数据科学家：R语言

Hui Lin

January 26, 2016

# 序言

本书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，意在为实践者提供从事数据科学家这门职业的指导。读者可以从阅读中了解到数据科学家需要的技能，及背后的“分析哲学”。书中会对部分最常用，有效的模型加以展开。

数据科学家这个行业的本质是应用。市面上有很多文章，出版物介绍各种数据模型，大多数此类书籍并不能让读者重复书中所述的分析过程，对于书中介绍的知识，读者真正实践起来会遇到很多困难。本书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，这也用到了统计软件R的自动化报告功能。可能有读者会问，为什么要可重复？回答这个问题，我想引用Buckheit and Donoho (1995)的一段话：

计算机科学刊物上的学术论文本身并不是学术知识，而只是为学术知识打的广告。真正的知识在于背后整个软件开发环境，以及生成（论文中）结果的完整过程。

本书极力避免数学公式，但可能有少数例外。我们在少数地方涉及技术细节的初衷是为了帮助读者理解模型的长处和弱点，而非单纯的介绍数理统计知识。当然想要成为数据科学家，仅靠阅读本书是远远不够的，读者需要进一步学习书中提到的参考资料，或者选修相关课程。对于各种算法的实施，这里使用的是R语言，原因如下：

1. R免费，且可以在不同操作系统上使用。
2. R开源、可扩展：它在通用公共许可（General Public License）下发行，在此构架下任何人可以检查修改源程序。含有很多最新的模型。
3. R有强大图形可视化和自动化报告功能
4. 笔者10年使用R的经验证明无论在学术还是业界，这都是非常有效的工具。

不熟悉R的读者可以在网上找到大量的入门教程。

本书第2章-第4章主要介绍数据科学家这个行业和“分析哲学”和数据分析的一般流程。这是非技术的部分，但对于从业者来说非常重要，它帮助你对这个职业设定一个合理的预期。其中会讨论数据科学家需要的技能。之后的章节会对这里提到的部分我觉得重要的技能进一步展开讨论，由于篇幅所限，不可能详细讨论开始这几章中提到的所有技能。 第5章开始进入技术部分，本章讲的是分析环节的第一步——数据预处理，这一步虽然不是正式建模，但却是整个分析环节中最花之间的一个步骤。这步没有到位将严重影响模型的质量。也正是因为预处理重要，所以单独作为一个章节，没有和第6章其它建模技术合并起来。第6章介绍的是一些在建模过程中需要的辅助性的技术以及建模需要注意的问题。第7章到第10章正式介绍各种笔者在从业过程中经常用到的模型。第7章，第9章和第10章末尾分别有个比较完整的案例，案例放在这些章节是因为其中主要的模型方法是该章节的主题，但这3个案例都综合性的用到之前提到的许多方法，对读者理解模型应用和建模背后的思路很有帮助。

书中的代码和数据可以在这个github页面上找到：<https://github.com/happyrabbit/DataScientistR>

# 介绍

* 什么是数据科学家？

数据科学和数据科学家成为了流行词汇。当有人问你干什么，你回答说数据科学家。对方会恍然大悟，觉得特别高大上，奥，数据科学家啊，听说过。是啊，没听说过数据科学家那就out了。如果接着问，数据科学家具体干什么的？然后就没有然后了。不知道你们有没有听过这样一则轶事，美国最高法院法官Potter Stewart被问到什么是淫秽时，他回答：“看下才知道。”这和数据科学很类似，很多概念，在大而化之的时候都可以存在，大家口耳相传，聊的不亦乐乎，但一追究细节，立即土崩瓦解。那么什么是数据科学家呢？回答

1. 数据准备和探索
2. 数据表达变换
3. 数据建模
4. 数据可视化和结果展示
5. 数据科学的科学（候选）

# 数据科学家干什么？

# 数据分析一般流程

## 问题到数据

## 数据到信息

在这个流程中，我要强调两个非常重要的技术环节，数据预处理和模型检验。数据预处理占用相当大比例的时间（图中每一个步骤方块中的百分比是大致占用的时间范围）。关于数据处理的常用方法，

## 信息到知识

业界普遍反映数据科学家欠缺将数据中的信息转化为商业知识的能力，大部分数据科学家止于从数据到信息这一步。如何做到从信息到知识呢？这个需要你对所在行业有较深入的了解，具有应用该行业说故事的能力。这是艺术的部分。这也是一个矛盾的地方，因为这意味着数据科学家不如很多人之前所设想的可以单独成为一个职能，应用在不同的领域。当从技术角度上说这是可行的，但是从实践的角度，仅是应用技术得到一个模型结果远远不够，将结果转化为真正的市场营销的决策建议并和相关营销人员交流，保证结果应用在市场营销中提高效率，是非常关键的收尾环节，这一步决定了你工作的价值。但这一步要求深入了解行业，做到这点需要在某个特定行业中从业较长时间，因此从实践上讲，一个数据分析师很难细致了解每一个行业，即使“从数据到信息”这个步骤可以一般化到不同领域，但从“信息到知识”将数据科学家从某种程度上限定于某一领域。

# 数据预处理

## Introduction

There are a number of reasons a model falls, such as:

1. Inadequate data pre-processing
2. Inadequate model validation
3. Unjustified extrapolation
4. Over-fitting In this section, I am going to summarize some common data pre-processing approaches.

## Centering and Scaling

It is the most straightforward data transformation. It centers and scales a variable to mean 0 and standard deviation 1. It ensures that the criterion for finding linear combinations of the predictors is based on how much variation they explain and therefore improves the numerical stability. Models involving finding linear combinations of the predictors to explain response/predictors variation need data centering and scaling, such as PCA and PLS. You can easily writing code yourself to conduct this transformation. Or the function preProcess() in package caret can apply this transformation to a set of predictors.

#install packages needed  
library(caret)

## Warning: package 'caret' was built under R version 3.2.3

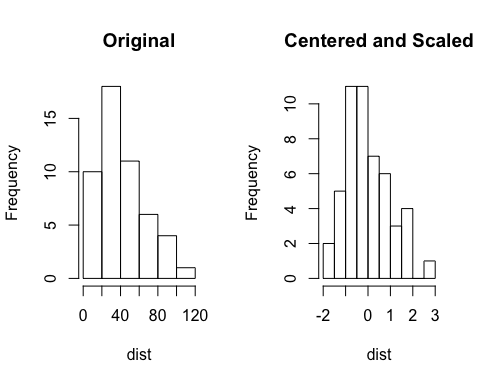
## Warning: package 'ggplot2' was built under R version 3.2.3

library(e1071)  
library(gridExtra)   
library(lattice)  
library(imputeMissings)  
library(RANN)  
library(corrplot)  
library(nnet)

head(cars)

## speed dist  
## 1 4 2  
## 2 4 10  
## 3 7 4  
## 4 7 22  
## 5 8 16  
## 6 9 10

trans<-preProcess(cars,method=c("center","scale"))  
transformed<-predict(trans,cars)  
par(mfrow=c(1,2))  
hist(cars$dist,main="Original",xlab="dist")  
hist(transformed$dist,main="Centered and Scaled",xlab="dist")

  
 Sometimes you only need to scale the variable. For example, if the model adds penalty to the parameter estimates (such as penalty is ridge regression and penalty in LASSO), the variables need to have similar scale to ensure a fair variable selection. I am heavy user of this kind of penalty-based model in my work and I used the following quantile transformation:

The reason to use 99% and 1% quantile instead of maximum and minimum values is to resist the impact of outliers.

It is easy to write a function to do it:

qscale<-function(dat){  
 for (i in 1:ncol(dat)){  
 up<-quantile(dat[,i],0.99)  
 low<-quantile(dat[,i],0.01)  
 diff<-up-low  
 dat[,i]<-(dat[,i]-low)/diff  
 }  
 return(dat)  
}

In order to illustrate, let’s simulate a data set with three variables: income, age and education.

set.seed(2015)  
income<-sample(seq(50000,150000,by=500),95)  
age<-income/2000-10  
noise<-round(runif(95)\*10,0)  
age<-age+noise  
income<-c(income,10000,15000,300000,250000,230000)  
age<-c(age,30,20,25,35,95)  
demo<-data.frame(income,age)  
demo$education<-as.factor(sample(c("High School","Bachelor","Master","Doctor"),100,replace = T,prob =c(0.7,0.15,0.12,0.03) ))  
summary(demo[,c("income","age")])

## income age   
## Min. : 10000 Min. :20.00   
## 1st Qu.: 76375 1st Qu.:30.25   
## Median : 98750 Median :44.25   
## Mean :103480 Mean :44.92   
## 3rd Qu.:126375 3rd Qu.:56.88   
## Max. :300000 Max. :95.00

It is clear that income and age are not on the same scale. Now apply the function qscale() on the simulated data demo.

transformed<-qscale(demo[,c("income","age")])  
summary(transformed)

## income age   
## Min. :-0.02101 Min. :-0.01904   
## 1st Qu.: 0.26077 1st Qu.: 0.17814   
## Median : 0.35576 Median : 0.44746   
## Mean : 0.37584 Mean : 0.46044   
## 3rd Qu.: 0.47304 3rd Qu.: 0.69033   
## Max. : 1.21015 Max. : 1.42375

## Resolve Skewness

## Resolve Outliers

## Missing Values

## Impute missing values with median/mode

## Impute missing values based on K-nearest neighbors

## Collinearity

## Sparse Variables

## Re-encode Dummy Variables

# 建模技术模块

## 数据划分和再抽样

## 变量选择

## 模型选择

## 建模需要注意的问题

### 因变量误差

### 自变量误差

### 失衡数据

# 线性回归极其衍生

## 普通线性回归

## 逻辑回归

## 罚函数模型

# 树模型

## 基本树模型

## 袋状树

## 随机森林

## 其它树话题

# 聚类判别分析

## 聚类分析

## 判别分析

关于100多种判别分析的选择。

## 案例：客户分组

# 关联法则分析

## 关联法则简介

关联法则挖掘的基本想法是：当若干事件共同发生的频率大于某人仅从它们各自单独发生的频率出发预期的共同发生率时，此共同发生的情况即为一个令人感兴趣的模式。

基本度量： - 支持 - 信用 - 提升

## 案例：商业购物篮分析

对于销售产品纪录，由于大部分东西不会和其它一些东西在一个交易中同时出现，因为可能的组合太多了。这就导致有大量的数据点但每个观测包含的信息相对较少。在这种情况下，许多分析都不起作用，如相关性分析和线性回归。关联法则挖掘关联法则分析试图从大量的稀疏数据中搜寻有信息量的不同模式。

## 关联法则可视化

# 数据可视化和结果展示

主要介绍可重复报告

# 数据科学的科学