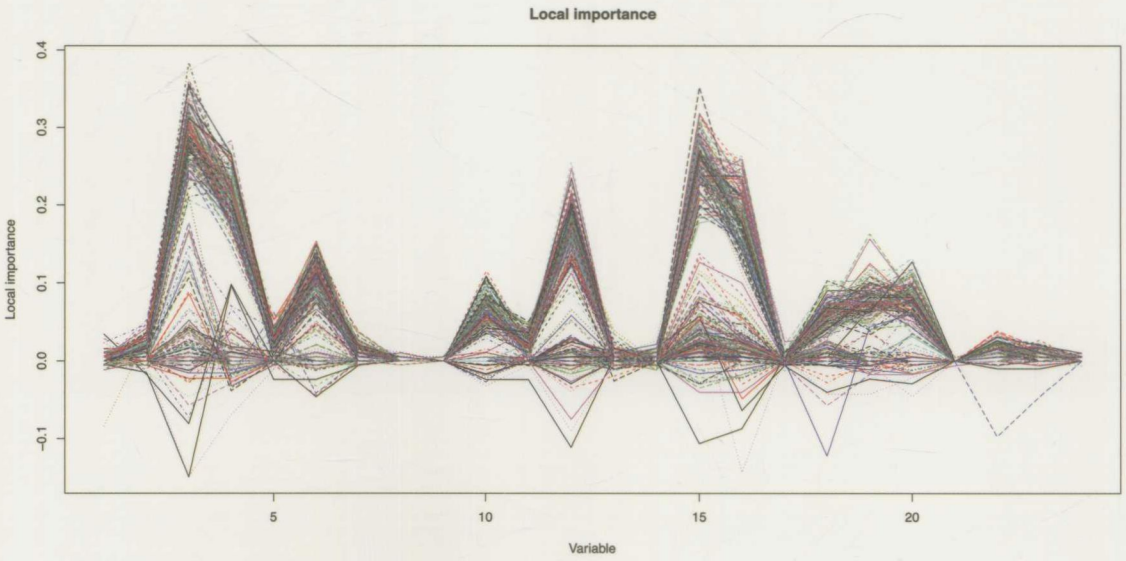
基于R应用的统计学丛书

应用回归及分类

——基于R

Applied Regression and Classification with R

吴喜之 编著



中国人民大学出版社

基于R应用的统计学丛书

应用回归及分类

——基于R

Applied Regression and Classification with R

吴喜之 编著

中国人民大学出版社
·北京·

图书在版编目(CIP)数据

应用回归及分类: 基于R / 吴喜之编著. —北京: 中国人民大学出版社, 2016.1
(基于R应用的统计学丛书)
ISBN 978-7-300-22287-5

I. ①应… II. ①吴… III. ①回归分析 – 高等学校 – 教材 IV. ①O212.1

中国版本图书馆CIP数据核字(2015) 第309285号

基于R应用的统计学丛书

应用回归及分类

—— 基于R

吴喜之 编著

Yingyong Huigui ji Fenlei

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010-62511242(总编室)	010-62511770(质管部)	
	010-82501766(邮购部)	010-62514148(门市部)	
	010-62515195(发行公司)	010-62515275(盗版举报)	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com (人大教研网)		
经　　销	新华书店		
印　　刷	北京鑫丰华彩印有限公司		
规　　格	185mm× 260mm　16开本	**版　　次**	2016年1月第1版
印　　张	15.75 插页1	**印　　次**	2016年1月第1次印刷
字　　数	326 000	**定　　价**	32.00元

前 言

本书不像很多教科书那样只讲80年之前的以数学假定和推导为主的内容, 而要强调最近20年最新和最有效的统计方法. 本书冠以"**分类**"二字, 是为了纠正由于只有"回归"而鲜有(如果不是没有)"分类"的教科书所造成的人们以为回归比分类更重要的偏见. 实际上, "**分类**"**一词很少出现在教科书的书名中的主要原因恐怕是长期以来数学主导的统计界缺乏除了判别分析之外的数学式的分类方法, 而引入近年来新发展的机器学习方法似乎又不合那些只认数学公式的统计学家的胃口**.

回归和分类的问题是相同的, 仅区别于因变量的形式. 在统计应用中, 最常见的是根据数据建立从自变量来预测因变量的模型, 也就是说, 用包含自变量和因变量的数据来训练一个模型, 然后用这个模型拟合新的自变量的数据来预测新的因变量的值.

上图为这样一个预测模型的示意图. 所谓因变量, 就是我们要预测的目标变量. 当因变量为数量变量时, 这种建模称为**回归**, 而当因变量为分类变量(定性变量)时, 则建模称为**分类**. 利用数据训练模型是一个学习过程, 因此, 统计建模过程也称为**统计学习**(statistical learning). 在有因变量的情况下, 无论是回归还是分类, 都属于**有指导学习**(supervised learning). 作为对照, 没有因变量的建模, 称为**无指导学习**(unsupervised learning).

目前有很多关于回归的教科书和课程, 但鲜有关于分类的教科书和课程. 而在回归中又以通常称为线性模型的线性最小二乘回归为主, 其原因是在前计算机时代, 线性模型是数学上最方便也最容易研究的模型, 关于线性模型的大量数学结果使其成为硕果累累的一大领域. 从线性模型又引申出非线性模型、广义线性模型、随机效应混合模型等新的建模方向, 使得回归领域不断扩大. 而在分类方面, 仅有在多元分析名下的"判别分析"可以做分类. 分类方面的研究在计算机出现前的很长一段时间远远不如回归那么普遍.

然而在实际工作中, 分类的需求并不比回归少, 但是, 由数学家所发明的经典方法无力解决如此多种多样的分类问题, 而又没有多少人愿意在文献中介绍他们不能解决的问题. 除此之外, 传统的回归方法也由于其对数据所限定的种种无法验证的假定而受到极大的限制和挑战. 计算机时代的到来彻底改变了这种局面. 各种机器学习方法的出现全面更新了传统回归领域的面貌和格局. 机器学习方法充分显示出在回归预测上的优越性能. 在分类领域, 机器学习方法在应用范围及预测精度上都普遍超过传统的诸如判别分析和二元时的logistic回归等参数方法.

本书的宗旨就是既要介绍传统的回归和分类方法, 又要引入机器学习方法, 并且通过实际例子, 运用R软件来让读者理解各种方法的意义和实践, 能够自主做数据分析并得到结论.

传统的回归分析教科书, 通常只讲所述方法能够做什么, 不讲其缺点和局限性, 并且很少涉及其他可用的方法, 而本书以数据为导向, 对应不同的数据介绍尽可能多的方法, 并且说明各种方法的优点、缺点及适用范围. 对于不同模型的比较, 本书将主要采用客观的交叉验证的方法. 对于每一个数据以及通过数据所要达到的目的, 都有许多不同的方法可用, 但具体哪种方法或模型最适合, 则依数据及目标而定, 绝不事先决定.

本书所有的分析都通过免费的自由软件R来实现.[1] 读者可以毫不困难地重复本书所有的计算. R网站[2]拥有世界各地统计学家贡献的大量最新程序包(package), 这些程序包以飞快的速度增加和更新, 已从2009年底的不到1000个增加到2015年8月中旬的7000多个. 它们代表了统计学家创造的针对各个统计方向及不同应用领域的崭新统计方法. 这些程序包的代码大多是公开的. 与此相对比, 所有商业软件远没有如此多的资源, 也不会更新得如此之快, 而且商业软件的代码都是保密的昂贵"黑匣子".

在发达国家, 不能想象一个统计研究生不会使用R软件. 那里很多学校都开设了R软件的课程. 今天, 任何一个统计学家想要介绍和推广其创造的统计方法, 都必须提供相应的计算程序, 而发表该程序的最佳地点就是R网站. 由于方法和代码是公开的, 这些方法很容易引起有关学者的关注, 这些关注对研究相应方法形成群体效应, 推动其发展. 不会编程的统计学家在今天是很难生存的.

在学校中讲授任何一种商业软件都是为该公司做义务广告, 如果没有相关软件公司的资助, 就没有学校愿意花钱讲授商业软件. 在教学中使用盗版软件是违法行为, 绝对不应该或明或暗地鼓励师生使用盗版商业软件, 使得师生通过盗版软件对其产生依赖性, 并抑制人们自由编程能力的发展.

对R软件编程的熟悉还有助于学习其他快速计算的语言, 比如C++, FORTRAN, Python, Java, Hadoop, Spark, NoSQL, SQL等, 这是因为编程理念的相似性, 这对于应对因快速处理庞大的数据集而面临的巨大的计算量有所裨益. 而熟悉一些傻瓜式商业软件, 对学习这些语言没有任何好处.

本书试图让读者理解世界是复杂的, 数据形式是多种多样的, 必须有超越书本、超越所谓权威的智慧和勇气, 才能充满自信地面对世界上出现的各种挑战.

由于统计正以前所未有的速度发展, R网站及其各个程序包也在不断更新, 因此, 笔者希望读者通过对本书的学习, 学会如何通过R不断学习新的知识和方法. "授人以鱼, 不如授之以渔", 成功的教师不是像百科全书那样告诉学生一些现成的知识, 而是

[1] R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.r-project.org/.

[2] 网址: http://www.r-project.org/.

让学生产生疑问和兴趣, 以促进其做进一步的探索.

本书所有的数据例子都可以从网上找到并且下载. 这些例子背后都有一些理论和应用的故事. 笔者并没有刻意挑选例子所在的领域, 统计方法对于各个实际领域是相通的. 我们想要得到的是在任何领域都能施展的能力, 而不是有限的行业培训. 如果你能够处理具有挑战性的数据, 那么无论该数据来自何领域, 你的感觉都会很好.

本书包括的内容有: 经典线性回归、广义线性模型、纵向数据(分层模型)、机器学习回归方法(决策树、bagging、随机森林、mboost、人工神经网络、支持向量机、k最近邻方法)、生存分析及Cox模型、经典判别分析与logistic回归分类、机器学习分类方法(决策树、bagging、随机森林、adaboost、人工神经网络、支持向量机、k最近邻方法). 其中, 纵向数据(分层模型)、生存分析及Cox模型的内容可根据需要选用, 所有其他的内容都应该在教学中涉及, 可以简化甚至忽略的内容为一些数学推导和某些不那么优秀的模型, 不可以忽略的是各种方法的直观意义及理念.

本书的适用范围很广, 其内容曾经在中国人民大学、首都经贸大学、中央财经大学、西南财经大学、云南财经大学、四川大学、哈尔滨理工大学、新疆财经大学、中山大学、内蒙古科技大学、云南师范大学及大理大学讲授过，对象包括数学、应用数学、金融数学、统计、精算、经济、旅游、环境等专业的本科生以及数学、应用数学、统计、计量经济学、生物医学、应用统计、经济学等专业的硕士和博士研究生. 作为成绩评定, 给每个学生分配若干网站上的实际数据, 并要求他们在学期末将分析处理这些数据的结果形成报告. 这些数据如何处理, 没有标准答案, 甚至有些必要的方法还超出了授课的范围, 需要学生做进一步的探索和学习.

笔者认为, 这本书可以作为本科生的回归分析及分类课程的教科书, 应用统计硕士的知识应该包括本书的全部内容. 希望本书对于各个领域的教师以及实际工作者都有参考价值.

本书的排版是笔者通过LaTeX软件实现的.

在任何国家及任何制度下都能够生存和发展的知识和能力, 就是科学, 是人们在生命的历程中应该获得的.

吴喜之

目 录

第一章　引 言

1.1　作为科学的统计

1.1.1　统计是科学

统计是**科学**(science), 而科学的基本特征是其方法论: **对世界的认识源于观测或实验所得的信息(或者数据), 总结信息时会形成模型(亦称假说或理论), 模型会指导进一步的探索, 直到遇到这些模型无法解释的现象, 这就导致对这些模型的更新和替代**. 这就是科学的方法. 只有用科学的方法进行的探索才能称为科学.

科学的理论完全依赖于实际, 统计方法则完全依赖于来自实际的数据. 统计可以定义为"收集、分析、展示和解释数据的科学", 或者称为**数据科学**(data science). 统计几乎应用于所有领域. 人们现在已经逐渐认识到, 作为数据科学的统计, 必须和实际应用领域结合, 必须和计算机科学结合, 才会有前途(见图1.1).

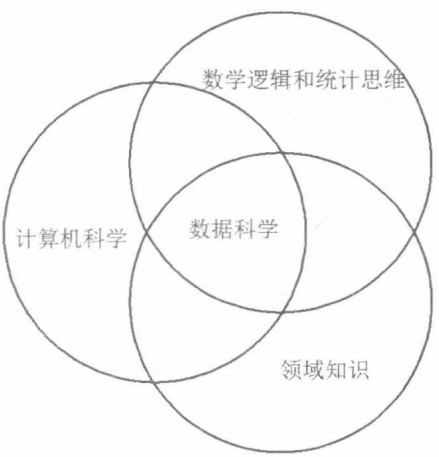


图 1.1　作为数学逻辑和统计批判性思维、计算机科学、实际领域知识之交的数据科学

统计的思维方式是**归纳**(induction), 也就是从数据所反映的现实得到比较一般的模型, 希望以此解释数据所代表的那部分世界. 这和以**演绎**(deduction)为主的数学思维方式相反, 演绎是在一些人为的**假定**(或者在一个公理系统)之下, 推导出各种结论.

1.1.2　模型驱动的历史及数据驱动的未来

在统计科学发展的前期, 由于没有计算机, 不可能应付庞大的数据量[1], 只能在对少量数据的背景分布做出诸如独立同正态分布之类的数学假定后, 建立一些假定的数学模型, 进行手工计算, 并推导出一些由这些模型所得结果的性质, 诸如置信区间、假

[1]请想象一下用纸和笔来计算简单线性回归所必须计算的预测矩阵$\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$, 假定$\boldsymbol{X}$为$30\times 5$的数值矩阵.

设检验的p值、无偏性及相合性等. 在数据与数学假定相差较远的情况下, 人们又利用中心极限定理或各种大样本定理得到当样本量趋于无穷时的一些类似性质. 统计的这种发展方式, 给统计打上了很深的数学烙印.

统计发展的历史痕迹体现在很多方面, 特别是流行"模型驱动"的研究及教学模式. 各统计院系的课程大多以数学模型作为课程的名称和主要内容, 一些数理统计杂志也喜欢发表没有数据背景的关于数学模型的文章. 很多学生毕业后只会推导一些课本上的公式, 却不会处理真实数据. 一些人对于有穷样本, 也假装认为是大样本, 并且堂而皇之地用大样本的性质来描述从有穷样本中得到的结论. 至于数据是否满足大样本定理的条件, 数据样本是不是"大样本"等关键问题尽量不谈或少谈. 按照模型驱动的研究方式, 一些学者不从数据出发, 而是想象出一些他们感觉很好的数学模型, 由于苦于世界上不存在"适合"他们模型的数据, 他们则可能按照自己的需要来模拟一些满足自己需要的数据来说明自己的模型"有价值". 这种自欺欺人的做法绝对是不科学的.

以模型而不是数据为主导的研究方式导致统计在某种程度上成为自我封闭、自我欣赏及自我评价的系统. 固步自封的后果是, 30多年来, 统计丢掉了许多属于数据科学的领域, 也失去了许多人才. 在存在大量现成数学模型无法处理的复杂数据的情况下, 计算机领域的研究人员和部分概率论及统计学家开发了许多计算方法, 处理了传统统计无法解决的大量问题. 诸如人工神经网络、决策树、boosting、随机森林、支持向量机等大量算法模型的相继出现宣告了传统数学模型主导(如果不是垄断的话)数据分析时代的终结. 这些研究最初根本无法刊登在传统统计杂志上, 因此大多出现在计算机及各应用领域的杂志上.

模型驱动的研究方法在前计算机时代有其合理性. 但是在计算机快速发展的今天, 仍然固守这种研究模式, 就不会有前途了. 人们在处理数据时, 首先寻求现有的方法, 当现有方法不能满足他们的需求时, 往往会根据数据的特征创造出新的可以计算的方法来满足实际需要, 这就是统计科学近年来飞速发展的历程. 创造模型的目的是适应现实数据. 统计研究应该是由问题或者数据驱动的, 而不是由模型驱动的.

随着时代的进步, 各个统计院系现在也开始设置诸如数据挖掘、机器学习等课程, 统计杂志也开始逐渐重视这些研究. 这些算法模型很多都不是用封闭的数学公式来描述的, 而是体现在计算机算法或程序上. 对于结果的风险也不是用假定的分布(或渐近分布)所得到的p值, 而是用没有参加建模训练的测试集的交叉验证的误差来描述的. 这些方法发展得很快, 不仅因为它们能够更加精确地解决问题, 还因为那些不懂统计或概率论的人也能够完全理解结果(这也是某些有"领域垄断欲"的传统统计学家不易接受的现实). 现在, 无论承认与否, 多数统计学家都明白, 如果不会计算机编程或者不与编程人员合作, 则不会产生任何有意义的成果.

1.1.3 数据中的信息是由观测值数目及相关变量的数目决定的

为了使得模型简单且具有可计算性, 传统的统计研究人员经常把很大精力投入到减少自变量数目的降维研究上, 而很多机器学习方法不但不降维, 而且希望有更多相

关变量的参与. 事实上, 所有的人都明白, 除了样本量之外, 变量越多, 信息量越大. 比如金融机构想要知道客户是否有信用, 就需要很多客户信息, 比如年龄、职业、收入、过去的信用记录等, 这些其实远远不够, 如果还能加上客户的行为、心理特征、朋友圈、理财效率和财产使用模式等则更好, 谁能够说应该为了计算方便而减去一些变量呢?

现代的计算机及算法对于信息量大的数据根本不惧怕, 它们欢迎巨大的样本量和变量维数, 因为维数是宝贵的资源, 从中可以得到低维状况无法得到的大量信息, 提高统计预测的准确性.

1.2 传统参数模型和机器学习算法模型

在回归和分类中, 模型都可以表示成下面抽象的形式:

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\epsilon}) \tag{1.1}$$

这里$\boldsymbol{y}$为因变量, 分类时$\boldsymbol{y}$是定性变量(亦称分类变量、属性变量等), 回归时$\boldsymbol{y}$是定量变量(亦称数量变量等), 它可能是向量; 而$\boldsymbol{x}$为自变量(可以是向量), 可以是定性变量或定量变量; 而$f()$是因变量和自变量之间的关系, 在参数模型中是一个公式, 在机器学习方法中是一个算法; 其中的$\boldsymbol{\theta}$在参数模型中可以代表参数(向量), 在算法模型中可以代表具体的模型; 最终, 所有的模型都是近似的, 这样就把模型和数据之间不吻合的地方都归到误差$\boldsymbol{\epsilon}$(可以是向量)上去. 有人把$\boldsymbol{\epsilon}$称为随机误差, 这是不妥的, 因为只有完全确定你的模型的准确性(不可能的), 才可以这样说, 但所有模型都是猜想, 不应该狂妄地说这误差是随机的.

从模型(1.1)可以引申出许多具体的模型, 比如误差可加模型

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{\theta}) + \boldsymbol{\epsilon}, \tag{1.2}$$

线性模型

$$\boldsymbol{y} = \boldsymbol{x}^{\top}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.3}$$

等, 这里$\boldsymbol{\beta}$是系数(参数).

1.2.1 参数模型比算法模型容易解释是伪命题

很多人认为机器学习的算法模型不如参数模型容易解释自变量对因变量的贡献. 这是因为他们对两种模型都缺乏了解.

以线性模型为例, 很多人认为拟合的系数代表自变量对因变量的贡献, 还说"当其他变量不变时, 一个自变量的系数代表该自变量对因变量的贡献". 其实, 这仅仅在所有变量都不相关时才成立, 而这在大多数回归中根本不成立(或至少无法验证).

但机器学习中的诸多方法可以从各个角度评价各个变量和观测值的关系. 比如随机森林不仅可以给出各个变量在回归和分类中从不同角度衡量的重要性以及对因变量的影响, 还可以给出每一个观测值和每一个变量之间的关系重要性, 以及所有观测

值之间在回归和分类中的关系. 目前还没有对任何一种经典方法有如此详尽的剖析. 这些结果比在沉重而又不可靠的数学假定下对系数的解释更加客观、合理.

1.2.2 参数模型的竞争模型的对立性和机器学习不同模型的协和性

对于每个数据, 总有一些不同的参数模型均被认为可以很好地解释数据所代表的现象, 它们互相竞争. 实际上, 这些模型的优劣仅仅是从不同的角度来刻画的, 除非用交叉验证, 否则很难比较. 但是机器学习方法可以把不同的竞争模型组合起来, 产生比单个模型更加精确的预测. 这如同俗语所说, "三个臭皮匠, 顶个诸葛亮".

1.2.3 评价和对比模型

很多传统回归分析教科书对于模型的评价是基于对数据及模型形式的数学假定, 且只用一个训练集本身对模型的拟合来判断模型是否合适. 这种用参与建模的数据加上主观假定来判断模型的方式不但很主观, 而且无法与其他模型做对比.

交叉验证的方法是在计算机时代才发展起来的, 它用训练数据集来训练模型, 然后用未参与建模的测试数据集来评价模型预测功能的优劣. 这对于在任何模型之间做预测比较都适用. 交叉验证不用对模型做任何假定, 因此是能够为各个领域的人所理解和接受的. 对于诸如回归和分类这样的有指导学习, 预测能力是反映模型好坏的最根本的标准.

交叉验证最常用的是N折交叉验证. 其要点为, 把数据随机分成N份, 轮流把其中1份作为测试集, 其余的$N-1$份合起来作为训练集; 然后用训练集拟合数据得到模型, 并用这样训练出来的模型来拟合未参加训练的测试集数据. 这种交叉验证共做N次. 对于分类, 就会得到在测试集中的N个误判率, 从而得到平均误判率; 而对于回归, 就可以得到在测试集中的N个**标准化均方误差**(normalized mean squared error, NMSE), 并得到其平均. 标准化均方误差NMSE定义如下:

$$NMSE = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{1.4}$$

这里的y_i是**测试集**的因变量观测值; $\hat{y}_i$是**利用训练集得到的模型拟合测试集得到的因变量拟合值**; $\bar{y}$是**测试集**的因变量观测值的均值. 其分母是不用任何模型, 而仅仅用因变量观测值的均值来作为拟合值的均方误差MSE; 而分子为运用模型的拟合结果. 如果标准化均方误差NMSE小于1, 说明用这个模型比不用模型要强, NMSE越小越好; 如果NMSE大于1, 则说明这个模型根本是垃圾, 不能用.

交叉验证可以做很多次, 每次的N个数据子集都不一样, 这样得到的结果更加客观. 由于数据子集的选择是随机的, 交叉验证结果不唯一, 但可以发现, 多次交叉验证的结果差别不会太大. 后面将会更加具体地介绍各种情况下交叉验证数据子集的选择过程.