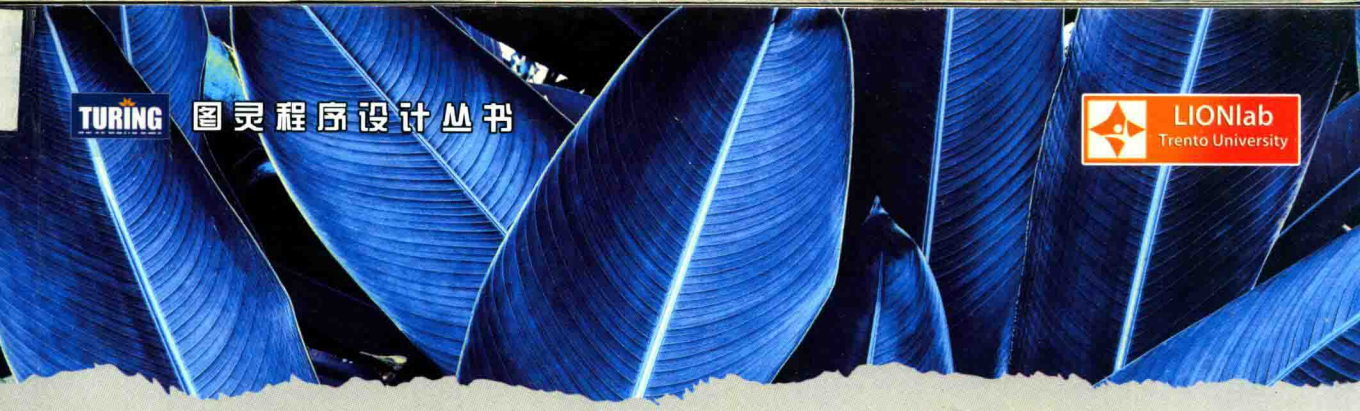


The LION Way

Machine Learning plus Intelligent Optimization

# 机器学习与优化

[意] 罗伯托·巴蒂蒂　毛罗·布鲁纳托　著

王彧弋　译

- 摒弃复杂的公式推导，从实践上手机器学习
- 人工智能领域先驱、IEEE会士巴蒂蒂教授领导的LION实验室多年机器学习经验总结

中国工信出版集团　人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵程序设计丛书

# The LION Way

## Machine Learning plus Intelligent Optimization

# 机器学习与优化

[意] 罗伯托·巴蒂蒂 毛罗·布鲁纳托 著
王彧弋 译

人民邮电出版社
北 京

**图书在版编目(CIP)数据**

机器学习与优化/（意）罗伯托·巴蒂蒂（Roberto Battiti），（意）毛罗·布鲁纳托（Mauro Brunato）著；王彧弋译. —北京：人民邮电出版社, 2018. 5
（图灵程序设计丛书）
ISBN 978-7-115-48029-3

Ⅰ. ①机… Ⅱ. ①罗… ②毛… ③王… Ⅲ. ①机器学习 Ⅳ. ①TP181

中国版本图书馆 CIP 数据核字（2018）第 044097 号

## 内 容 提 要

本书是机器学习实战领域的一本佳作，从机器学习的基本概念讲起，旨在将初学者引入机器学习的大门，并走上实践的道路。本书通过讲解机器学习中的监督学习和无监督学习，并结合特征选择和排序、聚类方法、文本和网页挖掘等热点问题，论证了“优化是力量之源”这一观点，为机器学习在企业中的应用提供了切实可行的操作建议。

本书适合从事机器学习领域工作的相关人员，以及任何对机器学习感兴趣的读者。

◆ 著 〔意〕罗伯托·巴蒂蒂 毛罗·布鲁纳托
译 王彧弋
责任编辑 朱 巍
执行编辑 温 雪 黄志斌
责任印制 周昇亮
◆ 人民邮电出版社出版发行 北京市丰台区成寿寺路 11 号
邮编 100164 电子邮件 315@ptpress.com.cn
网址 http://www.ptpress.com.cn
三河市君旺印务有限公司印刷
◆ 开本：800×1000 1/16
印张：17.5 彩插：2
字数：420 千字 2018 年 5 月第 1 版
印数：1-3 500 册 2018 年 5 月河北第 1 次印刷
著作权合同登记号 图字：01-2014-4553 号

定价：89.00 元

读者服务热线：(010)51095186 转 600 印装质量热线：(010) 81055316

反盗版热线：(010)81055315

广告经营许可证：京东工商广登字 20170147 号

## 作者简介


**罗伯托·巴蒂蒂**（Roberto Battiti）

人工智能领域先驱，IEEE会士。因在无功搜索优化（RSO）方向做出了开创性的工作而名震学界。目前为意大利特伦托大学教授，同时担任特伦托大学机器学习与智能优化实验室（LION lab）主任。

**毛罗·布鲁纳托**（Mauro Brunato）

意大利特伦托大学助理教授，LION研究团队成员。


## 译者简介


**王彧弋**

博士，现于瑞士苏黎世联邦理工学院从事研究工作，主要研究方向为理论计算机科学与机器学习。



封面设计 | 83720326@qq.com
自在书装设计

微信连接

回复“机器学习”查看相关书单

微博连接

关注@图灵教育 每日分享IT好书

QQ连接

图灵读者官方群I：218139230
图灵读者官方群II：164939616

**图灵社区**
**iTuring.cn**

在线出版，电子书，《码农》杂志，图灵访谈

站在巨人的肩上

Standing on Shoulders of Giants


iTuring.cn

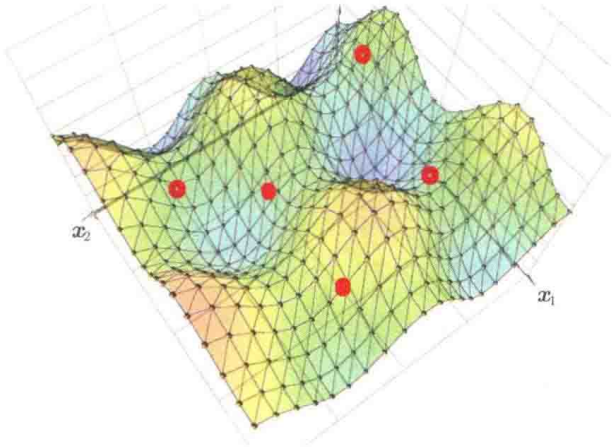


图 1-2　从样本中使用克里金法构造模型。一些样本在图中用点标示出来。表面的高度和颜色依赖于产金量

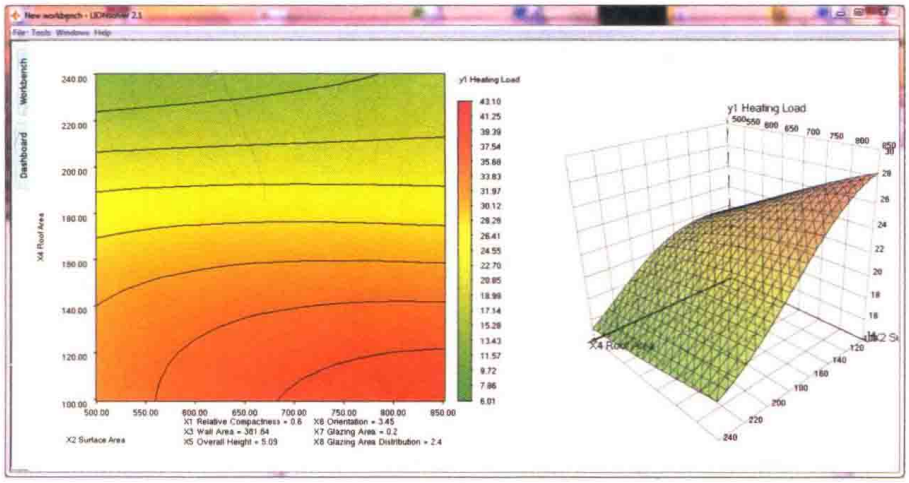


图 9-4　用 LION 软件 Sweeper 分析神经网络的输出。输出值和冬季加热房子消耗的能量，是输入参数的函数。图中展示了颜色编码的输出（左）和表面图（右）。非线性是可见的

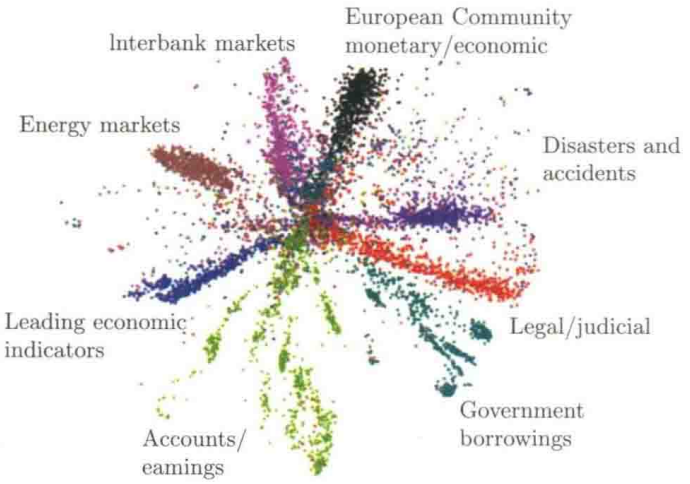


图 10-4　代码由一个 2000-500-250-125-2 自编码器根据路透社的新闻故事生成。图中用不同的颜色对应于不同主题的聚类，这是清晰可见的（详见参考文献 [57]）

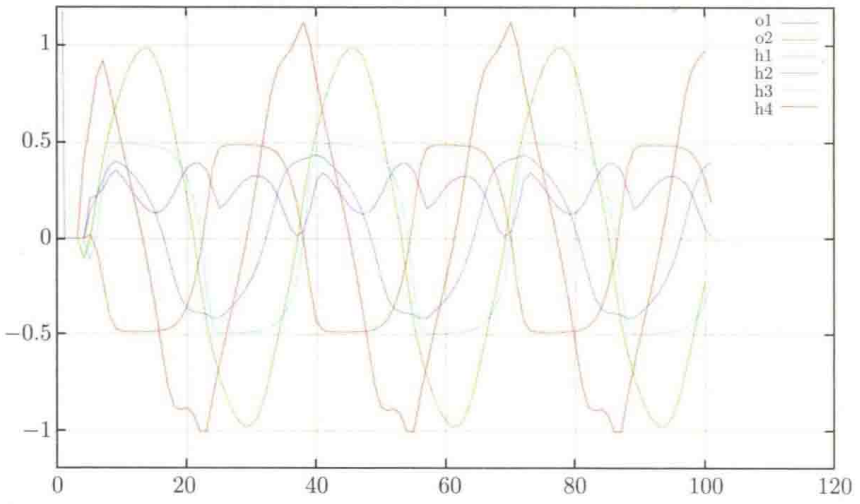


图 14-3　递归神经网络沿着环形训练：输出和隐藏层神经元

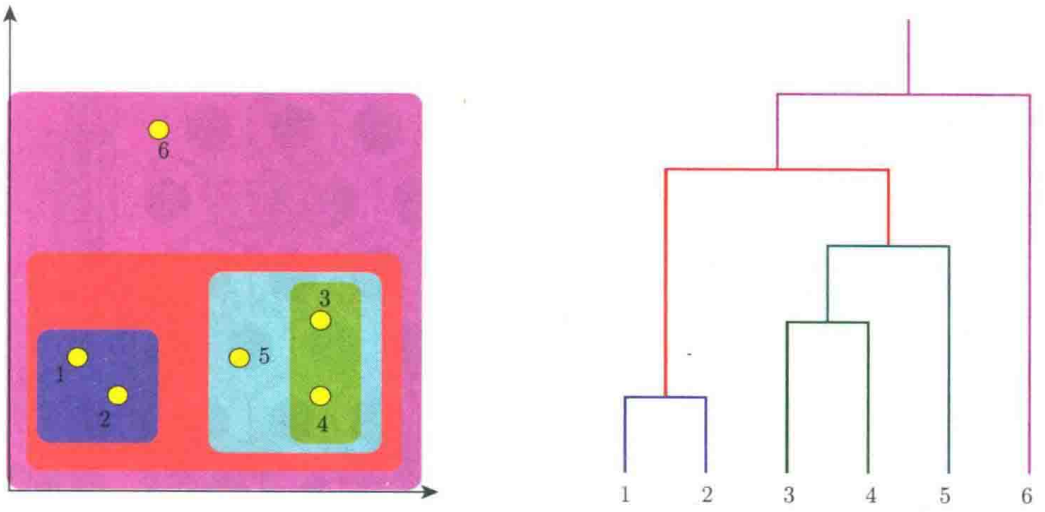


图 16-1　二维空间中数据点自底向上聚类示意图（使用标准欧几里得距离），每个数据点都由两个数值构成

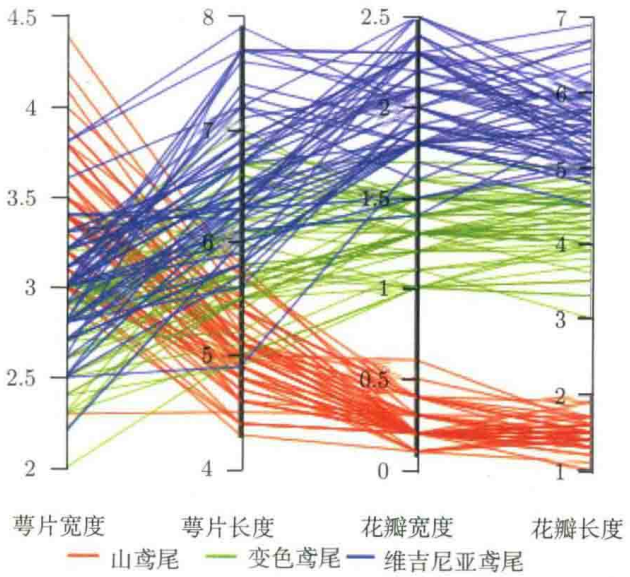


图 16-5　费希尔鸢尾花数据集（每朵花包含 4 个度量属性）的平行坐标展示，每个属性都用一个垂直轴表示，数据中的第 $i$ 项属性值表示为折线与对应的第 $i$ 个垂直轴的交点

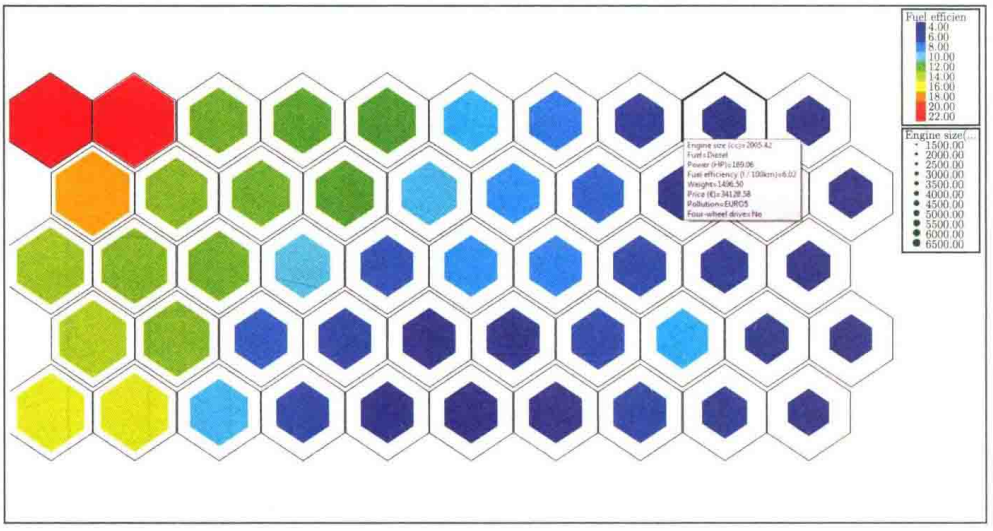


图 17-5　一个 SOM，颜色和大小取决于二维原型向量的两个坐标，可以将鼠标移到神经元上来显示原型的值（通过 LIONoso.org 提供的软件）

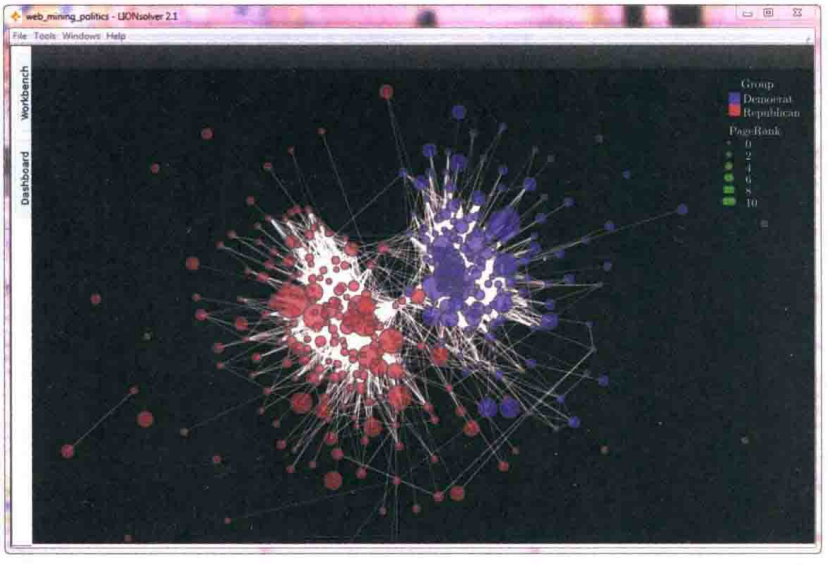


图 19-2　社交网络分析：美国议员的可视化网络。两个政党（从聚类软件无法得到）呈现出非常不同的两个类别

# 版 权 声 明

# 目　录


**第 1 章　引言** ····· 1
1.1 学习与智能优化：燎原之火 ····· 1
1.2 寻找黄金和寻找伴侣 ····· 3
1.3 需要的只是数据 ····· 5
1.4 超越传统的商业智能 ····· 5
1.5 LION 方法的实施 ····· 6
1.6 “动手”的方法 ····· 6
**第 2 章　懒惰学习：最近邻方法** ····· 9
**第 3 章　学习需要方法** ····· 14
3.1 从已标记的案例中学习：最小化和泛化 ····· 16
3.2 学习、验证、测试 ····· 18
3.3 不同类型的误差 ····· 21

**第一部分　监督学习**

**第 4 章　线性模型** ····· 26
4.1 线性回归 ····· 27
4.2 处理非线性函数关系的技巧 ····· 28
4.3 用于分类的线性模型 ····· 29
4.4 大脑是如何工作的 ····· 30
4.5 线性模型为何普遍，为何成功 ····· 31
4.6 最小化平方误差和 ····· 32
4.7 数值不稳定性和岭回归 ····· 34
**第 5 章　广义线性最小二乘法** ····· 37
5.1 拟合的优劣和卡方分布 ····· 38
5.2 最小二乘法与最大似然估计 ····· 42
5.2.1 假设检验 ····· 42
5.2.2 交叉验证 ····· 44
5.3 置信度的自助法 ····· 44
**第 6 章　规则、决策树和森林** ····· 50
6.1 构造决策树 ····· 52
6.2 民主与决策森林 ····· 56
**第 7 章　特征排序及选择** ····· 59
7.1 特征选择：情境 ····· 60
7.2 相关系数 ····· 62
7.3 相关比 ····· 63
7.4 卡方检验拒绝统计独立性 ····· 64
7.5 熵和互信息 ····· 64
**第 8 章　特定非线性模型** ····· 67
8.1 logistic 回归 ····· 67
8.2 局部加权回归 ····· 69
8.3 用 LASSO 来缩小系数和选择输入值 ····· 72
**第 9 章　神经网络：多层感知器** ····· 76
9.1 多层感知器 ····· 78
9.2 通过反向传播法学习 ····· 80
9.2.1 批量和 bold driver 反向传播法 ····· 81
9.2.2 在线或随机反向传播 ····· 82
9.2.3 训练多层感知器的高级优化 ····· 83
**第 10 章　深度和卷积网络** ····· 84
10.1 深度神经网络 ····· 85
10.1.1 自动编码器 ····· 86
10.1.2 随机噪声、屏蔽和课程 ····· 88
10.2 局部感受野和卷积网络 ····· 89
**第 11 章　统计学习理论和支持向量机** ····· 94
11.1 经验风险最小化 ····· 96
11.1.1 线性可分问题 ····· 98
11.1.2 不可分问题 ····· 100
11.1.3 非线性假设 ····· 100
11.1.4 用于回归的支持向量 ····· 101
**第 12 章　最小二乘法和健壮内核机器** ····· 103
12.1 最小二乘支持向量机分类器 ····· 104

12.2 健壮加权最小二乘支持向量机 …… 106
12.3 通过修剪恢复稀疏 …… 107
12.4 算法改进：调谐 QP、原始版本、无补偿 …… 108
**第 13 章 机器学习中的民主** …… 110
13.1 堆叠和融合 …… 111
13.2 实例操作带来的多样性：装袋法和提升法 …… 113
13.3 特征操作带来的多样性 …… 114
13.4 输出值操作带来的多样性：纠错码 …… 115
13.5 训练阶段随机性带来的多样性 …… 115
13.6 加性 logistic 回归 …… 115
13.7 民主有助于准确率–拒绝的折中 …… 118
**第 14 章 递归神经网络和储备池计算** …… 121
14.1 递归神经网络 …… 122
14.2 能量极小化霍普菲尔德网络 …… 124
14.3 递归神经网络和时序反向传播 …… 126
14.4 递归神经网络储备池学习 …… 127
14.5 超限学习机 …… 128

## 第二部分 无监督学习和聚类

**第 15 章 自顶向下的聚类：K 均值** …… 132
15.1 无监督学习的方法 …… 134
15.2 聚类：表示与度量 …… 135
15.3 硬聚类或软聚类的 K 均值方法 …… 137
**第 16 章 自底向上（凝聚）聚类** …… 142
16.1 合并标准以及树状图 …… 142
16.2 适应点的分布距离：马氏距离 …… 144
16.3 附录：聚类的可视化 …… 146
**第 17 章 自组织映射** …… 149
17.1 将实体映射到原型的人工皮层 …… 150
17.2 使用成熟的自组织映射进行分类 …… 153
**第 18 章 通过线性变换降维（投影）** …… 155
18.1 线性投影 …… 156
18.2 主成分分析 …… 158
18.3 加权主成分分析：结合坐标和关系 …… 160
18.4 通过比值优化进行线性判别 …… 161
18.5 费希尔线性判别分析 …… 163
**第 19 章 通过非线性映射可视化图与网络** …… 165
19.1 最小应力可视化 …… 166
19.2 一维情况：谱图绘制 …… 168
19.3 复杂图形分布标准 …… 170
**第 20 章 半监督学习** …… 174
20.1 用部分无监督数据进行学习 …… 175
20.1.1 低密度区域中的分离 …… 177
20.1.2 基于图的算法 …… 177
20.1.3 学习度量 …… 179
20.1.4 集成约束和度量学习 …… 179

## 第三部分 优化：力量之源

**第 21 章 自动改进的局部方法** …… 184
21.1 优化和学习 …… 185
21.2 基于导数技术的一维情况 …… 186
21.2.1 导数可以由割线近似 …… 190
21.2.2 一维最小化 …… 191
21.3 求解高维模型（二次正定型） …… 191
21.3.1 梯度与最速下降法 …… 194
21.3.2 共轭梯度法 …… 196
21.4 高维中的非线性优化 …… 196
21.4.1 通过线性查找的全局收敛 …… 197
21.4.2 解决不定黑塞矩阵 …… 198
21.4.3 与模型信赖域方法的关系 …… 199
21.4.4 割线法 …… 200
21.4.5 缩小差距：二阶方法与线性复杂度 …… 201
21.5 不涉及导数的技术：反馈仿射振荡器 …… 202
21.5.1 RAS：抽样区域的适应性 …… 203
21.5.2 为健壮性和多样化所做的重复 …… 205

**第 22 章　局部搜索和反馈搜索优化** ····· 211
22.1　基于扰动的局部搜索 ····· 212
22.2　反馈搜索优化：搜索时学习 ····· 215
22.3　基于禁忌的反馈搜索优化 ····· 217
**第 23 章　合作反馈搜索优化** ····· 222
23.1　局部搜索过程的智能协作 ····· 223
23.2　CoRSO：一个政治上的类比 ····· 224
23.3　CoRSO 的例子：RSO 与 RAS 合作 ····· 226
**第 24 章　多目标反馈搜索优化** ····· 232
24.1　多目标优化和帕累托最优 ····· 233
24.2　脑–计算机优化：循环中的用户 ····· 235

**第四部分　应用精选**

**第 25 章　文本和网页挖掘** ····· 240
25.1　网页信息检索与组织 ····· 241
25.1.1　爬虫 ····· 241
25.1.2　索引 ····· 242
25.2　信息检索与排名 ····· 244
25.2.1　从文档到向量：向量–空间模型 ····· 245
25.2.2　相关反馈 ····· 247
25.2.3　更复杂的相似性度量 ····· 248
25.3　使用超链接来进行网页排名 ····· 250
25.4　确定中心和权威：HITS ····· 254
25.5　聚类 ····· 256
**第 26 章　协同过滤和推荐** ····· 257
26.1　通过相似用户结合评分 ····· 258
26.2　基于矩阵分解的模型 ····· 260
**参考文献** ····· 263
**索引** ····· 269

# 第1章　引　　言

人不应该过着野兽般的生活，而是要追寻美德与知识。

—— 但丁

## 1.1　学习与智能优化：燎原之火

**优化是指为了找到更好的解决方案而进行的自动化搜寻过程**。可以说，流程、方案、产品和服务之所以能持续改进，正是缘于优化为之**提供**的**强大动力**。优化不仅关乎方案的确定（从一些给定的可行方案中，选出最好的一个），**它还能主动创造出新的解决方案**。

优化催生了**自动化的创造和革新**。这看起来非常矛盾，因为自动化通常不会和创造与革新联系起来。因此，那些相信机器只能用来处理单调的重复性工作的人们在阅读本书时，会觉得书中的观点简直是胡言乱语，甚至会感受到如同被挑衅一般的愤怒。

自伽利略（1564—1642）之后，人们希望用科学改变世界，而这不仅需要哲学上的阐释，还需要**测量和实验**的支持。“测量那些可测量的，并使那些不可测量的变得可测量。”测量一开始看起来并不起眼，但它允许人们用务实的方式逐渐改变世界，只要人们还关心生产方式和生活质量。

几乎所有的商业问题都可以归结为**寻找一个最优决策值** $\boldsymbol{x}$，这要通过**使某个收益函数** $\text{goodness}(\boldsymbol{x})$ **最大化**来实现。为了能形象地理解，我们假设有一个集合变量 $\boldsymbol{x}=(x_1,\cdots,x_n)$，

它描述的可以是一个或多个待调节的旋钮，也可以是将要做出的选择，还可以是待确定的参数。在市场营销中，$\boldsymbol{x}$ 可以是一个向量，其数值表示为各类宣传活动（电视、报纸、各种网站、社交媒体）分配的预算，goodness($\boldsymbol{x}$) 则可以是由这些宣传活动而产生的新客户数量。在网站优化中，$\boldsymbol{x}$ 可以涉及图片、链接、话题和不同大小文本的使用，goodness($\boldsymbol{x}$) 则可以是该网站的普通访客成为客户的转化率。在工程学中，$\boldsymbol{x}$ 可以是一个汽车发动机的设计参数集，goodness($\boldsymbol{x}$) 则可以是该发动机每加仑汽油所能行驶的英里数。

将问题归结为“优化一个收益函数”也**激励着决策者，使用量化的目标，就可以用可衡量的方式来领会宗旨，也就可以专注于方针的制定而非执行的细枝末节**。当人们深陷于执行的泥潭中，以至于遗忘了目标时，企业就染上了“疫病”，此时如果外界环境发生了变化，这种“疫病”将会使企业无法做出及时的应对。

**自动化**是解决这个问题的关键：将一个问题形式化地表述后，我们把得到的收益模型输入计算机，计算机将自动创造出并找到一个或多个最佳的选项。另外，当条件和重点发生改变时，只需要修改一下收益函数的量化目标，再重启优化过程就可以了。当然，CPU 时间会是个问题，也并非每次都能保证找到全局最优解决方案。但可以肯定的是，使用计算机来搜寻，无论是速度还是范围，都远远领先于人力搜寻，并且这一领先优势会越来越明显。

然而，在大多数现实场景中，优化的惊人力量仍遭到很大程度的压制。优化在现实中没有被广泛采纳的主要原因是，**标准的数学优化理论假设存在一个需要最大化的收益函数**，也就是说，有一个明确定义的模型 goodness($\boldsymbol{x}$) 为每个输入配置 $\boldsymbol{x}$ 匹配一个结果。而目前，在现实的商业情境里，这个函数通常是不存在的。即使存在，靠人力找到这个函数也是极其困难、极其昂贵的。试想，问一个 CEO“请您告诉我，优化您业务的数学公式是什么”，显然不是咨询工作中开始对话的最佳方式。当然，一个经理对于目标应该会有一些想法和权衡，但是这些目标并没有以数学模型的方式给定，它们是动态的、模糊的，会随着时间改变，并且受限于估计误差和人们的学习进程。直觉被用来替代那些明确给定的、量化的和数据驱动的决策过程。

如果优化是燃料，那么点燃这些燃料的火柴就是**机器学习**。机器学习通过摒弃那种明确定义的目标 goodness($\boldsymbol{x}$) 来拯救优化：我们**可以通过丰富的数据来建立模型**。

**机器学习与智能优化**（learning and intelligent optimization，LION）结合了学习和优化，它从数据中学习，又将优化用于解决复杂的、动态的问题。LION 方法提高了自动化水平，并将数据与决策、行动直接联系起来。描述性分析和预测性分析之后，LION 的第三阶段（也是最终阶段）是**规范性分析**（prescriptive analysis）。**在自助服务的方式中，决策者手中直接握有更多的权力**，而不必求助于中间层的数据科学家。就像汽车的发动机一样，LION 包含一系列复杂的机制，但是用户（司机）并不需要知道发动机的内部工作原理，就可以享用它带来的巨大好处。在未来的几十年内，LION 方法带来的创新，将会像野火那样，以燎原之势延伸到大多数行业。那么企业就像野火频发的生态系统中的植物一样，只有适应并拥抱 LION 技术才能生存下来，并繁荣昌盛；否则，无论之前如何兴盛，在竞争逐渐加剧的挑战面前，都可能

土崩瓦解。

**LION 范式关注的并不是数学上的收益模型，而是海量数据**，以及如何针对多种具体选择（包括实际的成功案例）进行专家决策，或者如何交互地定义成功的标准。当然，这些都是建立在让人们感觉轻松愉快的基础之上的。例如，在市场营销中，相关数据可以描述之前的资金分配和宣传活动的成效；在工程学中，数据可以描述发动机设计的实验（真实的或模拟的）和相应的油耗测量方式。

## 1.2 寻找黄金和寻找伴侣

用于优化的机器学习需要数据。数据来源可以是**以往的优化过程**，也可以是**决策者的反馈**。

要了解这两种情境，先来看两个具体的例子。丹尼尔·克里金（Danie G. Krige，见图 1-1）是一名南非的采矿工程师，他曾遇到一个问题：如何在一张地图上找到挖掘金矿的最佳坐标 [74]。大约在 1951 年，他开创性地将统计学的思想应用于新金矿的估值，而这一方法仅需用到有限的几个矿坑。需要优化的函数是 $\mathrm{Gold}(\boldsymbol{x})$，即坐标 $\boldsymbol{x}$ 处的金矿的金量。当然，在一个新的地方 $\boldsymbol{x}$ 评估 $\mathrm{Gold}(\boldsymbol{x})$ 是非常昂贵的。你可以想象，挖一个新矿没那么快，也没那么简单。但是在一些试探性的挖掘之后，工程师们会积累一些把坐标 $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3 \cdots$ 和金量 $\mathrm{Gold}(\boldsymbol{x}_1), \mathrm{Gold}(\boldsymbol{x}_2), \mathrm{Gold}(\boldsymbol{x}_3)$ 关联起来的**实例知识**。克里金的直觉告诉他，用这些实例（**来自以往优化过程的数据**）可以建立起函数 $\mathrm{Gold}(\boldsymbol{x})$ 的模型。这个称为 $\mathrm{GoldModel}(\boldsymbol{x})$ 的模型归纳以往的实验结果，为地图上的每个位置 $\boldsymbol{x}$ 给出金量的估计值。通过优化，这个模型找到使预计黄金产量 $\mathrm{GoldModel}(\boldsymbol{x})$ 最大化的地点 $\boldsymbol{x}_{\mathrm{best}}$，于是这个 $\boldsymbol{x}_{\mathrm{best}}$ 成为下一个挖掘的地点。

图 1-1 丹尼尔·克里金，克里金法的发明者

可以用如图 1-2 所示的模型来形象地说明这个过程。先在地图上为每个矿坑插一根针，每根针的高度取决于在该处发现的金量。克里金的模型可以看作基于这些针的“训练”信息