

统计学强调低维空间问题的统计推导；机器学习强调高维预测问题。

目 录

第 1 章 成功之路..... 1	
1.1 流程..... 1	
1.2 业务理解..... 2	
1.2.1 确定业务目标..... 3	
1.2.2 现状评估..... 4	
1.2.3 确定分析目标..... 4	
1.2.4 建立项目计划..... 4	
1.3 数据理解..... 4	
1.4 数据准备..... 5	
1.5 建模..... 5	
1.6 评价..... 6	
1.7 部署..... 6	
1.8 算法流程图..... 7	
1.9 小结..... 10	
第 2 章 线性回归：机器学习基础技术..... 11	
2.1 单变量回归..... 11	
2.2 多变量线性回归..... 18	
2.2.1 业务理解..... 18	
2.2.2 数据理解和数据准备..... 18	
2.2.3 模型构建与模型评价..... 21	
2.3 线性模型中的其他问题..... 30	
2.3.1 定性特征..... 30	
2.3.2 交互项..... 32	
2.4 小结..... 34	
第 3 章 逻辑斯蒂回归与判别分析..... 35	
3.1 分类方法与线性回归..... 35	
3.2 逻辑斯蒂回归..... 36	
3.2.1 业务理解..... 36	
3.2.2 数据理解和数据准备..... 37	
3.2.3 模型构建与模型评价..... 41	
3.3 判别分析概述..... 46	
3.4 多元自适应回归样条方法..... 50	
3.5 模型选择..... 54	
3.6 小结..... 57	
第 4 章 线性模型中的高级特征选择技术..... 58	
4.1 正则化简介..... 58	
4.1.1 岭回归..... 59	
4.1.2 LASSO..... 59	
4.1.3 弹性网络..... 60	
4.2 商业案例..... 60	
4.2.1 业务理解..... 60	
4.2.2 数据理解和数据准备..... 60	
4.3 模型构建与模型评价..... 65	
4.3.1 最优子集..... 65	
4.3.2 岭回归..... 68	
4.3.3 LASSO..... 71	
4.3.4 弹性网络..... 73	
4.3.5 使用 glmnet 进行交叉验证..... 76	
4.4 模型选择..... 78	
4.5 正则化与分类问题..... 78	
4.6 小结..... 81	
第 5 章 更多分类技术：K 最近邻与 支持向量机..... 82	
5.1 K 最近邻..... 82	
5.2 支持向量机..... 84	
5.3 商业案例..... 86	

5.3.1 业务理解	86	8.4 随机森林	151
5.3.2 数据理解和数据准备	87	8.5 业务理解	152
5.3.3 模型构建与模型评价	92	8.6 数据理解与数据准备	152
5.3.4 模型选择	98	8.7 模型构建与模型评价	155
5.4 SVM 中的特征选择	100	8.7.1 层次聚类	155
5.5 小结	101	8.7.2 K 均值聚类	162
第 6 章 分类回归树	103	8.7.3 果瓦系数和 PAM	165
6.1 本章技术概述	103	8.7.4 随机森林与 PAM	167
6.1.1 回归树	104	8.8 小结	168
6.1.2 分类树	104	第 9 章 主成分分析	169
6.1.3 随机森林	105	9.1 主成分简介	170
6.1.4 梯度提升	106	9.2 业务理解	173
6.2 商业案例	106	9.3 模型构建与模型评价	176
6.2.1 模型构建与模型评价	107	9.3.1 主成分抽取	176
6.2.2 模型选择	121	9.3.2 正交旋转与解释	177
6.2.3 使用随机森林进行特征选择	121	9.3.3 根据主成分建立因子得分	178
6.3 小结	123	9.3.4 回归分析	178
第 7 章 神经网络与深度学习	124	9.4 小结	184
7.1 神经网络介绍	124	第 10 章 购物篮分析、推荐引擎与 序列分析	185
7.2 深度学习简介	128	10.1 购物篮分析简介	186
7.3 业务理解	131	10.2 业务理解	187
7.4 数据理解和数据准备	132	10.3 数据理解和数据准备	187
7.5 模型构建与模型评价	136	10.4 模型构建与模型评价	189
7.6 深度学习示例	139	10.5 推荐引擎简介	192
7.6.1 H2O 背景介绍	139	10.5.1 基于用户的协同过滤	193
7.6.2 将数据上载到 H2O 平台	140	10.5.2 基于项目的协同过滤	194
7.6.3 建立训练数据集和测试 数据集	141	10.5.3 奇异值分解和主成分分析	194
7.6.4 模型构建	142	10.6 推荐系统的业务理解	198
7.7 小结	146	10.7 推荐系统的数据理解与数据准备	198
第 8 章 聚类分析	147	10.8 推荐系统的建模与评价	200
8.1 层次聚类	148	10.9 序列数据分析	208
8.2 K 均值聚类	149	10.10 小结	214
8.3 果瓦系数与围绕中心的划分	150	第 11 章 创建集成多类分类	215
8.3.1 果瓦系数	150	11.1 集成模型	215
8.3.2 PAM	151	11.2 业务理解与数据理解	216

11.3 模型评价与模型选择.....	217	第 13 章 文本挖掘.....	250
11.4 多类分类.....	219	13.1 文本挖掘框架与方法.....	250
11.5 业务理解与数据理解.....	220	13.2 主题模型.....	252
11.6 模型评价与模型选择.....	223	13.3 业务理解.....	254
11.6.1 随机森林.....	224	13.4 模型构建与模型评价.....	257
11.6.2 岭回归.....	225	13.4.1 词频分析与主题模型.....	257
11.7 MLR 集成模型.....	226	13.4.2 其他定量分析.....	261
11.8 小结.....	228	13.5 小结.....	267
第 12 章 时间序列与因果关系.....	229	第 14 章 在云上使用 R 语言.....	268
12.1 单变量时间序列分析.....	229	14.1 创建 AWS 账户.....	269
12.2 业务理解.....	235	14.1.1 启动虚拟机.....	270
12.3 模型构建与模型评价.....	240	14.1.2 启动 Rstudio.....	272
12.3.1 单变量时间序列预测.....	240	14.2 小结.....	274
12.3.2 检查因果关系.....	243	附录 R 语言基础.....	275
12.4 小结.....	249		