

套路！机器学习：北美数据科学家的私房课



[套路！机器学习：北美数据科学家的私房课_下载链接1](#)

著者:林荟

出版者:电子工业出版社

出版时间:

装帧:平装

isbn:9787121326585

数据科学家目前是北美最热门的职业之一，平均年薪突破10万美元。但数据科学并不是一个低门槛的行业，除了对数学、统计、计算机等相关领域的技术要求以外，还要相关

应用领域的知识。《套路！机器学习：北美数据科学家的私房课》的写作对象是那些现在从事数据分析相关行业，或者之后想从事数据分析行业的人，意在为实践者提供数据科学家这门职业的相关信息。读者可以从阅读中了解到数据科学能解决的问题，数据科学家需要的技能，及背后的“分析哲学”。对于新手而言，一开始就直奔艰深的理论，很容易因为困难而失去兴趣最终放弃。因此《套路！机器学习：北美数据科学家的私房课》倡导的是一种循序渐进的启发教学路径，着重在于数据科学的实际应用，让读者能够重复书中的结果，学习数据分析技能最好的方式是实践！为了平衡理论和应用，书中包括了一些选学小节，用来介绍更多的模型数理背景或给出必要的参考资料来源。抽丝剥茧介绍技术内核，帮助大家知其然，同时知其所以然。

希望笔者在北美从事数据科学工作多年踏遍大大小小不计其数的坑换来的经验，能够帮助读者更加顺利地成为数据科学家！

作者介绍:

目录: 第1章 白话数据科学 1

1.1 什么是数据科学 3

1.2 什么是数据科学家 5

1.2.1 数据科学家需要的技能 6

1.2.2 数据科学算法总结 10

1.3 数据科学可以解决什么问题 20

1.3.1 前提要求 20

1.3.2 问题种类 22

1.4 小结 25

第2章 数据集 26

2.1 服装消费者数据 26

2.2 航空公司满意度调查 33

2.3 生猪疫情风险预测数据 37

第3章 数据分析流程 41

3.1 从问题到数据 42

3.2 从数据到信息 44

3.3 从信息到行动 46

第4章 数据预处理 47

4.1 介绍 47

4.2 数据清理 50

4.3 缺失值填补 52

4.3.1 中位数或众数填补 53

4.3.2 K-近邻填补 54

4.3.3 装袋树填补 56

4.4 中心化和标量化 56

4.5 有偏分布 59

4.6 处理离群点 63

4.7 共线性 66

4.8 稀疏变量 70

4.9 编码名义变量 71

4.10 小结 73

第5章 数据操作 75

5.1 数据读写 76

5.1.1 取代传统数据框的tibble对象 76

5.1.2 高效数据读写：readr包 80

5.1.3 数据表对象读取 83

5.2 数据整合 91

- 5.2.1 base包: apply() 91
- 5.2.2 plyr包: dplyr()函数 93
- 5.2.3 dplyr包 96
- 5.3 数据整形 102
 - 5.3.1 reshape2包 102
 - 5.3.2 tidyr包 105
- 5.4 小结 107
- 第6章 基础建模技术 109
 - 6.1 有监督和无监督 109
 - 6.2 误差及其来源 111
 - 6.2.1 系统误差和随机误差 111
 - 6.2.2 因变量误差 117
 - 6.2.3 自变量误差 121
 - 6.3 数据划分和再抽样 122
 - 6.3.1 划分训练集和测试集 123
 - 6.3.2 重抽样 131
 - 6.4 小结 135
- 第7章 模型评估度量 136
 - 7.1 回归模型评估度量 136
 - 7.2 分类模型评估度量 139
 - 7.2.1 Kappa统计量 141
 - 7.2.2 ROC曲线 143
 - 7.2.3 提升图 145
 - 7.3 小结 146
- 第8章 特征工程 148
 - 8.1 特征构建 149
 - 8.2 特征提取 152
 - 8.2.1 初步探索特征 153
 - 8.2.2 主成分分析 158
 - 8.2.3 探索性因子分析 163
 - 8.2.4 高维标度化 167
 - 8.2.5 知识扩展: 3种降维特征提取方法的理论 171
 - 8.3 特征选择 177
 - 8.3.1 过滤法 178
 - 8.3.2 绕封法 188
 - 8.4 小结 195
- 第9章 线性回归及其衍生 196
 - 9.1 普通线性回归 197
 - 9.1.1 最小二乘线性模型 197
 - 9.1.2 回归诊断 201
 - 9.1.3 离群点、高杠杆点和强影响点 204
 - 9.2 收缩方法 205
 - 9.2.1 岭回归 205
 - 9.2.2 Lasso 209
 - 9.2.3 弹性网络 212
 - 9.3 知识扩展: LASSO的变量选择功能 213
 - 9.4 主成分和偏最小二乘回归 215
 - 9.5 小结 221
- 第10章 广义线性模型压缩方法 222
 - 10.1 初识GLMNET 223
 - 10.2 收缩线性回归 227
 - 10.3 逻辑回归 235
 - 10.3.1 普通逻辑回归 235
 - 10.3.2 收缩逻辑回归 236

10.3.3 知识扩展：群组lasso逻辑回归	239
10.4 收缩多项回归	243
10.5 泊松收缩回归	246
10.6 小结	249
第11章 树模型	250
11.1 分裂准则	252
11.2 树的修剪	256
11.3 回归树和决策树	260
11.4 装袋树	268
11.5 随机森林	273
11.6 助推法	277
11.7 知识扩展：助推法的可加模型框架	283
11.8 知识扩展：助推树的数学框架	286
11.8.1 数学表达	286
11.8.2 梯度助推数值优化	289
11.9 小结	290
第12章 神经网络	292
12.1 投影寻踪回归（PROJECTION PURSUIT REGRESSION）	293
12.2 神经网络（NEURAL NETWORKS）	296
12.3 神经网络拟合	299
12.4 训练神经网络	300
12.5 用CARET包训练神经网络	302
12.6 小结	311
参考文献	312
• • • • •	(收起)

[套路！机器学习：北美数据科学家的私房课_下载链接1](#)

标签

计算机

机器学习

数据分析

管理

R

图书馆k

图书馆

akb

评论

模型评估居然没提KS GINI一嘴，话说，谢益辉博士还在奥马哈吗

[套路！机器学习：北美数据科学家的私房课_下载链接1](#)

书评

[套路！机器学习：北美数据科学家的私房课_下载链接1](#)