_1, \dots, \mathbf{Z}_n, (\mathbf{X}_{n+1}^T, Y)^T$  上进行训练, 构造一个增广回归预测器  $\widehat{f}_Y$ 。接下来我们计算不符合分数, 即残差的绝对值。

$$\begin{aligned} R_{Y,i} &= \left| Y_i - \widehat{f}_Y(\mathbf{X}_i) \right|, \quad i = 1, \dots, n \quad \text{和} \\ R_{Y,n+1} &= \left| Y - \widehat{f}_Y(\mathbf{X}_{n+1}) \right|, \end{aligned}$$

并且我们基于  $n + 1$  个不符合分数对  $R_{Y,n+1}$  进行排序, 得到

$$\begin{aligned} \pi(Y) &= \frac{1}{n+1} \sum_{i=1}^{n+1} I\{R_{Y,i} \leq R_{Y,n+1}\} \\ &= \frac{1}{n+1} + \frac{1}{n+1} \sum_{i=1}^n I\{R_{Y,i} \leq R_{Y,n+1}\}, \end{aligned}$$

即  $\pi(Y)$  是增广样本中残差绝对值小于最后一个样本点的残差绝对值  $R_{Y,n+1}$  的样本所占的比例, 其中  $I\{\cdot\}$  表示示性函数。由于数据点的可交换性和  $\widehat{f}_Y$  的对称性, 当我们在估计  $Y_{n+1}$  时, 我们发现构造的次序统计量  $\pi(Y_{n+1})$  在集合  $\{1/(n+1), 2/(n+1), \dots, 1\}$  上是均匀分布的, 这意味着:

$$\text{Pr}\{(n+1)\pi(Y_{n+1}) \leq \lceil(1-\alpha)(n+1)\rceil\} \geq 1 - \alpha.$$

上述推理可以通过如下假设检验过程理解, 即对应原假设  $\mathbf{H}_0 : Y_{n+1} = Y$ 。在给定显著性水平  $\alpha$  下, 我们有如下结论:

- (1) 如果事件  $\{(n+1)\pi(Y) > \lceil(1-\alpha)(n+1)\rceil\}$  成立, 则拒绝原假设;
- (2)  $1 - \pi(Y)$  相当于  $P$  值, 如果  $1 - \pi(Y) < \alpha$ , 则拒绝原假设。

通过在  $\mathbf{Y}_{\text{trial}}$  遍历所有  $Y$  的取值，我们可以得到在  $\mathbf{X}_{n+1}$  处的保形预测区间，即：

$$\hat{C}_{\text{conf}}(\mathbf{X}_{n+1}) = \{Y \in \mathbf{Y}_{\text{trial}} : (n+1)\pi(Y) \leq \lceil(1-\alpha)(n+1)\rceil\}.$$

针对每一个新的自变量，我们都必须重复以上步骤产生一个预测区间。为了完整起见，我们总结为如下**算法 1**。

表 12.1 完全保形预测算法

---

**算法 1：完全保形预测**

---

**输入：** 数据  $(\mathbf{X}_i^T, Y_i)^T, i = 1, \dots, n$ , 显著性水平  $\alpha \in (0, 1)$ , 回归算法  $\mathcal{A}$ , 用于构造预测区间的新的点  $\mathbf{X}_{\text{new}} = \{\mathbf{X}_{n+1}, \mathbf{X}_{n+2}, \dots\}$ , 和作为试验集合  $\mathbf{Y}_{\text{trial}} = \{Y_1, Y_2, \dots\}$

**输出：**  $\mathbf{X}_{\text{new}}$  中每一个元素对应的预测区间

**for**  $x \in \mathbf{X}_{\text{new}}$  **do**

**for**  $Y \in \mathbf{Y}_{\text{trial}}$  **do**

$$\hat{f}_Y = \mathcal{A}(\{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n), (x, Y)\})$$

$$R_{Y,i} = |Y_i - \hat{f}_Y(\mathbf{X}_i)|, \quad i = 1, \dots, n,$$

以及  $R_{Y,n+1} = |Y - \hat{f}_Y(x)|$

$$\pi(Y) = \left(1 + \sum_{i=1}^n I\{R_{Y,i} \leq R_{Y,n+1}\}\right) / (n+1)$$

**end for**

$$\hat{C}_{\text{conf}}(x) = \{Y \in \mathbf{Y}_{\text{trial}} : (n+1)\pi(Y) \leq \lceil(1-\alpha)(n+1)\rceil\}$$

**end for**

**返回**  $\hat{C}_{\text{conf}}(x)$ , for each  $x \in \mathbf{X}_{\text{new}}$

---

上述保形预测得到的预测区间具有有效的有限样本覆盖，并且这个区间是精确的，意味着它没有过多地覆盖，以下我们对此进行说明：

如果  $(\mathbf{X}_i^T, Y_i)^T, i = 1, \dots, n$  为独立同分布的，那么对于一个新的独立同分布的点  $(\mathbf{X}_{n+1}^T, Y_{n+1})^T$ , 保形预测建立的预测区间  $\hat{C}_{\text{conf}}(\mathbf{X}_{n+1})$  有以下覆盖保证：

$$\Pr\left(Y_{n+1} \in \hat{C}_{\text{conf}}(\mathbf{X}_{n+1})\right) \geq 1 - \alpha,$$

如果我们另外假设对所有  $Y \in \mathcal{R}$ , 拟合的绝对残差  $R_{Y,i} = |Y_i - \hat{f}_Y(\mathbf{X}_i)|, i = 1, \dots, n$  有一个连续的联合分布，那么下式也成立：

$$\Pr\left(Y_{n+1} \in \hat{C}_{\text{conf}}(\mathbf{X}_{n+1})\right) \leq 1 - \alpha + \frac{1}{n+1}.$$

该结论的证明详见文章 [132]。

## 12.3 保形方法

### 12.3.1 分裂保形预测

上一节中讲述的完全保形预测方法计算量较大，因为它要遍历所有  $Y$  的实验集。即对于任何  $\mathbf{X}_{n+1}$  和未知的  $Y$ ，为了辨别某一个给定的  $Y$  是否包含在  $\widehat{C}_{\text{conf}}(\mathbf{X}_{n+1})$  中，我们在增广数据集（其中包括新的点  $(\mathbf{X}_{n+1}, Y)$ ）上重新训练模型，并重新计算和排序绝对残差。这个步骤需要在  $\mathbf{Y}_{\text{trial}}$  集合里面的所有元素上，重复进行，计算成本很高。在非线性回归问题中，核密度估计或样条方法去训练模型的较高复杂度都会影响完全保形预测。特别是在高维回归中，我们可能会使用相对复杂的回归预测器，如 Lasso 回归或深度学习等，模型的训练仍会花费大量时间，因此执行有效的完全保形预测仍然是一个有待解决的问题。

幸运的是，有一种完全通用的方法，我们称之为分裂保形预测，其计算成本大大低于完全保形预测方法。分裂保形方法采用样本分割将拟合步骤和排序步骤分离，其计算代价与拟合步骤相对简单。**算法 2** 总结了分裂保形预测算法，其他细节请参考 [133]。在这里，以及以后讨论分裂保形预测时，为了简单起见，我们假设样本容量  $n$  是偶数，当  $n$  是奇数时只需要非常小的改变。

表 12.2 分裂保形预测算法

---

#### 算法 2：分裂保形预测

---

**输入：** 数据  $(\mathbf{X}_i^T, Y_i)^T, i = 1, \dots, n$ , 显著性水平  $\alpha \in (0, 1)$ , 回归算法  $\mathcal{A}$ ,

**输出：**  $\mathbf{x} \in \mathcal{R}^p$  上的预测区间

将  $\{1, \dots, n\}$  随机分裂为两个大小相等的数据集  $\mathcal{I}_1, \mathcal{I}_2$

$$\widehat{f} = \mathcal{A}(\{(\mathbf{X}_i^T, Y_i)^T : i \in \mathcal{I}_1\})$$

$$R_i = |Y_i - \widehat{f}(\mathbf{X}_i)|, i \in \mathcal{I}_2$$

$$d = \{R_i : i \in \mathcal{I}_2\} \text{ 中第 } k \text{ 小的值, 其中 } k = \lceil (1 - \alpha)(\frac{n}{2} + 1) \rceil$$

返回  $\widehat{C}_{\text{split}}(\mathbf{x}) = [\widehat{f}(\mathbf{x}) - d, \widehat{f}(\mathbf{x}) + d]$ , 对于所有  $\mathbf{x} \in \mathcal{R}^p$

---

如果  $(\mathbf{X}_i^T, Y_i)^T, i = 1, 2, \dots, n$  是独立同分布的，对于一个新的数据点  $(\mathbf{X}_{n+1}^T, Y_{n+1})^T$ ，由**算法 2** 建立的分裂保形预测区间  $\widehat{C}_{\text{split}}(\mathbf{X}_{n+1})$ ，满足：

$$\Pr(Y_{n+1} \in \widehat{C}_{\text{split}}(\mathbf{X}_{n+1})) \geq 1 - \alpha.$$

另外，如果我们假设残差  $R_i, i \in \mathcal{I}_2$  具有连续的联合分布，那么

$$\Pr(Y_{n+1} \in \widehat{C}_{\text{split}}(\mathbf{X}_{n+1})) \leq 1 - \alpha + \frac{2}{n + 2}.$$

与完全保形预测方法相比，分裂保形预测除了具有极高的效率外，在内存需求方面也具有优势。例如，如果回归算法  $\mathcal{A}$  涉及变量选择，如 Lasso 或双向逐步回归等，当我们在估计新的点  $\mathbf{X}_i, i \in \mathcal{I}_2$  时，我们只需运用基于样本  $\mathcal{I}_1$  选择的变量，来拟合模型计算基于样本  $\mathcal{I}_2$  的残差。因此算法 2 可以大大节省内存。

分裂保形预测也可以使用不均衡分裂实现。采用  $\rho \in (0, 1)$ ，令  $|\mathcal{I}_1| = \rho n$  则  $|\mathcal{I}_2| = (1 - \rho)n$ 。例如，当回归算法很复杂的时候，可以选择  $\rho > 0.5$  使训练得到的回归预测器  $\hat{f}$  更准确。

### 12.3.2 Jackknife 保形预测

Jackknife 保形预测是一种计算复杂性介于完全保形法和分裂保形法之间的保形预测方法。该方法利用留一残差的分位数来定义预测区间，具体内容如算法 3 所示。

表 12.3 Jackknife 保形预测算法

---

#### 算法 3: Jackknife 保形预测

---

输入: 数据  $(\mathbf{X}_i^T, Y_i)^T, i = 1, \dots, n$ , 显著性水平  $\alpha \in (0, 1)$ , 回归算法  $\mathcal{A}$ ,

输出:  $\mathbf{x} \in \mathcal{R}^p$  上的预测区间

**for**  $i \in \{1, \dots, n\}$  **do**

$$\hat{f}^{(-i)} = \mathcal{A}(\{(\mathbf{X}_\ell, Y_\ell) : \ell \neq i\})$$

$$R_i = |Y_i - \hat{f}^{(-i)}(\mathbf{X}_i)|$$

**end for**

$d = \{R_i : i \in \{1, \dots, n\}\}$  中第  $k$  小的值, 其中  $k = \lceil n(1 - \alpha) \rceil$

返回  $\hat{C}_{\text{Jack}}(\mathbf{x}) = [\hat{f}(\mathbf{x}) - d, \hat{f}(\mathbf{x}) + d]$ , 对于所有  $\mathbf{x} \in \mathcal{R}^p$

---

与分裂保形法相比，Jackknife 法的一个优点是它在构造绝对残差时使用更多的训练数据，随后再构造分位数。这意味着它通常可以产生更短的间隔长度。我们注意到，由于对称性，Jackknife 法具有有限样本内覆盖性质：

$$\Pr(Y_i \in \hat{C}_{\text{Jack}}(\mathbf{X}_i)) \geq 1 - \alpha, \quad \text{对于所有 } i = 1, \dots, n$$

但是就样本外覆盖而言（真正的预测推断），它的属性比较脆弱。

接下来看几篇经典的 Jackknife 估计方法文章。Butler 和 Rothman [134] 表明，在低维线性回归中，Jackknife 法在足够强的正则条件下产生渐近有效区间，也意味着需要保证线性回归估计量的一致性。最近，Steinberger 和 Leeb [135] 在高维回归中建立了 Jackknife 区间的渐近有效性，他们虽然不需要基本估计量  $\hat{f}$  的一致性，但需要  $\hat{f}$  的一致渐近均方误差有界的条件。此外，Butler 和 Rothman [134] 和 Steinberger 和 Leeb [135] 的回归分析均基于线性模型，即回归函数是特征的线性函数，特征独

立于误差，误差是同方差的。但是这里介绍的 Jackknife 保形预测方法并不需要这些条件。

### 12.3.3 局部加权保形预测

当噪声是异方差的时候，我们可以用局部加权的思路替换完全保形预测方法或分裂保形方法中的绝对残差，即使用

$$V_i = \frac{|Y_i - \hat{f}(\mathbf{X}_i)|}{\hat{\sigma}(\mathbf{X}_i)}$$

替代

$$R_i = |Y_i - \hat{f}(\mathbf{X}_i)|$$

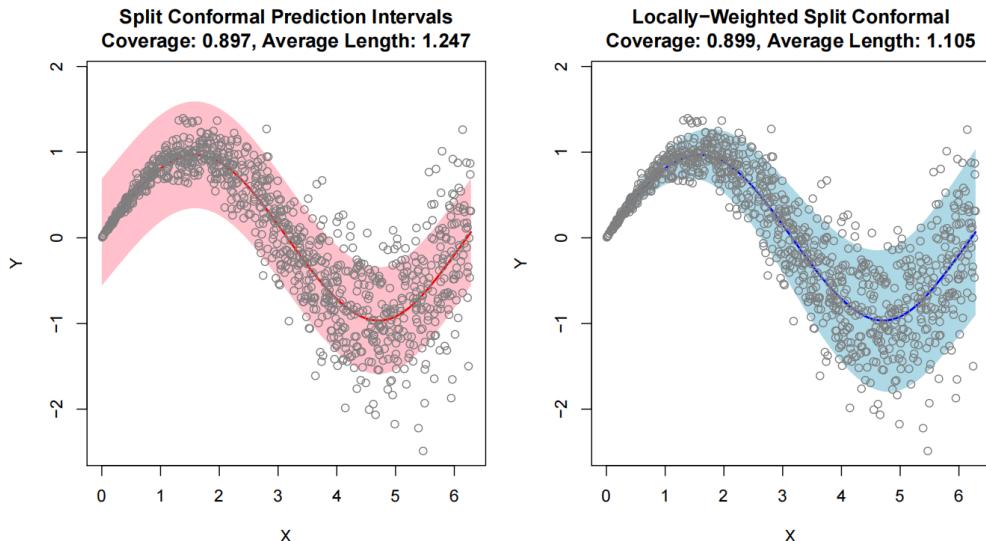
其中  $\hat{\sigma}^2(\mathbf{x})$  是绝对残差  $\text{var}(|Y - \hat{f}(\mathbf{X})| \mid \mathbf{X} = \mathbf{x})$  的方差函数的一个估计。(注意： $\hat{f}$  和  $\hat{\sigma}$  可以联合计算，也可以单独计算。)

局部加权对预测波段的局部宽度可能有很大变化。以下是在分裂保形预测区间形式：

$$\tilde{C}_{\text{local}}(\mathbf{x}) = [\hat{f}(\mathbf{x}) - \hat{\sigma}(\mathbf{x})\tilde{q}, \hat{f}(\mathbf{x}) + \hat{\sigma}(\mathbf{x})\tilde{q}]$$

其中  $\tilde{q} = \text{Quantile}(1 - \alpha; \{V_i\}_{i \in \mathcal{I}_2})$  表示  $V_i$  的  $1 - \alpha$  分位数。

### 异方差噪声示例



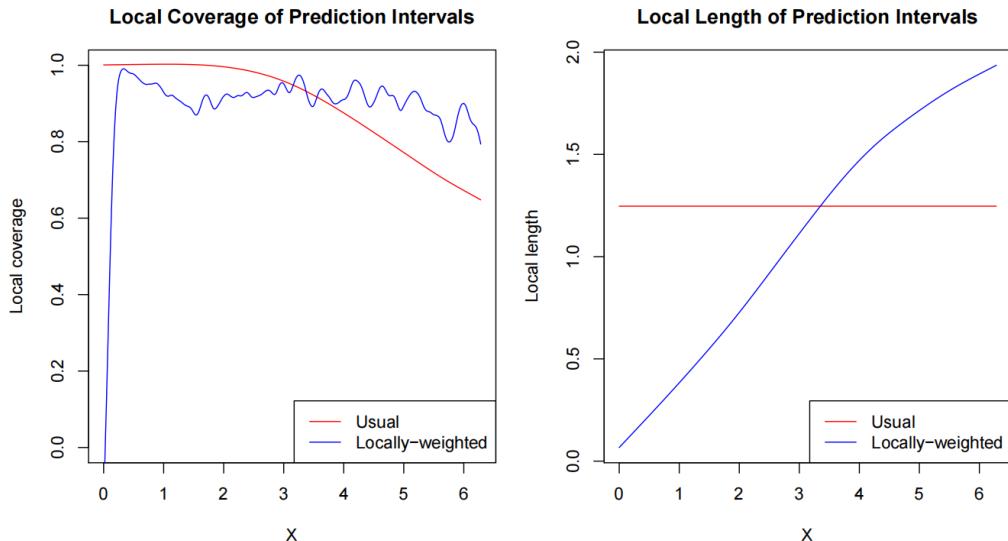


图 12.3 局部加权保形预测区间，其中  $n=100, p=1$ 。

从图12.3 可以看出，通过引入局部权重，我们可以获得能适应数据异质性的预测区域，在保证和之前的方法具有相同的有效性的同时，它具有更小的平均长度，局部的覆盖率也约等于我们想要的值（比如说 0.9）。

## 12.4 保形分类

选择合适的不符合度量是进行保形预测的重要步骤，我们在使用保形预测法进行分类时，应该根据数据类型和预测方法来选择不符合度量。下面我们将介绍两种不同的不符合度量下保形预测在分类中的应用。

### 12.4.1 Softmax 法

在机器学习尤其是深度学习中，Softmax 是个常用而且比较重要的函数，尤其在多分类的场景中使用广泛。假设预测目标是  $K$  个类别，则通过神经网络多层变换后得到一个  $K$  维的输入  $\mathbf{z} = (z_1, \dots, z_K)$ ，接下采用 Softmax 函数将  $\mathbf{z}$  映射为  $K$  个 0-1 之间的实数，并且归一化保证和为 1，因此多分类的概率之和也刚好为 1。Softmax 函数形式如下：

$$\text{Softmax}(\mathbf{z})_k = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}}, \quad k = 1, \dots, K.$$

在分类保形预测中，我们可以将 Softmax 函数作为我们的预测函数，Softmax 方法经常用于图片分类，下面我们简单介绍 Softmax 方法下的保形预测。

假设我们的预测任务是判断一个  $p$  维图像  $\mathbf{X}$  属于哪一个类别  $Y \in \{1, \dots, K\}$ 。首先我们使用训练数据和 Softmax 函数得到预测模型，也可称其为分类预测器  $\hat{f}$ 。分类预测器可以对于每个输入  $\mathbf{x}$  输出对应每个类别的 Softmax 分数： $\hat{f}(\mathbf{x}) = \text{Softmax}(\mathbf{z}) \in [0, 1]^K$ ，其中  $\hat{f}(\mathbf{x})_k = \text{Softmax}(\mathbf{z})_k$ 。然后，我们取适当数量的未用于训练的新数据点  $(\mathbf{X}_1^T, Y_1)^T, \dots, (\mathbf{X}_n^T, Y_n)^T$  作为校准数据集。利用  $\hat{f}$  和校准数据，我们试图构建一个可能的类别的置信集  $\mathcal{T}(\mathbf{x}) \subset \{1, \dots, K\}$ ，且这个置信集在下面意义下是有效的：

$$1 - \alpha \leq \Pr(Y_{n+1} \in \mathcal{T}(\mathbf{X}_{n+1})) \leq 1 - \alpha + \frac{1}{n+1}.$$

其中， $(\mathbf{X}_{n+1}^T, Y_{n+1})^T$  是来自同一分布的新测试点。换句话说，置信集包含正确类别的概率几乎正好是  $(1 - \alpha)$ ，我们称这种性质为边际覆盖。为了使用  $\hat{f}$  和校准数据构造  $\mathcal{T}$ ，我们将执行一个简单的校准步骤。首先，我们设置不符合分数为 1 减去正确类别的 Softmax 分数：

$$S_i = S(\mathbf{X}_i, Y_i) = 1 - \hat{f}(\mathbf{X}_i)_{Y_i},$$

如果预测器不好， $S_i$  有可能会很大。接下来定义  $\hat{q}$  为  $S_1, \dots, S_n$  的  $\lceil (1 - \alpha)(n + 1) \rceil$  分位数，这实际上是经过小小修正的  $n$  个样本的  $(1 - \alpha)$  分位数。对于一个新的输入  $\mathbf{X}_{n+1}$  ( $Y_{n+1}$  未知)，产生一个保形预测集：

$$\mathcal{T}(\mathbf{X}_{n+1}) = \{Y : S(\mathbf{X}_{n+1}, Y) < \hat{q}\} = \left\{Y : \hat{f}(\mathbf{X}_{n+1})_Y > 1 - \hat{q}\right\}$$

该预测集包含那些对应 Softmax 分数足够大的类别，如图12.4 所示。值得注意的是，不论我们应用什么样的预测模型，选择任何分布的数据集，这个算法得到的预测集都是具有覆盖保证的。

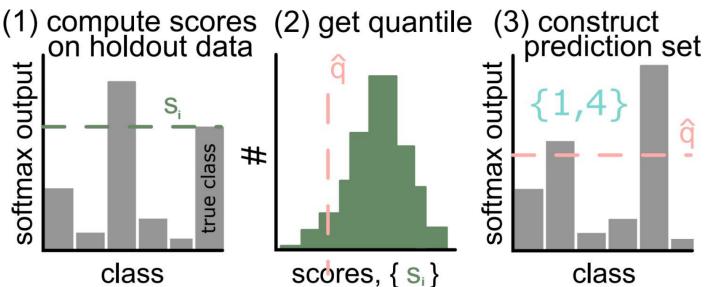


图 12.4 保形分类——Softmax 法

上述校准步骤的 PyTorch 代码如下：

```
#First, get conformal scores. n = calib_y.shape[0]
scores = 1-model(calib_x).softmax(dim=1)[:,calib_y]
```

```
# Second, get the adjusted quantile
qhat = torch.quantile(scores, np.ceil((n+1)(1-alpha))/n)
#Finally, deploy!
prediction_sets = (model(test_x) > (1-qhat)).nonzero()
```

### 12.4.2 最近邻法

最近邻法也是我们进行分类常用的方法之一。在保形预测中，我们使用最近邻法分类时，分类预测器将新的样本分类为与其距离最近的样本所对应的类别。

假设有  $n$  个样本  $\mathbf{Z}_1 = (\mathbf{X}_1^T, Y_1)^T, \mathbf{Z}_2 = (\mathbf{X}_2^T, Y_2)^T, \dots, \mathbf{Z}_n = (\mathbf{X}_n^T, Y_n)^T$ ，每个  $\mathbf{Z}_i$  包含特征向量  $\mathbf{X}_i$  和类别  $Y_i \in \{1, \dots, K\}$ 。然后我们对于一个新的样本点  $\mathbf{Z}_{n+1} = (\mathbf{X}_{n+1}^T, Y_{n+1})^T$ ，我们只能观察到  $\mathbf{X}_{n+1}$  而不知道  $Y_{n+1}$ 。最近邻法寻找距离  $\mathbf{X}_{n+1}$  最近的  $\mathbf{X}_i$ ，并使用它的类别  $Y_i$  作为  $Y_{n+1}$  的预测值。如果没有合适的预测器，我们很难度量这个预测的正确性。但是我们可以通过比较  $\mathbf{X}$  到与其有相同类别的旧例子的距离和  $\mathbf{X}$  到与其有不同类别的旧例子的距离，来得到不符合度量，因此我们用下式表示不符合分数：

$$S_i = S(\mathbf{X}_i, Y_i) = \frac{\min \{\|\mathbf{X}_j - \mathbf{X}_i\| : j \neq i, Y_j = Y_i\}}{\min \{\|\mathbf{X}_j - \mathbf{X}_i\| : j \neq i, Y_j \neq Y_i\}}$$

不符合分数  $S_i$  大小，也可以代表分类是否正确。接下来定义  $\hat{q}$  为  $S_1, \dots, S_n$  的  $\lceil (1 - \alpha)(n + 1) \rceil$  分位数。对于一个新的输入  $\mathbf{X}_{n+1}$ ，我们得到保形预测集：

$$\mathcal{T}(\mathbf{X}_{n+1}) = \{Y : S(\mathbf{X}_{n+1}, Y) < \hat{q}\}$$

## 12.5 R 语言实现

保形预测涉及的 R 包，可在该网址查阅：

<https://github.com/ryantibs/conformal/>

- 保形方法：全保形预测，分裂保形预测，Jackknife 保形预测，留一保形预测，局部加权法
- 基本的估计方法：线性回归，岭回归，套索回归，逐步回归，随机森林，或自定义
- 简单地运行模拟结果的框架在 [132] 中可以看到

### 12.5.1 分裂保形预测的简单代码

```

conformal.pred.split = function(x, y, x0, train.fun, predict.fun, alpha=0.1, rho=0.5,
w=NULL, mad.train.fun=NULL, mad.predict.fun=NULL, split=NULL, seed=NULL, verbose=FALSE) {
  #设置数据
  x = as.matrix(x)
  y = as.numeric(y)
  n = nrow(x)
  p = ncol(x)
  x0 = matrix(x0, ncol=p)
  n0 = nrow(x0)
  #检查输入参数
  check.args(x=x, y=y, x0=x0, alpha=alpha, train.fun=train.fun, predict.fun=predict.fun,
  mad.train.fun=mad.train.fun, mad.predict.fun=mad.predict.fun)
  #如果rho是空值，或者长度不等于一，或者不是数值，或者rho小于等于0或大于等于1，则停止
  if (is.null(rho) || length(rho) != 1 || !is.numeric(rho) || rho<=0 || rho>=1)
    stop("rho must be a number in between 0 and 1")#rho必须是0和1之间的数字
  if (is.null(w)) w = rep(1, n+n0)
  #用户可以用一个字符串作为verbose参数
  if (verbose == TRUE) txt = ""
  if (verbose != TRUE && verbose != FALSE) {
    txt = verbose
    verbose = TRUE
  }
  #如果用户为了分割传递索引，则
  if (!is.null(split)) i1 = split
  #否则进行随机分割
  else {
    if (!is.null(seed)) set.seed(seed)
    i1 = sample(1:n, floor(n*rho))
    #Floor接受一个数字参数x，并返回一个包含不大于x中相应元素的最大整数的数字向量
    #sample(x, size, replace = FALSE, prob = NULL)
    #Sample从x的元素中提取指定大小的样本，使用替换或不替换
  }
  #从样本1到n中提取(rho*n)个样本，如果(rho*n)不是整数，则取小于它的最大整数
  i2 = (1:n)[-i1]#i2表示1到n去掉i1已经提取过的值
  n1 = length(i1)
  n2 = length(i2)
  if (verbose) {
    cat(sprintf("%sSplitting data into parts of size %i and %i ...\\n", txt, n1, n2))
    cat(sprintf("%sTraining on first part ...\\n", txt))
  }
  #用第一部分的数据i1进行训练
  out = train.fun(x[i1,,drop=F], y[i1])#输入的是x的i1行，和Y的i1行，得到训练模型out
  #使用训练好的模型out和点X，得到X的预测值Y
  #并将其排列为n行的矩阵拟合，可以看到和真实值的差别
  fit = matrix(predict.fun(out, x), nrow=n)
  #使用训练好的模型out和新的点X0，得到新的X的预测值Y，并将其排列为n0行的矩阵
  pred = matrix(predict.fun(out, x0), nrow=n0)
  m = ncol(pred)#令m等于新预测值Y的列数
}

```

```

if (verbose) {
  cat(sprintf("%sComputing residuals and quantiles on second part ...\\n",txt))
}

#使用第二部分的数据i2计算残差和分位数
#计算校准集残差绝对值，得到的结果是一个n2*1的矩阵
res = abs(y[i2] - matrix(predict.fun(out,x[i2,,drop=F]),nrow=n2))
lo = up = matrix(0,n0,m)#暂且定义low和up均为n0*m的矩阵，用0填充
for (l in 1:m) {
  #局部评分
  if (!is.null(mad.train.fun) && !is.null(mad.predict.fun)) {
    res.train = abs(y[i1] - fit[i1,l])
    mad.out = mad.train.fun(x[i1,,drop=F],res.train)
    mad.x2 = mad.predict.fun(mad.out,x[i2,,drop=F])
    mad.x0 = mad.predict.fun(mad.out,x0)
    res[,l] = res[,l] / mad.x2
  }
  else {
    mad.x0 = rep(1,n0)#不加权时，mad.x0是n0维的向量，以1填充
  }
  #order将残差绝对值按升序排列，得到的是残差的位置序号
  o = order(res[,l]); r = res[o,]; ww = w[i2][o]#r是将残差按升序排列
  for (i in 1:n0) {
    q = weighted.quantile(c(r,Inf),1-alpha,w=c(ww,w[n+i]),sorted=TRUE)#取分位数
    lo[i,l] = pred[i,l] - q * mad.x0[i]#置信区间的下界是perd-q
    up[i,l] = pred[i,l] + q * mad.x0[i]#置信区间的上界是perd+q
  }
}
return(list(pred=pred,lo=lo,up=up,fit=fit,split=i1))
}

```

使用上述程序所构造的函数时，需要输入以下代码片段，定义预测函数和初始数据：

```

fun = lm.funs()# Or, e.g., lasso.funs(), rf.funs()
obj = conformal.pred(x, y, x0, alpha=0.1,
train.fun=funs$train, predict.fun=funs$predict)

```

## 12.5.2 局部加权保形预测的简单代码

局部加权保形带允许带的长度在  $X$  上以一种解释异方差的方式变化。其思想是对条件扩散 (conditional spread) 进行建模，并利用对条件扩散的估计来缩放波段。我们使用局部 (核加权) 多项式回归对绝对残差的条件扩展进行建模，然后使用它来缩放波段。我们将修改这个例子，使异方差确实存在：

```

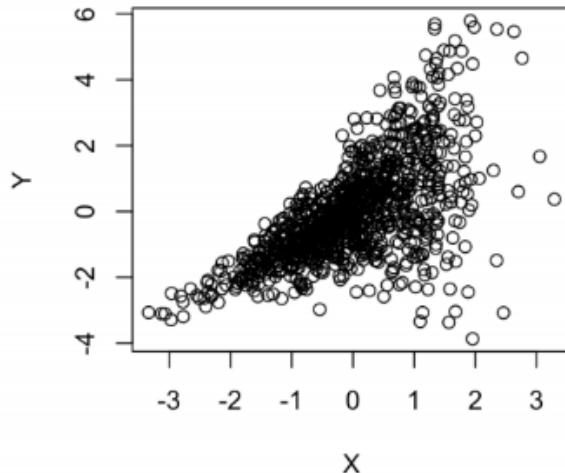
set.seed(1)
n <- 1000
X <- rnorm(n)#1000个服从标准正态分布的数据

```

```

Uhet <- rnorm(n, sd=exp(.5*X))#1000个服从正态分布, 标准差为exp(.5*X)的数据
Y <- X + Uhet#线性回归模型存在异方差性, 即随机误差项具有不同的方差
regData <- data.frame(X,Y)
plot(X, Y)

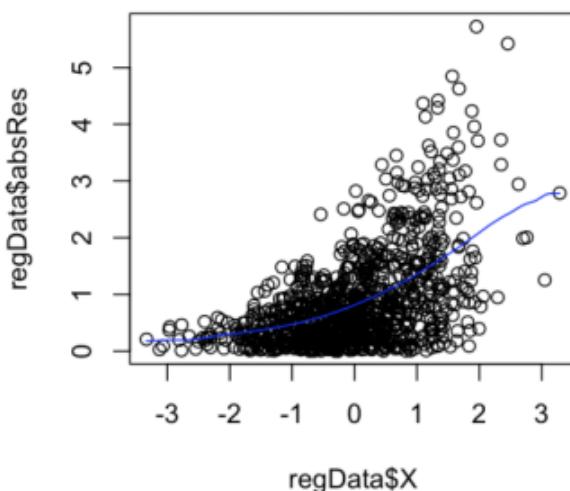
```



```

fitlm.het <- lm(Y~X, data=regData)#对X和Y做线性回归
regData$absRes <- abs(fitlm.het$residuals)#残差绝对值
library(locpol)#locpol(): 局部多项式估计, 可以用来做非参数回归
#对于每个X, 以该点为中心, 按照预定宽度构造一个区间, 以x为中心, 根据残差给带宽加权
locFit <- locpol(absRes~X,data=regData,bw=2)#bw 平滑参数, 带宽
plot(regData$X, regData$absRes)
#展示了X和线性回归残差的关系, 随着x的增大, 残差也越来越大
points(locFit$lpFit[,locFit$X],locFit$lpFit[,locFit$Y],type="l",col="blue")

```



现在我们应用局部加权保形预测：

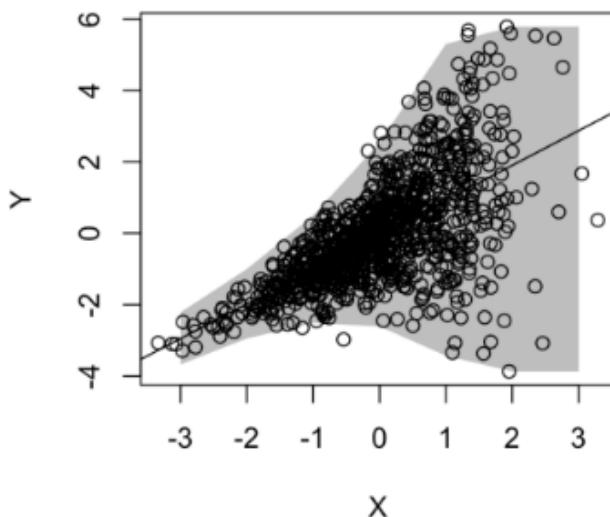
```

nEval <- 200
yCand <- seq(from=min(Y), to=max(Y), length=nEval)
#定义局部加权保形预测函数
confPredict.lw <- function(y, Xin, bwArg){#bwArg带宽参数 #Xin是一个新的点
  nData <- nrow(regData)#原数据行数
  regData.a <- rbind(regData,c(Xin, y))#将原数据和c(Xin, y)按行合并
  fitlm.a <- lm(Y~X, data=regData.a)#用加了一行的新数据做线性回归
  regData.a$resOut.sc <- abs(fitlm.a$residuals)#加一列新的线性回归的残差绝对值
  xevalVec <- c(rev(seq(from=Xin, to=min(X), by= -0.25)),seq(from=Xin+.25, to=max(X),
  by=.25))#产生新的预测点,选取新的预测点Xin,以其为原点,生成部分数据,按从小到大排列
  #用加了一个新预测点的数据,以X为中心,在给定带宽的情况下,根据残差绝对值对带宽加权
  #xeval: 预测点的向量 预测多个新的点
  locFit <- locpol(resOut.sc~X,data=regData.a,bw=bwArg,xeval=xevalVec)
  resScale <- locFit$lpFit[,2][locFit$lpFit[,1]==Xin]#新加的点在locpol中对应的残差值
  #新的点在局部加权中对应的残差乘以 所有的点在1001个数据线性回归后对应的残差
  resOut <- regData.a$resOut.sc*resScale
  #新的点在新的回归中对应的残差乘以新的点在局部加权中对应的残差
  resOut_new <- regData.a$resOut.sc[length(resOut)]
  #找分位数点 #在resout中找小于resOut_new的点, 并求其均值
  pi.y <- mean(apply(as.matrix(resOut), 1,function(x){x<=resOut_new}))
  #ceiling取不小于 () 的最小整数
  testResult <- pi.y*(nData+1) <= ceiling(.975*(nData+1))
  return(list(testResult,resScale))
}

aug.over.X.lw <- function(Xval){
  #输出预测区间的最小值和最大值
  Cout.lw <- range(yCand[unlist(sapply(yCand,confPredict.lw,Xin=Xval, bwArg=2)[1,])])
  return(Cout.lw)
}

Xvals <- -3:3
#lapply把指定的待应用的函数应用于列表的每一个元素, 并返回列表结构的输出
#输入Xvals, 输出预测区间的最大值和最小值, 最大值一列, 最小值一列
augBands.lw <- matrix(unlist(lapply(Xvals, aug.over.X.lw)),ncol=2, byrow=T)
plot(X,Y, type="n")
polygon(c(Xvals, rev(Xvals)),
c(augBands.lw[,1],rev(augBands.lw[,2])),#包含多边形顶点坐标的向量
border=F,col="gray")#绘制多边形
points(X,Y)
abline(fitlm.het)

```



以上是我们运用 R 语言来实现保形预测的简单操作,下面我们也可以通过 python 语言实现保形预测。

## 12.6 Python 语言实现

### 12.6.1 完全保形预测的简单代码

我们可以根据完全保形预测的原理,通过构造不符合分数,确定分位点,得到保形预测区间。首先导入必要的包,并定义相应函数

```
import random
import sys
import statistics
eps = 0.20 #显著性水平
M = 1000
#定义不符合分数
def A(B, z):
    z0 = statistics.mean(B) #z0为预测值
    ans = abs(z0 - z) #取预测值减真实值的绝对值为不符合分数
    return ans
#完全保形预测 判断分位数点
def pz(z, zn):
    an = A(z, zn) #an = z - zn
    z.append(zn) #z后加上zn
    cnt = 0
    n = len(z) #z的长度为n
```

```

for i in range(n):
    z0 = z[:i] + z[i+1:] #z0等于z去掉zi
    a = A(z0, z[i])
    if a >= an:
        cnt += 1
z.pop() #删除z中最后一个值，即加进去的zn
return cnt / n #返回分位数点
def G(z):
    ans = []
    for zn in range(M):
        p = pz(z, zn) #样本分位数点
        #如果样本分位数点大于显著性水平，则预测值zn包含在eps置信水平下的预测中
        if p > eps:
            ans.append(zn)
    return ans
#定义gen为符合期望为mu，标准差为sigma的正态分布的随机数
def gen():
    return int(random.gauss(mu, sigma))

```

将数据代入，得到保形预测区间

```

Z = [17, 20, 10, 17, 12, 15, 19, 22, 17, 19, 14, 22, 18, 17, 13, 12, 18, 15, 17]
mu = 500
sigma = 100
n = 2000
z = []
z.append(gen())
correct = total = 0
for i in range(n - 1):
    x = gen()
    print("%03d-%03d" % (min(G(z)), max(G(z))))
    if x in G(z):
        correct += 1
    total += 1
    print("%03d/%03d = %.0f" % (correct, total, correct / total * 100))
    z.append(x)
print(correct / total)

```

## 12.6.2 保形分类和保形回归的简单代码

另外我们也可以使用 python 调用 nonconformist 包，直接进行保形预测。下面是我们用该方法分别进行保形分类和保形回归的代码：

首先是保形分类，在这里，我们以使用支持向量机法对鸢尾花数据集进行分类为例：

```

from sklearn.datasets import load_iris
import numpy as np

```

```

from sklearn.svm import SVC
from nonconformist.cp import IcpClassifier
from nonconformist.nc import NcFactory
iris = load_iris()
idx = np.random.permutation(iris.target.size)      #对鸢尾花数据数量进行随机排列
# 将数据分为适当的训练集、校准集和测试集
idx_train, idx_cal, idx_test = idx[:50], idx[50:100], idx[100:]
model = SVC(probability=True)    # 建立预测模型
nc = NcFactory.create_nc(model)   # 建立不符合度量
icp = IcpClassifier(nc)          # 建立一个归纳保形分类预测器
# 用训练集拟合分类预测器
icp.fit(iris.data[idx_train, :], iris.target[idx_train])
# 用校准集校准分类预测器
icp.calibrate(iris.data[idx_cal, :], iris.target[idx_cal])
# 用测试集生成在95%置信度下的预测区间
prediction = icp.predict(iris.data[idx_test, :], significance=0.05)
# 打印预测结果
print(prediction[:5, :])

```

该预测结果是一个布尔型 numpy 数组，行数是测试集的数量，列数是类别数，每一行是一个代表在特定显著性水平下的类别是否包含在预测区间中的布尔向量。在这个例子中，对于给定的测试对象，我们也许会得到一个 [True True False] 的向量，这意味着有 95% 的可能，True 对应的两个类别中有一个是正确的。

其次是保形回归，我们以使用随机森林法对波士顿数据集进行回归为例：

```

from sklearn.datasets import load_boston
import numpy as np
from sklearn.ensemble import RandomForestRegressor
from nonconformist.cp import IcpRegressor
from nonconformist.nc import NcFactory
boston = load_boston()
idx = np.random.permutation(boston.target.size)
# 将数据分为适当的训练集、校准集和测试集
idx_train, idx_cal, idx_test = idx[:300], idx[300:399], idx[399:]
model = RandomForestRegressor()    # 建立预测模型
nc = NcFactory.create_nc(model)    # 建立不符合度量
icp = IcpRegressor(nc)           # 建立一个归纳保形回归器
# 用训练集拟合回归模型
icp.fit(boston.data[idx_train, :], boston.target[idx_train])
# 用校准集校准模型
icp.calibrate(boston.data[idx_cal, :], boston.target[idx_cal])
# 用测试集生成在95%置信度下的预测区间
prediction = icp.predict(boston.data[idx_test, :], significance=0.05)
# 打印预测结果
print(prediction[:5, :])

```

该预测结果是一个数值型 numpy 数组，行数是测试集的数量，列数是 2，每一行是一个代表区间下界和上界的向量，表示在特定显著性水平下的预测区间。在这个例子中，对于给定的测试对象，我们也许会得到一个数值向量 [26.47 39.45]，这

意味着在 95% 置信度下得到的预测区间为 [26.47, 39.45]。

# 第十三章 人工神经网络

1943年，心理学家W·McCulloch和数理逻辑学家W·Pitts在分析、总结神经元(Neuron)基本特性的基础上，首先提出了神经元的数学模型[136]。到目前为止，人工神经网络(后面简称为神经网络)已经发展成了一个相当大的、多学科交叉的学科领域。对神经网络的定义已有很多，本书采用了Kohonen在1988年提出的定义，即“[神经网络](#)是由具有适应性的简单单元组成的广泛并行互连的网络，它的组织能够模拟生物神经系统对真实世界物体所作出的交互反应”[137]。在机器学习中谈论神经网络时指的是“神经网络学习”，或者说，是机器学习与神经网络这两个学科领域的交叉部分。

## 13.1 人工神经元

神经网络起源于对生物神经元的研究，如图13.1所示[生物神经元](#)包括细胞体、树突、轴突等部分，其中树突是用于接受输入信息，输入信息经过突触处理，当达到一定条件时通过轴突传出，此时神经元处于激活状态；反之没有达到相应条件，则神经元处于抑制状态。

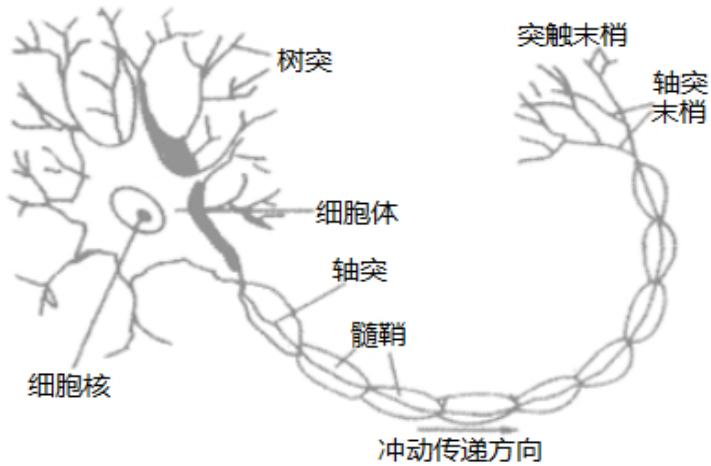


图 13.1 生物神经元

神经网络的基本节点是[人工神经元](#)(后面简称为[神经元](#))，其工作原理是仿照生

物神经元提出的，如图13.2所示。输入值经过加权和偏置后，由激活函数处理后输出。

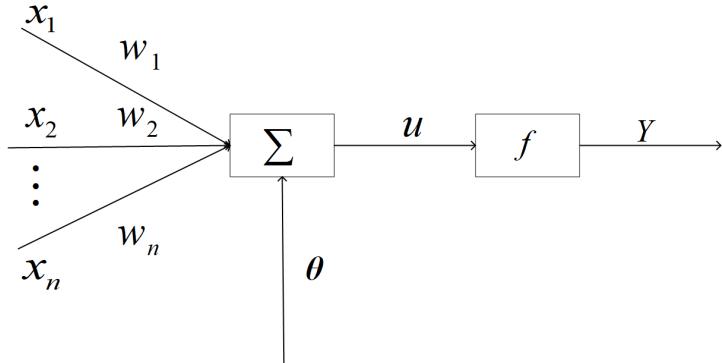


图 13.2 人工神经元模型

从图13.2所示的神经元示意图来看，对某一个神经元来说，它可以同时接受  $n$  个输入信号，分别用  $x_1, x_2, x_3, \dots, x_n$  来表示。每个输入端与神经元之间有联接权值  $w_1, w_2, w_3, \dots, w_n$ ，神经元的总输入为对每个输入进行加权求和，同时加上偏置  $\theta$ ，即：

$$u = \sum_{i=1}^n w_i x_i + \theta \quad (13.1.1)$$

神经元的输出  $Y$  是对  $u$  的映射，

$$Y = f(u) = f\left(\sum_{i=1}^n w_i x_i + \theta\right) \quad (13.1.2)$$

其中， $f$  为激活函数。

**激活函数**能够给神经元引入非线性因素，使得神经网络可以任意逼近任何非线性函数，这样神经网络就可以应用到更多的非线性模型中。而理想的激活函数应该是像图13.3(a) 中  $\text{sgn}(x)$  函数那样的跃阶函数，可以将输入神经元的值映射为“抑制”或“兴奋”的输出，也就是 0 和 1 [138]。然而跃阶函数的性质比较差，它是不连续、不光滑的，因此实际常用 sigmoid 函数作为激活函数，如图13.3(b)，它把输入神经元的值映射到了  $(0,1)$  的范围内进行输出。

## 13.2 前馈神经网络

在实际问题中，不可能只使用一个神经元，会使用非常多的神经元进行计算。一般情况下，神经元分层排列，每个神经元只接受前一层的输入，并输出到下一层，网络没有下一层到上一层的回路信号，这种网络称为**前馈神经网络**（Feedforward Neural

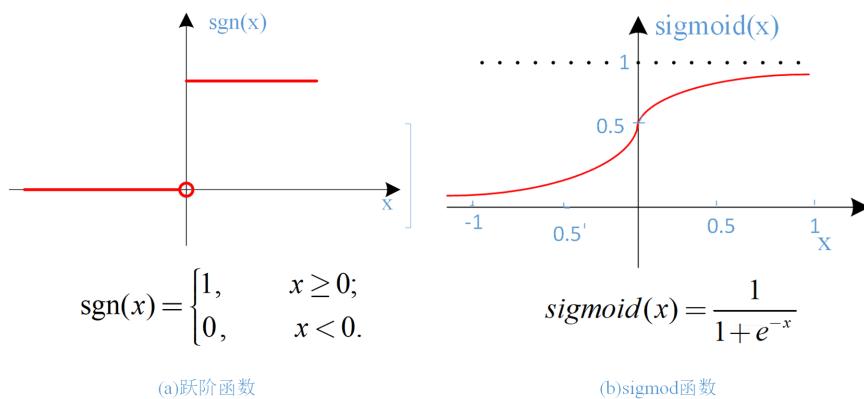


图 13.3 神经元激活函数

Networks)。前馈并不意味着网络中信号不能向后传，而是指网络拓扑结构上不存在环或回路。前馈神经网络的第一层为输入层 (Input Layer)，最后一级为输出层 (Output Layer)，输入层与输出层之间的各层统一称为隐含层 (Hidden Layers)。隐含层和输出层神经元都是拥有激活函数的功能神经元。一个网络可以只包含一个隐含层，也可以包含多个隐含层。在这个意义上感知器可以理解为一种常见的前馈神经网络。

### 13.2.1 单层感知器

**感知器**，也称感知机，是 Frank Rosenblatt 提出的一种神经网络 [139]。如图13.4所示，这是一个简单的感知器逻辑图，该感知器具有两层，第一层为输入层，将输入的值传递给下一层；第二层为计算单元，并将结果输出。因此，该网络也被称为**单层感知器**，它能容易实现与、或、非等逻辑运算，也常用于线性分类问题。

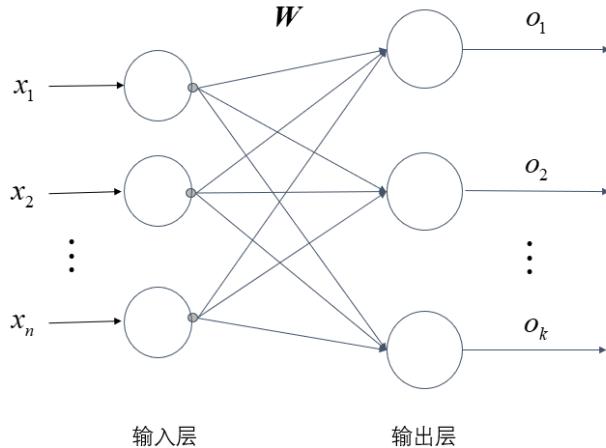


图 13.4 单层感知器模型

假设输入模式为 \$n\$ 维特征向量  $\mathbf{X} = (x_1, \dots, x_n)^T$ , 则感知器的输入层应有 \$n\$ 个神经元。若输出类别有 \$k\$ 个, 则输出层应包含 \$k\$ 个神经元, 即  $O = (o_1, o_2, \dots, o_k)^T$ 。输入层的第 \$j\$ 个神经元与输出层的第 \$i\$ 个神经元的连接权值为 \$w\_{ji}\$, 则第 \$i\$ 个神经元的输出为:

$$o_i = f \left( \sum_{j=1}^n (w_{ji} x_j + \theta_i) \right) \quad (13.2.1)$$

如果感知器模型是一种针对二分类问题中的神经网络模型, 其激活函数为符号函数  $\text{sgn}(x)$ , \$k\$ 值为 2, 感知机模型通过输入层接受输入信息  $\mathbf{X} = (x_1, \dots, x_n)^T$ , 其输出可以表示为:

$$o_i = \text{sgn} \left( \sum_{j=1}^n (w_{ji} x_j + \theta_i) \right) \quad (13.2.2)$$

一般地, 给定训练数据集和感知器后, 权重  $\mathbf{W}$  以及阈值  $\theta$  可通过学习得到。当把阈值  $\theta$  可看作一个固定输入为 1.0 的“哑结点” (Dummy Node) 并将所对应的连接权也记为  $\mathbf{W}$  时, 权重和阈值的学习就可统一为权重的学习。

### 13.2.2 多层感知器

单层感知器只能解决线性可分的问题, **多层感知器**可以突破这一局限, 实现输入和输出之间的非线性映射。多层感知器的神经元层级之间采用全连接的方式, 上层的神经元的输出作为输入推送给下一层的所有神经元。多层感知器中, 除了输入输出层, 中间可以有多个隐层。最简单的情况时只含一个输入层、一个隐含层和一个输出层, 即三层感知器, 也叫**单隐层前馈神经网络**, 如图13.5所示。

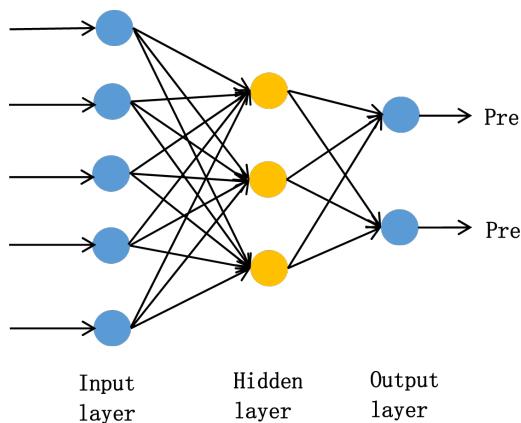


图 13.5 多层感知器示意图

若第  $l$  层的第  $i$  个神经元的输出为  $a_i^{(l)}$ ，其与上一层第  $j$  个神经元的连接权值为  $w_{ji}^{(l-1)}$ ，偏置为  $\theta_i^{(l)}$ ；与下一层的第  $k$  个神经元的连接权值为  $w_{ik}^{(l)}$ ，偏置为  $\theta_k^{(l+1)}$ 。则该神经元的输入为： $z_i^{(l)} = \sum_j w_{ji}^{(l-1)} a_j^{(l-1)} + \theta_i^{(l)}$ ，其中  $a_j^{(l-1)}$  为上一层第  $j$  个神经元的输出，则第  $l$  层第  $i$  个神经元的输出为：

$$a_i^{(l)} = f(z_i^{(l)}) = f\left(\sum_j w_{ji}^{(l-1)} a_j^{(l-1)} + \theta_i^{(l)}\right) \quad (13.2.3)$$

其中， $f$  为激活函数。

从上面可以看出，人工神经网络的学习过程，就是根据训练数据来调整神经元之间的连接权值（Connection Weight）和每个功能神经元的偏置；换言之，神经网络学到的东西，蕴含在连接权值与偏置中。

### 13.3 神经网络的正向与反向传播算法

在前几节中，我们已经学习了神经网络的基本知识。由于神经网络的结构是多种多样的，在本节中将以单隐层神经网络和双隐层神经网络为例，介绍神经网络具体的学习过程是怎样的。

#### 13.3.1 神经网络的正向传播

先以简单的单隐层网络为例（图13.6），其中第一层是输入层单元，输入原始数据。第二层是隐藏层单元，对数据进行处理，然后呈递到下一层。最后是输出单元，计算后，输出计算结果。

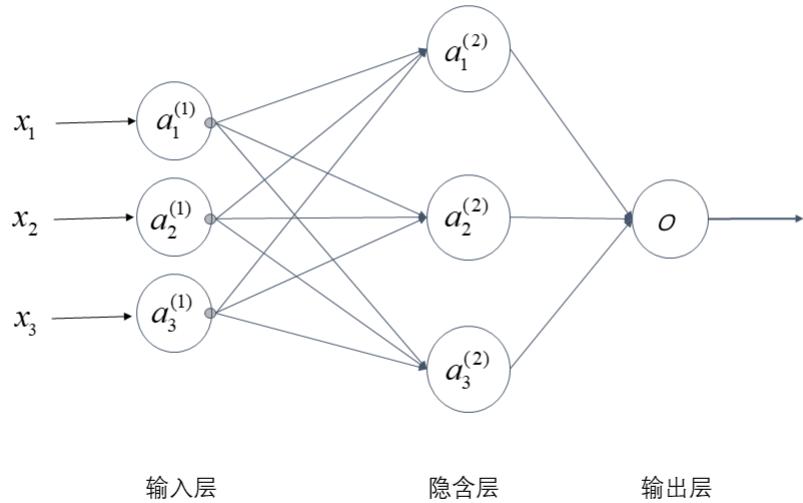


图 13.6 简单的单隐层神经网络

一般情况下，为了提高拟合能力，需要增加偏置项，如图13.7为一个 3 层的神经网络，其中第一层为输入层，最后一层为输出层，中间一层为带偏置的隐藏层。

下面引入一些标记法来帮助描述模型：

$a_i^{(j)}$  代表第  $j$  层的第  $i$  个单元， $w^{(j)}$  代表从第  $j$  层映射到第  $j+1$  层时的权重的矩阵，例如  $w^{(1)}$  代表从第一层映射到第二层的权重的矩阵。其尺寸为：以第  $j+1$  层的单元数量为行数，以第  $j$  层的激活单元数加一为列数的矩阵。例如：图13.7所示的神经网络中如  $w^{(1)}$  的尺寸为  $3 \times 4 = 12$  。

对于图13.7所示的模型，激活单元和输出分别表达为：

$$a_1^{(2)} = f \left( w_{10}^{(1)} x_0 + w_{11}^{(1)} x_1 + w_{12}^{(1)} x_2 + w_{13}^{(1)} x_3 \right) \quad (13.3.1)$$

$$a_2^{(2)} = f \left( w_{20}^{(1)} x_0 + w_{21}^{(1)} x_1 + w_{22}^{(1)} x_2 + w_{23}^{(1)} x_3 \right) \quad (13.3.2)$$

$$a_3^{(2)} = f \left( w_{30}^{(1)} x_0 + w_{31}^{(1)} x_1 + w_{32}^{(1)} x_2 + w_{33}^{(1)} x_3 \right) \quad (13.3.3)$$

$$O = f \left( w_{10}^{(2)} a_0^{(2)} + w_{11}^{(2)} a_1^{(2)} + w_{12}^{(2)} a_2^{(2)} + w_{13}^{(2)} a_3^{(2)} \right) \quad (13.3.4)$$

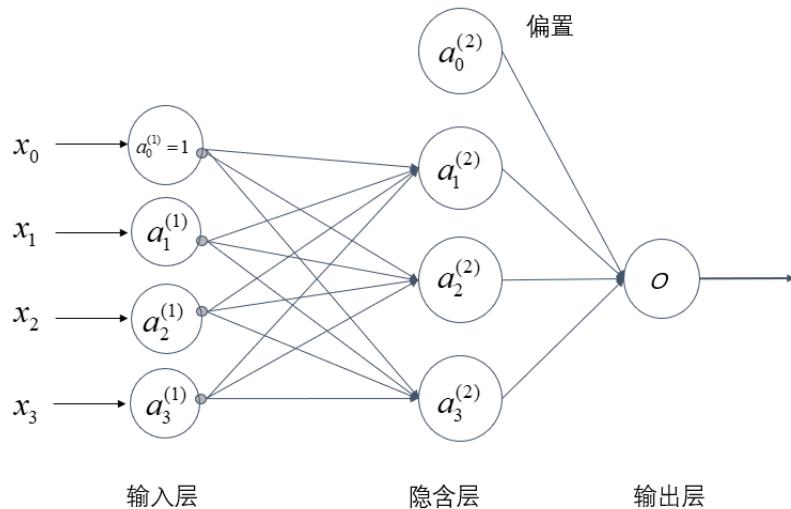


图 13.7 加入偏置的神经网络模型

上述情况是一种两分类的神经网络模型，但当有更多种分类时，比如以下这种情况，该怎么办？

假如要训练一个神经网络算法来识别路人、汽车、摩托车和卡车，在输出层我们应该有 4 个值。例如，第一个值为 1 或 0 用于预测是否是行人，第二个值用于判断是否为汽车，以此类推。输入向量  $x$  有三个维度，两个中间层，输出层 4 个神经元分别用来表示 4 类，也就是每一个数据在输出层都会出现  $(a, b, c, d)^T$ ，且  $(a, b, c, d)^T$  中仅有 1 个为 1，其余为 0 时，表示当前类。图 13.8 是该神经网络的结构示例。

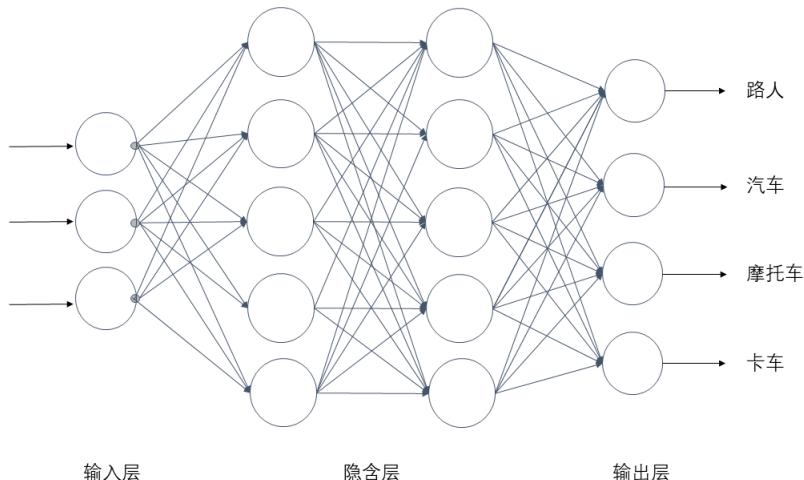


图 13.8 神经网络结构示意图

神经网络算法的输出结果为以下四种可能情形之一：

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

### 13.3.2 神经网络的损失函数

首先引入一些便于稍后讨论的符号：

假设神经网络的训练样本有  $m$  个，第  $i$  个样本的输入为  $\mathbf{X}_i = (x_{i0}, x_{i1}, \dots, x_{in})^T$ ，第  $i$  个样本的标签为  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ik})^T$ ，其整体映射函数为  $G(\mathbf{X})$ ，若权重  $W$  为变量时，则整体映射函数记为  $G_W(\mathbf{X})$ 。 $L$  表示神经网络层数， $S_l$  表示每层的神经元个数 ( $l = 1, \dots, L$ )，则  $S_L$  代表最后一层中神经元的个数。若该神经网络是  $k$  分类问题，那么显然  $S_L = k$ 。

神经网络学习的过程，实际上就是最小化损失函数的过程。而损失函数是用来估计模型的预测值  $G(\mathbf{X})$  与真实值  $\mathbf{Y}$  的不一致程度，通常用误差函数进行衡量。对于连续函数，当  $G(\mathbf{X})$  和  $\mathbf{Y}$  为维度为一时，其损失函数常定义为：

$$J(W) = \frac{1}{2m} \left[ \sum_{i=1}^m (G_W(\mathbf{X}_i) - \mathbf{Y}_i)^2 \right] + \lambda \|W\|_2^2 \quad (13.3.5)$$

公式13.3.5右侧的前半部分为经验风险函数，后半部分为正则化项， $\lambda$  为正则化因子，由于是 2-范数的平方项，称为  $L_2$  正则化项，也可以是其它形式的正则化项。

对于离散函数，比如，逻辑回归中常用的整体损失函数为：

$$J(W) = -\frac{1}{m} \left[ \sum_{i=1}^m \mathbf{Y}_i \log G_W(\mathbf{X}_i) + (1 - \mathbf{Y}_i) \log (1 - G_W(\mathbf{X}_i)) \right] + \lambda \|W\|_2^2 \quad (13.3.6)$$

在逻辑回归中，只有一个输出变量，但是在神经网络中，可以有很多输出变量。如果  $G_W(\mathbf{X})$  是一个维度为  $k$  的向量， $G_W(\mathbf{X})_j$  表示  $G_W(\mathbf{X})$  的第  $j$  个维度，并且训练集中的因变量也是同样维度的一个向量，因此损失函数会比逻辑回归更加复杂一些，则损失函数为：

$$\begin{aligned} J(W) = & -\frac{1}{m} \left[ \sum_{i=1}^m \sum_{j=1}^k Y_{ij} \log (G_W(\mathbf{X}_i))_j + (1 - Y_{ij}) \log (1 - (G_W(\mathbf{X}_i))_j) \right] \\ & + \frac{\lambda}{2m} \sum_{l=1}^{L-1} \sum_{i=1}^{S_l} \sum_{j=1}^{S_{l+1}} (w_{ij}^{(l)})^2 \end{aligned} \quad (13.3.7)$$

这个看起来复杂很多的损失函数，其背后的思想还是一样的，即希望通过损失函数来观察算法预测的结果与真实情况的误差有多大。唯一不同的是，对于每一行特征，都会给出  $k$  个预测，基本上可以利用循环，对每一行特征都预测  $k$  个不同结果，然后再利用循环在  $k$  个预测中选择可能性最高的一个，将其与  $\mathbf{Y}$  中的实际数据进行比较。

在确定了模型与损失函数后，通过梯度下降算法就可以让神经网络进行学习了，但是，使用梯度下降算法有一个前提：该算法需要提前知道当前点的梯度，然而事实并非如此，如果仔细观察过此模型的损失函数之后，我们会发现这个函数相当的复杂，其导函数难以计算，而在更加复杂的神经网络模型中，对损失函数求导往往是不可能的，因此我们需要通过其他方法来进行实现。

### 13.3.3 反向传播算法

1974 年，Paul Werbos [140] 首次给出了训练神经网络的学习算法——[反向传播算法](#)（BP，Back Propagation），这个算法可以高效的计算每一次迭代过程中的梯度，它是迄今最成功的神经网络学习算法。在实际使用神经网络时，大多是在使用 BP 算法进行训练。值得指出的是，BP 算法不仅可用于多层前馈神经网络，还可用于其他类型的神经网络，例如训练递归神经网络 [141]。但通常说“BP 网络”时，一般是指用 BP 算法训练的多层前馈神经网络。

下面举一个简单的例子来说明 BP 算法是如何运作的：

假设训练集只有一个样本  $(\mathbf{X}, \mathbf{Y})$ ，使用四层神经网络进行训练，其中  $k = 4$ ,  $S_L = 4$  (如图13.8)，激活函数为 sigmoid 函数，损失函数为式13.3.5，该网络前向传播算法为：

一般从最后一层的误差开始计算，定义  $\delta^{(i)} = \frac{\partial J(W)}{\partial z^i}$  来表示误差，则  $\delta^{(4)} = \frac{\partial J(W)}{\partial z^4} = (a^{(4)} - \mathbf{Y}) * f'(z^{(4)})$ ，在利用这个误差来计算前一层的误差  $\delta^{(3)} = (W^{(3)})^T \delta^{(4)} * f'(z^{(3)})$ ，其中 \* 表示矩阵对应元素相乘， $f'(z^{(3)})$  是激活函数的导数， $f'(z^{(3)}) = a^{(3)} * (1 - a^{(3)})$ 。而  $(W^{(3)})^T \delta^{(4)}$  则是权重导致的误差和。下一步是计算第二层的误差  $\delta^{(2)} = (W^{(2)})^T \delta^{(3)} * f'(z^{(2)})$ 。第一层是输入变量，不存在误差。

有了所有误差的表达式后，便可以计算损失函数的偏导数了，假设  $\lambda = 0$ ，即不做任何正则化处理时有：

$$\frac{\partial J(W)}{\partial W_{ji}^{(l)}} = \frac{\partial J(W)}{\partial z^{(l+1)}} \frac{\partial z^{(l+1)}}{\partial W_{ji}^{(l)}} = a_i^{(l)} \delta_i^{(l+1)}$$

其中， $a^l = (a_1^l, a_2^l, \dots, a_k^l)$  且  $\delta^l = (\delta_1^l, \delta_2^l, \dots, \delta_k^l)$

BP 算法的学习过程实际上由正向传播过程和反向传播过程组成。在正向传播过程中，输入信息通过输入层经隐含层，逐层处理并传向输出层。如果在输出层得不到

**Algorithm 1**

$$a^{(1)} = \mathbf{X} \quad (13.3.8)$$

$$z^{(2)} = W^{(1)} a^{(1)} \quad (13.3.9)$$

$$a^{(2)} = f(z^{(2)}) \left( \text{add } a_0^{(2)} \right) \quad (13.3.10)$$

$$z^{(3)} = W^{(2)} a^{(2)} \quad (13.3.11)$$

$$a^{(3)} = f(z^{(3)}) \left( \text{add } a_0^{(3)} \right) \quad (13.3.12)$$

$$z^{(4)} = W^{(3)} a^{(3)} \quad (13.3.13)$$

$$a^{(4)} = f(z^{(4)}) \quad (13.3.14)$$

期望的输出值，则取输出与期望的误差的平方和作为目标函数，转入反向传播，逐层求出目标函数对各神经元权值的偏导数，构成目标函数对权值向量的梯量，作为修改权值的依据，网络的学习在权值修改过程中完成。误差达到所期望值时，网络学习结束。

### 13.3.4 全局最小与局部极小

若用  $J$  表示神经网络在训练集上的误差，则它显然是关于连接权值和偏置的函数。此时，神经网络的训练过程可看作一个参数寻优过程，即在参数空间中，寻找一组最优参数使得  $J$  最小。

我们常会谈到两种“最优”：“局部极小”(Local Minimum) 和“全局最小”(Global Minimum)。对  $\omega^*$  和  $\theta^*$ ，若存在  $\epsilon > 0$  使得

$$\forall (\omega; \theta) \in \{(\omega; \theta) \mid \|(\omega; \theta) - (\omega^*; \theta^*)\| \leq \epsilon\}$$

都有  $J(\omega; \theta) \geq J(\omega^*; \theta^*)$  成立，则  $(\omega^*; \theta^*)$  为**局部极小解**；若对参数空间中的任意  $(\omega; \theta)$  都有  $J(\omega; \theta) \geq J(\omega^*; \theta^*)$ ，则  $(\omega^*; \theta^*)$  为**全局最小解**。直观地看，局部极小解是参数空间中的某个点，其邻域点的误差函数值均不小于该点的函数值；全局最小解则是指参数空间中所有点的误差函数值均不小于该点的误差函数值。两者对应的分别称为误差函数的**局部极小值**和**全局最小值**。

显然，参数空间内梯度为零的点，只要其误差函数值小于邻点的误差函数值，就是局部极小点；可能存在多个局部极小值，但却只会有一个全局最小值。也就是说，“全局最小”一定是“局部极小”，反之则不成立。例如，图13.9中有两个局部极小，但只有其中之一是全局最小。显然，在参数寻优过程中是希望找到全局最小。

基于梯度的搜索是使用最为广泛的参数寻优方法。在此类方法中，我们从某些初始解出发，迭代寻找最优参数值。每次迭代中，先计算误差函数在当前点的梯度，然后根据梯度确定搜索方向。例如，由于负梯度方向是函数值下降最快的方向，因此梯度下降法就是沿着负梯度方向搜索最优解。若误差函数在当前点的梯度为零，则已达到局部极小，更新量将为零，这意味着参数的迭代更新将在此停止。显然，如果误差函数仅有一个局部极小，那么此时找到的局部极小就是全局最小；然而，如果误差函数具有多个局部极小，则不能保证找到的解是全局最小。对后一种情形，称参数寻优陷入了局部极小，这显然不是我们所希望的。

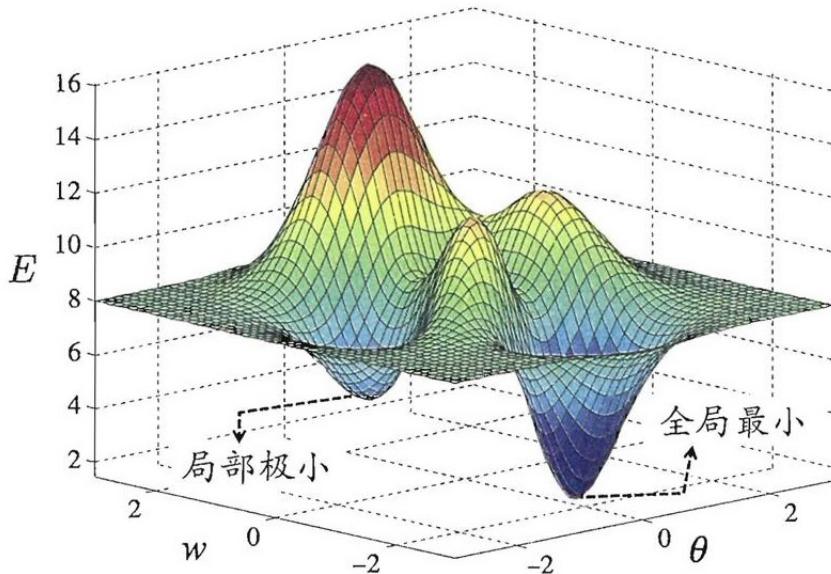


图 13.9 全局最小与局部极小

在现实任务中，常采用以下策略来试图“跳出”局部极小，从而进一步接近全局最小：

1、以多组不同参数值初始化多个神经网络，按标准方法训练后，取其中误差最小的解作为最终参数。这相当于从多个不同的初始点开始搜索，这样就可能陷入不同的局部极小，从中进行选择有可能获得更接近全局最小的结果。

2、使用“[模拟退火](#)”（SA，Simulated Annealing）技术 [142]。模拟退火在每一步都以一定的概率接受比当前解更差的结果，从而有助于“跳出”局部极小。在每步迭代过程中，接受“次优解”的概率要随着时间的推移而逐渐降低，从而保证算法稳定。

3、使用随机梯度下降。与标准梯度下降法精确计算梯度不同，随机梯度下降法在计算梯度时加入了随机因素。于是，即便陷入局部极小点，它计算出的梯度仍可能不为零，这样就有机会跳出局部极小继续搜索。

此外，[遗传算法](#)（GA，Genetic Algorithms）[143] 也常用来训练神经网络以更好地逼近全局最小。需注意的是，上述用于跳出局部极小的技术大多是启发式智能算法，理论上尚缺乏保障。

## 13.4 径向基网络

[径向基函数神经网络](#)（简称为径向基（RBF，Radial Basis Function）网络）是一种局部逼近型网络模型。所谓[局部逼近型网络](#)，是指网络输出仅与少数几个连接权重相关，对于每个参与模型训练的样本，通常仅有少数与其相关的权重需要进行更新。这种局部性的参数更新方式有利于加速模型的训练。

机器学习中回归任务的本质是根据已知离散数据集求解与之相符的连续函数，基本求解思路是对已知的离散数据进行拟合，使得拟合函数与已知离散数据的误差在某种度量意义下达到最小。RBF 网络对于此类问题的求解思路则是通过对已知离散数据进行插值的方式确定网络模型参数。

RBF 神经网络一共分为三层，如图13.10所示。第一层为输入层，由信号源节点组成；第二层为隐藏层，隐藏层中神经元的激活函数，即[径向基函数](#)是对中心点径向对称且衰减的非负线性函数，该函数是局部响应函数。局部响应函数一般要根据具体问题设置相应的隐藏层神经元个数；第三层为输出层，是对输入模式做出的响应，输出层根据线性权重进行调整，采用的是线性优化策略，因而学习速度相对较快。该网络将径向基函数作为隐单元的“基”构成隐含层空间，将输入直接映射到隐空间。

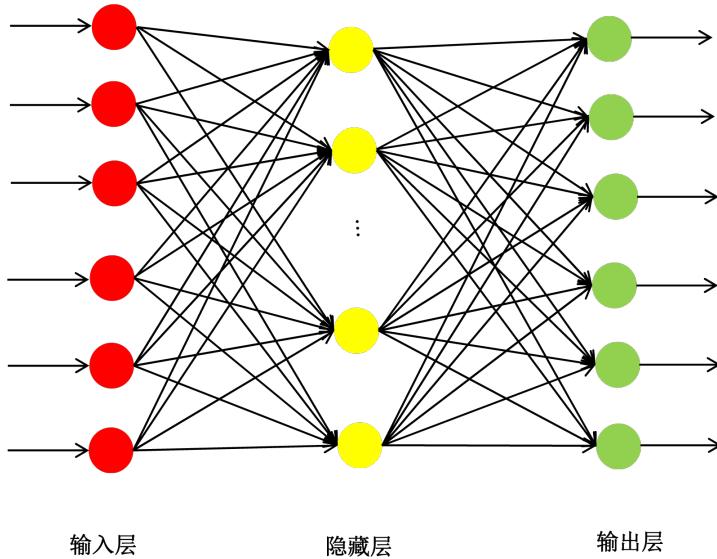


图 13.10 RBF 神经网络模型

RBF 网络的基本思想是：用 RBF 作为隐单元的“基”构成隐含层空间，这样就可以将输入直接映射到隐空间，而不需要通过“权”连接。当 RBF 的中心点确定以后，这种映射关系也就确定了。而隐含层空间到输出空间的映射是线性的，即网络的输出是隐单元输出的线性加权和，此处的权即为网络可调参数。隐含层的作用是把向量从低维度映射到高维度，这样低维度线性不可分的情况到高维度就可以变得线性可分了，这即是核函数的思想。这样网络由输入到输出的映射是非线性的，而网络输出对可调参数而言却又是线性的。网络的权就可由线性方程组直接解出，从而大大加快学习速度并避免局部极小问题。

对于给定训练样本集  $D = \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_m, Y_m)\}$ ，其中

$$\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$$

为输入向量， $Y_i$  为  $\mathbf{X}_i$  所对应的数值型真实值，插值的目标是找到某个函数  $f$  使得：

$$f(\mathbf{X}_i) = Y_i \quad i = 1, \dots, m \quad (13.4.1)$$

通常  $f$  为非线性函数，故其函数图像为一个插值曲面，该曲面经过数据集  $D$  中所有样本点。为求解函数  $f$ ，可考虑选择  $m$  个与样本点对应的基函数  $\Phi_1, \Phi_2, \dots, \Phi_m$ ，并将  $f$  进行近似表示为这组基函数的线性组合，即有：

$$f(\mathbf{X}) = \sum_{i=1}^m w_i \Phi_i \quad (13.4.2)$$

只需确定上式中权重参数  $w_1, w_2, \dots, w_m$  取值，就可确定插值函数  $f$  的具体形

式，进而将上述插值问题转换为如下非线性模型的参数问题：

$$\Phi_i = \Phi(\|\mathbf{X} - \mathbf{X}_i\|) \quad (13.4.3)$$

其中  $\Phi_i$  为非线性函数，其函数自变量为输入数据  $\mathbf{X}$  到中心点  $\mathbf{X}_i$  的距离。

由于从中心点  $\mathbf{X}_i$  到相同半径的球面上任意点的距离相等，即距离具有径向相同性，故通常将这种以距离作为自变量的基函数，称为径向基函数。将径向基函数的具体形式带入式中，则可得到插值函数  $f$  如下形式：

$$f(\mathbf{X}) = \sum_{i=1}^m w_i \Phi(\|\mathbf{X} - \mathbf{X}_i\|) \quad (13.4.4)$$

将数据集  $D$  中任意样本  $(\mathbf{X}_j, Y_j)$  带入可得：

$$Y_j = \sum_{i=1}^m w_i \Phi(\|\mathbf{X}_j - \mathbf{X}_i\|) \quad (13.4.5)$$

将数据集  $D$  中数据均代入式 13.4.5 可得到由  $m$  个线性方程组成的线性方程组。记径向基函数的取值  $\Phi(\|\mathbf{X}_j - \mathbf{X}_i\|)$  为  $\Phi_{ij}$ ，则可将该线性方程组表示为如下矩阵形式：

$$\Phi \mathbf{w} = \mathbf{Y} \quad (13.4.6)$$

其中  $\Phi = (\Phi_{ij})_{n \times n}$ ,  $\mathbf{w} = (w_1, w_2, \dots, w_n)^T$ ,  $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ 。

径向基函数可以有多种选择，如高斯径向基函数、反演  $S$  型径向基函数等。

(1) 高斯径向基函数：

$$G(\mathbf{X}, \mathbf{X}_i) = \exp\left(-\frac{1}{2\delta_i^2}\|\mathbf{X} - \mathbf{X}_i\|^2\right) \quad (13.4.7)$$

(2) 反演  $S$  型径向基函数：

$$R(\mathbf{X}, \mathbf{X}_i) = \frac{1}{1 + \exp\left(\frac{\|\mathbf{X} - \mathbf{X}_i\|^2}{\delta_i^2}\right)} \quad (13.4.8)$$

其中  $\delta_i$  为扩展常数，其取值越大，数据分布范围越宽。

在实际应用中，高斯径向基函数表示为：

$$G(\mathbf{X}, c_i) = \exp\left(-\beta_i \|\mathbf{X} - c_i\|^2\right)$$

其中  $c_i$  称为中心。通常采用两步过程来训练该网络：第一步，确定神经元中心  $c_i$ ，常用的方式包括随机采样、聚类等；第二步，利用 BP 算法等来确定参数  $w_i$  和  $\beta_i$ 。

## 13.5 其它常见的神经网络

神经网络模型中包含了大量的不同模型，就算相同模型也可以通过改变网络的结构变得不同，因此本书不能详尽的列举所有的模型，所以选择了最常见的几种模型进行了简单的介绍。

### ART 网络模型

**竞争型学习** (Competitive Learning) 是神经网络中一种常用的无监督学习策略，在使用该策略时，网络的输出神经元相互竞争，每一时刻仅有一个竞争获胜的神经元被激活，其他神经元的状态被抑制。这种机制亦称“胜者通吃” (Winner-take-all) 原则。

**自适应谐振理论 (ART, Adaptive Resonance Theory) 网络** [144] 是竞争型学习的重要代表。该网络由比较层、识别层、识别阈值和重置模块构成。其中，比较层负责接收输入样本，并将其传递给识别层神经元。识别层每个神经元对应一个模式类，神经元数目可在训练过程中动态增长以增加新的模式类。

在接收到比较层的输入信号后，识别神经元之间相互竞争以产生获胜神经元。竞争的最简单方式是，计算输入向量与每个识别层神经元所对应的模式类的代表向量之间的距离，距离最小者胜。获胜神经元将向其他识别层神经元发送信号，抑制其激活。若输入向量与获胜神经元所对应的代表向量之间的相似度大于识别阈值，则当前输入样本将被归为该代表向量所属类别，同时，网络连接权将会更新，使得以后在接收到相似输入样本时该模式类会计算出更大的相似度，从而使该获胜神经元有更大可能获胜；若相似度不大于识别阈值，则重置模块将在识别层增设一个新的神经元，其代表向量就设置为当前输入向量。

ART 比较好地缓解了竞争型学习中的“**可塑性-稳定性窘境**” (Stability-plasticity Dilemma)，可塑性是指神经网络要有学习新知识的能力，而稳定性则是指神经网络在学习新知识时要保持对旧知识的记忆。这就使得 ART 网络具有一个很重要的优点：可进行增量学习 (Incremental Learning) 或在线学习 (Onlinelearning)。早期的 ART 网络只能处理布尔型输入数据，此后 ART 发展成了一个算法族，包括能处理包括实值输入的 ART2 网络、结合模糊处理的 FuzzyART 网络，以及可进行监督学习的 ARTMAP 网络等。

### 自组织映射网络模型

**自组织映射 (SOM, Self-Organizing Map) 网络** [145] 是一种竞争学习型的无监督神经网络，它能将高维输入数据映射到低维空间（通常为二维），同时保持输入数据在高维空间的拓扑结构，即将高维空间中相似的样本点映射到网络输出层中的邻近神经元。

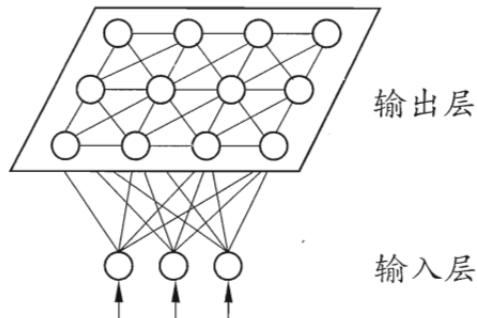


图 13.11 SOM 网络结构

如图13.11所示，SOM 网络中的输出层神经元以矩阵方式排列在二维空间中，每个神经元都拥有一个权向量，网络在接收输入向量后，将会确定输出层获胜神经元，它决定了该输入向量在低维空间中的位置。SOM 的训练目标就是为每个输出层神经元找到合适的权向量，以达到保持拓扑结构的目的。

SOM 的训练过程很简单：在接收到一个训练样本后，每个输出层神经元会计算该样本与自身携带的权向量之间的距离，距离最近的神经元成为竞争获胜者，称为**最佳匹配单元**（Best Matching Unit）。然后，最佳匹配单元及其邻近神经元的权向量将被调整，以使得这些权向量与当前输入样本的距离缩小。这个过程不断迭代，直至收敛。

## 级联相关网络

一般的神经网络模型通常假定网络结构是事先固定的，训练的目的是利用训练样本来确定合适的连接权、阈值等参数。与此不同，**结构自适应网络**则将网络结构也当作学习的目标之一，并希望能在训练过程中找到最符合数据特点的网络结构。**级联相关（Cascade-Correlation）网络** [146] 是结构自适应网络的重要代表。

级联相关网络有两个主要成分：“级联”和“相关”。**级联**是指建立层次连接的层级结构。在开始训练时，网络只有输入层和输出层，处于最小拓扑结构；随着训练的进行，如图13.12所示，新的隐层神经元逐渐加入，从而创建起层级结构。当新的隐层神经元加入时，其输入端连接权值是冻结固定的。**相关**是指通过最大化新神经元的输出与网络误差之间的相关性（Correlation）来训练相关的参数。

与一般的前馈神经网络相比，级联相关网络无需设置网络层数、隐层神经元数目，且训练速度较快，但其在数据较小时易陷入过拟合。

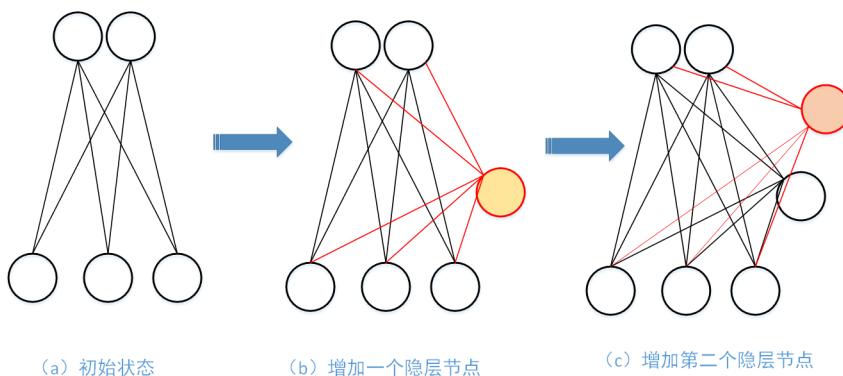


图 13.12 级联相关网络的训练过程

### Elman 网络

与前馈神经网络不同，“**递归神经网络**”（Recurrent Neural Networks）允许网络中出现环形结构，从而可让一些神经元的输出反馈回来作为输入信号。这样的结构与信息反馈过程，使得网络在  $t$  时刻的输出状态不仅与  $t$  时刻的输入有关，还与  $t-1$  时刻的网络状态有关，从而能处理与时间有关的动态变化。

**Elman 网络** [147] 是最常用的递归神经网络之一，其结构如图13.12所示，它的结构与多层前馈网络很相似，但隐层神经元的输出被反馈回来，与下一时刻输入层神经元提供的信号一起，作为隐层神经元在下一时刻的输入。隐层神经元通常采用 sigmoid 激活函数，而网络的训练则常通过推广的 BP 算法进行 [141]。

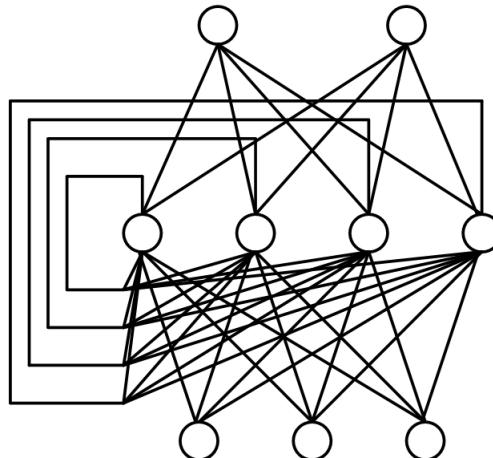


图 13.13 Elman 网络结构

## 13.6 神经网络实践

本小节采用 Iris 数据集。可通过花萼长度，花萼宽度，花瓣长度，花瓣宽度 4 个属性预测鸢尾花卉属于 (Setosa, Versicolour, Virginica) 三个种类中的哪一类。为了更好显示计算结果，设置 2 个隐层，分别是 5 个单元和 4 个单元。基于 R 语言实现 BP 神经网络构建，具体实现代码如下：

```
library(neuralnet)
data("iris")
index <- sample(x = 2, size = nrow(iris), replace=TRUE, prob = c(0.7,0.3))
#按比例抽样形成训练集和测试集
trainset<-iris[index==1,]
testset<-iris[index==2,]
#为训练集新增 versicolor, virginica, setosa 数据列 (类型编码 100,010,001)
trainset$setosa<-trainset$Species=="setosa"
trainset$virginica<-trainset$Species=="virginica"
trainset$versicolor<-trainset$Species=="versicolor"
#构建网络结构
net_model=neuralnet(setosa+versicolor+virginica~
Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, #输入输出关系公式
data = trainset, #训练集
hidden=c(5,4)) #隐藏层神经元数量
print(net_model) #输出模型
plot(net_model)
test_output<-compute(net_model,covariate=testset) #预测模型
result_output=as.data.frame(test_output$net.result)
cid=apply(result_output,1,which.max) #得到每行中最大值所在列
results=c("setosa","versicolor","virginica")[cid] #使用品种名替换最大值
table(results,testset$Species) #使用混淆矩阵判断结果
#输出结果为：
results      setosa versicolor virginica
setosa        17       0       0
versicolor     0      14       0
virginica      0       1      13
```

图13.14给出了神经网络的结构，包括权重。另外，从混淆矩阵结果上看，只有一个分类错误，正确率还是很高的。

### 13.6.2 Python 语言实践

#### 多层感知器网络构建

本小节同样采用 Iris 数据集预测鸢尾花卉种类，首先进行数据导入和预处理。

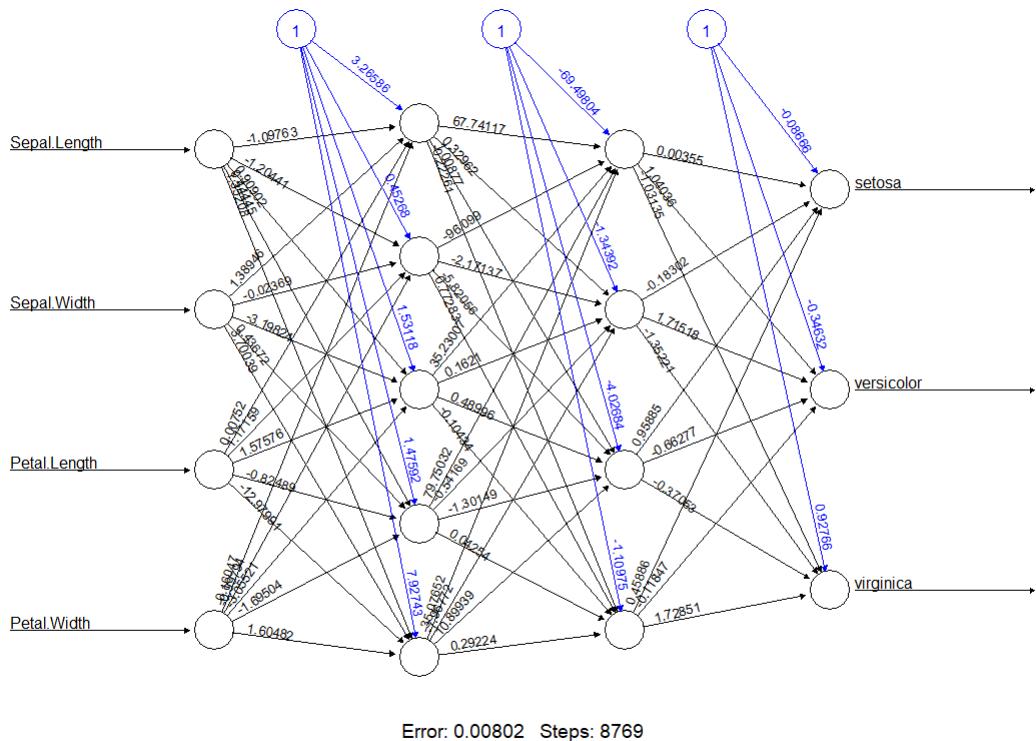


图 13.14 神经网络结构

```

import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
df_Iris = pd.read_csv('data\iris.csv')
X = df_Iris[['sepal_length','sepal_width','petal_length','petal_width']];
y = df_Iris['species'];
#将数据按照8:2的比例随机分为训练集、测试集
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

```

再建立多层感知器模型，设置隐含层为两层，第一层 5 个神经元，第二层 4 个神经元。

```

dt = MLPClassifier(solver='lbfgs', alpha=1e-5, hidden_layer_sizes=(5,4),
random_state=1)
#采用L-BFGS方法求解MLP; alpha为L2参数; 隐含层有2层, 第一层5个神经元, 第二层4个神经元
dt.fit(X_train, y_train)#训练模型
cengindex = 0
for wi in dt.coefs_:
    cengindex += 1
    print('第%d层网络层:' % cengindex)
    print('权重矩阵维度:',wi.shape)
    print('系数矩阵:\n',wi)

```

```
dt.score(X_test, y_test) #用测试集评估模型的好坏
```

## BP 神经网络构建

本小节以常见的 BP 神经网络为例，讲述神经网络的 python 实现方法。

我们使用的案例是 1974 个化合物的小肠上皮细胞渗透性数据，其自变量是 729 个分子描述符，因变量为小肠上皮细胞渗透性，属于分类问题，而我们要做的就是根据已有数据构建一个小肠上皮细胞渗透性预测模型。实现代码如下：首先导入相应包并设置随机种子

```
import pandas as pd
from keras.models import Model, Sequential
from keras.layers import Conv2D, Conv2DTranspose, Input, MaxPooling2D, add, Dropout,
core, Dense, Activation, BatchNormalization
from sklearn.model_selection import train_test_split
from keras.optimizers import Adam
from keras.callbacks import ModelCheckpoint
from sklearn.metrics import accuracy_score
import keras
import numpy as np
from numpy import random
random.seed(1)
```

再导入数据，进行预处理

```
xiangmu = 'Caco-2'
data1 = pd.read_csv("./data/1975-729.csv") #读取数据
data2 = pd.read_csv("./data/1975-5.csv")
data1 = data1.values
y_data = data2[xiangmu]
y_data = y_data.values
X = data1
Y = y_data
#将数据集划分为训练集与测试集，并对因变量进行 Onehot 编码
X_train, X_test, y_train, y_test = train_test_split(X, Y,
test_size=0.1, random_state=1, stratify =Y)
Y_train = keras.utils.to_categorical(y_train, 2)
Y_test = keras.utils.to_categorical(y_test, 2)
```

接下来搭建 BP 神经网络模型

```
model = Sequential() #逐层搭建神经网络模型
model.add(Dense(600, input_dim=729, init='uniform', activation='relu'))
model.add(BatchNormalization())
model.add(Dense(100, input_dim=600, init='uniform', activation='relu'))
model.add(BatchNormalization())
model.add(Dense(50, input_dim=100, init='uniform', activation='relu'))
model.add(BatchNormalization())
model.add(Dense(10, input_dim=50, init='uniform', activation='relu'))
model.add(BatchNormalization())
```

```

model.add(Dense(2, input_dim=10, activation='softmax'))
model.summary() #查看模型摘要
model.compile(loss='categorical_crossentropy', optimizer=Adam(learning_rate=0.1),
metrics=['acc'])#选择类别交叉熵作为损失函数，Adam作为训练的优化器，学习率为0.1
model_checkpoint = ModelCheckpoint('./data/bp-'+xiangmu+'-1.hdf5',
monitor='loss', verbose=1, save_best_only=True)#设定模型的ckpt储存方式

```

最后进行模型拟合并进行预测

```

model.fit(X_train, Y_train, epochs = 1000, batch_size = 1000, callbacks=[model_checkpoint])#训练模型
model.load_weights('./data/bp-'+xiangmu+'-1.hdf5')#读取ckpt文件
y_pred = model.predict(X_test)#进行预测
y_pred = np.argmax(y_pred, axis=1)
acc = accuracy_score(y_test, y_pred) * 100#对预测值的精度进行估计
print("\nTesting Accuracy: {:.3f} %".format(acc))

```

## 13.7 总结

本章主要讨论了前馈神经网络的数学模型、正向和反向计算方法，RBF 神经网络的数学模型和计算方法等，并对其它神经网络形式做了简单介绍。但在理论和实践中是针对分类问题进行讲解和编程实现的。对于回归问题，限于篇幅，并未涉及，读者可参考马丁 T. 哈根的书《神经网络设计（原书第 2 版）》进行学习。

## 13.8 练习

- 1、请推导 BP 算法 2 的过程，并用 Python 或 R 语言编程实现。
- 2、根据 RBF 神经网络计算原理，请使用 Python 或 R 语言编写程序实现。
- 3、在基于 Python 中 sklearn 的多层感知器网络构建中，请对参数进行消融实验，探索参数的影响情况和最佳组合。若使用 R 语言，完成同样的工作。
- 4、在基于 Python 中 sklearn 的多层感知器网络构建中，请使用前面章节所讲过的决策树、贝叶斯、支持向量机等模型进行方法对比实验，比较各模型对于此问题解决的效果。若使用 R 语言，完成同样的工作。

## 第十四章 深度学习

理论上来说，参数越多的模型复杂度越高、容量（Capacity）越大，这意味着它能完成越复杂的学习任务。但一般情形下，复杂模型的训练效率低，易陷入过拟合，因此难以受到人们青睐。而随着云计算、大数据时代的到来，计算能力的大幅提高可缓解训练低效性，训练数据的大幅增加则可降低过拟合风险，因此，以深度学习（Deep Learning）为代表的复杂模型开始受到人们的关注。

典型的深度学习模型就是有许多层的神经网络。显然，对神经网络模型，提高容量的一个简单办法是增加隐层的数目。隐层多了，相应的神经元连接权、阈值等参数就会更多。模型复杂度也可通过单纯增加隐层神经元的数目来实现，从上一章的学习中可以了解到，单隐层的多层前馈网络已具有很强大的学习能力；但从增加模型复杂度的角度来看，增加隐层的数目显然比增加隐层神经元的数目更有效，因为增加隐层数不仅增加了拥有激活函数的神经元数目，还增加了激活函数嵌套的层数。然而，多隐层神经网络难以直接用经典算法（例如标准 BP 算法）进行训练，因为误差在多隐层内反向传播时，往往发散而不能收敛到稳定状态。

无监督逐层训练（Unsupervised Layer-wise Training）是多隐层网络训练的有效手段，其基本思想是每次训练一层隐节点，训练时将上一层隐节点的输出作为输入，而本层隐节点的输出作为下一层隐节点的输入，这称为预训练（Pre-training）；在预训练全部完成后，再对整个网络进行微调（Fine-tuning）训练。在使用无监督逐层训练时，首先训练第一层，这是关于训练样本的受限玻尔兹曼机（RBM, Restricted Boltzmann machine）模型，可按标准的 RBF 训练；然后，将第一层预训练好的隐结点视为第二层的输入结点，对第二层进行预训练；……。各层预训练完成后，再利用 BP 算法等对整个网络进行训练。

事实上，“预训练 + 微调”的做法可视为将大量参数分组，对每组先找到局部看来比较好的设置，然后再基于这些局部较优的结果联合起来进行全局寻优。这样就在利用了模型大量参数所提供的自由度的同时，有效地节省了训练开销。

另一种节省训练开销的策略是权共享（Weight Sharing），即让一组神经元使用相同的连接权。这个策略在卷积神经网络（CNN, Convolutional Neural Network）[148] [149] 中发挥了重要作用。

以 CNN 进行手写数字识别任务为例 [149]，如图14.1所示，网络输入是一个  $32 \times 32$  的手写数字图像，输出是其识别结果，CNN 复合多个“卷积层”和“采样层”对输入信号进行加工，然后在连接层实现与输出目标之间的映射。每个卷积层都包含多个特征映射，每个特征映射（Feature Map）是一个由多个神经元构成的“平面”，通过一种卷积滤波器提取输入的一种特征。例如，图14.1中第一个卷积层由 6

个特征映射构成，每个特征映射是一个  $28 \times 28$  的神经元阵列，其中每个神经元负责从  $5 \times 5$  的区域通过卷积滤波器提取局部特征。采样层亦称为**池化层** (Pooling)，其作用是基于局部相关性原理进行亚采样，从而在减少数据量的同时保留有用信息。例如图14.1中第一个采样层有 6 个  $14 \times 14$  的特征映射，其中每个神经元与上一层中对应特征映射的  $2 \times 2$  邻域相连，并据此计算输出。通过复合卷积层和采样层，图14.1中的 CNN 将原始图像映射成 120 维特征向量，最后通过一个由 84 个神经元构成的连接层和输出层连接完成识别任务。CNN 可用 BP 算法进行训练，但在训练中，无论是卷积层还是采样层，其每一组神经元（即图14.1中的每个“平面”）都是用相同的连接权，从而大幅减少了需要训练的参数数目。

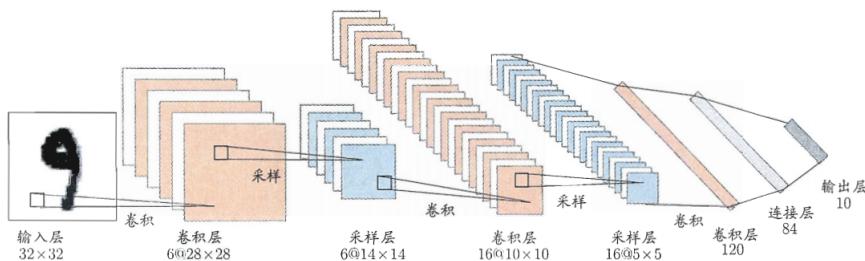


图 14.1 卷积神经网络用于手写数字识别 [149]

也可以从另一个角度来理解深度学习。无论是**深度信念网络** (DBN, Deep Belief Network) 还是 CNN，其多隐层堆叠、每层对上一层的输出进行处理的机制，可看作是在对输入信号进行逐层加工，从而把初始的、与输出目标之间联系不太密切的输入表示，转化成与输出目标联系更密切的表示，使得原来仅基于最后一层输出映射难以完成的任务成为可能。换言之，通过多层处理，逐渐将初始的“低层”特征表示转化为“高层”特征表示后，用“简单模型”即可完成复杂的分类等学习任务。由此可将深度学习理解为进行**特征学习** (Feature Learning) 或**表示学习** (Representation Learning)。

以往在机器学习用于现实任务时，描述样本的特征通常需由人类专家来设计，这称为**特征工程** (Feature Engineering)。众所周知，特征的好坏对泛化性能有至关重要的影响，人类专家设计出好特征也并非易事；特征学习则通过机器学习技术自身来产生好特征，这使机器学习向“全自动数据分析”又前进了一步。

## 14.1 卷积神经网络

神经网络在得到广泛应用的同时，参数过多、容易发生过拟合和训练时间长等缺点也暴露出来。能否减少神经网络中参数的数目，并进一步提升神经网络的性能？

CNN 应运而生。

**卷积神经网络** (CNN, Convolutional Neural Network) 又称为卷积网络，是在图像处理和计算机视觉领域应用较为广泛的一种神经网络。与全连接神经网络同属于神经网络模型，相对于全连接神经网络而言，卷积神经网络的不同之处在于加入了卷积层 (Convolution) 和池化层 (Pooling) 两种结构，这两种结构是 CNN 必不可少的结构。前面已经介绍了它们基本的计算原理。

CNN 最早可以追溯到 1986 年 BP 算法的提出 [150]，1989 年 LeCun [151] 将 BP 算法用到多层神经网络中，1998 年 LeCun [149] 提出 LeNet-5 模型，卷积神经网络的雏形完成。在接下来近十年的时间里，卷积神经网络的相关研究趋于停滞，原因有两个：一是研究人员意识到多层神经网络在进行 BP 训练时的计算量极其之大，当时的硬件计算能力完全不可能实现；二是包括支持向量机在内的浅层机器学习算法也渐渐开始展露头角。直到 2006 年，Hinton [152] 在 Science 发文，指出“多隐层神经网络具有更为优异的特征学习能力，并且其在训练上的复杂度可以通过逐层初始化来有效缓解”，深度学习开始觉醒，并逐渐走入人们的视线。2009 年，李飞飞牵头建立了 ImageNet 数据集，该数据集包含图片的种类和数量远远超过以往所有的数据集。自 2010 年以来，每年都举行 ImageNet 大规模视觉识别挑战赛 (ILSVRC)，研究团队在给定的数据集上评估其算法，并在几项视觉识别任务中争夺更高的准确性。在刚开始的两年，冠军被支持向量机取得，而在 2012 年，AlexNet 取得了比赛的冠军，识别的正确率得到明显提升，这之后占据高位的一直是卷积神经网络，包括我们现在所熟知的 VGGNet，GoogLeNet，ResNet 等。2016 年，AlphaGo 战胜了围棋世界冠军、职业九段棋手李在石，更是将深度学习和卷积神经网络推向了高潮。

图14.2展示了近些年来 CNN 的发展历程。

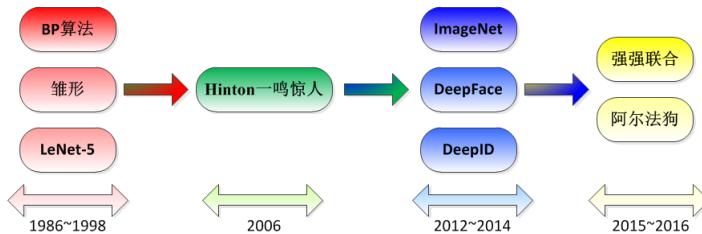


图 14.2 CNN 大事件

### (一) 卷积和卷积神经网络

**卷积** (Convolution) 运算有连续和离散两种定义，一维情形下有如下所示：

$$(f * g)(n) = \int_{-\infty}^{\infty} f(t)g(n-t) dt$$

$$(f * g)(n) = \sum_{t=-\infty}^{\infty} f(t)g(n-t)$$

从公式中可以发现，卷积运算先对  $g$  函数进行翻转，相当于在数轴上把  $g$  函数从右边折到左边去，也就是卷积中的“卷”。然后再把  $g$  函数平移到  $n$ ，在这个位置对两

个函数的对应点相乘，然后相加，这个过程是卷积中的“积”。因此，所谓“卷积”，可直观理解为“卷”和“积”两个过程。

在卷积神经网络中“卷”和“积”如何体现呢？下面通过图14.3说明卷积神经网络的计算过程。

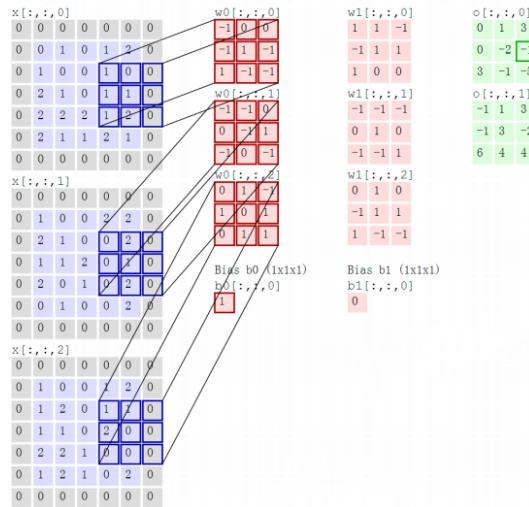


图 14.3 卷积运算

最左侧为输入，中间红色部分表示卷积核，最右侧为输出。卷积核移动到输入左上角开始计算，以此向右、向下滑动。例如，第一个卷积核移动到第一个输入的左上角时，此时计算过程为： $0 \times (-1) + 0 \times 0 + 0 \times 0 + 0 \times (-1) + 0 \times 1 + 1 \times (-1) + 0 \times 1 + 1 \times (-1) + 0 \times (-1) = -2$ 。类似地，可以继续后面的运算。

通过计算的过程可以发现，“卷”并没有发挥作用，卷积神经网络仅仅应用了“积”。卷积神经网络具备两大特点：局部连接和权值共享。**局部连接**指的是卷积层的节点仅仅和其前一层的部分节点相连接；**权值共享**指的是同一张图使用相同的卷积核。局部连接和权值共享减少了参数数量，加快了神经网络的学习速率，同时也在一定程度上减少了过拟合的可能。

## (二) 基本参数

了解了卷积神经网络的基本过程后，接下来学习卷积神经网络中的几个基本参数。

### 1. 填充 (Padding)

在进行卷积操作的时候， $6 \times 6$  的图像经过  $3 \times 3$  的 filter 卷积得到  $4 \times 4$  的卷积结果，如图14.4所示。用更加一般的形式表达：如果我们有一个  $n \times n$  的图像，用  $f \times f$  的过滤器做卷积，那么输出的维度就是  $(n - f + 1) \times (n - f + 1)$ 。在这个例子里是  $6 - 3 + 1 = 4$ ，因此得到了一个  $4 \times 4$  的输出。

但是这样做存在明显缺点：每次卷积操作后图像尺寸都会缩小，同时角落边缘的像素点，在卷积计算的时候只被一个输出所触碰或者使用，但是中间的像素点会

被多次卷积计算，所以那些在角落的像素点参与卷积计算较少，意味着丢掉一些图像边缘位置的信息。

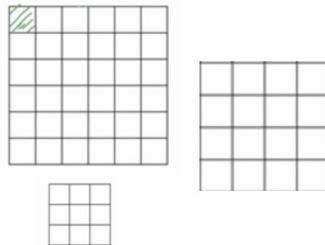


图 14.4 图像由  $6 \times 6$  到  $3 \times 3$

卷积操作中的Padding，可以自行指定  $p$  的值。一般有如下两种 Padding 方式。

- (1) Valid: 不填充，即  $p = 0$ ；
- (2) Same: 填充后输出矩阵的大小与原矩阵保持一致，即  $p = \frac{f - 1}{2}$ ，其中卷积核的大小为  $f \times f$ 。通常将  $f$  设置为奇数，这样我们方便我们对称填充。

## 2、步幅 (Stride)

滑动卷积核时，我们会先从输入的左上角开始，每次往左滑动一列或者往下滑动一行逐一计算输出，我们将每次滑动的行数和列数称为Stride。

## 3、池化 (Pooling)

Pooling是实现下采样的一种运算，能在提取图像关键特征的基础上减小图片尺寸，以减少训练中的参数数量，达到减少计算量、增大感受野并防止过拟合的目的。

如图14.5所示，Pooling主要有最大值池化(Max-Pooling)和平均值池化(Average-Pooling)两种。最大值池化从所选区域的矩阵元素中取最大值，而平均值池化是将所选区域的矩阵元素求和后取平均值。

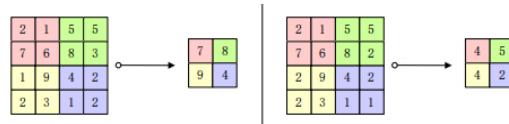


图 14.5 最大值池化和平均值池化示意图

## 4、欠拟合和过拟合

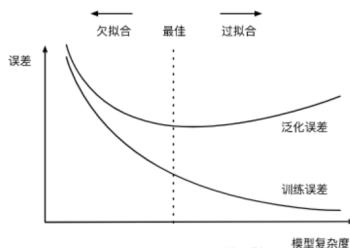


图 14.6 欠拟合和过拟合

### (1) 欠拟合

如图14.6，**欠拟合**是指模型拟合程度不高，数据距离拟合曲线较远，或指模型没有很好地捕捉到数据特征，不能够很好地拟合数据。解决办法是提高模型的复杂度，如增加神经网络的层数、宽度，减少正则化的参数等。

### (2) 过拟合

如图14.6，一个假设在训练数据上能够获得比其他假设更好的拟合，但是在训练数据外的数据集上却不能很好地拟合数据，表现为泛化能力差，称为**过拟合**。导致过拟合的原因有很多，主要包括以下几点：模型过于复杂、参数过多；数据太少；训练集和测试集的数据分布不同；样本里的噪音数据干扰过大，大到模型过分记住了噪音特征，反而忽略了真实的输入输出间的关系。解决办法包括  $L_1$ 、 $L_2$  正则化，扩增数据，Dropout；Early Stopping 等。

### $L_1$ 、 $L_2$ 正则化

在损失函数上添加正则化项，其中  $L_1$  正则化为参数  $w$  的绝对值（岭回归）、 $L_2$  正则化为参数  $w$  的平方值（Lasso 回归）。通过对  $w$  值的修正，使其偏离不会太大，从而减少过拟合的产生。

### 扩增数据

更多的数据集，能够让搭建的网络在更多的数据中不断修正调整，进而训练出更好的模型。然而数据量都是有限的，所以可以从已有数据出发，对其进行调整以得到更多的数据集。如图14.8，可以针对已有图像应用随机图像转换，如旋转、对称和放大等，来增加图像数量。



图 14.7 原图



图 14.8 翻转、镜面、缩放变换

### Dropout

不同于  $L_1$ 、 $L_2$  正则化通过修改代价函数防止过拟合，Dropout 修改的对象是神经网络。如图14.9所示，其思想主要是随机使得神经网络隐藏层中的一些神经元失活，在简化网络结构的同时防止了过拟合，而且随机失活迫使每一个神经元学习到有效的特征，增强了搭建网络的泛化能力。

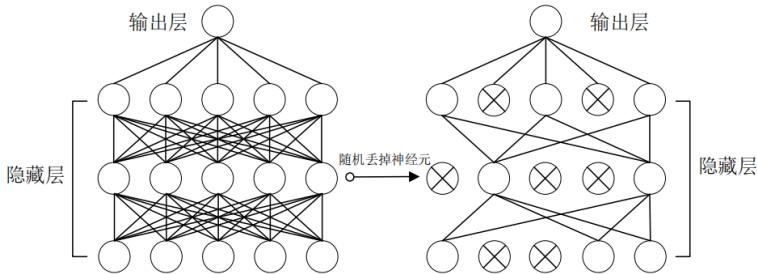


图 14.9 Dropout

在神经网络中使用 Dropout 时，一般以概率 0.5 选择神经元是否有效，即每层仅留下半数的神经元。因此使用 Dropout 相当于训练了多个只有半数隐层神经元的神经网络，每一个这样的网络，都可以给出一个分类结果，这些结果有对有错。随着训练的进行，大部分网络可以给出正确的分类结果，此时少数的错误分类结果不会对最终结果造成大的影响。

### Early Stopping

当训练有足够的表示能力甚至会过拟合的大模型时，经常观察到训练误差会随着时间的推移逐渐降低但验证集的误差会再次上升。这意味着如果返回使验证集误差最低的参数设置，就可以获得更好的模型（有希望获得更好的测试误差）。在每次验证集误差有所改善后，我们存储模型参数的副本。当训练算法终止时，返回这些参数而不是最新的参数。当验证集上的误差在事先指定的循环次数内没有进一步改善时，算法就会终止。需要注意的是，通过提前终止自动选择超参数的显著代价是训练期间要定期评估验证集。

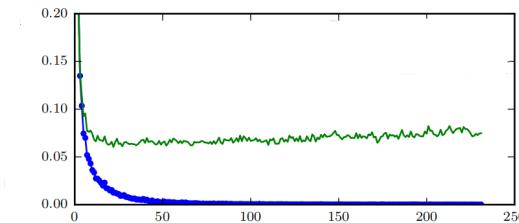


图 14.10 Early Stopping

图14.10中横轴为 epoch 次数，纵轴为损失。蓝线为训练集上的损失，随着 epoch 次数的不断增加明显减少，绿线为验证集上的误差。上述策略称为提前终止（Early Stopping）。

### (三) 卷积神经网络的一般结构

在卷积神经网络中，输入图像通过多个卷积层和池化层进行特征提取，逐步由低层特征变为高层特征；高层特征再经过全连接层和输出层进行特征分类，产生一维向量，表示当前输入图像的类别。因此，根据每层的功能，卷积神经网络可以划分为两个部分：由输入层、卷积层和池化层构成特征提取器，以及由全连接层和输出层构成分类器。

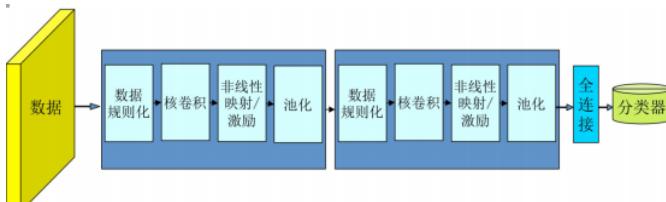


图 14.11 卷积神经网络架构

在实际操作中，卷积神经网络的架构可以如图14.11所示。

## 1、数据规则化：Batch Normalization

随着网络的深度增加，每层特征值分布会逐渐地向激活函数的输出区间的上下两端（激活函数饱和区间）靠近，这样下去会导致梯度消失。[Batch Normalization \(BN\)](#) 通过将该层特征值分布重新拉回标准正态分布，特征值将落在激活函数对于输入较为敏感的区间，输入的小变化可导致损失函数较大的变化，使得梯度变大，避免梯度消失，同时也可加快收敛。

BN 在实际工程中被证明了能够缓解神经网络难以训练的问题。主要来说有如下优点：

- (1) BN 使得神经网络中每层输入数据的分布相对稳定，加速了模型学习速度；
- (2) BN 使得模型对网络中的参数不那么敏感，简化调参过程，使得网络学习更加稳定；
- (3) BN 允许网络使用饱和性激活函数（例如 sigmoid, tanh 等），缓解梯度消失问题；
- (4) BN 具有一定的正则化效果。

## 2、卷积池化层：特征提取

如图14.12，给定图片，如何确定图中包含哪些物体？可以检测图片中的边缘来达到识别的目的。比如说，图片中的栏杆和行人的轮廓线都可以看作是垂线，同样，栏杆就是很明显的水平线，它们也能被检测到。图片中的大部分物体都能用垂直和水平边缘线来刻画，那么如何在图像中检测这些边缘呢？下面通过一个简单例子来说明。

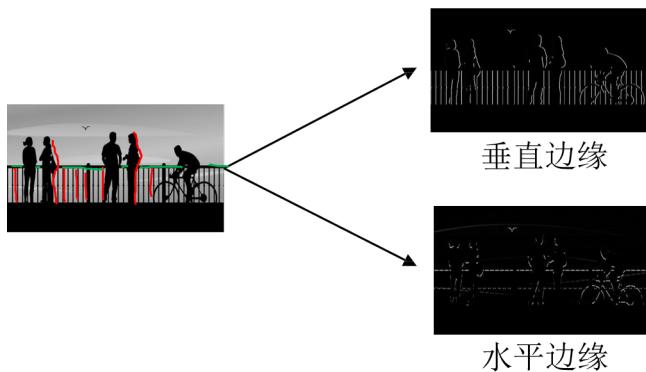


图 14.12 边缘检测

一个  $6 \times 6$  的灰度图，可以表示为左亮右暗的形式，如图14.13左所示。很明显，中间存在明显的分界线，因此我们试图将该垂直边缘识别出来。选择卷积核为  $3 \times 3$ ，相应的输出为  $4 \times 4$ 。仍旧通过颜色的亮暗来表示卷积核和输出矩阵，如图14.13中和图14.13右所示。不难看出，输出矩阵对应的图中间有段发亮的区域，对应原图中的垂直边缘。

$$\begin{array}{|c|c|c|c|c|c|c|} \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 10 & 10 & 10 & 0 & 0 & 0 \\ \hline
 \end{array} * \begin{array}{|c|c|c|} \hline
 1 & 0 & -1 \\ \hline
 1 & 0 & -1 \\ \hline
 1 & 0 & -1 \\ \hline
 \end{array} = \begin{array}{|c|c|c|c|c|c|} \hline
 0 & 30 & 30 & 0 \\ \hline
 0 & 30 & 30 & 0 \\ \hline
 0 & 30 & 30 & 0 \\ \hline
 0 & 30 & 30 & 0 \\ \hline
 \end{array}$$
  

$$\begin{array}{|c|c|} \hline
 \text{white} & \text{gray} \\ \hline
 \end{array} * \begin{array}{|c|c|c|} \hline
 \text{white} & \text{gray} & \text{black} \\ \hline
 \end{array} \quad \begin{array}{|c|c|c|} \hline
 \text{gray} & \text{white} & \text{gray} \\ \hline
 \end{array}$$

图 14.13 卷积提取特征示意图

显然这里检测到的边缘宽度远大于原图中的边缘宽度，这和原图像素大小有关，如果改用像素值为  $500 \times 500$  的图像或者更大，检测效果将会有显著提升。垂直边缘通过这样的方式可以很容易检测出，水平边缘的检测类似。推广到一般情况，卷积核就是通过这样的方式，提取到输入图像的特征。

### 3、全连接层

把分布式特征映射到样本标记空间，即将特征整合到一起输出为一个值，减少特征位置对分类带来的影响。

例如：假设你是一只小蚂蚁，你的任务是找小面包。你的视野比较窄，因此只能看到很小的一片区域。当你找到一片小面包之后，你不确定你找到的是不是全部的小面包，所以你和其余的蚂蚁一块儿开了个会，把所有的小面包都拿出来以确定是否已经找到所有的小面包。某种程度上可以认为，全连接层就是这个蚂蚁大会。需

要注意的是，在数据输入到全连接层前还需要进行拉直操作，将所有数据拉成一维向量。或者可以认为是某种卷积操作。由于全连接层参数过多，正逐渐被全局平均池化（GAP，Global Average Pooling）的方法代替。

#### (四) 常见的卷积神经网络

##### 1、LeNet

LeNet由 LeCun [149] 在 1998 年提出，用于解决手写数字识别的视觉任务。

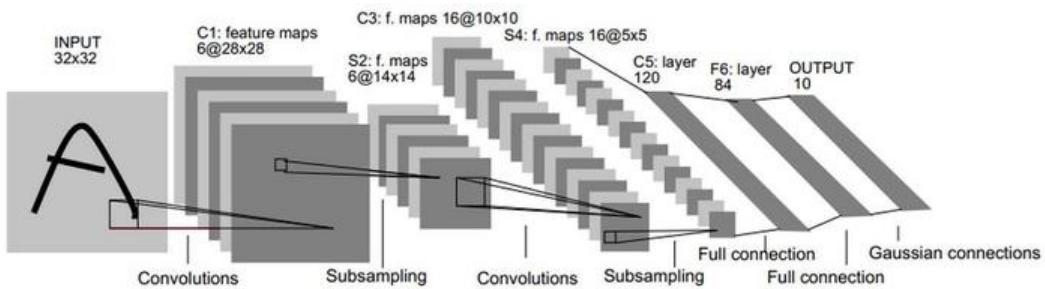


图 14.14 LeNet

图14.14中，Convolutions 表示卷积，Subsampling 表示下采样，即池化，Full Connection 表示全连接，Gaussian Connection 表示高斯连接。具体来看，LeNet 共有五层，包括 2 层卷积层和 3 层全连接层。

LeNet 中每个卷积层均使用尺寸为  $5 \times 5$  的卷积核，步长为 1，并使用 sigmoid 激活函数。两个最池化层中池化核均为  $2 \times 2$ ，且步长为 2，这里选择的是平均池化。由于池化核尺寸与步长相同，因此池化后在输出上每次滑动所覆盖的区域互不重叠，池化不改变通道数量。当卷积层块的输出传入全连接层块时，全连接层块会将小批量中每个样本变平。全连接层块含 3 个全连接层。它们的输出个数分别是 120、84 和 10，其中 10 为输出类别的个数。

##### 2、AlexNet

AlexNet是 2012 年 ImageNet 竞赛的冠军，自此越来越多越来越复杂的神经网络被提出并得到广泛应用。

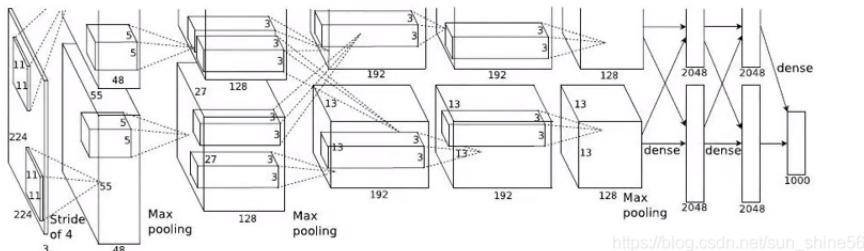


图 14.15 AlexNet

限于单个 GPU 的计算能力, AlexNet 在两个 GPU 上实现卷积神经网络的搭建和计算。图14.15中 Stride 表示步幅, Max Pooling 表示池化选择最大池化, Dense 为全连接。搭建和计算的大致如下:

AlexNet 共有八层, 包括 5 层卷积层和 3 层全连接层。输入尺寸为  $227 \times 227 \times 3$  的彩色图片, 使用  $11 \times 11$  的卷积核, 步长选择为 4, 输出图像的尺寸为  $55 \times 55 \times 96$ , 其中输出图像长(宽)度 =  $\frac{\text{输入图像长(宽)度} + 2 \times \text{Padding} - \text{卷积核长(宽)度}}{\text{步幅}} + 1$ , 因此  $\frac{227 + 2 \times 0 - 11}{4} + 1 = 55$ 。Pooling 选择步长为  $2, 3 \times 3$  的池化核, 输出图像的尺寸变为  $27 \times 27 \times 96$ 。

再次使用  $5 \times 5$  的卷积核, 步长选择为 1, 并且使用 padding=2, 输出得到  $\frac{27 + 2 \times 2 - 5}{1} + 1 = 27$ , 通道数为 256。Pooling 选择步长为  $2, 3 \times 3$  的池化核, 输出图像的尺寸为  $13 \times 13 \times 256$ 。

第三、四、五层均选择  $3 \times 3$  的卷积核, 且步幅均为 1。Pooling 仍旧选择步幅为  $2, 3 \times 3$  的池化核, 输出图像的尺寸变为  $6 \times 6 \times 256$ 。

AlexNet 具有如下四个特点:

- (1) 使用ReLU 函数作为激活函数, 很好地增强了网络的非线性表达能力。
- (2) 引入局部响应归一化 (LRN, Local Response Normalization)。通过 ReLU 函数得到的值域没有固定区间, 因此要对得到的结果进行归一化。具体来说放大对分类贡献较大的神经元, 抑制对分类贡献较小的神经元。
- (3) 使用重叠的最大池化。即池化层中卷积核的尺寸大于步长, 这样池化层的输出间出现重叠和覆盖, 提升了特征的丰富性。
- (4) 通过 Dropout 和数据扩增等方式来防止神经网络出现过拟合现象。

### 3、ZFNet

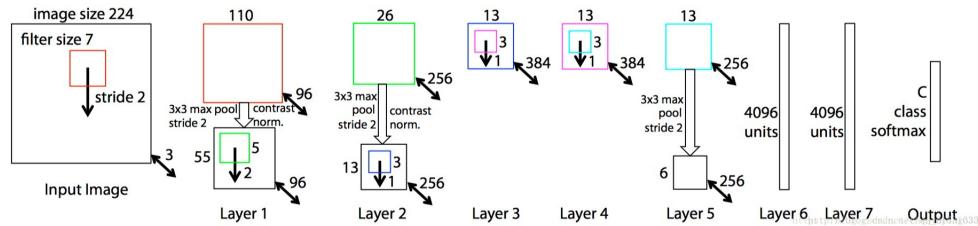


图 14.16 ZFNet

图14.16中 image size 为输出图片的尺寸,  $224 \times 224$ , filter size 为卷积核的大小, stride 为步幅, max pool 表示池化选择最大池化, softmax 为最终用于分类的函数, Layer 标明了 ZFNet 的每一层。ZFNet是 AlexNet 的改进版本, 共有八层, 仍旧是 5 层卷积层和 3 层全连接层。改进主要有以下两点:

(1) 使用一块 GPU 搭建稠密连接结构; (2) 第一个卷积层中卷积核的尺寸从  $11 \times 11$  变为  $7 \times 7$ , 同时步长从 4 减小为 2; 为了让后续输出图像的尺寸与 AlexNet 中的输出图像尺寸保持一致, 第 2 个卷积层的步长从 1 变为 2。

整体来看, ZFNet 的改进似乎并不明显。那么为什么要这样做这样的修改? 修改的动机又是什么呢? 这是基于卷积神经网络深层特征的可视化提出的。可视化操作, 针对的是已经训练好的网络, 或者训练过程中的网络快照。可视化操作不会改变网络的权重, 只是用于分析和理解在给定输入图像时网络观察到了什么样的特征, 以及训练过程中特征发生了什么变化。

可视化主要有如下三个操作:

(1) **Rectification**: 因为使用的是 ReLU 激活函数, 前向传播时只将正值原封不动输出, 负值置 0, “反激活”过程与激活过程没什么分别, 直接将来自上层的输出再次输入到 ReLU 激活函数中即可。

(2) **Unpooling**: 在前向传播时, 记录相应 max pooling 层每个最大值来自的位置, 在 Unpooling 时, 根据来自上层的 map 直接填在相应位置上, 其余位置为 0。

(3) **Transposed Convolution**: 卷积操作输出图像的尺寸一般小于等于输入图像的尺寸, Transposed Convolution 可以将尺寸恢复到与输入相同, 相当于上采样过程, 该操作的做法是, 与 Convolution 共享同样的卷积核, 但需要将其左右上下翻转(即中心对称), 然后作用在来自上层的输出图像进行卷积, 结果继续向下传递。

不断进行上述操作, 可以将特征映射回输入所在的像素空间, 进而可以呈现出人眼可以理解的特征。给定不同的输入图像, 看看每一层关注到最显著的特征是什么。

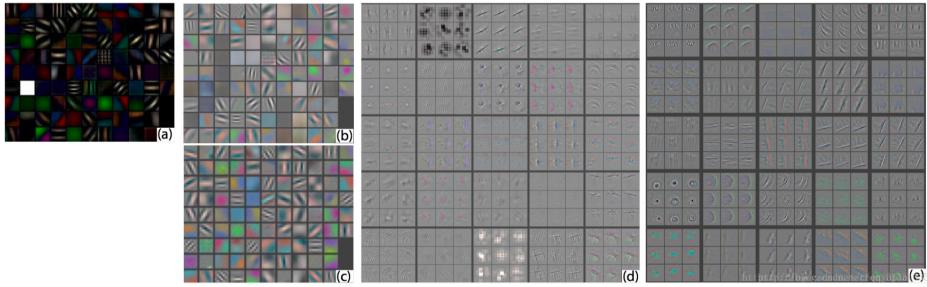


图 14.17 可视化特征图

图14.17 (a) 为没有经过裁剪的图片经过第一个卷积层后的特征可视化图，注意到有一个特征全白，(b) 为 AlexNet 中第一个卷积层特征可视化图，(c) 为 ZFNet 中第一个卷积层可视化图，可以看到相比前面有更多的独特的特征以及更少的无意义的特征，如第 3 列的第 3 到 6 行，(d) 为 AlexNet 中第二个卷积层特征可视化图，(e) 为 ZFNet 中的第二个卷积层特征可视化图，可以看到 (e) 中的特征更加干净，清晰，保留了更多的第一层和第二层中的信息。

通过对 AlexNet 的特征进行可视化，Zeiler 等人发现 AlexNet 第一层中有大量的高频和低频信息的混合，却几乎没有覆盖到中间的频率信息；且第二层中由于第一层卷积用的步长为 4 太大了，导致了有非常多的混叠情况。为了解决这个问题，Zeiler 等人提高采样频率，将步长从 4 调整为 2，与之相应的将卷积核尺寸也缩小（可以认为步长变小了，卷积核没有必要看那么大范围了），修改后第一层呈现了更多更具区分力的特征，第二层的特征也更加清晰 [153]。

#### 4、VGGNet

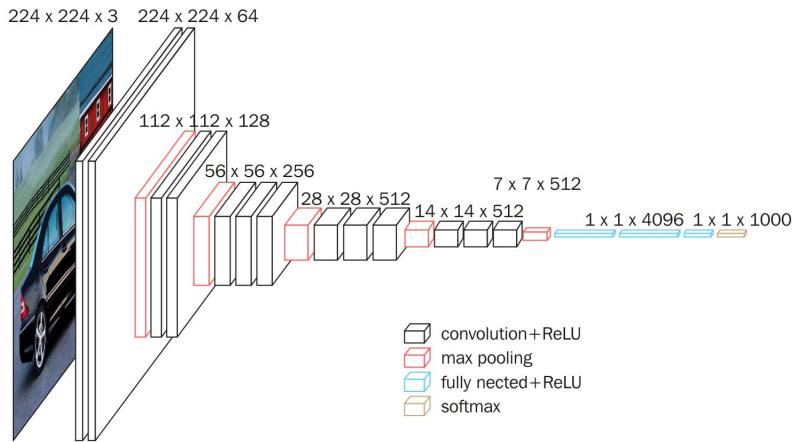


图 14.18 VGGNet

VGGNet (Visual Geometry Group Net)是由牛津大学计算机视觉组合和 Google DeepMind 公司研究员 [154] 一起研发的深度卷积神经网络。它探索了卷积神经网络的深度和其性能之间的关系，通过反复的堆叠  $3 \times 3$  的卷积核和  $2 \times 2$  的最大池化层，

成功地构建了 16 到 19 层的卷积神经网络。因此，VGGNet 是指一系列网络，下面以 VGG-16 为例进行介绍。

图14.18中 convolution+Relu 表示卷积后选择 Relu 激活函数, max pooling 表示池化层选择最大池化, full nected+Relu 表示全连接层后选择 Relu 激活函数, softmax 激活函数用于最后的分类。

VGGNet 是一种专注于构建卷积层的简单网络，相比 AlexNet 和 ZFNet 参数量大大减少，一个很重要的原因是用到了卷积核的堆叠。例如通过 2 个  $3 \times 3$  的卷积核代替 1 个  $5 \times 5$  的卷积核，3 个  $3 \times 3$  的卷积核代替 1 个  $7 \times 7$  的卷积核。通过这样的方式不仅节省了大量参数，还使得网络获得了更大的感受野，同时增强了卷积神经网络的非线性能力。

## 5、GoogLeNet

在通过卷积神经网络识别处理图像时，经常遇到图像突出部分的大小差别很大的情况。如图14.19和14.20，猫的图像可以是以下任意情况。每张图像中猫所占区域是不同的。



图 14.19 不同位置的猫 1



图 14.20 不同位置的猫 2

由于信息位置的巨大差异，为卷积操作选择合适的卷积核大小就比较困难。信息分布更全局性的图像偏好较大的卷积核，信息分布比较局部的图像偏好较小的卷积核。同时，非常深的网络更容易过拟合。将梯度更新传输到整个网络是很困难的。再加上简单地堆叠较大的卷积层非常消耗计算资源，因此考虑在同一层级上运行具备多个尺寸的卷积核呢？网络本质上会变得稍微宽一些，而不是更深。2015 年，串并联网络架构的 GoogLeNet 应运而生 [155]。GoogLeNet 最基本的网络块是 Inception，它是一个并联网络块，经过不断的迭代优化，发展出了 Inception-v1、Inception-v2、Inception-v3、Inception-v4、Inception-ResNet 共 5 个版本。Inception 家族的迭代逻辑是通过结构优化来提升模型泛化能力、降低模型参数。

Inception 就是把多个卷积或池化操作，放在一起组装成一个网络模块，设计神经网络时以模块为单位去组装整个网络结构。在未使用这种方式的网络里，我们一层往往只使用一种操作，比如卷积或者池化，而且卷积操作的卷积核尺寸也是固定大小的。但是，在实际情况下，在不同尺度的图片里，需要不同大小的卷积核，这样才能使性能最好，或者对于同一张图片，不同尺寸的卷积核的表现效果是不一

样的，因为他们的感受野不同。所以，我们希望让网络自己去选择，Inception 便能够满足这样的需求，一个 Inception 模块中并列提供多种卷积核的操作，网络在训练的过程中通过调节参数自己去选择使用，同时，由于网络中都需要池化操作，所以此处也把池化层并列加入网络中。

## 14.2 简单循环神经网络

循环神经网络（RNN，Recurrent Neural Network）是一类以序列（Sequence）数据为输入，在序列的演进方向进行递归（Recursion）且所有节点（循环单元）按链式连接的递归神经网络（Recursive Neural Network）。

在日常生活和学习中，有许多数据的不同样本之间是有关联的，比如气象数据、经济数据，而之前讲述的神经网络并不能很好的学习这一关联，而循环神经网络就能很好的解决这一问题。

要了解 RNN，首先从一个简单的循环神经网络开始（如图14.21）。其中， $\mathbf{X}$  是一个向量，它表示输入层的值（图中并未画出表示神经元节点的圆圈）； $\mathbf{S}$  是一个向量，它表示隐藏层的值； $\mathbf{U}$  是输入层到隐藏层的权重矩阵； $\mathbf{O}$  也是一个向量，它表示输出层的值； $\mathbf{V}$  是隐藏层到输出层的权重矩阵；而权重矩  $\mathbf{W}$  就是隐藏层上一时间的值对这一时间的输入的权重，它表示上一时间隐藏层的值对这一时间的影响。这一简单的循环神经网络与全连接神经网络的不同就在于他多出了权重矩阵  $\mathbf{W}$ 。

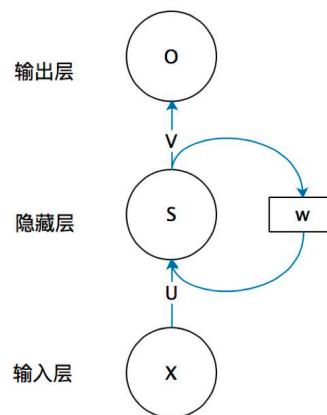


图 14.21 简单的循环神经网络

图14.22展示了这个神经网络展开的样子，可以发现权重矩阵  $\mathbf{W}$  实际上就相当于  $t - 1$  时刻对  $t$  时刻的一个影响。

这个网络在  $t$  时刻接收到输入值为  $\mathbf{X}_t$ ，隐藏层的值是  $\mathbf{S}_t$ ，输出值是  $\mathbf{O}_t$ 。关键一点是， $\mathbf{S}_t$  的值不仅仅取决于  $\mathbf{X}_t$ ，还取决于  $\mathbf{S}_{t-1}$ 。可以用下面的公式来表示循

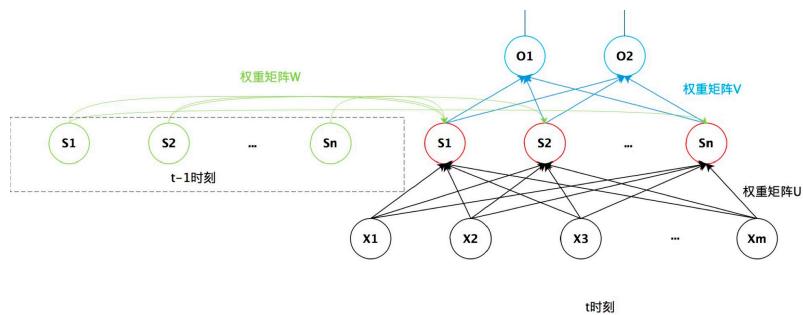


图 14.22 简单的循环神经网络

环神经网络的计算方法：

$$\mathbf{O}_t = g(\mathbf{V} \cdot \mathbf{S}_t)$$

$$\mathbf{S}_t = f(\mathbf{U} \cdot \mathbf{X}_t + \mathbf{W} \cdot \mathbf{S}_{t-1})$$

循环神经网络的参数同样可以利用前文提到的反向传播算法进行求解。但同时，需要注意的是：反向传播算法按照时间逆序将信息一步步向前传递，如果输入序列较长，会存在梯度爆炸与梯度消失问题，也称为**长程依赖问题**，长程依赖问题的具体表现为：很久以前的输入，对当前时刻的网络影响较小，在反向传播时，梯度也很难影响到很久以前的输入。

### 14.3 长短时记忆神经网络

为了解决长程依赖问题，人们将基于图14.23中所展示神经网络进行了进一步改进，引入了**门控机制** (Gating Mechanism)。门 (Gate) 是一种让信息有选择地通过的方式，它由一个 sigmoid 函数和一个点乘运算组成。利用 sigmoid 函数返回 0 到 1 一个之间的数字的特性，令这个数字描述每个元素有多少信息可以通过，0 表示不通过任何信息，1 表示全部通过。

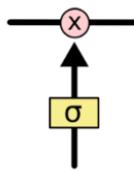


图 14.23 门

门机制控制了对于当前时刻的有哪些信息将被保留，有哪些信息将被遗忘。也将控制有哪些新信息将被加入记忆信息，有哪些信息又将从记忆信息中被提取出作为

当前时刻的输出。下面介绍一个加入了门控机制后的循环神经网络——[长短时记忆神经网络](#) (LSTM, Long Short-Term Memory Neural Network)。LSTM 的结构有很多多种形式，但是都大同小异。这里介绍一种较为流行的结构 GRU (Gated Recurrent Unit)。结构图14.24如下：

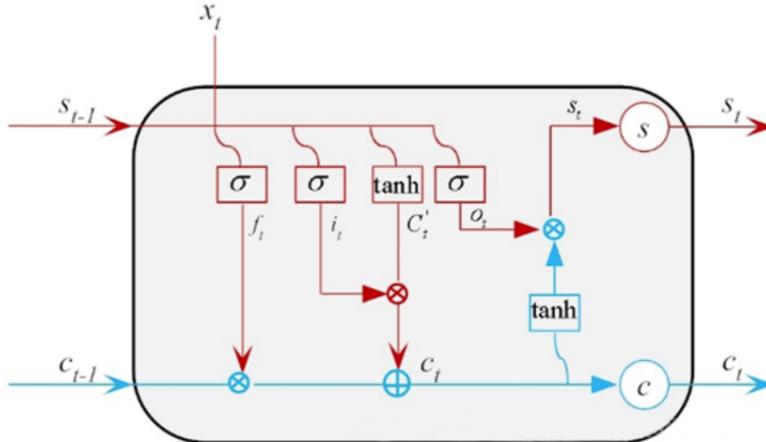


图 14.24 LSTM 完整结构逻辑图

LSTM 模型的关键是引入了一组记忆单元  $c_t$ ，以及三个控制“门”：输入门、输出门、控制门。它们允许网络可以学习何时遗忘历史信息，何时用新信息更新记忆单元。在时刻  $t$  时，记忆单元  $c_t$  记录了到当前时刻为止的所有历史信息，并受三个“门”控制，三个门中元素的数值取值范围是  $(0, 1)$ ：

$$\text{Input Gate : } i_t = \sigma(U_i x_t + W_i s_{t-1})$$

$$\text{Forget Gate : } f_t = \sigma(U_f x_t + W_f s_{t-1})$$

$$\text{Output Gate : } o_t = \sigma(U_o x_t + W_o s_{t-1})$$

下面具体理解 LSTM 模型。首先了解 LSTM 模型中的核心机制：记忆细胞状态。

细胞状态就像一根传送带，直接在整个链上运行，运行过程中只有一些少量的线性交互，以保证信息在上面流传保持不变。这些信息将会随着网络一直传递下去，保证了在学习当前时刻输入信息的同时，不会遗忘掉之前学习到的信息。

在 LSTM 模型中引入的门控机制，目的就是为了控制记忆细胞状态，向其中增添与删除信息。首先是遗忘门  $f_t$  的更新机制：

$$f_t = \sigma(U_f x_t + W_f s_{t-1})$$

其中， $W_f$  是关于隐藏状态  $s_t$  的参数矩阵， $U_f$  是关于当前输入  $x_t$  的参数矩阵。由于 sigmoid 函数的存在， $f_t$  的取值范围在  $(0, 1)$  之间。

遗忘门  $f_t$  的具体功能是根据  $x_t$  (当前输入) 和  $s_{t-1}$  (前一个隐藏状态), 为记忆状态  $C_{t-1}$  ( $t-1$  时刻状态) 中的每个元素输出  $0 \sim 1$  之间的数字, 以决定每一个  $C_{t-1}$  元素保留多少信息。1 代表完全保留, 0 代表彻底删除。

下一步是确定向记忆细胞中增添多少新信息。这一步骤分为两部分, 首先需要确定当前输入信息, 再决定输入信息的保留程度。对于当前输入信息  $\tilde{C}_t$  的计算方式如下:

$$\tilde{C}_t = \tanh(U_c x_t + W_c s_{t-1})$$

这里同样利用上一时刻隐藏层向量  $s_{t-1}$  与当前输入向量  $x_t$  分别乘上参数矩阵  $W_c$  与  $U_c$ , 进行一个线性变化。不同的是由于  $\tilde{C}_t$  不是一个门, 不需要其取值范围限定在  $(0, 1)$  之间, 于是选择用  $\tanh()$  函数代替 sigmoid 函数进行非线性激活。

得到  $\tilde{C}_t$  后, 使用输入门  $i_t$  来决定保留  $\tilde{C}_t$  中的信息至记忆细胞状态中:

$$i_t = \sigma(U_i x_t + W_i s_{t-1})$$

与遗忘门相同,  $W_i$  是关于前一个隐藏状态  $s_{t-1}$  的参数矩阵,  $U_i$  是关于当前输入  $x_t$  的参数矩阵。同理, 输入门  $i_t$  的取值范围在  $(0, 1)$  之间。

最后是输出门  $o_t$ : 基于当前的  $c_t$ , 输出那些作为当前时刻隐藏层的信息  $h_t$ 。 $o_t$  的更新方法依旧类似于之前的遗忘门与输出门:

$$o_t = \sigma(U_o h_{t-1} + W_o x_t + b_o)$$

其中,  $U_o$  是关于前一个隐藏状态  $h_{t-1}$  的参数矩阵,  $W_o$  是关于当前输入,  $o_t$  的参数矩阵,  $b_o$  是一个偏置向量。同理, 由于 sigmoid 函数, 输入门  $o_t$  的取值范围也在  $(0, 1)$  之间。在获得  $o_t$  之后, 利用之前更新好的当前时刻历史记忆信息  $C_t$ , 进行当前时刻隐藏层  $h_t$  信息的更新:

$$h_t = o_t(C_t)$$

在更新  $h_t$  的时候, 首先用非线性函数  $\tanh()$  对  $C_t$  进行激活, 然后再与  $o_t$  进行点乘运算, 保证向量维度保持不变。 $h_t$  会作为下一个时刻的输入之一, 再参与未来新的一系列更新。

最后对 LSTM 结构进行小结, 但在此之前, 有两点需要强调的是:

(1) 尽管三个门, 以及当前输入信息  $\tilde{C}_t$  的更新都具有相同的形式, 但是每各更新所用的参数都是独立不同的, 这保证了反向传播时对三个门控单元即当前记忆信息独立地进行更新。

(2) 以上公式中的所有参数都是可学习的参数。它们在最初是自动随机生成的随机数, 但是随着模型的不断训练, 这些参数会随着反向传播算法不断的进行更新, 最终使得损失函数数值尽可能的小以提高模型的表达能力。

LSTM 结构回顾:

- (1) LSTM 有单独的细胞状态，该状态贯穿整个 LSTM 网络。
- (2) 用遗忘门  $\tilde{C}_t$  和输入门  $i_t$  决定历史记忆信息与新记忆信息的保留或放弃。
- (3) 当前输入信息  $\tilde{C}_t$  来源于  $s_{t-1}$  和  $x_t$ 。
- (4) 当前记忆信息由  $f_{tt-1} + i_t \times \tilde{C}_t$  计算得出。
- (5) 输出门控制细胞状态的输出  $s_t = o_t(C_t)$ 。

以上所介绍的循环神经网络主要大量应用在自然语言处理场合。自然语言处理 (NLP) 是指利用人类交流所使用的自然语言与机器进行交互通讯的技术。通过人为的对自然语言的处理，使得计算机对其能够可读并理解。在本章末尾，引入了一个在自然语言处理情境应用循环神经网络的代码示例，可供参考与学习。

## 14.4 自编码器

在机器学习任务中经常需要采用某种方式对数据进行有效编码，例如对原始数据进行特征提取便是几乎所有机器学习问题均需解决的编码任务。除此之外，对数据进行降维处理或稀疏编码也是常见的编码任务。通常对数据进行编码时需要按照编码要求将原始数据转化为特定形式的编码数据，并要求编码数据尽可能多地保留原始数据信息，对这样的编码数据进行分析处理不仅会更加方便，而且可保证分析处理的结果较为准确。对数据的编码过程事实上是一个数据映射过程。若将某类原始数据  $\mathbf{X}$  编码为特定形式的数据  $y$ ，则需找到某个合适的映射  $f$  使得  $y = f(\mathbf{X})$ 。通常称映射  $f$  为编码器，并称将原始数据  $\mathbf{X}$  映射为编码数据  $y$  的过程为编码过程。若编码数据  $y$  保留了大部分原始数据  $\mathbf{X}$  中的信息，则从理论上说一定存在某种方式将数据  $y$  映射为与  $\mathbf{X}$  相近的数据  $\mathbf{X}'$ 。假设存在映射  $g$  满足：

$$\mathbf{X}' = g(y) = g[f(\mathbf{X})] \quad (14.4.1)$$

则认为编码器  $f$  可对原始数据  $\mathbf{X}$  进行有效编码，此时称映射  $g$  为解码器。

通常将用于实现自编码器的神经网络模型称为自编码器。由于神经网络输入层不具备数据处理能力，故自编码器中除输入层之外还包括两个分别用于实现编码器  $f$  和解码器  $g$  的模块，每个模块的神经元层数既可为单层也可为多层。一个单隐层的自编码器网络结构如图14.25所示：

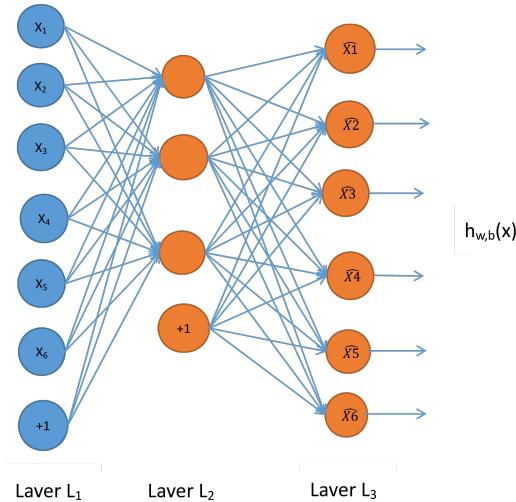


图 14.25 单隐层的自编码器

由图14.25可以看到，自编码器输出层的节点与输入层相等，训练这个网络以期望得到近似恒等函数。该模型首先使用编码器对输入数据进行编码，然后使用解码器尽量将编码数据还原为输入数据，即模型输出  $\mathbf{X}'$  尽量保持与模型输入  $\mathbf{X}$  一致。显然，若编码器  $f$  和解码器  $g$  均采用恒等映射，即  $f(a) = a$ ， $g(a) = a$ ，则有如下关系：

$$y = f(\mathbf{X}) = \mathbf{X}, \mathbf{X}' = g(y) = g[f(\mathbf{X})] = \mathbf{X} \quad (14.4.2)$$

此时恒有  $\mathbf{X}' = \mathbf{X}$ 。但这样的自编码器显然毫无意义，因为自编码器的主要功能是对原始数据进行编码而非对编码数据进行还原，即更加注重自编码器中的编码器模块。

若自编码器的神经元均使用激活函数  $\sigma$ ，则对于输入数据  $\mathbf{X} = (X_1, X_2, \dots, X_m)^T$ ，由前向计算方法可知自编码器隐含层第  $j$  个神经元的输出应为：

$$f_j(\mathbf{X}) = \sigma \left( \sum_{i=1}^m w_{ij}^{(1)} X_i + b_j^{(2)} \right) \quad (14.4.3)$$

由于隐含层数据处理结点个数为  $s$ ，故自编码器的隐含层可将  $m$  维数据输入转化为  $s$  维数据  $\mathbf{y} = (f_1(\mathbf{X}), f_2(\mathbf{X}), \dots, f_s(\mathbf{X}))^T$ 。若  $s < m$ ，则是对数据  $\mathbf{X}$  进行降维；若  $s > m$ ，则是对数据  $\mathbf{X}$  进行升维。自编码器隐含层到输出层的数据处理过程与此类似，最终输出  $\mathbf{X}'$  为：

$$\mathbf{X}' = (g_1(\mathbf{y}), g_2(\mathbf{y}), \dots, g_{m-1}(\mathbf{y}))^T \quad (14.4.4)$$

上式输出层第  $j$  个神经元输出  $g_j(\mathbf{y})$  的具体取值为：

$$g_j(y) = \sigma \left( \sum_{i=1}^s w_{ij}^{(2)} f_i(\mathbf{X}) + b_j^{(3)} \right) \quad (14.4.5)$$

若已经完成模型训练过程，则  $\mathbf{X}'$  应与数据输入  $\mathbf{X}$  差别不大。自编码器的模型训练通常使用不带标注信息的示例样本，故这是一种无监督学习方式。但由于要求自编码器的输入数据与输出数据尽可能接近，故对于训练样本  $\mathbf{X}_k$ ，若模型参数均已知，则可直接通过对比模型输入  $\mathbf{X}_k$  与输出  $\mathbf{X}'_k$  的差异确定损失函数  $L(\mathbf{X}_k, \mathbf{X}'_k)$ 。这相当于将自编码器的训练样本集看作监督学习的训练样本集  $\{(\mathbf{X}_1, \mathbf{X}_1), (\mathbf{X}_2, \mathbf{X}_2), \dots, (\mathbf{X}_n, \mathbf{X}_n)\}$ ，即根据自编码器的特点将无监督学习方式转化为监督学习方式。

可得自编码器模型优化目标函数：

$$J(\mathbf{W}) = \frac{1}{n} \sum_{k=1}^n L(\mathbf{X}_k, \mathbf{X}'_k) \quad (14.4.6)$$

其中  $\mathbf{W}$  为自编码器的参数向量。

损失函数  $L$  通常采用  $\mathbf{X}_k, \mathbf{X}'_k$  之间欧式距离的平方，即  $L(\mathbf{X}_k, \mathbf{X}'_k) = \|\mathbf{X}_k - \mathbf{X}'_k\|_2^2$ 。由此可得目标函数的具体形式为：

$$J(\mathbf{W}) = \frac{1}{n} \sum_{k=1}^n \|\mathbf{X}_k - \mathbf{X}'_k\|_2^2 \quad (14.4.7)$$

确定了目标函数之后，可采用适当模型优化算法并结合反向传播算法对模型参数进行优化，得到所求的自编码器。具体过程与反向传播神经网络优化过程类似。

如上所述，若自编码器隐含层神经元数目  $s$  大于输入数据维度  $m$ ，则可实现对输入数据的升维，此时只需确保编码后数据向量中取值为 0 的分量较多即可实现对原始数据的稀疏编码。

由于自编码器隐含层输出向量  $y = (f_1(\mathbf{X}), f_2(\mathbf{X}), \dots, f_s(\mathbf{X}))^\top$  即为编码数据，故只需限制  $y$  中取值为 0 的分量较多便可实现对原始数据的稀疏编码。令自编码器隐含层神经元采用 Sigmoid 激活函数，则对于训练样本集  $D = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  自编码器隐含层第  $j$  个神经元的平均激活程度为：

$$\bar{f}_j(\mathbf{X}) = \frac{1}{n} \sum_{i=1}^n f_j(\mathbf{X}_i) \quad (14.4.8)$$

为将隐含层第  $j$  个神经元的输出值限制为 0，可令  $\bar{f}_j(\mathbf{X}) = \varepsilon$ 。这里  $\varepsilon$  为某个接近于 0 的正数。由于  $f_j(\mathbf{X}_i) > 0$  且  $f_j(\mathbf{X}_i)$  在数据集 D 上期望接近于 0，故  $f_j(\mathbf{X}_i)$  的取值也接近于 0。若对隐含层中大部分神经元输出均施加此约束条件，即  $\bar{f}_j(\mathbf{X}) = \varepsilon, j = 1, 2, \dots, s$ ，则可保证  $y$  中大部分元素取值均接近于 0，实现对原始数据的稀疏编码。为将该约束条件纳入训练过程，需调整模型优化的目标函数，即在原始目标函数  $J(W)$  基础上添加约束  $\bar{f}_j(\mathbf{X}) = \varepsilon, j = 1, 2, \dots, s$  的惩罚项  $\lambda(\bar{f}_j(\mathbf{X}))$ ，将目标函数转化为如下形式：

$$\begin{aligned}
 J(\mathbf{W}) &= J'(\mathbf{W}) = J(\mathbf{W}) + \lambda(\bar{f}_j(\mathbf{X})) \\
 &= \frac{1}{n} \sum_{k=1}^n L(\mathbf{X}_k, \mathbf{X}'_k) + \alpha \lambda(\bar{f}_j(\mathbf{X}))
 \end{aligned} \tag{14.4.9}$$

其中， $\alpha$  为惩罚项的权重。当  $\alpha$  取值较大时，所求自编码器的编码数据具有较好的稀疏性，但会舍弃较多的原始数据信息；当  $\alpha$  取值较小时，所求自编码器的编码数据会保留较多的原始数据信息，但稀疏性较差。

一般，惩罚项  $\lambda(\bar{f}_j(\mathbf{X}))$  具有如下形式：

$$\lambda(\bar{f}_j(\mathbf{X})) = \frac{1}{s} \sum_{j=1}^s (\bar{f}_j(\mathbf{X}) - \varepsilon)^2 \tag{14.4.10}$$

除此之外，还可使用  $(\bar{f}_j(\mathbf{X}))$  与  $\epsilon$  之间的 K-L 散度作为单个隐含层结点的惩罚项，由此得到如下惩罚项：

$$\lambda(\bar{f}_j(\mathbf{X})) = \frac{1}{s} \sum_{j=1}^s \text{KL}(\varepsilon, \bar{f}_j(\mathbf{X})) = \frac{1}{s} \sum_{j=1}^s \left[ \varepsilon \log \frac{\varepsilon}{\bar{f}_j(\mathbf{X})} + (1 - \varepsilon) \log \frac{1 - \varepsilon}{1 - \bar{f}_j(\mathbf{X})} \right] \tag{14.4.11}$$

其中  $\text{KL}(\varepsilon, \bar{f}_j(\mathbf{X}))$  为  $\bar{f}_j(\mathbf{X})$  与  $\epsilon$  之间的 K-L 散度。

上述自编码器模型均要求模型的输入与输出尽可能相一致。这种自编码器得到的编码数据有时未必是对原始数据的最优表示。

如果某种自编码器能对被破坏的数据  $\mathbf{X}^b$  进行编码并将其解码为真实原始数据  $\mathbf{X}$ ，则该自编码器的编码方式显然更为有效。通常称这种通过引入噪声来增加编码鲁棒性的自编码器为降噪自编码器，如图14.26所示。

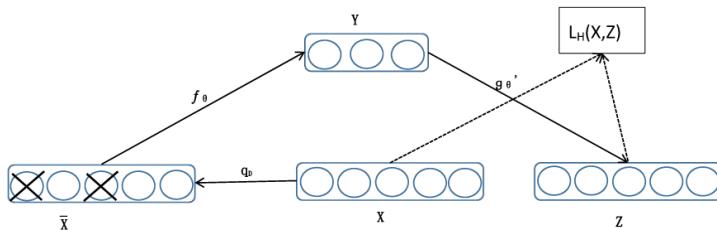


图 14.26 降噪自编码器模型

对于输入数据  $\mathbf{X}$ ，按照  $q_D$  分布进行加噪“损坏”，由图14.26可以看到，加噪过程是按照一定的概率将输入层的某些节点清零，然后将其作为自编码器的输入进行训练，除了输入层数据的处理不同，其余部分都和自编码器相同。

## 14.5 玻尔兹曼机

玻尔兹曼机 (BM, Boltzmann Machines) 是一种反馈随机神经网络，其具有两种输出状态，即 0 或 1。输出状态的取值根据概率统计方法取得的结果而决定，这种概率统计方法类似于一个 Boltzmann 分布。在神经元状态变化中引入了统计概率，网络的平衡状态服从 Boltzmann 分布，其运行机制基于模拟退火算法，可以理解为：离散 Hopfield 神经网络 + 模拟退火 + 隐单元 =BM。

### 14.5.1 随机神经网络

如果将 BP 算法中的误差函数看作一种能量函数，则 BP 算法通过不断调整网络参数使其能量函数按梯度单调下降，而反馈网络 (Hopfield 神经网络) 通过动态演变过程使网络的能量函数沿着梯度单调下降，在这一点上两类网络的指导思想是一致的。正因如此，常常导致网络落入局部极小点而达不到全局最小点，对于 BP 网，局部极小点意味着训练可能不收敛；对于 Hopfield 网络，则得不到期望的最优解。导致这两类网络陷入局部极小点的原因是：网络的误差函数或能量函数是具有多个极小点的非线性空间，而所用的算法却一味追求网络误差或能量函数的单调下降。也就是说，算法赋予网络的是只会“下山”而不具备“爬山”的能力。如果为具有多个局部极小点的系统打一个形象的比喻，设想托盘上有一个凸凹不平的多维能量曲面，若在该曲面上放置一个小球，它在重力作用下，将滚入最邻近的一个低谷（局部最小点）而无法跳出。但该低谷不一定就是曲面上最低的那个低谷（全局最小点）。因此，局部极小问题只能通过改进算法来解决。[随机神经网络](#)可赋予网络既能“下坡”也能“爬山”的本领，因而能有效地克服上述缺陷。随机网络与其他神经网络相比有两个主要区别：

(1) 在学习阶段，随机网络不像其他网络那样基于某种确定性算法调整权值，而是按某种概率分布进行修改。

(2) 在运行阶段，随机网络不是按某种确定性的网络方程进行状态演变，而是按某种概率分布决定其状态的转移。神经元的净输入不能决定其状态取 1 还是取 0，但能决定其状态取 1 还是取 0 的概率。这就是随机神经网络算法的基本概念，图14.27为随机网络算法与梯度下降算法区别的示意图。

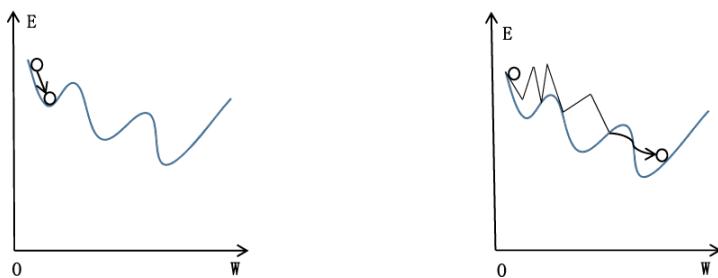


图 14.27 梯度下降算法与随机网络算法区别的示意图

### 14.5.2 模拟退火算法

**模拟退火算法**是随机网络中解决能量局部极小问题的一个有效方法，其基本思想是模拟金属退火过程，金属退火过程大致是，先将物体加热至高温，使其原子处于高速运动状态，此时物体具有较高的内能；然后，缓慢降温，随着温度的下降，原子运动速度减慢，内能下降；最后，整个物体达到内能最低的状态。模拟退火过程相当于沿水平方向晃动托盘，温度高则意味着晃动的幅度大，小球肯定会从任何低谷中跳出，而落入另一个低谷。这个低谷的高度（网络能量）可能比小球原来所在低谷的高度低（网络能量下降），但也可能反而比原来高（能量上升）。后一种情况的出现。从局部和当前来看，这个运动方向似乎是错误的；但从全局和发展的角度看，正是由于给小球赋予了“爬山”的本事，才使它有可能跳出局部低谷而最终落入全局低谷。当然，晃动托盘的力度要合适，并且还要由强至弱（温度逐渐下降），小球才不致因为有了“爬山”的本领而越爬越高。

在随机网络学习过程中，先令网络权值作随机变化，然后计算变化后的网络能量函数。网络权值的修改应遵循以下准则：若权值变化后能量变小，则接受这种变化；否则也不应完全拒绝这种变化，而是按预先选定的概率分布接受权值的这种变化。其目的在于赋予网络一定的“爬山”能力。实现这一思想的一个有效方法就是 Metropol 等 [156] 提出的模拟退火算法。

### 14.5.3 BM

神经网络模型的训练构造通常使用目标函数最小化的优化方式实现。前述各类前馈神经网络模型的训练构造均从误差最小化的角度设计目标函数，对此类目标函数进行优化后可保证在训练集上的整体预测误差达到最小且具备一定的泛化能力。事实上，还可从系统稳定性角度出发设计目标函数。由于系统越稳定则其能量越低，故为得到一个稳定的模型输出，可设计与网络模型相关的能量函数作为网络模型优

化的目标函数，由此实现对神经网络模型的优化求解。BM 便是此类神经网络的代表模型之一。BM 包含可视层与隐含层（如图14.28）所示，通过可视层神经元完成与外部的信息交互且可视层与隐含层的所有神经元均参与信息处理过程。该模型的理想效果是获得训练集  $D = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n\}$  中样本在模型稳定状态下的输出值。

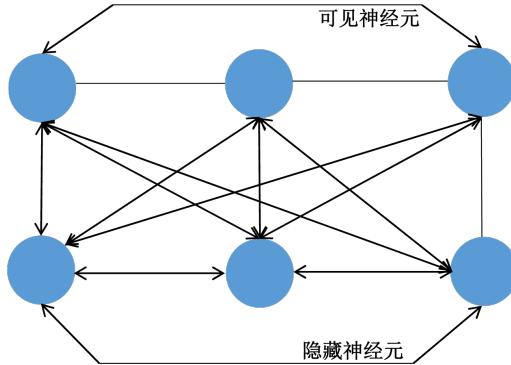


图 14.28 BM 网络运行图

BM 中所有神经元两两之间均存在信息传递且任意两个神经元之间的连接权重均相等，若使用  $w_{ij}$  表示 BM 中的  $i$  个神经元到第  $j$  个神经元的连接权重，则有  $w_{ij} = w_{ji}$ ，当  $i=j$  时有  $w_{ii} = w_{ii} = 0$ 。BM 中每个神经元的输出信号均限制为 0 或 1，并且每个神经元的状态取值具有一定的随机性，即以一定概率输出 0 或 1，这个概率与该神经元的输入相关。

具体地说，对于包含  $k$  个神经元的 BM，由于其第  $j$  个神经元的输入数据为其他所有神经元的输出信号，故该结点的总输入为：

$$I_j = \sum_{i=1}^k w_{ij} O_i + \theta_j \quad (14.5.1)$$

其中  $O_i$  为第  $i$  个神经元的输出信号， $\theta_j$  为第  $j$  个神经元所对应的偏置项。在网络中添加一个取值恒为 1 的第 0 个神经元作为偏置项，则可将偏置项  $\theta_j$  表示为连接权重  $w_{0j}$ ，则有：

$$I_j = \sum_{i=0}^k w_{ij} O_i \quad (14.5.2)$$

此时可将第  $j$  个神经元的输出信号  $O_j$  取值为 1 的概率定义为

$$P(O_j = 1) = \frac{1}{1 + e^{-\frac{I_j}{T}}} \quad (14.5.3)$$

相应地，该神经元输出为 0 的概率为：

$$P(O_j = 0) = \frac{e^{-\frac{I_j}{T}}}{1 + e^{-\frac{I_j}{T}}} \quad (14.5.4)$$

通常称参数  $T$  称为温度。

BM 所采用能量函数具体形式如下：

$$J(w_{ij}, O_i, O_j) = -\frac{1}{2} \sum_{i=0}^k \sum_{j=0}^k w_{ij} O_i O_j \quad (14.5.5)$$

对于第  $j$  个结点的输出  $O_j$ ：

若  $O_j = 0$ , 则对于  $i \leq j$ , 有:  $J(w_{ij}, O_i, 0) = 0$ ;

若  $O_j = 1$ , 则对于  $i \leq j$ , 有:  $J(w_{ij}, O_i, 1) = -\frac{1}{2} \sum_{i=0}^k w_{ij} O_i = -\frac{1}{2} I_j$ 。

若满足  $J(w_{ij}, O_i, 1) > J(w_{ij}, O_i, 0) = 0$ , 则说明  $O_j$  取 1 时的能量高于  $O_j$  取 0 时能量且  $I_j < 0$ , 可知  $P(O_j = 1) < 0.5 < P(O_j = 0)$ , 此时第  $j$  个神经元以较大概率输出使得网络能量降低的取值, 但仍有可能选择使得网络能量升高的取值。

若满足  $J(w_{ij}, O_i, 1) < J(w_{ij}, O_i, 0) = 0$ , 则说明  $O_j$  取 1 时的能量低于  $O_j$  取 0 时能量且  $I_j > 0$ , 故有  $P(O_j = 1) > 0.5 > P(O_j = 0)$ , 此时第  $j$  个神经元倾向于选择使得网络能量更低的取值作为输出。

由以上分析可知, BM 的各神经元均倾向于选择使得网络能量降低的输出值, 故该网络模型的能量函数取值呈现总体下降趋势, 但亦存在能量函数取值上升的可能性。这样可有效避免网络模型的优化计算陷入局部最优。

用 BM 网络进行联想时, 可通过学习用网络稳定状态的概率来模拟训练集样本的出现概率。根据学习类型, BM 网络可分为自联想和异联想两种情况 (如图14.29所示)。自联想 BM 网络中的可见节点既是输入节点又是输出节点, 隐节点的数目由学习的需要而定, 最少可以为 0。异联想 BM 网络中的可见节点需按功能分为输入节点组和输出节点组。

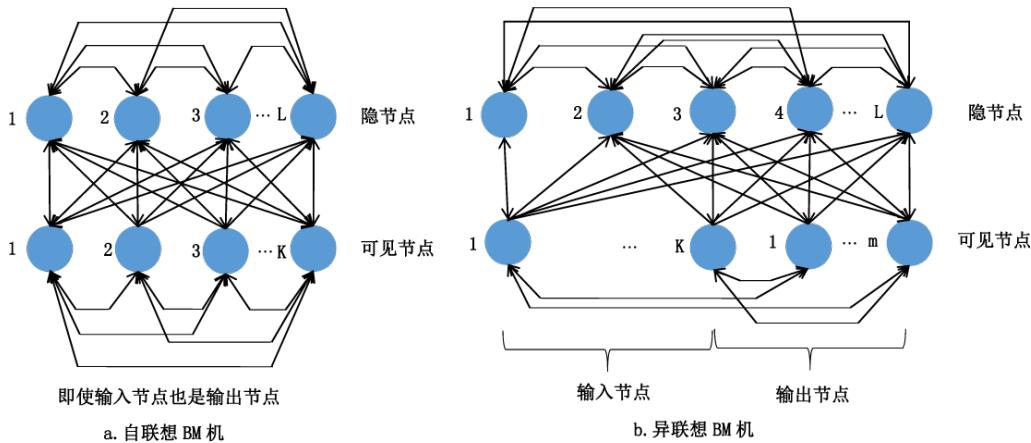


图 14.29 BM 网络拓扑结构

## 14.6 深度学习实践

### 14.6.1 R 语言实践

在本小节中，我们将利用 R 中的 keras 包，以 IMDB 数据集为例展示深度学习中的循环神经网络。IMDB 数据集包含了 50000 条偏向明显的评论，其中 25000 条作为训练集，25000 作为测试集。label 为 pos (positive) 和 neg (negative)，属于分类问题。

详细构骤实现代码如下：

```
library(keras)
install_keras(tensorflow = 'nightly')#导入kears包与tensorflow后端
max_unique_word <- 2500#设定取频率靠前的2500单词
max_review_len <- 100#设定影评最大长度
my_imdb <- dataset_imdb(num_words = max_unique_word)
str(my_imdb)#检查加载对象的结构
x_train <- my_imdb$train$x#数据集加载
y_train <- my_imdb$train$y
x_test <- my_imdb$test$x
y_test <- my_imdb$test$y
x_train <- pad_sequences(x_train, maxlen = max_review_len)
x_test <- pad_sequences(x_test, maxlen = max_review_len)#执行填充和截断过程
#使用简单循环神经网络模型
rnn_model <- keras_model_sequential()
rnn_model %>%
layer_embedding(input_dim = max_unique_word, output_dim = 128) %>%
layer_simple_rnn(units = 64, dropout = 0.2, recurrent_dropout = 0.2) %>%
layer_dense(units = 1, activation = 'sigmoid')
rnn_model %>% compile(loss = 'binary_crossentropy',optimizer = 'adam',
metrics = c('accuracy'))#编译 RNN 模型
batch_size = 128
epochs = 5
validation_split = 0.2
rnn_history <- rnn_model %>% fit(x_train, y_train,batch_size = batch_size,
epochs = epochs,validation_split = validation_split)
plot(rnn_history)
rnn_model %>% evaluate(x_test, y_test)

#部分输出结果截取如下：
24512/25000 [=====.>.] - ETA: 0s - loss: 0.6390 - acc: 0.6187
24704/25000 [=====.>.] - ETA: 0s - loss: 0.6387 - acc: 0.6189
25000/25000 [=====] - 7s 292us/sample - loss: 0.6385 - acc:
0.6188
$loss
[1] 0.6385054
$acc
[1] 0.61884
```

简单的 RNN 循环性能可能不够优秀，因此我们尝试使用 LSTM 模型进行优化。

```

lstm_model <- keras_model_sequential()
lstm_model %>%
layer_embedding(input_dim = max_unique_word, output_dim = 128) %>%
layer_lstm(units = 64, dropout = 0.2, recurrent_dropout = 0.2) %>%
layer_dense(units = 1, activation = 'sigmoid')
lstm_model %>% compile(loss = 'binary_crossentropy',
optimizer = 'adam',metrics = c('accuracy'))
batch_size = 128
epochs = 5
validation_split = 0.2
lstm_history <- lstm_model %>% fit(x_train, y_train,batch_size = batch_size,
epochs = epochs,validation_split = validation_split)
plot(lstm_history)
lstm_model %>%
evaluate(x_test, y_test)

#部分输出结果截取如下：
24640/25000 [=====]>.] - ETA: 0s - loss: 0.3774 - acc: 0.8390
24768/25000 [=====]>.] - ETA: 0s - loss: 0.3772 - acc: 0.8391
24992/25000 [=====]>.] - ETA: 0s - loss: 0.3766 - acc: 0.8393
25000/25000 [=====] - 19s 774us/sample - loss: 0.3767 - acc:
0.8392
$loss
[1] 0.3766563
$acc
[1] 0.83924

```

从这个简单的例子中，我们可以看到 LSTM 模型的性能比简单的 RNN 模型有了显著的提高，但与此同时，LSTM 模型的计算时间大约翻了一倍。

### 14.6.2 Python 语言实践

#### 基于 PyTorch 的深度神经网络的构建

本节使用的数据集为 CIFAR-10 数据集，该数据集共有 60000 张彩色图像，这些图像是  $32 \times 32$ ，分为 10 个类，每类 6000 张图。这里面有 50000 张用于训练，构成了 5 个训练批，每一批 10000 张图；另外 10000 用于测试，单独构成一批。

实现代码如下：首先使用 PyTorch 构建 CNN

```

import torch.nn
from torch import nn
from torch.nn import Conv2d,Sequential,MaxPool2d,Flatten,Linear
#构造网络结构
class NetModel(nn.Module):
    def __init__(self):
        super(Module, self).__init__()
        self.model=Sequential(Conv2d(3,32,5,padding=2),

```

```
#输入通道数为3，卷积核数为32，方形卷积核长度为5，两边填充2个维度的0
    MaxPool2d(2),#池化视野大小为2
    Conv2d(32,32,5,padding=2), MaxPool2d(2),
    Conv2d(32,64,5,padding=2), MaxPool2d(2),
    Flatten(),#在0维展开
    Linear(1024,64),#定义两个全连接层
    Linear(64,10))
def forward(self,x):
    x=self.model(x)
    return x #使用默认激活函数
```

接着读取数据，进行数据划分

```
#读取数据
from torch.utils.data import DataLoader
import torchvision
import torch.nn
train_data = torchvision.datasets.CIFAR10#读入训练集
(root = './cifar10data', train=True, transform=torchvision.transforms.ToTensor())
test_data=torchvision.datasets.CIFAR10#读入测试集
(root='./cifar10data', train=False, transform=torchvision.transforms.ToTensor())
train_len = len(train_data)
test_len = len(test_data)
print("训练集的长度为: {}".format(train_len))
print("测试集的长度为: {}".format(test_len))
```

再设置相应参数，代入数据进行模型训练

```
train_loader = DataLoader(train_data,batch_size=64)#加载数据模型
test_loader = DataLoader(test_data,batch_size=64)
net = NetModel()#加载网络模型
loss_fn = nn.CrossEntropyLoss()#定义损失函数
learing_rate = 1e-2#定义优化器
optimzer = torch.optim.SGD(net.parameters(),lr=learing_rate)
total_train_step=0#记录训练的次数
total_test_step=0#记录测试的次数
epoch=10#训练的轮数
net.train()#训练模型
for i in range(epoch):
    print("-----第{}轮数训练开始-----".format(i+1))
    for data in train_loader:
        img,target = data
        output = net(img)
        loss=loss_fn(output,target)
        optimzer.zero_grad()#对模型进行优化
        loss.backward()
        optimzer.step()
        total_train_step=total_train_step+1
        if total_train_step % 100 ==0:
            print("训练次数{},Loss:{}" .format(total_train_step,loss))
    net.eval()#测试步骤开始
    total_test_loss=0
    total_test_accuracy=0
```

```

with torch.no_grad():
    for data in test_loader:
        img,target = data
        output = net(img)
        loss = loss_fn(output,target)
        total_test_loss+=loss
        accuracy = (output.argmax(1)==target).sum() #1代表横向
        total_test_accuracy+=accuracy.item()
print("在整个测试集上的损失率为:{}".format(total_test_loss))
print("在整个测试集上的正确率为:{}".format(total_test_accuracy/test_len))

```

在输出结果中可以看到，经过 10 轮训练，正确率有了大幅度的提升。

## 基于 keras 的深度神经网络的构建

下面将以 MNIST 手写数字识别数据集为例展示深度学习神经网络的构建方法。

LeNet-5 是由有着卷积神经网络之父美誉的 Yann LeCun (中文译为杨立昆) 于 1998 年提出的一种经典的卷积网络结构 [149]。它是第一个成功应用于数字识别问题的卷积神经网络。在 MNIST 数据集上，LeNet-5 模型可以达到大约 99.2% 的正确率。作为早期的一种卷积神经网络结构，LeNet-5 的提出极大的推动了后续卷积神经网络的发展，它通常被认为是 CNN 的开山之作。其构建过程代码如下：首先展示 MNIST 数据集，并完成数据准备

```

from keras.datasets import mnist
from keras.utils import np_utils
from matplotlib import pyplot as plt
(X0,Y0),(X1,Y1) = mnist.load_data()#读取数据
print(X0.shape)
plt.figure()
fig,ax = plt.subplots(2,5)
ax = ax.flatten()
for i in range(10):
    Im = X0[Y0==i][0]
    ax[i].imshow(Im)
plt.show();
N0 = X0.shape[0];N1 = X1.shape[0]
print([N0,N1])
X0 = X0.reshape(N0,28,28,1) / 255
X1 = X1.reshape(N1,28,28,1) / 255
YY0 = np_utils.to_categorical(Y0)
YY1 = np_utils.to_categorical(Y1)
print(YY1)

```

再基于 keras 进行模型搭建

```

from keras.layers import Conv2D,Dense,Flatten,Input,MaxPooling2D
from keras import Model
input_layer = Input([28,28,1])#构建模型
x = input_layer
x = Conv2D(6,[5,5],padding = "same", activation = 'relu')(x)

```

```

x = MaxPooling2D(pool_size = [2,2], strides = [2,2])(x)
x = Conv2D(16,[5,5],padding = "valid", activation = 'relu')(x)
x = MaxPooling2D(pool_size = [2,2], strides = [2,2])(x)
x = Flatten()(x)
x = Dense(120,activation = 'relu')(x)
x = Dense(84,activation = 'relu')(x)
x = Dense(10,activation = 'softmax')(x)
output_layer = x
model = Model(input_layer, output_layer)
model.summary()

```

模型的编译通过 model.compile 实现。它的损失函数为 categorical\_crossentropy，这等价于要优化的是一个对数似然函数。选择 Adam 优化算法，需要监控预测精度，因此定义 metrics 为 accuracy。详细代码如下：

```

model.compile(loss='categorical_crossentropy',optimizer='adam',metrics=['accuracy'])
model.fit(X0, YY0, epochs=10, batch_size=200, validation_data=[X1, YY1])
#模型的输出结果为：
Train on 60000 samples, validate on 10000 samples
Epoch 1/10
60000/60000 [=====] - 5s 82us/step - loss: 0.3830 - accuracy
: 0.8897 - val_loss: 0.0941 - val_accuracy: 0.9709
...
Epoch 10/10
60000/60000 [=====] - 2s 33us/step - loss: 0.0199 - accuracy
: 0.9935 - val_loss: 0.0331 - val_accuracy: 0.9896
<keras.callbacks.callbacks.History at 0x22dc0ce7c88>

```

在输出结果中可以看到，LeNet-5 在这个手写数字识别数据集上已经达到了一个很高的精度，其预测结果的正确率已经达到了 98.96%。

## 14.7 总结

本章节针对图像数据，讨论了卷积神经网络、循环神经网络、长短时记忆网络、自编码器等网络的结构和计算过程。但限于篇幅，并未讨论视频、文本、自然语言等数据类型中的问题，也未讨论生成对抗网络、迁移学习、强化学习等比较新的深度学习模型，请参考邱锡鹏所著《神经网络与深度学习》。

## 14.8 练习

- 总结 CNN 算法的计算过程，并指出其存在的问题。
- 当图像尺寸变为 2 倍，CNN 的参数数量变为几倍？为什么？

3. 解释 Batch Normalization 的意义。
4. 模型的超参数是什么？它和参数有什么不同？针对上一节中的程序，请进行超参数优化。
5. 对于课本中第一个例子，进行网络结构优化，以期望达到最佳分类正确率。
6. 根据课本中描述的常见卷积神经网络，请用 PyTorch 或 Keras 构建网络模型并编程实现。

# 第十五章 并行计算

随着数据的增加，机器学习中的计算瓶颈越发凸显，并行计算就成为解决这个瓶颈的关键技术。根据计算设备的不同，一般分为基于 CPU 的并行计算和基于 GPU 的并行计算，本章基于 Python 和 R 编程环境，给出了基于 CPU 和基于 GPU 的并行计算的相关概念、原理和计算流程。

## 15.1 并行计算相关概念

### 15.1.1 进程

进程的概念是 60 年代初首先由麻省理工学院的 MULTICS 系统和 IBM 公司的 CTSS/360 系统引入的。[进程](#)（Process）是具有一定独立功能的程序在某个数据集合上的一次运行活动，是系统进行资源分配和调度的一个独立单位。程序只是一组指令的有序集合，它本身没有任何运行的含义，只是一个静态实体。而进程则不同，它是程序在某个数据集上的执行，是一个动态实体。它因创建而产生，因调度而运行，因等待资源或事件而被处于等待状态，因完成任务而被撤销，反映了一个程序在一定的数据集上运行的全部动态过程。图15.1为 Windows 任务管理器中进程一览表，每个进程表现为一个独立的执行程序。

#### 1、进程的特征

根据进程的特点，它具有以下特征：

- (1) 动态性：进程的实质是程序在多道程序系统中的一次执行过程，进程是动态产生，动态消亡的。
- (2) 并发性：任何进程都可以同其他进程一起并发执行。
- (3) 独立性：进程是一个能独立运行的基本单位，同时也是系统分配资源和调度的独立单位。
- (4) 异步性：由于进程间的相互制约，使进程具有执行的间断性，即进程按各自独立的、不可预知的速度向前推进。
- (5) 结构特征：进程由程序、数据和进程控制块三部分组成。

多个不同的进程可以包含相同的程序：一个程序在不同的数据集里就构成不同的进程，能得到不同的结果；但是执行过程中，程序不能发生改变。

## 2、进程的状态

进程执行时的间断性，决定了进程可能具有多种状态。事实上，运行中的进程可能具有如图15.2所示的三种基本状态：运行态、就绪态和阻塞态。

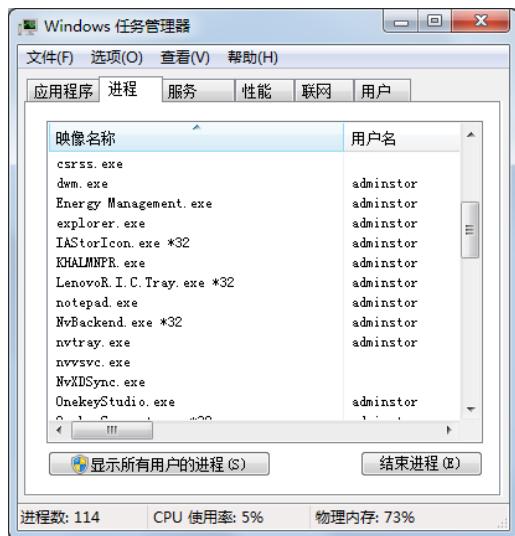


图 15.1 Windows 任务管理器之进程显示

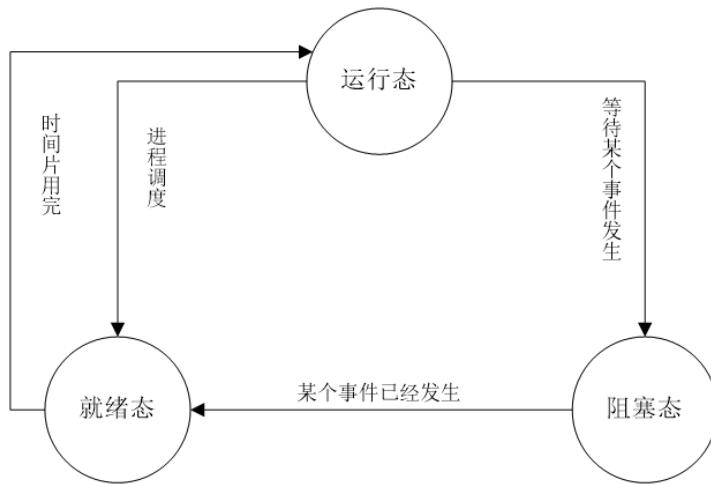


图 15.2 进程状态图

### (1) 就绪态 (Ready)

进程已获得除处理器外的所需资源，等待分配处理器资源；只要分配了处理器进程就可执行。就绪进程可以按多个优先级来划分队列。例如，当一个进程由于时间片用完而进入就绪状态时，排入低优先级队列；当进程由 I/O 操作完成而进入就绪状态时，排入高优先级队列。

### (2) 运行态 (Running)

进程占用处理器资源，处于此状态的进程的数目小于等于处理器的数目。在没有其他进程可以执行时（如所有进程都在阻塞状态），通常会自动执行系统的空闲进程。

### (3) 阻塞态 (Blocked)

由于进程等待某种条件（如 I/O 操作或进程同步），在条件满足之前无法继续执行。该事件发生前即使把处理机分配给该进程，也无法运行。

## 15.1.2 线程

线程 (Thread) 是进程的一个实体，是 CPU 调度和分派的基本单位。线程不能够独立执行，必须依存在进程中，由进程提供多个线程执行控制。从内核角度讲线程是活动体对象，而进程只是一组静态的对象集，进程必须至少拥有一个活动线程才能维持运转。图15.3表明一个进程最多可以调用 8 个线程。

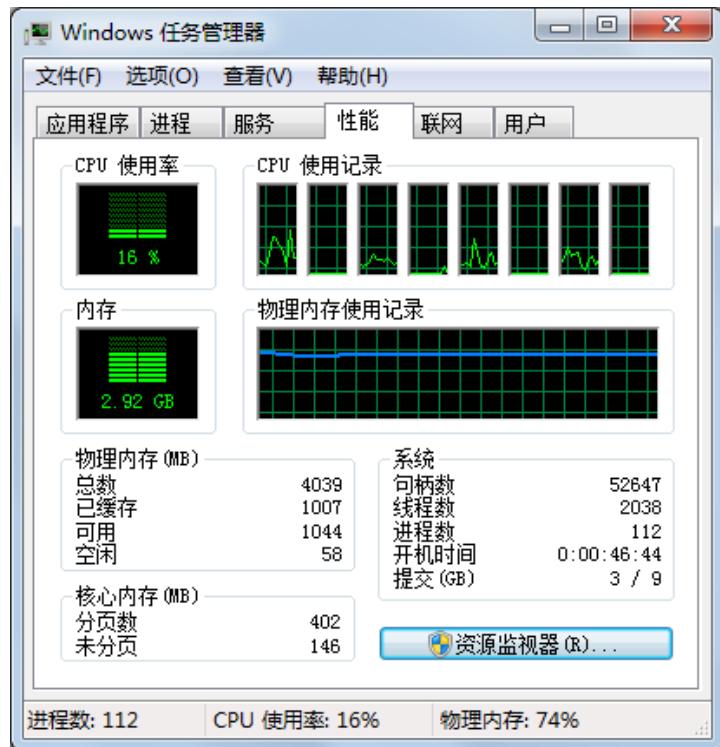


图 15.3 进程状态图

线程和进程的关系是：线程是属于进程的，线程运行在进程空间内，同一进程所产生的线程共享同一内存空间，当进程退出时该进程所产生的线程都会被强制退

出并清除。线程可与属于同一进程的其它线程共享进程所拥有的全部资源，但是其本身基本上不拥有系统资源，只拥有一点在运行中必不可少的信息（如程序计数器、一组寄存器和栈）。

在操作系统中引入线程带来的主要好处是：

- (1) 在进程内创建、终止线程比创建、终止进程要快。
- (2) 同一进程内的线程间切换比进程间的切换要快，尤其是用户级线程间的切换。

另外，线程的出现还因为以下几个原因：

(1) 并发程序的并发执行，在多处理环境下更为有效。一个并发程序可以建立一个进程，而这个并发程序中的若干并发程序段就可以分别建立若干线程，使这些线程在不同的处理机上执行。

(2) 每个进程具有独立的地址空间，而该进程内的所有线程共享该地址空间。这样可以解决父子进程模型中，子进程必须复制父进程地址空间的问题。

(3) 线程对解决客户/服务器模型非常有效。

不同的平台对线程的状态定义不同，大致可以定义为运行、挂起、睡眠、阻塞、就绪、终止这六种，如图15.4所示。

**运行**：就是线程获得了 CPU 的控制权，正在执行计算。

**挂起**：一般是指被挂起，因为同一时刻，需要“同步”运行的线程不止他一个，所以基于时间片轮转的原则，它在独占了一段时间的 CPU 后，被挂起，线程环境被压栈。

**睡眠**：一般是指主动挂起，这种情况在 WINDOWS 平台不存在。

**阻塞**：与挂起和睡眠类似，都是失去 CPU 的控制权。与挂起更相像，也是被挂起的。不同之处在于，被挂起的线程没有额外的表示，而被阻塞的线程会被记录下来，当等待的因素就绪后，线程会转为就绪状态。例如你在线程中调用一些系统服务函数，会引起线程控制权的一次裁决，从而挂起本线程，造成本线程的阻塞。挂起、睡眠、阻塞看起来差不多，但其实本质上还是有以上所述区别的。

**就绪**：顾名思义，就是指它准备好了，一旦轮到它，它就可以转为运行状态。

**终止**：线程结束。

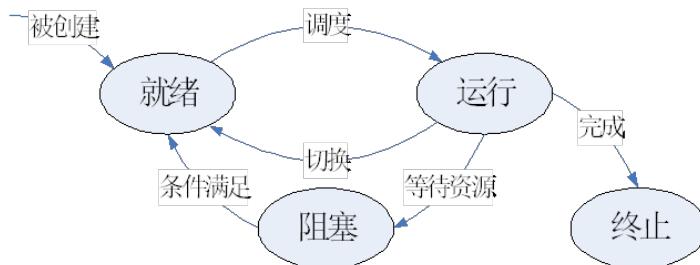


图 15.4 线程状态图

### 15.1.3 并行计算与分布式计算

**并行计算** (Parallel Computing) 是指同时使用多种计算资源解决计算问题的过程，是提高计算机系统计算速度和处理能力的一种有效手段。它的基本思想是用多个处理器（多个核）来协同求解同一问题，即将被求解的问题分解成若干个部分，各部分均由一个独立的处理器来并行计算。并行计算系统既可以是专门设计的、含有多个处理器的超级计算机，也可以是以某种方式互连的若干台的独立计算机构成的集群。通过并行计算集群完成数据的处理，再将处理的结果返回给用户。

并行计算可分为时间上的并行和空间上的并行。时间上的并行是指流水线技术。而空间上的并行是指多个处理器并发的执行计算，即通过网络将两个以上的处理器连接起来，达到同时计算同一个任务的不同部分，或者单个处理器无法解决的大型问题。但这个空间一般是指集中放在一起的集群计算机。并行计算中主要研究的是空间上的并行问题。从程序和算法设计人员的角度来看，并行计算又可分为数据并行和任务并行。空间上的并行导致了两类并行机的产生，按照 Flynn (弗林) 分类法为：单指令流多数据流 (SIMD) 和多指令流多数据流 (MIMD)。我们常用的串行机也叫做单指令流单数据流 (SISD)。

而**分布式计算** (Distributed computing) 也是一种并行计算，主要是指通过网络相互连接的两个以上的处理器相互协调，各自执行相互依赖的不同应用，从而达到协调资源访问，提高资源使用效率的目的。但是，它无法达到并行计算所倡导的提高求解同一个应用的速度，或者提高求解同一个应用的问题规模的目的。对于一些复杂的应用系统，分布式计算和并行计算通常相互配合，既要通过分布式计算协调不同应用之间的关系，又要通过并行计算提高求解单个应用的能力。

因此，并行计算一般在企业内部进行，而分布式计算可能会跨越局域网，或者直接部署在互联网上，节点之间几乎不互相通信。很多公益性的项目，就是使用分布式计算的方式在互联网上实现，比如以寻找外星人为目的的 SETI 项目。

### 15.1.4 同步与异步

**同步** (Synchronization)：进程之间的关系不是相互排斥临界资源的关系，而是相互依赖的关系。进一步的说明：就是前一个进程的输出作为后一个进程的输入，当第一个进程没有输出时第二个进程必须等待。具有同步关系的一组并发进程相互发送的信息称为消息或事件。

**异步** (Asynchronization)：异步和同步是相对的一个概念，同步就是顺序执行，执行完一个再执行下一个，需要等待、协调运行。异步就是彼此独立，在等待某事件的过程中继续做自己的事，不需要等待这一事件完成后再工作。线程就是实现异步

的一个方式。异步让调用方法的主线程不需要同步等待另一线程的完成，从而可以让主线程干其它的事情。

### 15.1.5 通信

在并行计算中，**通信**是指进程或线程之间的数据传输或内存访问。一般分为两种：共享内存和消息传递。

如图15.5，**共享内存**这种方式在多核并行计算中比较常见，通常会设置一个共享变量，然后多个线程去操作同一个共享变量，从而达到线程通讯的目的。这种通讯模式中，不同的线程之间是没有直接联系的。都是通过共享变量这个“中间人”来进行交互。而这个“中间人”必要情况下还需被保护在临界区内（加锁或同步）。由此可见，一旦共享变量变得多起来，并且涉及到多种不同线程对象的交互，这种管理会变得非常复杂，极容易出现数据竞争、死锁等问题。

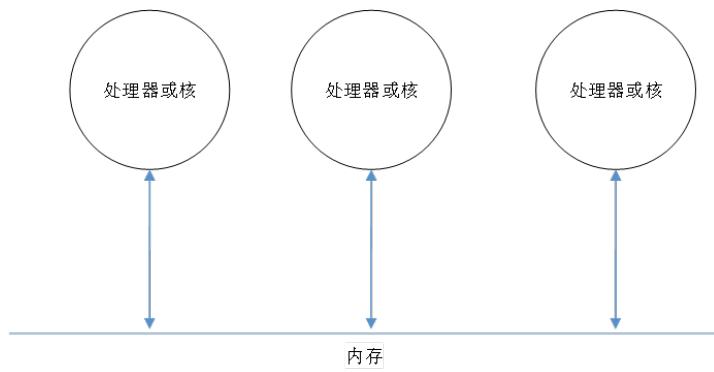


图 15.5 共享内存

而**消息传递**方式主要针对多处理器或多节点并行计算，采取的是进程或线程之间的直接通信，不同的进程或线程之间通过显式的发送消息来达到交互目的。如图15.6，消息传递模型有以下特征：

- (1) 计算时任务集可以用它们自己的内存。多任务可以在相同的物理处理器上，同时可以访问任意数量的处理器。
- (2) 任务之间通过接收和发送消息来进行数据通信。
- (3) 数据传输通常需要每个处理器协调操作来完成。例如，发送操作有一个接受操作来配合。

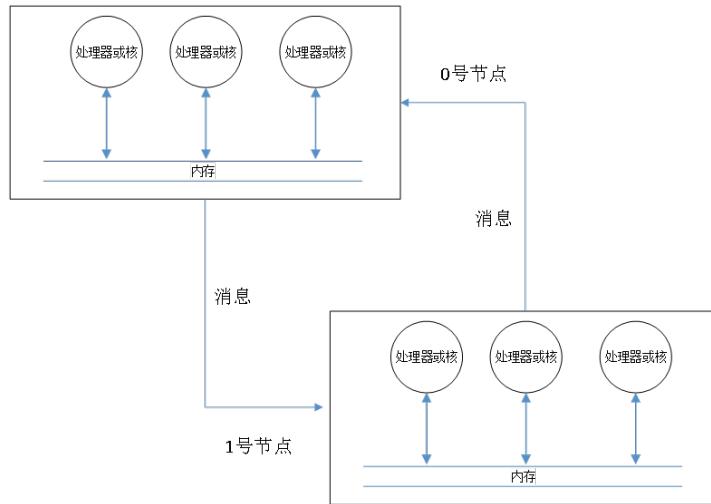


图 15.6 消息传递

表15.1总结了这两种通讯模式的异同情况。

并发模型	通信机制	同步机制
共享内存	线程之间共享程序的公共状态，线程之间通过写-读内存中的公共状态来隐式进行通信。	同步是显式进行的。程序员必须显式指定某个方法或某段代码需要在线程之间互斥执行。
消息传递	线程之间没有公共状态，线程之间必须通过明确的发送消息来显式进行通信。	由于消息的发送必须在消息的接收之前，因此同步是隐式进行的。

表 15.1 共享内存与消息传递的异同

### 15.1.6 加速比

为了更方便地描述并行计算的性能，一般采用加速比指标来进行度量。**加速比**定义如下：

$$S = \frac{T_s}{T_p} \quad (15.1.1)$$

其中， $T_s$  表示单处理器上最优串行化算法计算的时间， $T_p$  表示使用  $p$  个 CPU 处理器并行计算的时间。

当加速比  $S < 1$  意味着并行计算的时间比串行计算的时间还长，并行计算效率反而降低； $S < p$  时，表示次线性加速； $S \approx p$  时，表示线性加速； $S > p$  时，表示超线

性加速。一般来说，加速比通常都小于 CPU 核数，只有极少数并行算法可以获得超线性加速比，例如并行搜索工作量少于串行搜索工作量等算法，另外，也有可能是由于高速缓存产生的额外加速效果所导致。因此，在并行算法设计中，其加速比一般要求向 CPU 核数靠近，加速比越接近线性加速，那么程序性能就越好。

但是，对于某些串行程序，并不是所有部分都可以用并行程序替代，有一部分必须要串行执行。令  $W_s$  为程序中的串行部分， $W_p$  为程序中的并行部分，则  $W = W_s + W_p$ . 根据 Amdahl 定律，在计算规模一定的情况下，加速比定义为：

$$S = \frac{W}{W_s + W_p/p} \quad (15.1.2)$$

串行部分比例为  $f = \frac{W_s}{W}$ ，则式 (15.1.2) 为

$$S = \frac{1}{f + (1 - f)/p} \quad (15.1.3)$$

随着 CPU 核数  $p$  的增加，这是一个递增的函数，但这个函数有上限，当  $p \rightarrow \infty$  时，有

$$S = \frac{1}{f}$$

图15.7为串行程序比例分别是 0.0, 0.1, …, 0.2 时的加速比随 CPU 核数  $p$  变化（这里核数取 2、4、8、16、24、48）的情形。从图中可以看出，当一定的情况下，加速比随核数的增加而增加，但是在 16 核之前增速较快，之后增速较为缓慢；当 CPU 核数一定时，加速比随着串行程序占比的减少而增加。因此，要想提高并行计算效率，需要从 CPU 核数和串行程序占比两个因素综合考虑。

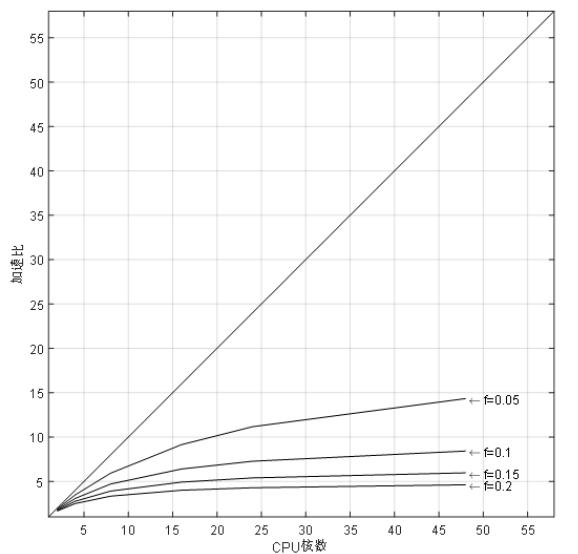


图 15.7 加速比随 CPU 核数  $p$  变化的情形

如果考虑并行计算时的通信、同步和归约等操作所花费的额外开销  $W_0$ ，则

$$S = \frac{W}{W_s + W_p/p + W_0}, \quad (15.1.4)$$

$$S = \frac{1}{f + (1-f)/p + W_0/W}. \quad (15.1.5)$$

当并行计算时的通信、同步和归约等操作所花费的额外开销比重较大时，平行效率降低，严重时加速比小于 1。在后面例子将遇到这种情况。对于很多大型计算，精度要求高，而计算时间要求不能增加。在这种计算规模增加的情况下，要想保持原有计算时间，必须增加处理器才能完成计算任务。Gustafson [157] 提出了变问题规模的加速比模型如下：

$$S = \frac{f + (1-f)p}{1 + W_0/W}. \quad (15.1.6)$$

从式 (15.1.6) 可以看出，当处理器个数增加时，必须控制额外开销的增加，才能达到线性加速。因此，在并行计算中，如何优化程序，是一个值得思考的问题。考虑到加速比计算的简便性，在本书中采用加速比公式 (15.1.1)。

## 15.2 基于 CPU 线程的并行计算

在 Python 中，提供了基于线程的并行模块 `threading`，`threading` 模块除了 `Thread` 类之外，还包括其它很多的同步机制，这些同步机制可以避免数据竞争问题的发生。

### 15.2.1 创建线程

使用 `Thread` 类主要有两种方法创建线程：

- 创建 `Thread` 类的实例，传递一个函数；
- 派生 `Thread` 类的子类，并创建子类的实例。

本教材主要采用第一种方式创建线程。对于 `Thread` 类，其主要属性和方法见表15.2。

属性/方法	描述
Thread 类属性	
name	线程名
ident	线程的标识符
daemon	布尔值，表示这个线程是否是守护线程
Thread 类方法	
init(group,...)	实例化一个线程对象，需要一个可调用的 target 对象，以及参数 args 或者 kwargs。还可以传递 name 和 group 参数。daemon 的值将会设定 thread.daemon 的属性。
start()	线程启动函数。开始执行该线程
run()	定义线程的方法。(通常在子类中重写)
join(timeout=None)	等待函数。直至启动的线程终止之前一直挂起，除非给出了 timeout(单位秒)，否则一直被阻塞。

表 15.2 Thread 类主要属性和方法

注意：守护线程一般是一个等待客户端请求的服务器。如果没有客户端请求，守护线程就是空闲的。如果把一个线程设置为守护线程，就表示这个线程是不重要的，进程退出时不需要等待这个线程执行完成。使用下面的语句：thread.daemon=True 可以将一个线程设置为守护线程，同样的也可以通过这个值来查看线程的守护状态。对于主线程，将在所有的非守护线程退出之后才退出。

如图15.8，当线程对象被创建，其活动可通过调用线程的 start() 方法开始。一旦线程活动开始，该线程会被启动。主线程可以调用一个线程的 join() 方法，这会阻塞调用该方法的线程，直到被调用 join() 方法的线程终结。

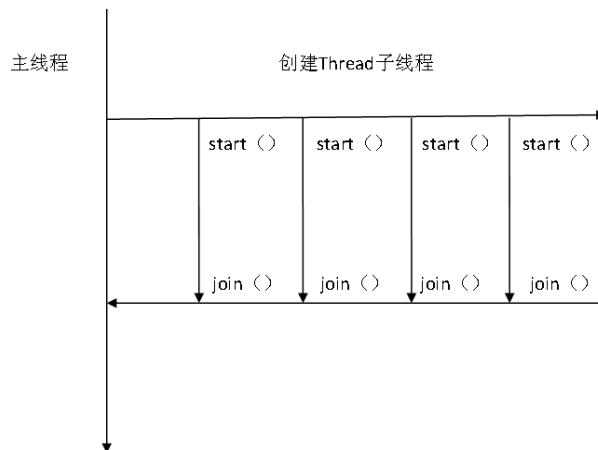


图 15.8 线程创建过程

下面给出具体的 Thread 类构造函数: class threading.Thread(group=None, target=None, name=None, args=(), kwargs={}, \*, daemon=None)

调用这个构造函数时, 必需带有关键字参数。参数如下:

group 应该为 None, 为了日后扩展 ThreadGroup 类实现而保留。

target 是用于 run() 方法调用的可调用对象, 通常是启动一个线程活动时要执行的目标函数。默认是 None, 表示不需要调用任何方法。

name 是线程名称。默认情况下, 由 “Thread-N” 格式构成一个唯一的名称, 其中 N 是小的十进制数。

args 是用于调用目标函数的参数元组。默认是 ()。

kwargs 是用于调用目标函数的关键字参数字典, 默认是 {}。

daemon 如果不是 None, 将显式地设置该线程是否为守护模式。如果是 None (默认值), 线程将继承当前线程的守护模式属性。

**例 15.1** 定义线程函数 HelloWord, 其参数为自定义的线程号 (从 0 开始), 其内容为延迟 1 秒后, 输出形如 Hello world! thread id:0 这样的字符串。使用 Thread 创建当前计算机最大 CPU 或核数的线程数, 并用 start 函数启动该线程, 同时使用计时函数来了解使用 join 函数的作用。

```
import threading
from time import sleep
from time import perf_counter
import multiprocessing
N_core = multiprocessing.cpu_count() # 获取当前计算机最大CPU或核数
def HelloWord(thread_id):#定义线程函数
    sleep(1)#延迟1秒
    print('Hello world! thread id:{} \n'.format(thread_id))
    return
if __name__ == "__main__":
    start = perf_counter()#计时开始
    for i in range(N_core):
        t = threading.Thread(target=HelloWord,args=(i,))#创建子线程
        t.start()#启动线程
        t.join()#阻塞
    end = perf_counter()#计时结束
    print('运行时间为: %s Seconds'%(end-start))
```

输出结果为:

```
Hello world! thread id:2
Hello world! thread id:1
Hello world! thread id:0
Hello world! thread id:5
Hello world! thread id:6
Hello world! thread id:4
Hello world! thread id:3
Hello world! thread id:7
运行时间为: 0.002644299998792121 Seconds
```

在主线程中创建了 8 个子线程，每个子线程都调用了线程函数 HelloWord，计时函数对主线程创建和启动子线程进行计时，由于不需要等待子线程结束，所以先输出了主线程运行时间，然后乱序输出各子线程的输出内容。如果在程序中加入 t.join() 等待函数，输出结果如下：

```
Hello world! thread id:0
Hello world! thread id:1
Hello world! thread id:2
Hello world! thread id:3
Hello world! thread id:4
Hello world! thread id:5
Hello world! thread id:6
Hello world! thread id:7
运行时间为：8.120153000000073 Seconds
```

很明显，主线程等子线程一个个执行完，再输出运行时间，该时间包含了等待时间。从时间上看与串行计算无异，也就是说并没有真正按照图15.8运行。因此，对于 join 函数，多数情况下根本不需要调用它。一旦线程启动，就会一直执行，直到给定的函数完成后退出。如果主线程还有其它事情要做（并不需要等待这些线程完成），可以不调用 join 函数。join 函数只有在你需要等待线程完成时候才是有用的。

### 15.2.2 同步

在多线程程序中，一般有一些特定的函数或代码块不希望被多个线程同时执行，这就需要使用同步了。threading 模块的同步机制，如表15.3所示，主要有锁机制、事件、信号量、栅栏。

对象	描述
Lock	锁对象
RLock	递归锁，是一个线程可以再次拥有已持有的锁对象
Condition	条件变量对象，使一个线程等待另一个线程满足特定的条件触发
Event	事件对象，普通版的 Condition
Semaphore	信号量，为线程间共享的资源提供一个“计数器”，计数开始值为设置的值，默认为 1
BoundedSemaphore	与 Semaphore 相同，有边界，不能超过设置的值
Timer	定时运行的线程对象，定时器
Barrier	栅栏，当达到某一栅栏后才可以继续执行

表 15.3 同步机制

### 1、锁（Lock）

锁仅有锁定和非锁定两种状态。它被创建时为非锁定状态。它有两个基本方法：acquire 和 release。当状态为非锁定时，acquire 将状态改为锁定并立即返回。当状态是锁定时，acquire 将阻塞至其它线程，调用 release 将其改为非锁定状态，然后 acquire 调用重置其为锁定状态并返回。release 只在锁定状态下调用，它将状态改为非锁定并立即返回。如果尝试释放一个非锁定的锁，则会引发 RuntimeError 异常。

当多个线程在 acquire 等待状态转变为未锁定被阻塞，然后 release 重置状态为未锁定时，只有一个线程能继续执行；至于哪个等待线程继续执行没有定义，并且会根据实现而不同。

具体方法如下：

`acquire(blocking=True, timeout=-1)`

可以阻塞或非阻塞地获得锁。当调用时参数 blocking 设置为 True（缺省值），阻塞直到锁被释放，然后将锁锁定并返回 True。在参数 blocking 被设置为 False 的情况下调用，将不会发生阻塞。如果调用时 blocking 设为 True 会阻塞，并立即返回 False，否则，将锁锁定并返回 True。

`release()`

释放一个锁。这个方法可以在任何线程中调用，不单指获得锁的线程。当锁被锁定，将它重置为未锁定，并返回。如果其他线程正在等待这个锁解锁而被阻塞，只允许其中一个允许。在未锁定的锁调用时，会引发 RuntimeError 异常。没有返回值。

`locked()`

如果获得了锁则返回真值。

**例 15.2** 定义计数函数 CountNum(thread\_id)，使用两个线程调用该函数，其中第一个线程延迟 1 秒，而第二个线程延迟 2 秒，在主线程中输出总计数。定义全局变量 counts，程序设计如下：

```
import threading
from time import sleep
counts=0
N_threads=2
def CountNum(thread_id):#定义线程函数
    global counts
    sleep(thread_id)
    for i in range(1,101):
        counts=counts+1
    print('thread id:{} , its counts is {}'.format(thread_id,counts))
    return
if __name__ == "__main__":
    for i in range(N_threads):
        t = threading.Thread(target=CountNum,args=(i+1,))#创建子线程
        t.start()#启动线程
    print('counts=%d Seconds'%counts)
```

运行结果如下：

```
counts=0 Seconds
thread id:1,its counts is 100
thread id:2,its counts is 200
```

很明显，主线程并没有得到子线程的计数结果。如果采用锁，当子线程各自计数时，主线程等待其完成后再进行输出，其结果就是总计数了。程序修改为：

```
import threading
from time import sleep
counts=0
N_threads=2
lock = threading.Lock() #创建锁
def CountNum(thread_id):#定义线程函数
    lock.acquire()#请求锁
    global counts
    sleep(thread_id)
    for i in range(1,101):
        counts=counts+1
    print('thread id:{} ,its counts is {}'.format(thread_id,counts))
    lock.release()#释放锁
    return
if __name__ == "__main__":
    for i in range(N_threads):
        t = threading.Thread(target=CountNum,args=(i+1,))#创建子线程
        t.start()#启动线程
    lock.acquire()#请求锁
    print('counts=%d Seconds'%counts)
    lock.release()#释放锁
```

运行结果为：

```
thread id:1,its counts is 100
thread id:2,its counts is 200
counts=200 Seconds
```

## 2. 事件 (Event)

**事件**用于不同进程间的相互通信。事件对象管理一个内部标识，调用 set 方法可将其设置为 true。调用 clear 方法可将其设置为 false。调用 wait 方法将进入阻塞直到标识为 true，这个标识初始时为 false。

具体方法如下：

**is\_set()**

当且仅当内部标识为 true 时返回 True。

**set()**

将内部标识设置为 true，所有正在等待这个事件的线程将被唤醒。当标识为 true 时，调用 wait() 方法的线程不会被阻塞。

**clear()**

将内部标识设置为 false，之后调用 wait() 方法的线程将会被阻塞，直到调用 set() 方法将内部标识再次设置为 true。

#### wait(timeout=None)

阻塞线程直到内部变量为 true。如果调用时内部标识为 true，将立即返回。否则将阻塞线程，直到调用 set() 方法将标识设置为 true 或者发生可选的超时。当提供了 timeout 参数且不是 None 时，它应该是一个浮点数，代表操作的超时时间，以秒为单位（可以为小数）。

**例 15.3** 针对例15.2，使用事件输出线程函数各自的计数。

```
import threading
from time import sleep
counts = 0
N_threads = 2
evGetData = threading.Event() #创建事件对象，内部标志默认为 False
evOutput = threading.Event()
def CountNum(thread_id): #定义线程函数
    evGetData.wait() #等待主线程激活
    global counts
    counts = 0
    sleep(thread_id)
    for i in range(1, 100*thread_id+1):
        counts = counts + 1
    print('thread id:{} , its counts is {}'.format(thread_id, counts))
    evOutput.set() #激活主线程
    return
if __name__ == "__main__":
    for i in range(N_threads):
        t = threading.Thread(target=CountNum, args=(i+1,)) #创建子线程
        t.start() #启动线程
        evGetData.set() #激活从线程
        evOutput.wait() #等待从线程激活
        print('counts=%d Seconds'%counts)
        evOutput.clear()
```

输出结果为：

```
thread id:1,its counts is 100
counts=100 Seconds
thread id:2,its counts is 200
counts=200 Seconds
```

## 15.3 基于 CPU 进程的并行计算

在 Python 程序中，代码执行由 Python 虚拟机（解释器主循环）来控制。对 Python 虚拟机的访问由 GIL (global interpreter lock, 全局解释器锁) 控制，GIL 保证同一时刻只有一个线程在执行。由于 GIL 的限制，python 多线程实际只能运行在单核 CPU，所以 15.2 节中的 threading 模块并不能真正实现多核 CPU 并行计算。如要实现多核 CPU 并行，只能通过多进程的方式实现。而 multiprocessing 模块是最常用的多进程模块，它同时提供了本地和远程并发操作，通过使用子进程而非线程有效地绕过了 GIL。因此，multiprocessing 模块允许程序员充分利用给定机器上的多个处理器或核进行并行计算。

### 15.3.1 创建进程

使用 multiprocessing 模块创建进程主要有两种方式：一是使用 Process 对象，二是采用 Pool 进程池。

#### 1. Process 对象

在 multiprocessing 中，通过创建一个 Process 对象然后调用它的 start 方法来生成进程，创建过程同线程创建过程。

Process 对象表示在单独进程中运行的活动，具体的 Process 构造函数形式为：

```
class multiprocessing.Process(group=None, target=None, name=None, args=(),  
    kwargs=*, daemon=None)
```

其中，group 应该是 None，它仅用于兼容 threading.Thread。

target 是由 run 方法调用的可调用对象，它默认为 None，意味着什么都没有被调用。

name 是进程名称。

args 是目标调用的参数元组。

kwargs 是目标调用的关键字参数字典。

daemon 可以设置为 True 或 False。如果是 None (默认值)，则该标志将从创建的进程继承。

Process 对象拥有和 threading.Thread 等价的大部分方法。具体说明如下：

`run()` 表示进程活动的方法，可以在子类中重载此方法。标准 run() 方法调用传递给对象构造函数的可调用对象作为目标参数 (如果有)，分别从 args 和 kwargs 参数中获取顺序和关键字参数。

`start()` 为启动进程活动，这个方法每个进程对象最多只能调用一次，它会将对象的 run() 方法安排在一个单独的进程中调用。

`join([timeout])` 表示如果可选参数 timeout 是 None (默认值)，则该方法将阻塞，直到调用 join() 方法的进程终止。如果 timeout 是一个正数，它最多会阻塞 timeout

秒。请注意，如果进程终止或方法超时，则该方法返回 None。检查进程的 exitcode 以确定它是否终止。一个进程可以被 join 多次。进程无法 join 自身，因为这会导致死锁。尝试在启动进程之前 join 进程是错误的。

`name`为进程的名称。该名称是一个字符串，仅用于识别目的，它没有语义，可以为多个进程指定相同的名称。初始名称由构造器设定，如果没有为构造器提供显式名称，则会构造一个形式为“Process-N1:N2:…:Nk”的名称。

`daemon`为进程的守护标志，一个布尔值。这必须在 start() 被调用之前设置。注意：在 Windows 上要想使用进程模块，就必须把有关进程的代码写在当前.py 文件的 if `_name_ == '_main_'` 语句的下面，才能正常使用 Windows 下的进程模块。Unix/Linux 下则不需要。

**例 15.4** 定义函数 HelloWord，其内容为延迟 1 秒后，输出形如 Hello world! current\_process name=Process-1 这样的字符串。在主进程中创建 4 个进程，同时在主进程中输出形如 Hello world! current\_process name=MainProcess 这样的字符串。

在 py 文件中输入程序：

```
from time import sleep
import multiprocessing as mp
def HelloWord():
    name=mp.current_process().name
    sleep(1)#延迟1秒
    print('Hello world! current_process name=%s \n'%name)
    return
if __name__ == "__main__":
    name=mp.current_process().name
    print('Hello world! current_process name=%s \n'%name)
    for i in range(4):
        p = mp.Process(target=HelloWord)
        p.start()
        p.join()
        name=mp.current_process().name
    print('Hello world! current_process name=%s \n'%name)
打开 Anaconda Prompt，键入以下命令（文件位置要换成自己的位置）：
python E:\MyPython\并行计算\eg4_1.py
```

输出结果如下：

```
Hello world! current_process name=MainProcess
Hello world! current_process name=Process-1
Hello world! current_process name=Process-2
Hello world! current_process name=Process-3
Hello world! current_process name=Process-4
Hello world! current_process name=MainProcess
```

一般来讲，在主程序中创建并启动子进程，然后主进程有处理数据、与子进程通信等工作，并等待子进程完成任务，基于这种框架，上述程序修改为：

```

from time import sleep
import multiprocessing as mp
N_core=4#进程数
def HelloWord(ID):
    name=mp.current_process().name
    sleep(1)#延迟1秒
    print('Hello world! current_process name=%s, ID=%d \n'%(name, ID))
    return
if __name__ == "__main__":
    name=mp.current_process().name#输出主进程名
    print('Hello world! current_process name=%s \n'%name)
processes = []#进程列表
for id in range(N_core):#创建子进程
    p = mp.Process(target=HelloWord, args=(id+1, ))
    p.start()
    processes.append(p)
name=mp.current_process().name#输出主进程名
print('Hello world! current_process name=%s \n'%name)
#主进程处理
#.....
#主进程处理结束
#等待各子进程处理结果
for p in processes:
    p.join()

```

## 2. Pool 进程池

Pool 进程池可以提供指定数量的进程供用户调用，当有新的请求提交到 Pool 中时，如果池还没有满，就会创建一个新的进程来执行请求。如果池满，请求就会告知先等待，直到池中有进程结束，才会创建新的进程来执行这些请求。

Pool 进程池常见的方法有：

[apply 函数](#)，原型：apply(func[, args=(), kwds={}])

该函数用于传递不定参数，使用阻塞方式调用 func，执行完一个进程后再去执行其它进程。

[apply\\_async 函数](#)，原型：apply\_async(func[, args=(), kwds={}, callback=None])

该函数与 apply 用法一致，但它是非阻塞的且支持结果返回后进行回调。能够多个线程同时异步执行。

[map\(\) 函数](#)，原型：map(func, iterable[, chunksize=None])

该函数与内置的 map 函数用法行为基本一致，它会使进程阻塞直到结果返回。注意：虽然第二个参数是一个迭代器，但在实际使用中，必须在整个队列都就绪后，程序才会运行子进程。

[map\\_async\(\) 函数](#)，原型：map\_async(func, iterable[, chunksize[, callback]])

该函数与 map 用法一致，但是它是非阻塞的。其有关事项见 apply\_async。

[close 函数](#)：关闭进程池，使其不再接受新的任务。

[terminal](#)：结束工作进程，不再处理未处理的任务。

**join 函数**: 表示主进程阻塞等待子进程的退出, join 要在 close 或 terminate 之后使用。

**例 15.5** 用 Pool 进程池完成例15.4。

```
from time import sleep
import multiprocessing as mp
N_core=4#进程数
def HelloWord(ID):
    name = mp.current_process().name
    sleep(1)#延迟1秒
    print('Hello world! current_process name=%s, ID=%d \n'%(name, ID))
    return
if __name__ == "__main__":
    name = mp.current_process().name#输出主进程名
    print('Hello world! current_process name=%s \n'%name)
    pools = []#进程池列表
    pool = mp.Pool(processes=N_core)#创建进程池
    for id in range(N_core):
        pools.append(pool.apply_async(HelloWord, args = (id+1,)))
    pool.close()
    name=mp.current_process().name#输出主进程名
    print('Hello world! current_process name=%s \n'%name)
    #主进程处理
    #.....
    #主进程处理结束
    #等待各子进程处理结果
    pool.join()
打开 Anaconda Prompt, 键入以下命令 (文件位置要换成自己的位置) :
python E:\MyPython\并行计算\eg5.py
```

输出结果如下:

```
Hello world! current_process name=MainProcess
Hello world! current_process name=MainProcess
Hello world! current_process name=SpawnPoolWorker-1, ID=1
Hello world! current_process name=SpawnPoolWorker-2, ID=2
Hello world! current_process name=SpawnPoolWorker-3, ID=3
Hello world! current_process name=SpawnPoolWorker-4, ID=4
```

### 15.3.2 进程间通信

在 multiprocessing 中, 进程间通信主要有**数据共享**和**数据传递**, 本书只介绍数据传递, 而数据传递有**队列** (Queue) 和**管道** (Pipe) 两种, 主要区别为:

- (1) Queue 使用 put 和 get 维护队列, Pipe 使用 send 和 recv 维护队列。
- (2) Pipe 只提供两个端点, 而 Queue 没有限制。

- (3) Queue 的封装比较好，Queue 只提供一个结果，可以被多个进程同时调用；而 Pipe 返回两个结果，分别由两个进程调用。
- (4) Queue 的实现基于 Pipe，所以 Pipe 的运行速度比 Queue 快很多。
- (5) 当只需要两个进程时，Pipe 更快，当需要多个进程同时操作队列时，使用 Queue。

### 1. 队列 (Queue)

队列是一个先进先出 (FIFO) 的数据结构，很多场景需要按先来后到的顺序进行处理。其形式为：

```
class multiprocessing.Queue([maxsize])
```

Queue 返回一个使用一个管道和少量锁和信号量实现的共享队列实例。当一个进程将一个对象放进队列中时，一个写入线程会启动并将对象从缓冲区写入管道中。

除了 task\_done() 和 join() 之外，Queue 实现了标准库类 queue.Queue 中所有的方法。常见方法有：

```
put(obj[, block[, timeout]])
```

将 obj 放入队列。如果可选参数 block 是 True (默认值)，而且 timeout 是 None (默认值)，将会阻塞当前进程，直到有空的缓冲槽。如果 timeout 是正数，将会在阻塞了最多 timeout 秒之后还是没有可用的缓冲槽时抛出 queue.Full 异常。反之 (block 是 False 时)，仅当有可用缓冲槽时才放入对象，否则抛出 queue.Full 异常 (在这种情形下 timeout 参数会被忽略)。

```
get([block[, timeout]])
```

从队列中取出并返回对象。如果可选参数 block 是 True (默认值) 而且 timeout 是 None (默认值)，将会阻塞当前进程，直到队列中出现可用的对象。如果 timeout 是正数，将会在阻塞了最多 timeout 秒之后还是没有可用的对象时抛出 queue.Empty 异常。反之 (block 是 False 时)，仅当有可用对象能够取出时返回，否则抛出 queue.Empty 异常 (在这种情形下 timeout 参数会被忽略)。

```
close()
```

指示当前进程将不会再往队列中放入对象。一旦所有缓冲区中的数据被写入管道之后，后台的线程会退出。

```
join_thread()
```

等待后台线程。这个方法仅在调用了 close() 方法之后可用。这会阻塞当前进程，直到后台线程退出，确保所有缓冲区中的数据都被写入管道中。默认情况下，如果一个不是队列创建者的进程试图退出，它会尝试等待这个队列的后台线程。这个进程可以使用 cancel\_join\_thread() 让 join\_thread() 方法什么都不做直接跳过。

```
cancel_join_thread()
```

防止 join\_thread() 方法阻塞当前进程。具体而言，这防止进程退出时自动等待后台线程退出。

**例 15.6** 定义函数 HelloWord，形参为队列和 ID 号，获得当前进程的 name，把形如 [‘Hello world!’, ID, name] 这样的列表加入队列中。在主进程中创建 4 个子进程，同时在主进程中得到并输出子进程的列表。

```
import multiprocessing as mp
N_core=4#进程数
def HelloWord(q, ID):
    name=mp.current_process().name
    q.put([ 'Hello world!', ID, name])
    return
if __name__ == '__main__':
    q = mp.Queue()
    processes = []#进程列表
    for id in range(N_core):
        p = mp.Process(target=HelloWord, args=(q, id+1))
        p.start()
        processes.append(p)
    print(q.get())
    for p in processes:#等待各子进程处理结果
        p.join()
打开Anaconda Prompt，键入以下命令（文件位置要换成自己的位置）：
python E:\MyPython\并行计算\eg6.py
```

输出结果如下：

```
['Hello world!', 1, 'Process-1']
['Hello world!', 2, 'Process-2']
['Hello world!', 3, 'Process-3']
['Hello world!', 4, 'Process-4']
```

## 2. 管道 (Pipe)

管道返回一个由管道连接的连接对象，默认情况下是双向。每个连接对象都有 send() 和 recv() 方法（相互之间的）。请注意，如果两个进程（或线程）同时尝试读取或写入管道的同一端，则管道中的数据可能会损坏。当然，在不同进程中同时使用管道的不同端的情况下不存在损坏的风险。其形式为：

`multiprocessing.Pipe([duplex])`

返回一对 Connection 对象 (conn1, conn2)，分别表示管道的两端。如果 duplex 被置为 True（默认值），那么该管道是双向的。如果 duplex 被置为 False，那么该管道是单向的，即 conn1 只能用于接收消息，而 conn2 仅能用于发送消息。

**例 15.7** 定义函数 HelloWord，形参为队列和 ID 号，获得当前进程的 name，把形如 [‘Hello world!’, ID, name] 这样的列表发送到主进程中。在主进程中创建 4 个子进程，同时在主进程中接收子进程传输的数据。

```
import multiprocessing as mp
N_core=4#进程数
def HelloWord(conn, ID):
```

```
name=mp.current_process().name
conn.send(['Hello world!', ID, name])
conn.close()
return

if __name__ == '__main__':
    parent_conn, child_conn = mp.Pipe()
    processes = []#进程列表
    for id in range(N_core):
        p = mp.Process(target=HelloWord, args=(child_conn,id+1))
        p.start()
        processes.append(p)
    print(parent_conn.recv())
    for p in processes:#等待各子进程处理结果
        p.join()
```

运行情况同例15.6。

### 15.3.3 同步

在 threading 模块中的同步机制，在多进程 multiprocessing 模块中也有，比如锁机制、事件、信号量、栅栏。

## 15.4 基于 GPU 线程的并行计算

前面两节是基于 CPU 的并行计算，实际上，在机器学习和深度学习中，使用更多的是基于 GPU 线程的并行计算，其优势在于核数众多，计算速度更快，特别对于矩阵乘法和卷积具有极大的计算优势。另外，GPU 还拥有大量且快速的寄存器以及 L1 缓存的易于编程性，使得 GPU 非常适合用于深度学习。目前，常用的 GPU 计算范式是来自英伟达的 CUDA 并行计算框架，这种架构可以通过 GPU 加速使得机器学习并行化。

### 15.4.1 CUDA 基本概念

CUDA 并不是一个独立运行的计算平台，而需要与 CPU 协同工作，可以看成是 CPU 的协处理器，因此当我们在说 CUDA 并行计算时，其实是指的基于 CPU+GPU 的异构计算架构。在异构计算架构中，GPU 与 CPU 通过 PCIe 总线连接在一起协同工作，如图15.9所示。

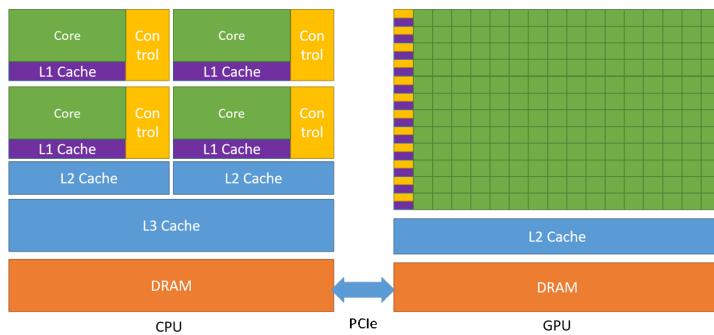


图 15.9 基于 CPU+GPU 的异构计算

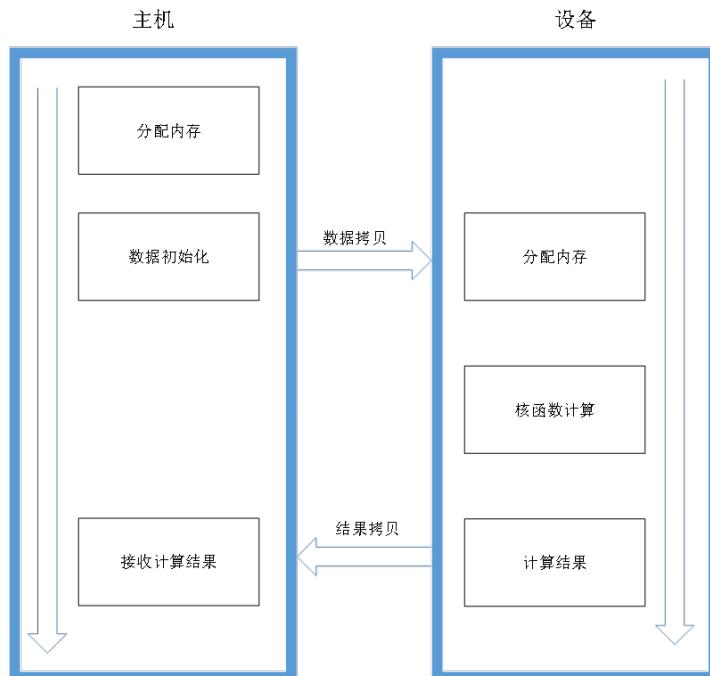


图 15.10 CUDA 执行流程

其中，CPU 所在位置称为为主机 (Host)，又称为主处理器，在并行计算中，主要涉及 CPU 及其内存，而 GPU 所在位置称为设备 (Device)，又称为协处理器，在并行计算中，主要涉及 GPU 及其内存。在 CUDA 程序中既包含主机程序，又包含设备程序，它们分别在 CPU 和 GPU 上运行。同时，主机与设备之间可以进行通信，即可以相互进行数据拷贝。在设备上的计算通过核函数 (Kernel) 形式执行，该函数以 `_global_` 为前缀作为标记。如图 15.10 所示，典型的 CUDA 程序的执行流程如下：

- (1) 分配主机内存，并进行数据初始化；
- (2) 配设备内存，并从主机将数据拷贝到设备上；
- (3) 调用 CUDA 的核函数在设备上完成指定的运算；

- (4) 将设备上的运算结果拷贝到主机上;
- (5) 释放设备和主机上分配的内存。

### 15.4.2 CUDA 线程组织

如图15.11所示, CUDA 二维线程组织从逻辑层面来讲, 主要分两层, 外层称为网格 (Grid), 同一个网格上的线程共享相同的全局内存空间, 里层称为线程块 (Block)。对于网格, 是由若干线程块组成, 而每个线程块里面包含很多线程。线程坐标的原点在左上角, 向右为  $x$  方向, 向左为  $y$  方向。

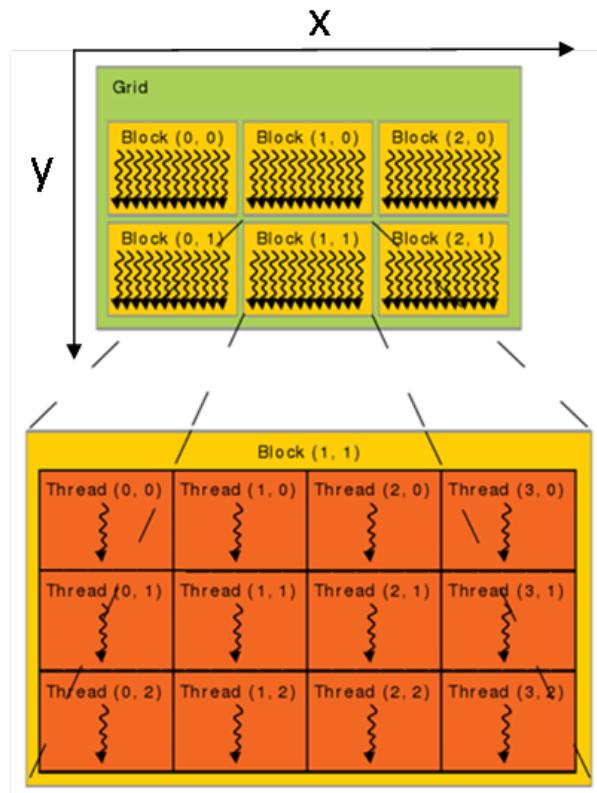


图 15.11 CUDA 二维线程组织

网格和线程块都是定义为 `dim3` 类型的变量, `dim3` 可以看成是包含三个无符号整数  $(x, y, z)$  成员的结构体变量, 在定义时, 缺省值初始化为 1。因此, 网格和线程块可以灵活地定义为 1-dim, 2-dim 以及 3-dim 结构。核函数在调用时也必须通过执行配置 `<<<grid, block>>>` 来指定核函数所使用的线程数及结构。图15.11中的网格和线程块可以这样定义:

```
dim3 grid(3, 2,1);
dim3 block(4, 3,1);
kernel_fun<<< grid, block >>>(prams...);
```

所以，一个线程需要两个内置的坐标变量 (blockIdx,threadIdx) 来唯一标识，它们都是 dim3 类型变量，其中 blockIdx 指明线程所在 grid 中的位置，而 threadIdx 指明线程所在 block 中的位置，如图中的 Thread(1,1) 满足：threadIdx.x = 1, threadIdx.y = 1, blockIdx.x = 1, blockIdx.y = 1。

对于每个线程，以一维为例，在实际计算中，坐标  $x$  的计算方式如图15.12，箭头所在位置的线程的坐标  $x = 5$ .

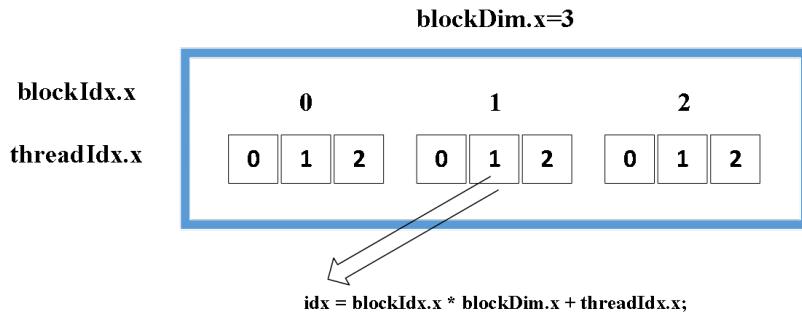


图 15.12 线程坐标  $x$  的计算方式

同理，可以得到二维的索引坐标： $idx = blockIdx.x * blockDim.x + threadIdx.x;$   
 $idy = blockIdx.y * blockDim.y + threadIdx.y$ 。

对于核函数来讲，一般情况下，一个核函数运行在一个网格上，如图15.13所示。

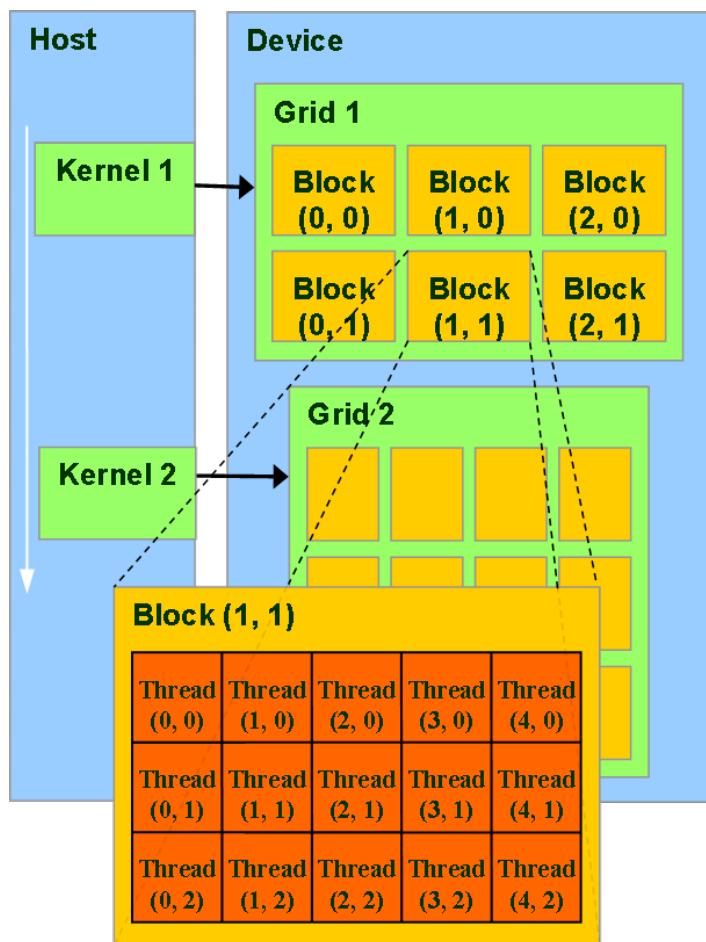


图 15.13 核函数与网格

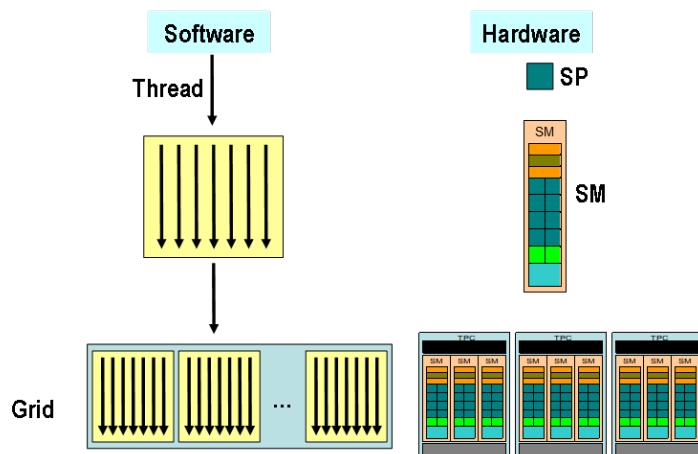


图 15.14 CUDA 编程的逻辑层和物理层

上面讲的是从软件逻辑层方面的线程组织方式。如图15.14所示，从显卡的物理层面来看，核心组件是流式多处理器（Streaming Multiprocessor），简称SM，包括CUDA核心（也称为Streaming Processor（SP））、共享内存、寄存器等，SM可以并发地执行数百个线程，并发能力就取决于SM所拥有的资源数。当一个核函数被执行时，它的网格中的线程块被分配到SM上，一个线程块只能在一个SM上被调度，而SM一般可以调度多个线程块，基本的执行单元是线程束（Warps），线程束包含32个线程，这些线程同时执行相同的指令，但是每个线程都包含自己的指令地址计数器和寄存器状态，也有自己独立的执行路径。

### 15.4.3 CUDA 内存组织

CUDA 内存组织如图15.15所示。可以看到，每个线程有自己的私有本地内存（Local Memory），而每个线程块有包含共享内存（Shared Memory），可以被线程块中所有线程共享，其生命周期与线程块一致。此外，所有的线程都可以访问全局内存（Global Memory）。还可以访问一些只读内存块：常量内存（Constant Memory）和纹理内存（Texture Memory）。

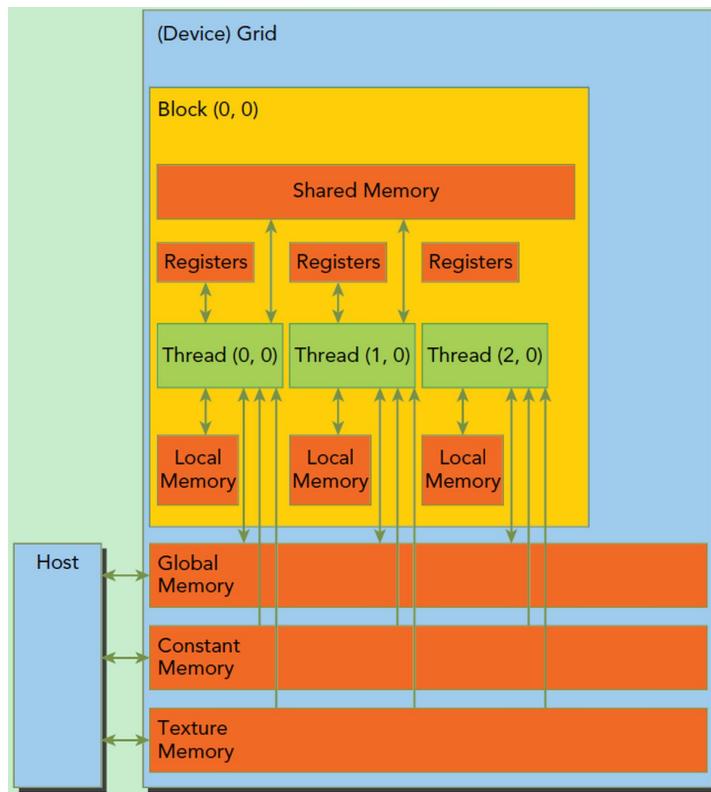


图 15.15 CUDA 内存组织

#### 15.4.4 PyCUDA

PyCUDA 作为 Python 的第三方库，可以访问 NVIDIA 的 CUDA 并行计算 API。它的异构编程模式和前面 CUDA 一致，它自己具有的特点是：

(1) 对象的自动清理。这使编写正确、无泄漏和无崩溃的代码变得更加容易。PyCUDA 在其分配的所有内存被释放之前，不会与上下文分离。

(2) 编程方便。提供了 pycuda.compiler.SourceModule 和 pycuda.gpuarray.GPUArray 等接口，让 CUDA 编程甚至比 NVIDIA 的基于 C 语言的运行时更方便。

(3) 完整支持 CUDA。PyCUDA 能够完全支持 CUDA 的 API。并能够自动进行错误检查，把所有 CUDA 错误自动转换为 Python 异常。

(4) 加速效果好。PyCUDA 的基础层是用 C++ 编写的，因此计算速度比较快。

##### 1. pycuda 安装

到官网下载 CUDA 包并安装，到微软官网下载 Visual Studio 2015 以上的 C++ 开发工具并安装。到 Python 环境使用 pip install pycuda 安装 pycuda。

##### 2. 实现过程

下面以随机生成的两个向量的点积计算为例，讲解实现过程。其中，driver 是导入 CUDA 的 API，autoinit 是用来启动和自动初始化 GPU 系统，SourceModule 是 NVIDIA 编译器（nvcc）的指令，指示要编译并上传到设备的对象。很明显，以下核函数是 C 语言形式。

```
#导入相关库
import pycuda.driver as cuda
import pycuda.autoinit
import numpy
from pycuda.compiler import SourceModule
#在CPU生成随机向量
a = numpy.random.randn(5).astype(numpy.float32)
b = numpy.random.randn(5).astype(numpy.float32)
#定义核函数
mod = SourceModule("""
__global__ void vector_dot(float *dest, float *a, float *b)
{const int i = threadIdx.x;dest[i] = a[i] * b[i];}
""")
func = mod.get_function("vector_dot")
#线程设置和执行核函数
dest = numpy.zeros_like(a)
func(drv.Out(dest), drv.In(a), drv.In(b), block=(400,1,1), grid=(1,1))
print(dest)#输出结果
```

### 15.4.5 TensorFlow

TensorFlow 是一个用于设计和部署数值计算的软件库，主要专注于机器学习中的应用程序。借助这个库，可以将算法描述为相关运算的图形，这些运算可以在各种支持 GPU 的平台上执行，包括便携式设备、台式机和高端服务器。

安装好 GPU 版的 TensorFlow 后，使用下面程序来测试 Tensorflow 是否支持 GPU：

```
import tensorflow as tf
tf.config.list_physical_devices()
```

如下输出结果表示支持 GPU：

```
[PhysicalDevice(name='/physical_device:CPU:0', device_type='CPU'),
PhysicalDevice(name='/physical_device:GPU:0', device_type='GPU')]
```

可以用以下程序测试 CPU 和 GPU 下运行时间：

```
import tensorflow as tf
import timeit
def cpu_run():
    with tf.device('/cpu:0'):
        c = tf.matmul(cpu_a, cpu_b)
    return c
def gpu_run():
    with tf.device('/gpu:0'):
        c = tf.matmul(gpu_a, gpu_b)
    return c
cpu_time = timeit.timeit(cpu_run, number=10)
gpu_time = timeit.timeit(gpu_run, number=10)
print('run time:', cpu_time, gpu_time)
```

## 15.5 并行计算实践

以上内容介绍了并行计算的相关概念以及例题的讲解也是在 Python 框架讲解的。在本节我们除了继续介绍 Python 语言实践之外，也会补充 R 语言实践。

### 15.5.1 R 语言实践

在 R 环境下的并行计算主要通过它内置的任务并行包来实现，包括 parallel 包和 foreach 包，其中，parallel 包是针对任务的粗粒度并行计算，而 foreach 包则是针对 for 循环的细粒度并行计算。接下来分别介绍这两个包。

## parallel 包

parallel 包是建立在 multicore 和 snow 包的基础上，涵盖了这两个包的大多数函数，并且加入了随机数的生成。由于是粗粒度并行计算，parallel 包处理比较大的计算块，一个典型的例子是在许多不同的数据集上计算相同的 R 函数。关键的一点是，这些计算块是不相关的，不需要以任何方式进行通信。通常情况下，这些块的时间长度大致相同，也就是负载基本均衡。其计算模型如下：

- (1) 启动  $M$  个 worker 进程（计算进程），并在 worker 进程上做任何需要的初始化。
- (2) 将每项任务所需的任何数据发送给 worker 进程。
- (3) 将任务分成大小大致相等的  $M$  个块，并将这些块（包括所需的 R 代码）发送给 worker 进程。
- (4) 等待所有的 worker 进程完成它们的任务，并获得它们的结果。
- (5) 对其它任务重复步骤 (2) - (4)。
- (6) 关闭 worker 进程。

在并行计算前，先要获得该计算设备的核数，可以使用函数 detectCores 获取，一般用 detectCores(logical=F) 获取实际物理核心数。然后初始化并行计算环境的工作是由 cl=makeCluster(cores) 完成，它可以申请 cores 个核参与计算。对于并行任务分配主要有两种方法：

一是针对表达式进行并行处理，由 clusterEvalQ(cl,expr) 函数利用创建的 cl 执行 expr。这里利用刚才创建的 cl 核并行计算 expr。expr 是执行命令的语句，不过如果命令太长的话，一般写到文件里比较好。比如把想执行的命令放在 Rcode.r 里：

```
clusterEvalQ(cl,source(file="Rcode.r"))
```

二是采用 par 开头的 apply 函数族进行任务并行计算。这族函数和 apply 的用法基本一样，不过要多加一个参数 cl。一般如果 cl 创建如上面 cl <- makeCluster(cl.cores) 的话，这个参数可以直接用作 parApply(cl=cl,...)。当然 Apply 也可以是 Sapply,Lapply 等等。注意 par 后面的第一个字母是要大写，而一般的 apply 函数族第一个字母不大写。

另外，parallel 包可以通过 library(help = “parallel”) 来查找相关信息。

**例 15.8** 定义 HelloWorld 函数，使其输出 Hello World! 和所在核的 ID 号。利用 parLapply 函数完成并行计算。

```
library(parallel)
HelloWorld <- function(i){
  sprintf('Hello World! The number:%d core ID is %d',i,Sys.getpid())
}
#cores <- detectCores()#获取计算机线程数
cores <- detectCores(logical=F)#获取实际物理核心数
cl <- makeCluster(cores)#初始化并行计算环境，申请cores个核参与计算
```

```
parLapply(cl,1:cores,HelloWorld)
stopCluster(cl)#终止并行计算
#输出结果如下:
[[1]]
[1] "Hello World!The number:1 core ID is 15472"
[[2]]
[1] "Hello World!The number:2 core ID is 18184"
[[3]]
[1] "Hello World!The number:3 core ID is 9168"
[[4]]
[1] "Hello World!The number:4 core ID is 3332"
```

注意每次计算的 ID 地址不一定一样。

**例 15.9** 利用 runif 函数生成 [0,1] 之间的随机向量，并进行向量加法运算。分别采用串行和并行计算，给出计算时间。

```
a <- runif(1000000, 0, 1)#生成[0,1]的随机向量
b <- runif(1000000, 0, 1)#生成[0,1]的随机向量
#串行计算
serial_logical_compute <-function(a,b){
  if(length(a) != length(b)){
    print("向量长度不一致")
    return(0)
  }
  c<-integer(length(a))#生成初始化为0的向量
  for(i in 1:length(a)){
    c[i] <- a[i] * b[i]
  }
  sum_c <- sum(c)
  print(sum_c)
}

start_time <- Sys.time()#起始时间
serial_logical_compute(a,b)
end_time <- Sys.time()#结束时间
s_time <- end_time-start_time
print(s_time)

#并行计算
library(parallel)
parallel_logical_compute <-function(id)
{
  if(length(a) != length(b)){
    print("向量长度不一致")
    return(0)
  }
  c<-integer(length(a))#生成初始化为0的向量
  for(i in seq(id,length(a),cores)){
    c[i] <- a[i] * b[i]
  }
  sum_c <- sum(c)
  print(sum_c)
```

```

}

start_time <- Sys.time() #起始时间
cores <- detectCores(logical=F) #获取实际物理核心数
cl <- makeCluster(cores,type="FORK") #初始化并行计算环境，申请cores个核参与计算
clusterExport(cl,c("a","b","cores")) #传入外部变量
result <- parLapply(cl,1:cores,parallel_logical_compute)
sum_c <- do.call(sum,result)
print(sum_c)
stopCluster(cl) #终止并行计算
end_time <- Sys.time() #结束时间
p_time <- end_time - start_time
print(p_time)
#输出结果如下：
[1] 250111
Time difference of 0.120677 secs
[1] 250111
Time difference of 1.94301 secs

```

从计算时间上看，并行计算的时间反而更长。当计算量少的时候，使用 parLapply 执行任务并行时，创建并行计算环境的时间远大于计算本身的时间，计算效率并不高。

**例 15.10** 利用 rnorm 生成 100 行 500000 列的随机矩阵，对矩阵的每列进行从小到大排序，并获取前 5 行数据。分别用串行计算和并行计算实现，计算加速比。

```

M <- matrix(rnorm(50000000), 100, 500000) #随机初始化矩阵
mysort <- function(x){
  return(sort(x)[1:5]) #排序并返回前5数据
}
do_apply <- function(M) {
  return(apply(M, 2, mysort)) #对M矩阵每列从小到大排序
}
do_parallel <- function(M, ncl){ #使用ncl个核并行处理M矩阵每列从小到大排序
  cl <- makeCluster(getOption("cl.cores", ncl))
  ans <- parApply(cl, M, 2, mysort)
  stopCluster(cl)
  return(ans)
}
#串行计算
start_time <- Sys.time() #起始时间
ans1 <- do_apply(M)
end_time <- Sys.time() #结束时间
s_time <- as.double(difftime(end_time, start_time), units="secs") #按秒计算时间差
sprintf('Serial computing time:%f secs', s_time)
#并行计算
library(parallel)
start_time <- Sys.time() #起始时间
cores <- detectCores(logical=F) #获取实际物理核心数
cl <- makeCluster(cores) #初始化并行计算环境，申请cores个核参与计算
ans2 <- do_parallel(M, cores)
end_time <- Sys.time() #结束时间

```

```

p_time <- as.double(difftime(end_time,start_time),units="secs")#按秒计算时间差
sprintf('Parallel computing time:%f secs',p_time)
sprintf('Acceleration ratio:%f',s_time/p_time)
#输出结果如下:
[1] "Serial computing time:27.547455 secs"
[1] "Parallel computing time:17.326461 secs"
[1] "Acceleration ratio:1.589907"

```

从结果可以看出，数据计算之间无依赖性，且计算量非常大的时候，并行计算可以提高计算效率。

### foreach 包

foreach 包是一个支持在 R 语言中调用多进程功能的第三方包，但必须与 doParallel 包一同使用，doParallel 包使 foreach 并行计算成为可能。foreach 的功能类似于 for 或者 lapply 函数，其最大的好处在于代码简单而且容易采用并行方式进行计算。foreach 函数形式为：

```

foreach(..., .combine, .init, .final=NULL, .inorder=TRUE,
       .multicombine=FALSE,
       .maxcombine=if (.multicombine) 100 else 2,
       .errorhandling=c('stop', 'remove', 'pass'),
       .packages=NULL, .export=NULL, .noexport=NULL,
       .verbose=FALSE)

when(cond)
e1 %: % e2
obj %do% ex
obj %dopar% ex
times(n)

```

各参数说明如下：

- (1) 第 1 个参数是必备参数，即必须有输入参数，结果默认返回 list；
- (2) .combine：运算之后结果的显示方式，default 是 list，“c” 返回 vector，cbind 和 rbind 返回矩阵，“+”和“\*”可以返回 rbind 之后的“+”或者“\*”，相当于归约操作；
- (3) .init：.combine 函数的第一个变量；
- (4) .final：返回最后结果；
- (5) .inorder：TRUE 则返回和原始输入相同顺序的结果（对结果的顺序要求严格的时候），FALSE 返回没有顺序的结果（可以提高运算效率）。这个参数适合于设定对结果顺序没有需求的情况；
- (6) .multicombine：设定.combine 函数的传递参数，default 是 FALSE 表示其参数是 2，TRUE 可以设定多个参数；
- (7) .maxcombine：设定.combine 的最大参数；
- (8) .errorhandling：如果循环中出现错误，对错误的处理方法；

- (9) .packages: 指定在 %dopar% 运算过程中依赖的 package(%do% 会忽略这个选项), 用于并行一些机器学习算法;
- (10) .export: 在编译函数的时候需要预先加载一些内容进去, 类似 parallel 的 clusterExport;
- (11) e1 %:% e2, 对象合并;
- (12) %do% 严格按照顺序执行任务(串行计算), %dopar% 并行执行任务, %do% 时候就像 sapply 或 lapply, %dopar% 就是并行启动器;
- (13) times(n), 返回计算时间, 不过几乎不用。

**例 15.11** 对于级数  $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$ , 令  $s(n) = \sum_{i=1}^n \frac{1}{i^2}$ , 计算当  $n=10000$  时的值。

```
library("foreach")
library("doParallel")
N <- 10000
pi <- 3.1415926535
#串行计算
start_time <- Sys.time()#起始时间
sum <- foreach(i=1:N,.combine="+") %do% {1.0/i^2}
end_time <- Sys.time()#结束时间
s_time <- as.double(difftime(end_time,start_time),units="secs")#按秒计算时间差
sprintf('Serial computing time:%f secs',s_time)
sprintf('Serial computing result:%f,the error:%f',sum,abs(sum-pi^2/6))
#并行计算
start_time <- Sys.time()#起始时间
cores <- detectCores(logical=F)#获取实际物理核心数
cl <- makeCluster(cores)#初始化并行计算环境, 申请cores个核参与计算
registerDoParallel(cl)#进行进程注册
sum <- foreach(i=1:N,.combine="+") %dopar% {1.0/i^2}
stopCluster(cl)
end_time <- Sys.time()#结束时间
p_time <- as.double(difftime(end_time,start_time),units="secs")#按秒计算时间差
sprintf('Parallel computing time:%f secs',p_time)
sprintf('Parallel computing result:%f,the error:$\%$f',sum,abs(sum-pi^2/6))
sprintf('Acceleration ratio:%f ',s_time/p_time)
#输出结果如下:
[1] "Serial computing time:4.662530 secs"
[1] "Serial computing result:1.644834,the error:0.000100"
[1] "Parallel computing time:6.940762 secs"
[1] "Parallel computing result:1.644834,the error:0.000100"
[1] "Acceleration ratio:0.671761"
```

从结果上看, 计算值是一样的, 但并行计算没有发挥出作用来, 原因和前面例题中的情况一样, 也是因为创建并行计算的环境的时间比计算本身还耗时, 实际计算时应避免这种情况。

### 蒙特卡罗算法计算 $\pi$ 值

**例 15.12** 单位圆的  $1/4$  面积是一个扇形，它是边长为 1 单位正方形的一部分，如图15.16所示。只要能求出扇形面积  $S_1$  在正方形面积  $S$  中占的比例  $K = S_1/S$  就立即能得到  $S_1$ ，从而得到  $\pi$  的值。怎样求出扇形面积在正方形面积中占的比例  $K$  呢？一个办法是在正方形中随机投入很多点，使所投的点落在正方形中每一个位置的机会相等看其中有多少个点落在扇形内。将落在扇形内的点数  $m$  与所投点的总数  $n$  的比  $m/n$  作为  $K$  的近似值。点落在扇形内的充要条件是  $x^2 + y^2 \leq 1$ 。则

$$\frac{m}{n} = \frac{S_1}{S} = \frac{4}{\pi}, \pi = \frac{4m}{n}$$

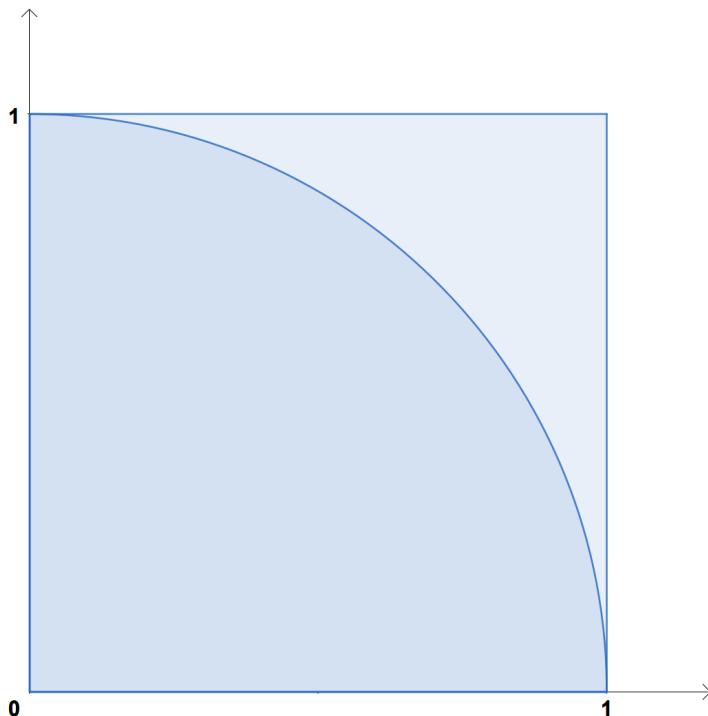


图 15.16 四分之一圆域

基于 R 语言的计算方法如下：

```
#串行计算
Serial_pi <- function(){
  N <- 1000000
  cnt <- 0
  for (i in 1:N){
    x <- runif(1, 0, 1)
    y <- runif(1, 0, 1)
    if ((x * x + y * y) < 1)
```

```

        cnt = cnt + 1
    }
    return (4 * cnt / N)
}

start_time <- Sys.time()#起始时间
result <- Serial_pi()
end_time <- Sys.time()#结束时间
s_time <- as.double(difftime(end_time,start_time),units="secs")#按秒计算时间差
sprintf('Serial computing time:%f secs',s_time)
sprintf('Serial computing result:%f',result)

#并行计算
library(parallel)
parallel_pi <- function(id){
    N <- 10000000
    cnt <- 0
    for (i in seq(id,N,cores)){
        x <- runif(1, 0, 1)
        y <- runif(1, 0, 1)
        if ((x * x + y * y) < 1)
            cnt = cnt + 1
    }
    return (4 * cnt / N)
}

start_time <- Sys.time()#起始时间
cores <- detectCores(logical=F)#获取实际物理核心数
cl <- makeCluster(cores)#初始化并行计算环境，申请cores个核参与计算
clusterExport(cl,c("cores"))#传入外部变量
result <- parLapply(cl,1:cores,parallel_pi)
result <- do.call(sum,result)
end_time <- Sys.time()#结束时间
p_time <- as.double(difftime(end_time,start_time),units="secs")#按秒计算时间差
sprintf('Parallel computing time:%f secs',p_time)
sprintf('Parallel computing result:%f',result)
sprintf('Acceleration ratio:%f',s_time/p_time)

#输出结果如下:
[1] "Serial computing time:11.167637 secs"
[1] "Serial computing result:3.141252"
[1] "Parallel computing time:2.148666 secs"
[1] "Parallel computing result:3.143928"
[1] "Acceleration ratio:5.197474"

```

从结果看，并行计算效果非常好，已经是超线性加速了。

### 15.5.2 Python 语言实践

#### 蒙特卡罗算法计算 $\pi$ 值

在上一小节中用 R 语言给出了具体计算过程，问题的详细说明也见上一小节，这里再用 Python 语言给出计算方法。容易写出其串行程序，具体程序如下：

```

from random import random
from math import sqrt
from time import perf_counter
times=10000000
counts=0
start = perf_counter()#计时开始
for i in range(1,times):
    x,y=random(),random()
    dist=sqrt(x**2+y**2)
    if dist <= 1.0:
        counts=counts+1
pi = 4*(counts/times)
end = perf_counter()#计时结束
print('pi=%f'%pi)
print('运行时间为: %s Seconds'%(end-start))

```

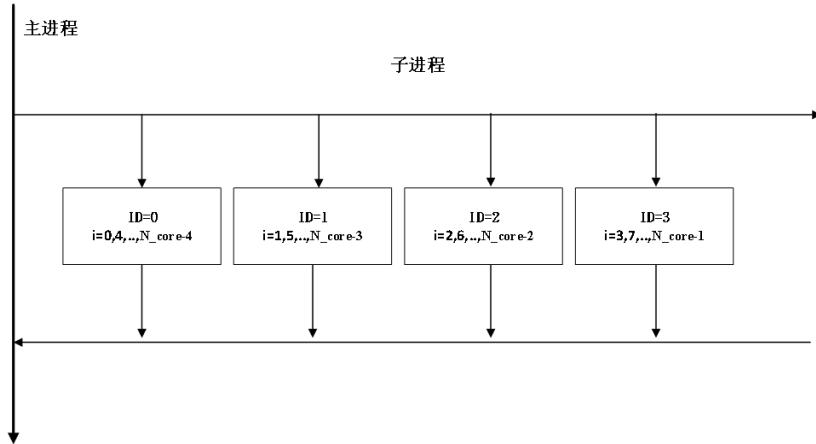


图 15.17 用 ID 号划分计算任务

从上面的程序中可以看出，计算是独立的，也就不存在数据竞争，完全可以进行并行计算。如果利用 multiprocessing 进行多核并行计算，只需要合理设计进程函数，每个进程通过调用该进程函数完成相应的计算，这就需要在进程函数内部合理分配计算任务，最好的方式就是均分，从而达到任务均衡。自定义进程 ID，若总进程为  $N_{core}$ ，则标记为  $0, 1, \dots, N_{core}-1$ 。如图15.17所示，根据线程 ID 号来分配不同的任务，把循环语句改造为：`for i in range(ID, times, N_core)` 这样线程负载均衡，可用 multiprocessing 中 map 函数完成。并行程序实现如下：

```

##蒙特卡洛算法并行计算Pi值
from random import random
from math import sqrt
from time import perf_counter
import multiprocessing as mp
times=10000000
N_core=4#进程数

```

```

def com_pi(ID):
    counts=0
    for i in range(ID,times,N_core):
        x,y=random(),random()
        dist=sqrt(x**2+y**2)
        if dist <= 1.0:
            counts=counts+1
    name=mp.current_process().name
    print('Current process name=%s, ID=%d \n'%(name, ID))
    return counts
if __name__ == "__main__":
    ids = list(range(0,N_core))#id列表
    start = perf_counter()#计时开始
    pool = mp.Pool(processes=N_core) #创建进程池
    pool_result=pool.map(com_pi,ids)
    pool.close()
    pool.join()
    counts=sum(pool_result)#把各子进程计算结果汇总
    pi = 4*(counts/times)
    end = perf_counter()#计时结束
    name=mp.current_process().name#主进程输出结果
    print('Hello world! current_process name=%s \n'%name)
    print('pi值=%f, 并行运行时间为: %s Seconds' %(pi,(end-start)))

```

打开 Anaconda Prompt，键入以下命令（文件位置要换成自己的位置），即可完成计算：python E:\MyPython\ 并行计算 \eg8.py，

为有利于计算加速比，把串行和并行程序放在一个文件 eg9.py 中，完整程序如下：

```

from random import random
from math import sqrt
from time import perf_counter
import multiprocessing as mp
times=10000000
N_core=4#进程数
def serial_com_pi():
    counts=0
    for i in range(1,times):
        x,y=random(),random()
        dist=sqrt(x**2+y**2)
        if dist <= 1.0:
            counts=counts+1
    return counts
def parallel_com_pi(ID):
    counts=0
    for i in range(ID,times,N_core):
        x,y=random(),random()
        dist=sqrt(x**2+y**2)
        if dist <= 1.0:
            counts=counts+1
    name=mp.current_process().name

```

```

print('Current process name=%s, ID=%d \n'%(name, ID))
return counts
if __name__ == "__main__":
    #串行计算
    start = perf_counter()#计时开始
    counts=serial_com_pi()
    pi = 4*(counts/times)
    end = perf_counter()#计时结束
    serial_time=end-start
    print('pi值=%f,串行运行时间为: %s Seconds' %(pi,serial_time))
    #并行计算
    ids = list(range(0,N_core))#id列表
    start = perf_counter()#计时开始
    pool = mp.Pool(processes=N_core)#创建进程池
    pool_result=pool.map(parallel_com_pi,ids)
    pool.close()
    pool.join()
    counts=sum(pool_result)#把各子进程计算结果汇总
    pi = 4*(counts/times)
    end = perf_counter()# 计时结束
    parallel_time = end - start
    name = mp.current_process().name#主进程输出结果
    print('Hello world! current_process name=%s \n'%name)
    print('pi值=%f,并行运行时间为: %s Seconds' %(pi,parallel_time))
    speedup = serial_time / parallel_time
    print('加速比=%f' %speedup)

```

打开 Anaconda Prompt，键入以下命令（文件位置要换成自己的位置），即可完成计算：python E:\MyPython\ 并行计算 \eg9.py，

## 15.6 总结

本章节通过给出进程、线程等并行计算概念，讨论了在 Python 和 R 语言环境下基于 CPU 的多核并行计算方式，并讨论了基于 CUDA 的 GPU 并行计算原理，但并未给出分布式并行计算的机器学习方法，请参考刘铁岩所著的《分布式机器学习：算法、理论与实践》。

## 15.7 练习

使用各种并行计算方法计算下列各题，并进行性能比较。

1. 使用级数或蒙特卡洛方法计算底数  $e$ 。
2. 使用级数或蒙特卡洛方法计算  $\pi$ 。

3. 计算级数  $\sum_{i=0}^{\infty} \frac{(-2)^i}{(2i+1)!}$ 。
4. 使用蒙特卡洛方法计算三重积分  $\int_1^2 \int_1^2 \int_1^2 (x^2 + y^2 + z^2) dx dy dz$ , 算法描述如下:

取 0 到 1 之间均匀分布的随机数点列

$$(t_0, t_1, \dots; t_{n-1}), i = 0, 1 \dots, m-1$$

令

$$x_j^j = a + (b_j - a)_j j^j j = 0, 1 \dots, n-1$$

取  $m$  足够大, 则有

$$s = \frac{1}{m} \sum_{i=0}^{n-1} f(x_0^i, x_1^i, \dots, x_n^i) \prod_{j=0}^{n-1} (b_j - a_j)$$

5. 使用本文中并行计算方法对前面各章涉及的机器学习算法 (可选择一两个) 进行并行化, 比较其计算效率, 并给出实验报告。

6. 对于例15.2和15.3, 请用 multiprocessing 模块中同名的锁和事件重新完成。

## 参考文献

- [1] Huang Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3), 283-304.
- [2] Pelleg D, & Moore A. (2000). X-means: extending K-means with efficient estimation of the number of the clusters[C].*Proceedings of the 17th International Conference on Machine Learning*. NewYork: ACM, 111-117.
- [3] Madan S, & Dana KJ. (2016). Modified balanced iterative reducing and clustering using hierarchies (m-BIRCH) for visual clustering. *Pattern Analysis and Applications*, 19(4), 1023-1040.
- [4] Trisminingsih R, & Shaztika SS. (2016). ST-DBSCAN clustering module in SpagoBI for hotspots distribution in Indonesia[C]//*Proceedings of the 3rd International Conference on Information Technology*. New York: ACM, 60-67.
- [5] Ankerst M, Breunig M, Kriegel HP, & Sander J. (1999). OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod record*, 28(2), 49-60.
- [6] McParland D, & Gormley IC. (2016). Model based clustering for mixed data: clustMD. *Advances in Data Analysis and Classification*, 10(2), 155-169.
- [7] Foss A, Markatou M, Ray B, & Heching A. (2016). A semiparametric method for clustering mixed data. *Machine Learning*, 105(3), 419-458.
- [8] Dhillon IS. (2001). Co-clustering documents and words using bipartite spectral graph partitioning [A]. In: *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* [C] ACM, 269-274.
- [9] Dhillon I, Mallela S, & Modha D. (2003). Information-theoretic co-clustering [A]. In: *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* [C]. ACM, 89-98.
- [10] Chi EC, Allen GI, & Baraniuk RG. (2016). Convex biclustering . *Biometrics*, 73(1): 10-19.
- [11] Pearson K. (1901). Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2), 559.

- [12] Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.
- [13] Van der Maaten L, & Hinton G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).
- [14] Hinton GE, & Roweis ST. (2003). Advances in Neural Information Processing Systems, Vol. 15.
- [15] Zou H, Hastie T, & Tibshirani R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2), 265-286.
- [16] Tucker LR. (1963). Implications of factor analysis of three way matrices for measurement of change. *Problems in measuring change*. Wisconsin:University of Wisconsin Press.
- [17] Levin J. (1963). Three-mode factor analysis[D]. Urbana: University of Illinois.
- [18] Wright J, Peng YG, Ma Y, et al. (2009). Robust principal component analysis: exact recovery of corrupted low-rank matrices by convex optimization[C]/ /Proceedings of Neural Information Processing Systems. Whistler: MIT Press.
- [19] 李航. (2019). *统计学习方法* (第 2 版). 北京: 清华大学出版社.
- [20] 范金城, 梅长林. (2002). *数据分析*. 北京: 科学出版社.
- [21] 高惠璇. (2005). *应用多元统计分析*. 北京: 北京大学出版社.
- [22] 何晓群. (2015). *多元统计分析*. 北京: 中国人民大学出版社.
- [23] 任雪松, 于秀林. (2011). *多元统计分析*. 北京: 中国统计出版社.
- [24] 王斌会. (2014). *多元统计分析及 R 语言建模*. 广州: 暨南大学出版社.
- [25] 王学民. (2017). *应用多元统计分析*. 上海: 上海财经大学出版社.
- [26] 袁志发, 周静芋. (2002). *多元统计分析*. 北京: 科学出版社.
- [27] 张润楚. (2006). *多元统计分析*. 北京: 科学出版社.
- [28] 何晓群. (2017). *应用回归分析* (第 5 版). 北京: 电子工业出版社.
- [29] 王黎明. (2019). *回归分析-概念、方法和应用*. 上海: 上海财经大学出版社.
- [30] 刘超. (2019). *回归分析-方法、数据与 R 的应用*. 北京: 高等教育出版社.
- [31] 唐年胜, 李会琼. (2014). *应用回归分析*. 北京: 科学出版社.

- [32] 王衡军. (2020). 机器学习. 北京: 清华大学出版社.
- [33] Horel AE. (1962). Applications of ridge analysis to regression problems. *Chem. Eng. Progress.*, 58, 54-59.
- [34] Golub GH, Heath M, & Wahba G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2), 215-223.
- [35] Bonat WH. (2018). Multiple response variables regression models in R: The mcglm package. *Journal of Statistical Software*, 84, 1-30.
- [36] Bonat WH, Walmes Marques Zeviani, & Fernando de Pol Mayer. (2018). Package ‘mcglm’ . <https://github.com/wbonat/mcglm>.
- [37] Bonat WH. (2018). mcglm: Multivariate Covariance Generalized Linear Models. <https://CRAN.R-project.org/package=mcglm>.
- [38] Anderson TW. (1973). Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure. *The Annals of Statistics*, 1(1), 135–141.
- [39] Pourahmadi M. (2000). Maximum Likelihood Estimation of Generalised Linear Models for Multivariate Normal Covariance Matrix. *Biometrika*, 87(2), 425–435.
- [40] Bonat WH, & Jørgensen B. (2016). Multivariate Covariance Generalized Linear Models. *Journal of the Royal Statistical Society C*, 65(5), 649–675.
- [41] Martinez-Beneito MA. (2013). A General Modelling Framework for Multivariate Disease Mapping. *Biometrika*, 100(3), 539–553.
- [42] Liang KY, & Zeger SL. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73(1), 13–22.
- [43] Jørgensen B, & Knudsen SJ. (2004). Parameter Orthogonality and Bias Adjustment for Estimating Functions. *Scandinavian Journal of Statistics*, 31(1), 93–114.
- [44] Bates D, & Maechler M. (2017). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.2-12. <https://CRAN.R-project.org/package=Matrix>.
- [45] Bivand R. (2017). spdep: Spatial Dependence: Weighting Schemes, Statistics and Models. R package version 0.7-4. <https://CRAN.R-project.org/package=spdep>.

- [46] Bonat WH. (2018). mcglm: Multivariate Covariance Generalized Linear Models. R package version 0.4.0. <https://CRAN.R-project.org/package=mcglm>.
- [47] Carey VJ. (2015). gee: Generalized Estimation Equation Solver. R package version 4.13-19. <https://CRAN.R-project.org/package=gee>.
- [48] Hadfield JD. (2010). MCMC Methods for Multi-Response Generalized Linear Mixed Models: The MCMCglmm R Package. *Journal of Statistical Software*, 33(2), 1–22.
- [49] Højsgaard S, Halekoh U, & Yan J. (2006). The R Package geepack for Generalized Estimating Equations. *Journal of Statistical Software*, 15(2), 1–11.
- [50] Martin AD, Quinn KM, & Park JH. (2011). “MCMCpack: Markov Chain Monte Carlo in R.” *Journal of Statistical Software*, 42(9), 1–22.
- [51] Masarotto G, & Varin C. (2017). Gaussian Copula Regression in R. *Journal of Statistical Software*, 77(8), 1–26.
- [52] McDaniel LS, Henderson NC, & Rathouz PJ. (2013). Fast Pure R Implementation of GEE: Application of the Matrix Package. *The R Journal*, 5(1), 181–187.
- [53] Pinheiro J, Bates D, DebRoy S, Sarkar D, & R Core Team. (2017). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-131. <https://CRAN.R-project.org/package=nlme>.
- [54] R Core Team. (2017). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>.
- [55] Ribeiro Jr PJ, & Diggle PJ. (2016). geoR: Analysis of Geostatistical Data. R package version 1.7-5.2. <https://CRAN.R-project.org/package=geoR>.
- [56] Rønnegård L, Shen X, & Alam M. (2010). Fitting Hierarchical Generalized Linear Models. *The R Journal*, 2(2), 20–28. <https://journal.R-project.org/archive/2010/RJ-2010-009>.
- [57] Wolak ME. (2012). nadiv: An R Package to Create Relatedness Matrices for Estimating Non-Additive Genetic Variances in Animal Models. *Methods in Ecology and Evolution*, 3(5), 792–796.
- [58] Vapnik V. (1963). Pattern recognition using generalized portrait method. *Automation and remote control*, 24, 774-780.

- [59] Tang Y, Jin B, Zhang Y, et al. (2004). Granular support vector machines for medical binary classification problems[C]//Proceedings of the IEEE CIBIB. Piscataway, HJ: IEEE Computationl Intelligence Society, 73-78.
- [60] Tang Y, Jin B, & Zhang Y. (2005). Granular support vector machines with association rules mining for protein homology prediction[J]. Artificial Intelligence in Medicine, 35: 121-134.
- [61] Lin C, & Wang S. (2002). Fuzzy support vector machines[J]. IEEE Transactions on Neural Networks, 3(2): 464-471.
- [62] Herbrich R, Graepel T, & Obermayer K. (2000). Large margin rank boundaries for ordinal regression[C]//Advances in Large Margin Classifiers. Cambridge, MA: MIT Press, 7:115-132.
- [63] Jayadcva R, & Khemchandani SC. (2007). Twin support vector machines for pattern classification[J]. IEEE Trans On Pattern Analysis and Machine Intelligence,29(5): 905-910.
- [64] Kumar MA, & Gopal M. (2009). Least squares twin support vector machines for pattern classification[J]. Expert Systems with Applications, 36(4):7535-7543.
- [65] Ye Q, Zhao C, Gao S, et al. (2012). Weighted twin support vector machines with local information and its application[J]. Neural Networks the Official Journal of the International Neural Net work Society, 35(11):31-39.
- [66] Vapnik V N. (2000). 统计学习理论的本质 [M]. 张学工, 译. 北京: 清华大学出版社.
- [67] Breiman L, Friedman J, Olshen R, & Stone C. (1984). Classification and Regression Trees. Chapman and Hall, New York.
- [68] Shannon CE. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3), 379-423.
- [69] Quinlan JR. (1986). Induction of decision trees. Machine learning, 1(1), 81-106.
- [70] Hastie T, Tibshirani R, & Friedman J. (2008). The Elements of Statistical Learning: Data Mining, Inferene and Prediction. Springer, New York.
- [71] Kuhn M, & Johnson K. (2013). Applied Predictive Modeling. Springer, New York.
- [72] Friedman JH. (1991). Multivariate adaptive regression splines. The annals of statistics, 19(1), 1-67.

- [73] Breiman L. (1996). Heuristics of Instability and Stabilization in Model Selection. *The Annals of Statistics*, 26, 2350-2383.
- [74] Loh WY, & Shih YS. (1997). Split Selection Methods for Classification Trees. *Statistica Sinica*, 7, 815-840.
- [75] Caron C, Boulesteix A, & Augustin T. (2007). Unbiased Split Selection for Classification Trees Based on the Gini Index. *Computational Statistics & Data Analysis*, 52, 483-501.
- [76] Loh WY. (2010). Tree-Structured Classifiers. *Computational Statistics*, 2, 364-369.
- [77] Breiman L. (2001). Random Forests. *Machine Learning*, 45, 5-32.
- [78] Tibshirani R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58, 267-288.
- [79] Quinlan R. (1992). Learning with Continuous Classes. *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, 343-348.
- [80] Zeileis A, Hothorn T, & Hornik K. (2008). Model-Based Recursive Partitioning. *Journal of Computational and Graphical Statistics*, 14, 185-205.
- [81] Quinlan JR. (2014). C4. 5: programs for machine learning[M]. Elsevier.
- [82] Breiman L. (1996). Bagging Predictors. *Machine Learning*, 24, 123-140.
- [83] Valiant LG. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134-1142.
- [84] Kearns M, & Valiant LG. (1988). Learning Boolean formulae or finite automata is as hard as factoring. Technical Report TR-14-88, Harvard University Aiken Computation Laboratory.
- [85] Schapire RE. (1990). The strength of weak learnability. *Machine learning*, 5(2), 197-227.
- [86] Freund Y. (1995). Boosting a weak learning algorithm by majority. *Information and computation*, 121(2), 256-285.
- [87] Drucker H, Schapire R, & Simard P. (1993). Boosting performance in neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4), 705-719.

- [88] Zhou ZH, Yang Y, Wu XD, & Kumar V. (2009). The Top Ten Algorithms in Data Mining. New York, USA: CRC Press, 127-149.
- [89] Friedman J, Hastie T, & Tibshirani R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, 28(2), 337-407.
- [90] Freund Y, & Schapire RE. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), 119-139.
- [91] Schapire RE, & Singer Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297-336.
- [92] Schapire RE, & Freund Y. (2012). Boosting: Foundations and Algorithms. MIT Press, Cambridge, MA.
- [93] Schapire RE, Freund Y, Bartlett P, & Lee WS. (1998). Boosting the margin: A new explanation for the effectiveness of votingmethods. *Annals of Statistics*, 26(5):1651-1686.
- [94] Breiman L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11(7): 1493-1517.
- [95] Reyzin L, & Schapire RE. (2006). How Boosting the margin can also boost classifier complexity. In: *Proceedings of the 23rd International Conference on Machine Learning*. New York,USA: ACM, 753-760.
- [96] Wang LW, Sugiyama M, Yang C, Zhou ZH, & Feng JF. (2008). On the margin explanation of Boosting algorithms. In: *Proceedings of the 21st Annual Conference on Learning Theory*. Madison, USA: Omni Press, 479-490.
- [97] Gao W, & Zhou ZH. (2013). On the doubt about margin explanation of boosting. *Artificial Intelligence*, 203:1-18.
- [98] Mason L, Baxter J, Bartlett P, & Frean M. (2000). Functional gradient techniques for combining hypotheses. *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 221-247.
- [99] Mason L, Baxter J, Bartlett P, & Frean M. (1999). Boosting algorithms as gradient descent. *Advances in Neural Information Processing Systems*, 12: 512-518.

- [100] Friedman J. (2001). Greedy function approximation: A gradient boosting machine, *Annals of Statistics*, 29(5): 1189-1232.
- [101] Q. Chen (陈强). (2020). 机器学习及 R 应用. 高等教育出版社.
- [102] R.G. Wang (汪荣贵). (2020). 机器学习简明教程. 机械工业出版社.
- [103] Akaike H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716-723.
- [104] Schwarz G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.
- [105] Hoerl AE, & Kennard RW. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.
- [106] Fan J, & Li R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456), 1348-1360.
- [107] Zhang CH. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2), 894-942.
- [108] Zou H, & Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2), 301-320.
- [109] Yuan M, & Lin Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1), 49-67.
- [110] Huang J, Ma S, Xie H, & Zhang CH. (2009). A group bridge approach for variable selection. *Biometrika*, 96(2), 339-355.
- [111] Breheny P, & Huang J. (2009). Penalized methods for bi-level variable selection. *Statistics and its interface*, 2(3), 369.
- [112] Simon N, Friedman J, Hastie T, & Tibshirani R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2), 231-245.
- [113] Fang K, Wang X, Zhang S, Zhu J, & Ma S. (2015). Bi-level variable selection via adaptive sparse group Lasso. *Journal of Statistical Computation and Simulation*, 85(13), 2750-2760.

- [114] Bates JM, & Granger CW. (1969). The combination of forecasts. *Journal of the operational research society*, 20(4), 451-468.
- [115] Fan J, & Lv J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5), 849-911.
- [116] Hall P, & Miller H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3), 533-550.
- [117] Li G, Peng H, Zhang J, & Zhu L. (2012). Robust rank correlation based screening. *The Annals of Statistics*, 40(3), 1846-1877.
- [118] Zhu LP, Li L, Li R, & Zhu LX. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496), 1464-1475.
- [119] Li R, Zhong W, & Zhu L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499), 1129-1139.
- [120] Székely GJ, Rizzo ML, & Bakirov NK. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6), 2769-2794.
- [121] Fan J, & Fan Y. (2008). High dimensional classification using features annealed independence rules. *Annals of statistics*, 36(6), 2605.
- [122] Mai Q, & Zou H. (2013). The Kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, 100(1), 229-234.
- [123] Cui H, Li R, & Zhong W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association*, 110(510), 630-641.
- [124] Xie J, Lin Y, Yan X, & Tang N. (2020). Category-adaptive variable screening for ultra-high dimensional heterogeneous categorical data. *Journal of the American Statistical Association*, 115(530), 747-760.
- [125] Fan J, & Song R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics*, 38(6), 3567-3604.
- [126] Fan J, Feng Y, & Song R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494), 544-557.

- [127] Zinkevich M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In Proceedings of the 20th international conference on machine learning (icml-03) (pp. 928-936).
- [128] Vovk V, Gammerman A, & Shafer G. (2005). Conformal prediction. Algorithmic learning in a random world, 17-51.
- [129] Vovk V, Nouretdinov I, & Gammerman A. (2009). On-line predictive linear regression. *The Annals of Statistics*, 1566-1590.
- [130] Lei J, Robins J, & Wasserman L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501), 278-287.
- [131] Lei J, & Wasserman L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 71-96.
- [132] Lei J, G' Sell M, Rinaldo A, Tibshirani RJ, & Wasserman L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523), 1094-1111.
- [133] Lei J, Rinaldo A, & Wasserman L. (2015). A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74(1), 29-43.
- [134] Butler R, & Rothman ED. (1980). Predictive intervals based on reuse of the sample. *Journal of the American Statistical Association*, 75(372), 881-889.
- [135] Steinberger L, & Leeb H. (2016). Leave-one-out prediction intervals in linear regression models with many variables. arXiv preprint arXiv:1602.05801.
- [136] McCulloch WS, & Pitts W. (1943). Bulletin of Mathematical Biophysics.
- [137] Kohonen T. (1988). An introduction to neural computing. *Neural Networks*.
- [138] 周志华.(2016). 机器学习. 北京: 清华大学出版社.
- [139] Rosenblatt F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6), 386.
- [140] Werbos P. (1974). Beyond regression:" new tools for prediction and analysis in the behavioral sciences. Ph. D. dissertation, Harvard University.
- [141] Pineda FJ. (1987). Generalization of Back-Propagation to recurrent neural-networks. *Physical Review Letters*.

- [142] Aarts E, & Korst J. (1989). Simulated Annealing and Boltzmann Machines:A Stochastic Approach to Combinatorial Optimization and Neural Computing. New York: John Wiley & Sons.
- [143] Goldberg DE. (1989). Genetic Algorithms in Search, Optimizaiton and Machine Learning. Boston: Addison-Wesley.
- [144] Carpenter GA, & Grossberg S. (1987). A massively parallel architecturefor a self-organizing neural pattern recognition machine. Computer Vision,Graphics, and Image Processing.
- [145] Kohonen T. (1982). Self-organized formation of topologically correct featuremaps. Biological Cybernetics.
- [146] Fahlman SE, & Lebiere C. (1990). The cascade-correlation learning architecture. Pittsburgh:chool of Computer Sciences,Carnergie Mellon University.
- [147] Elman JL. (1990). Finding structure in time. Cognitive Science.
- [148] LeCun Y, & Bengio Y. (1995). Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks, 3361(10).
- [149] LeCun Y, Bottou L, Bengio Y, & Haffner P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
- [150] Rumelhart DE, Hinton GE, & Williams RJ. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.
- [151] LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, & Jackel LD. (1989). Backpropagation applied to handwritten zip code recognition. Neural computation, 1(4), 541-551.
- [152] Hinton GE, & Salakhutdinov RR. (2006). Reducing the dimensionality of data with neural networks. science, 313(5786), 504-507.
- [153] Zeiler MD, & Fergus R. (2014). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.
- [154] Simonyan K, & Zisserman A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- [155] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, ... & Rabinovich A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

- [156] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, & Teller E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6), 1087-1092.
- [157] Gustafson JL. (1988). Reevaluating Amdahl's law. *Communications of the ACM*, 31(5), 532-533.
- [158] threading-基 于 线 程 的 并 行.<https://docs.python.org/zh-cn/3/library/threading.html>.
- [159] multiprocessing-基 于 进 程 的 并 行.<https://docs.python.org/zh-cn/3/library/multiprocessing.html>.
- [160] CUDA C++ Programming Guide.<https://docs.nvidia.com/cuda/cuda-c-programming-guide/index.html>.
- [161] 樊哲勇. (2020). CUDA 编程: 基础与实践 [M]. 北京: 清华大学出版社.
- [162] 詹卡洛 · 扎克内. (2021). Python 并行编程实战 (第二版) [M]. 北京: 中国电力出版社.
- [163] 陈华. (2018). 多核并行计算 [M]. 青岛: 中国石油大学出版社.