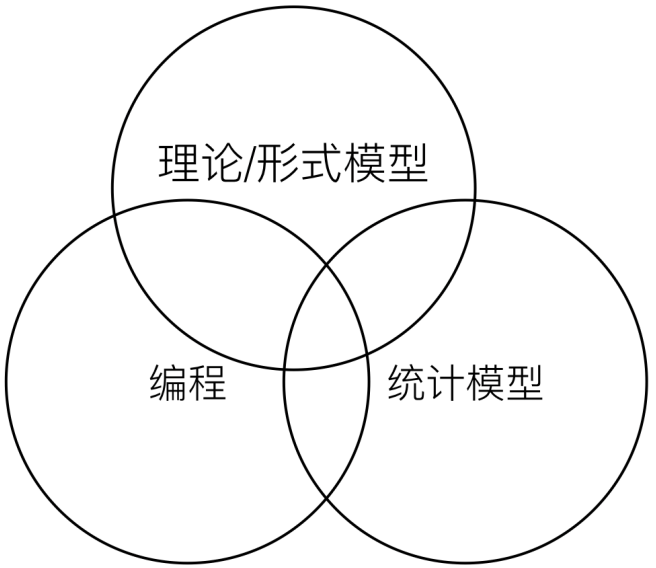
数据工程与R语言基础

拯救小明

史冬波
2020年3月2日


理论/形式模型
编程
统计模型

数据科学三要素

1. 理论模型（背景知识）: 决定了发现的方向，可以指导数据分析
2. 数据（编程、实验）: 验证理论假设，或发现新的现象。决定研究的下限。
3. 统计模型: 从数据中检验假设的工具，与理论一起决定了研究的上限，在一定程度上与数据是替代关系。

编程+理论+统计模型：文科生的编程不一样

- 具备中级段位的编程能力：在收集、预处理数据方面的竞争力。尤其注重数据的透明性与可复现性
- 具备过硬的统计模型基础：保证实证研究的可靠性

AI 的冲击与挑战

AI智能发展的三要素

人工智能目前实现的就是个分类器功能，最简单的分类器就是线性回归。人工智能的大发展的得益于其算料、算力以及算法的突破。

1. 算料，即数据。传感器技术的发展，互联网的发展使得数据的采集越来越及时，积累越来越多，数据储存和运输的成本越来越低。大数据是一个商业概念，而不是一个学术概念，因此大数据很难界定，也存在太多误用。但是，无论如何大数据必须要大，一台笔记本电脑可以打开的一定不是大数据。

2. 算力，计算机集群。真正的突破在于分布式运算的突破，一百万个臭皮匠，肯定打死一个诸葛亮的水平。欧洲核子中心 (CERN) 一年用电约 1.3 TWh;北京东城区西城区2016 年用电总量10.1TWh，人口200万;史老师的实验室不需要暖气。

3. 算法，深度学习、神经网络。机器学习算法，以Alpha GO为代表。神经网络是一个黑箱，但是特别管用。

数据工程

数据工程

数据工程顾名思义就是将数据科学的工作流程化，用软件工程的思路优化数据科学流程，具体实现中，遵循以下四个原则。

数据工程四原则

1. 可复现，以人类语言和计算机语言的形式，详细记录每一步计算。这是科学的基本精神，与可证伪性一起，是区分科学与伪科学的标志。

2. 自动化，Single Point of Truth, Don't Repeat Yourself. 不可在分析做任何重复，任何有意义的信息都应该被共享，

3. 正交分工，将数据分析任务切分为相互不影响的组成部分，分工是现代社会的基础。

4. 最佳工具，尽量使用高级语言和语法糖，为每个子任务选择合适的工具。只有在性能分析之后，才在必要时使用低级语 言进行性能加速，最佳工具会随时间变化。

R语言

1. R语言始终是不错编程语言之一
2. R语言是学习数据科学的最佳选择，没有之一
3. R语言可以独立完成整套数据工程流程
4. R就像蝙蝠侠：侦探工作、智慧、狡黠、使用工具、动脑多于蛮力
5. Python就像超人：肌肉力量、超级力量、优雅、全面、蛮力多于用脑

课程目标

- 安装R与RStudio
- RStudio的环境配置
- 像计算器一样使用R
- 数据类型与数据结构
- 控制流程
- 输入输出

解锁成就：批量读入文件

小红是一名热爱学习、乐于助人的同学，很快就被同学选举为班里的团支书。虽然小红同学非常给班级和班级的同学做服务，但是当大块的时间都被用来完成琐碎的事务性工作的时候，依然会有一些黯然神伤。小红也想把更多的时间用于学习，尤其是他还选修了数学辅修专业。本就紧张的学业，更加亚历山大。

马上青年节要到了，班上准备组织一次春游活动。小红作为团支书（为了剧情需要，不知道班长跑哪了），需要全班1200名同学那一天有时间。同时，小红为班级同学设计一件时尚的班衫，有红色、灰色、黑色和蓝色四种颜色，小红统计了同学们需要哪种颜色的版衫。现在小红收到了1200名同学回复的报名表。不过看到这么多报名表，小红犯难了把这些表整合成一张表又得多少时间呀！

救救小红吧

让他把有限的时间投入到无限的学习当中去吧！

在哪里获取资源

订阅https://www.r-bloggers.com/

RStudio官方教程

CRAN官方手册

教科书与手册https://bookdown.org/

搜索引擎、视频网站、Canvas

如何求助?

1. 使用R.version汇报你当前的环境
2. 描述遇到的错误
3. 提交可复现错误的代码（最好是使用公开数据）
4. 补充最佳答案