《R 语言数据分析》自学指南

一、概述

本学习指南主要针对学堂在线 MOOC 课程《R 语言数据分析》。

本学习指南主要用以指导该课程的自学。若采用翻转课堂或其它方式讲授、学习该课程的部分或全部内容，且希望获得相关的教学建议、学习指南，请直接与该课程主讲老师取得联系（axb@bupt.edu.cn）

本指南最新版本，可在 https://github.com/byaxb/RDataAnalytics 获取。

二、课程脉络

《R 语言数据分析》课程分为三个部分，分别是：

上部：论道

主要讲述机器学习/数据挖掘方法论，包括机器学习的本质、学习的过程、学习的结果、推理方式、数学与工程的关系等。提出了本课程的一些核心理念，如：

- 气象万千、数以等观
- 所谓学习、归类而已
- 实事求是的讲，机器学习就是实事求是的过程
- 一切都是关系结构
- 源于数学、归于工程
- ……

中部：执具

主要讲述 R 语言这个工具。在阐述 R 语言编程环境、运行机制的基础上，重点阐述两个方面的内容，即 R 语言的基础编程和数据对象。

- 基础编程——R 语言编程的核心思想是利用别人的包和函数，讲述自己的故事。因此，在基础编程部分，首先阐述如何找到合适的扩展包，然后阐述控制流、函数，也就是如何展开自己的逻辑、讲好自己的故事。
- 数据对象——面向数据对象学习 R 语言，可能是掌握 R 这一工具的有效途径（注意，此处是面向数据对象，而非面向对象）。要掌握 R 语言，主要在于掌握其中的三组六类数据对象，即：向量/因子、矩阵/数值、列表/数据框。从某种意义上讲，掌握了 R 语言的主要数据对象的应用场景、主要操作，也就掌握了 R 语言的核心。这一部分内容，是掌握 R 语言的关键。

在阐述完上述两方面的主体内容之后，重点讲解了 tidyverse 这么一种编码风格。让同学们体会 R 语言的简洁、优雅。

执具的最后一个章节，是《最美不过数据框》。这是一个承上启下的章节。数据框，是最美好的数据对象（承上）；在数据框里，蕴含着各种各样的关系结构，包括关联、分类、聚类等（启下）。

下部：博术

如果说上部是“道”，那这一部分就是机器学习/数据挖掘之“术”。

遵循数据科学的一般方法论，从探索性数据分析、关联、分类、聚类四个方面来展开。

- 观数以形——开展探索性数据分析，主要是通过少量数字来刻画数据、通过直观的图形来展示数据，揭示变量之间的关系和数据空间的形态。
- 相随相伴、谓之关联——关联规则的挖掘。阐述关联规则的情境、Apriori 算法解决的核心问题、算法原理、规则评估、R 语言实现、规则可视化等。
- 既是世间法、自当有分别——分门别类，即有监督学习。此部分内容较为庞杂，主要是围绕变量之间的关系和和数据空间的结构，以此为主线，将 7 种经典算法模型串珠成链。针对每一种算法模型，既阐述其直觉、也揭示其原理、更注重其实践。

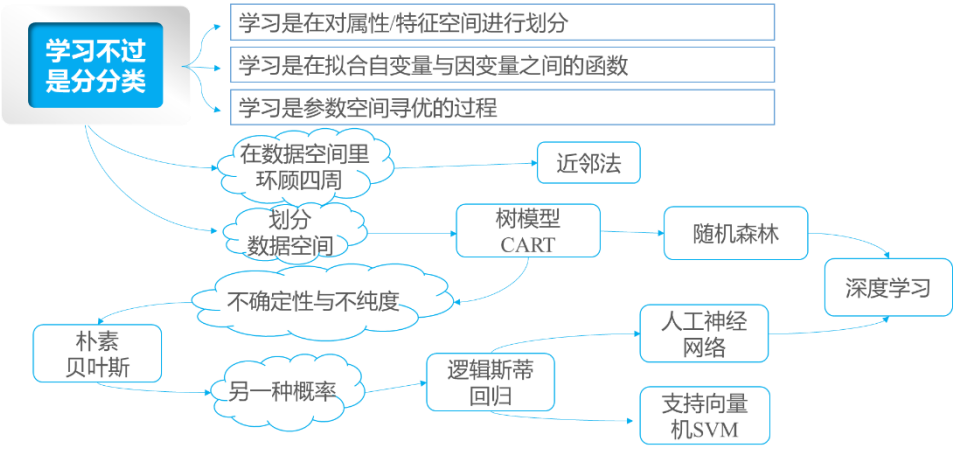

图 1 有监督学习脉络图

- 方以类聚、物以群分——聚类分析，即无监督学习。此部分主要阐述两类经典的聚类方法，即基于划分的 kMeans 方法以及层次聚类方法。突出了易被忽视的空间均匀度（即是否适合聚类）、聚类结果的评估（模型评估不只是有监督学习才有）、聚类算法原理的直观展示、聚类算法的 R 语言实现、聚类结果的可视化等。

在阐述完聚类分析的核心内容后，从学术创新、算法创新的角度，提出“嫁接”、“遇见”这类学术创新观念，并以层次聚类在异常侦测中的创新应用为例，阐述如何在学习现有算法模型的基础上，开展学术创新。同时，也比较自然地涵盖了机器学习/数据挖掘的另一主题，即：异常侦测。

最后，课程以“庐山烟雨浙江潮”诗句结尾，以司空见惯的成语连连看，来重温、总结课程的核心理念，并以“庐山烟雨浙江潮”来隐喻 R 语言数据分析的简单而又美好。

课程的具体脉络，通过思维导图展示如下：

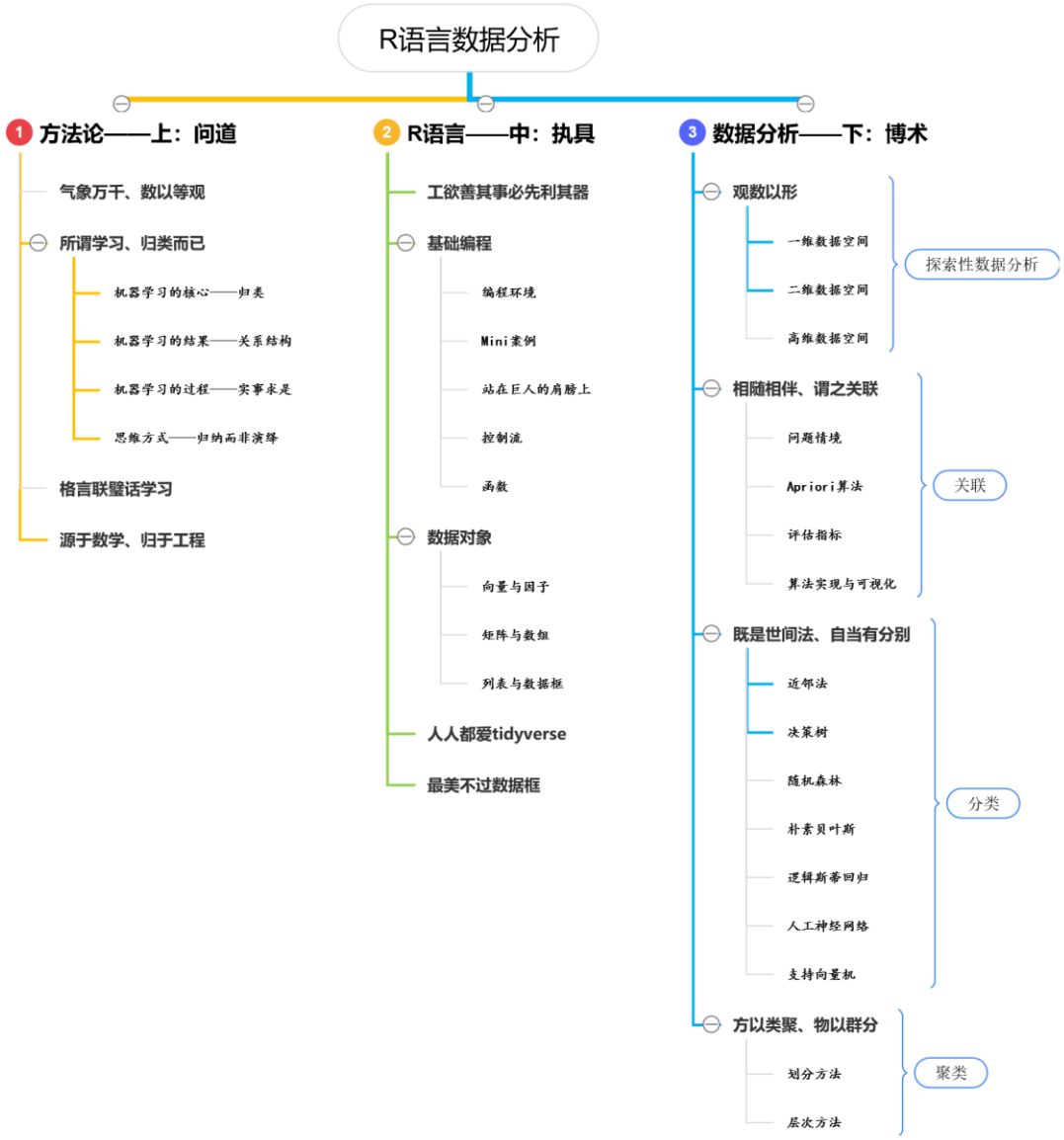

图 2 课程内容体系图

三、课时分布

本 MOOC 课程，总视频时长约 950 分钟（含片头和片尾）。即：完整地播放一遍视频，大约需要 16 个小时，其中：

- 《上部：问道》总计 104 分钟，约 1.5 学时（为了方便大家制定学习计划，此处不精确计算，仅给出一个大致学时。比如此处的 1.5 学时，也就意味着一个半小时能看完全部视频）。
- 《中部：执具》总计 325 分钟，约 5.5 学时。其中基础编程环境（含 Mini 小案例）约 2.5 学时，数据对象部分（含 tidyverse）约 3 学时。
- 《下部：博术》总计 500 分钟，约 8.5 学时。其中，探索性数据分析 1.5 学时；关联部分 1.5 学时；分类部分 4 学时；聚类部分 1.5 学时。

视频时长明细如下，同学们可据此制定具体的学习计划：

视频序号	章节名称			时长	累计时长
	部	章	节		
1	上部：问道	第 1 章 气象万千、数以等观		11	1.5 学时（104）
		第 2 章 所谓学习、归类而已			
2			2.1 所谓学习、归类而已（I）	15	
3			2.2 所谓学习、归类而已（II）	8	
4			2.3 所谓学习、归类而已（III）	14	
5			2.4 所谓学习、归类而已（IV）	15	
6		第 3 章 格言联璧话学习		17	
7		第 4 章 源于数学、归于工程		25	
8	中部：执具	第 5 章 工欲善其事、必先利其器		13	2.5 学时（142）
		第 6 章 基础编程——用别人的包和函数讲述自己的故事			
9			6.1 编程环境	16	
10			6.2Mini 案例	19	
11			6.3 站在巨人的肩膀上	17	
12			6.4 控制流	39	
13			6.5 函数（I）	19	
14			6.6 函数（II）	20	
		第 7 章 数据对象——面向数据对象学习 R 语言			3 学时（182）
15			7.1 向量与因子（I）	40	
16			7.2 向量与因子（II）	26	
17			7.3 矩阵与数组（I）	18	
18			7.4 矩阵与数组（II）	15	
19			7.5 列表与数据框（I）	13	
20			7.6 列表与数据框（II）	27	
21		第 8 章 人人都爱 tidyverse		28	
22		第 9 章 最美不过数据框		16	
	下部：博术	第 10 章 观数以形			1.5 学时（92）
23			10.1 一维数据空间（I）	21	
24			10.2 一维数据空间（II）	19	
25			10.3 二维数据空间	24	
26			10.4 高维数据空间	27	

27		第 11 章 相随相伴、谓之关联		12	1.5 学时（86）
28			11.1 关联规则（I）	30	
29			11.2 关联规则（II）	14	
30			11.3 关联规则（III）	30	
31		第 12 章 既是世间法、自当有分别		24	4 学时（228）
32			12.1 近邻法（I）	13	
33			12.2 近邻法（II）	15	
34			12.3 决策树（I）	31	
35			12.4 决策树（II）	24	
36			12.5 随机森林	14	
37			12.6 朴素贝叶斯	12	
38			12.7 逻辑斯蒂回归	30	
39			12.8 人工神经网络（I）	24	
40			12.9 人工神经网络（II）	13	
41			12.10 支持向量机	28	
42		第 13 章 方以类聚、物以群分		20	1.5 学时（94）
43			13.1 划分方法	28	
44			13.2 层次方法	36	
45		第 14 章 庐山烟雨浙江潮		10	

四、建议学习时长

从上述视频时长可以看出，要完整的播放本 MOOC 课程的全部视频，大致需要 16 个小时。但这绝不意味着两天就可以把 R 语言数据分析课程学完（假如按 8 小时工作制计的话）。

建议学习时长是一个学期。

实际上，根据以往同学们的学习经历，要入门 R 语言数据分析，应该是两至三个月。方法论、基本概念、算法模型、代码编写，都有一个消化吸收的过程，并无速成之法。

五、建议学习方式

边学边做：在干中学、在学中干

可能有部分同学觉得方法论部分是本课程的特色内容之一。实际上，本课程实践性很强。如果只是一味地看视频，而不去写代码、具体上手实践，那方法论部分的理念，也难有真切的体会。对于 R 语言这一工具的掌握，更不是看看视频就可以。没有人可以通过看几段游泳视频，就学会游泳。R 语言同样如此。

所以，本课程的学习，最好是一边看视频，一边写代码，才能取得相对比较好的效果。在这里，也请同学们注意，千万不要简单的把代码拷贝过来，运行一遍，浅尝辄止。一定要亲手敲入每一行代码!!

同时，也强烈推荐同学们针对自己的问题，套用、修改、优化课程代码以及从其他渠道获得的代码（如 Github、Stackoverflow 等）。实际上，照着已有代码过一遍很容易，一旦针对自己的实际问题写代码，可能会寸步难行。在学习之初碰到这些问题再正常不过，但一定要迈开步子、大胆尝试。

六、推荐学习资源

本课程内容由任课教师自行编制，逻辑上相对自成体系，基本可以按照视频顺序开展学习，进而入门 R 语言数据分析。

当然，R 之所以强大，是因为它背后强大的生态。同学们在学习本课程的同时，也需要养成有效利用 R 各类资源的习惯。以下是部分推荐资源：

首先是官方网站 https://www.r-project.org/。

官方网站是一个容易被忽视的宝库。官方网站不只是提供了 R 的下载，其中也有诸多资源值得我们学习和关注，比如：

- TASK VIEWS。R 并没有一个官方的机构，对现有的扩展包进行全面、系统的分门别类和规范管理。但 TASK VIEWS 按照不同的主题，对经典扩展包进行了梳理。建议同学们对其中感兴趣的主题进行认真研读。当然，一些通用主题，比如 Machine Learning、Cluster、Graphics、MissingData 等，应该是每个人都应该看一看的。
- The R Journal。该期刊文章的作者，大部分也是扩展包的作者，对大量的扩展包进行了系统、深入的论述。

其次是 RStudio。

这里推荐的 RStudio，并不是指 IDE，而是指 https://rstudio.com/。

- Resources > Cheatsheet。不同主题的速查表。
- Blogs > RViews, tidyverse, tensorflow 等，都值得关注。比如 RViews 中的每月 Top 40 packages 介绍。

再次是 Github。

Github 不只是托管了全球众多的代码，也提供了丰富的学习资源，比如

- https://github.com/qinwf/awesome-R
- https://github.com/josephmisiti/awesome-machine-learning

最后是 Stackoverflow。

对于 R 编码碰到的绝大部分问题，应该都可以通过 Stackoverflow 找到答案。实际上，

这里有数十万个与 R 相关的问答，而 R 领域的很多大神，如 Hadley、谢益辉等，都活跃在这个论坛：

- https://stackoverflow.com/questions/tagged/r

在学习咱们课程的时候，有问题当然可以直接在交流互动区中发帖，但从长远来看，利用好 Stack Overflow，毫无疑问是必需的。

七、若干注意事项

- 注意前后呼应。课程虽然分为三个部分，但方法论、R 语言工具、算法模型三部分并不是割裂的。尤其是在学习算法模型部分时，需要注意温习方法论部分相关的理念。这样才能把书读薄，核心理念才能一以贯之。
- 关于答题。在每一章学习完之后，都给出来一定数量的习题进行测试。需要注意的是，这些测试只有一次机会。假如对于该章节的内容还不熟悉，建议先不要答题。因为测试得分，均以第一次答题情况为准，并没有二次答题的机会。另外，习题中有很多涉及到代码的题目，建议大家先自行思考。在正式解答之前，可以借助 R 跑一遍，看看代码执行的结果究竟是什么。如果和自己推断出来的结果不一致，可以思考一下原因是什么。R 跑出来的结果，当然都是正确答案^-^
- 养成良好的 R 语言编码习惯。比如按照项目的方式来组织管理代码、做好版本管理、利用好 RStudio 的代码分节功能（RStudio >> Code >> Insert Section）、规范注释、做好异常处理等。在养成良好的编码习惯的同时，其实也是在提升科学编程的素养。

八、本指南由授课老师艾新波编写并负责解释。若有任何疑问，请通过以下方式取得联系：axb@bupt.edu.cn。