<<数据之魅>>

图书基本信息

书名：<<数据之魅>>

13位ISBN编号：9787302290988

10位ISBN编号：7302290989

出版时间：2012-7

出版时间：清华大学出版社

作者：（美）Philipp K. Janert

页数：524

译者：黄权

前言

本书展现了我在高科技行业的各个公司中从事数据工作所获得的经验。
它汇聚了我所发现的许多最有用的概念和技术，包括我希望自己能够早点知道的主题——然而我没有。
我所学的专业是物理，但我也从事了多年的软件工程师工作。
本书将反映出我这种双重背景。
一方面，本书是为程序员以及软件领域的其他人士而写：我假定你和我一样，有能力通过自己动手编程来轻松自如地操纵数据。
另一方面，我思考数据的方式是由我的背景和教育决定的。
作为一个物理学家，我不会只满足于描述数据或者做黑盒式的预测：分析的目的总是为了深入理解我们所观察的数据是怎样产生的。
传达这种理解的工具就是模型：对所研究的系统进行描述(换句话说，不只是对数据的描述！)，必要时进行简化但要保留相关的信息。
一个模型可能很粗糙(就像一头球形的牛)，但如果它能够帮助我们更好地理解系统的工作原理，那么它就是一个成功的模型。
(精确度可以在之后获得，如果确实需要的话。)　我对模型和简化描述的强调并不具有普遍性：其他作者和从业人员可能持有不同的看法。
但是它们对于我的方法和观点来说是基本的。
这本书相当具有个人色彩。
尽管我努力使之合理全面，但我所选择的主题都是我认为在实践中相关和有用的——不管它们是否是“经典”。
本书还包含其他数据分析相关书中不涉及的主题。
尽管它们既不新颖也非独创，但在数据分析这一特定背景下通常并不使用或讨论它们——但我发现它们不可或缺。
在整本书中，我提供了大量明确而具体的建议、意见和评估。
这些评述反映了我的个人兴趣、经验和理解。
我不敢说我的观点一定是正确的，请根据具体需要对我所说的进行评估和取舍。
在我看来，一个充分论证的明确立场比列出所有待选的可能算法更有用——即使后来你决定不同意我的观点。
价值并不存在于观点中，而是存在于支持它的论据中。
如果你的论据比我的好，或者仅仅只是更适合你，那么我也认为自己已经达到了我的目的！

<<数据之魅>>

内容概要

《数据之魅：基于开源工具的数据分析》结合作者多年来从事数据分析工作的丰富经验，阐述了数据分析所涉及的概念和方法。
本书四部分19章，主题包括如何通过图表来观察数据，如何通过各种建模方法来分析数据，然后着重阐述如何进行数据挖掘，最后强调数据分析在商业和金融等领域的实际应用。
本书包含大量的模拟过程及结果展示，并通过实例来阐述如何使用开源工具来进行数据分析。
通过本书的阅读，读者可以清楚地了解这些方法的实际用法及用途。
本书结构合理，通俗易懂，适合数据分析爱好者和从业者阅读，也适合以科学计算为工具的科研人员参考。
同时，本书还适用于计算机科学、数学、工程技术和其他相关专业本科或研究生的数据分析课程，是一本不错的参考书。

作者简介

作者：（美国）雅奈特（Janert K.P.） 译者：黄权、陆昌辉、邹雪梅、费柳凤

<<数据之魅>>

书籍目录

第1 章导论1 数据分析1 本书内容2 关于讲习班 3 关于数学4 需要具备的知识 6 本书不涉及的内容6 第Ⅰ部分图表：观察数据 第2 章单一变量：形状和分布 11 数据点和抖动图 12 直方图和核密度估计 14 直方图15 核密度估计 19 (选学)如何选择最优带宽 22 累积分布函数23 (选学)概率图分布和qq 图分布的对比 25 秩序图和上升图 30 仅用于适当时机：汇总统计量和箱形图 33 汇总统计量 33 box-and-whisker 图 36 (讲习班)numpy 38 numpy 实践 38 numpy 详解 41 扩展阅读 45 第3 章两个变量：建立关系 47 散点图 47 克服噪声：平滑 48 样条 50 loess51 示例 52 残差 54 其他观点及提醒55 对数图 57 倾斜 61 线性回归以及诸如此类的方法 62 描述重要信息 66 图形分析与图形演示68 (讲习班)matplotlib 69 交互式使用matplotlib 70 案例学习：matplotlib 与loess73 控制属性 74 matplotlib 对象模型及结构 76 零碎知识 77 扩展阅读 78 第4 章以时间为变量： 时序分析 79 示例 79 任务 83 需求和现实 84 平滑处理 84 移动平均法 85 指数平滑法 86 不要忽视显而易见的东西 90 相关函数 91 示例 92 实现上的问题 93 (选学)过滤器和卷积 95 (讲习班)scipy.signal 96 扩展阅读 98 第5 章多变量：图形的多变量分析 99 假色图100 概览：多值图 105 散点图矩阵105 协作图 107 变种 108 组成问题 110 组成的改变110 多维组成：树形图和马赛克图112 新颖的曲线类型116 标识符116 平行坐标图117 交互式探索120 查询和缩放121 连接和涂层121 大游览与投影寻踪121 工具 122 工作坊：多变量图形工具123 R 123 实验工具124 python 的chaco 库124 扩展阅读 125 第6 章插曲：数据分析会话 127 数据分析会话127 工作坊：gnuplot 软件136 扩展阅读 138 第Ⅱ部分分析：数据建模 第7 章推算和粗略计算141 推算的原理 142 估计大小143 建立关联145 使用数字146 10 的幂146 小扰动147 对数148 更多示例149 我所知道的一些常见事(物) 的相关数字151 这些数字是否足够好？ 151 准备工作：可行性和成本 153 完成之后：引用和 呈现数字154 (选学)进一步探索摄动理论和误差传播 155 误差传播156 工作坊：Gnu科学库(GSL)158 扩展阅读 161 第8 章缩放参数模型163 模型163 建模 164 模型的运用和误用 164 参数的缩放 165 缩放参数165 示例：维度参数 167 示例：优化问题 169 示例：成本模型 170 (选学)缩放参数与 量纲分析172 其他理论174 平均场近似 175 背景知识和其他示例176 常见的时间演变方案 178 无限增长和衰减现象178 约束增长：逻辑斯谛方程180 振荡 181 案例学习：多少台服务器才是 最好的？ 182 为什么要建模？ 184 工作坊：Sage.184 扩展阅读188 第9 章关于概率模型的讨论 191 9.1 二项分布和伯努利试验191 精确的结果192 利用伯努利试验建立平均场模型194 9.2 高斯分布和中心极限定理195 中心极限定理 195 中心项与尾项 197 为什么高斯分布如此实用？ 198 (选学)高斯积分199 幂律分布和非常规统计学201 幂律分布的用法203 (选学)期望值为无限时的分布204 接下来的研究 206 其他分布 206 几何分布207 泊松分布207 对数正态分布209 特殊用途的分布211 (选学)案例学习——随时间变化的单一访问者数量 211 工作坊：幂律分布215 扩展阅读 219 第10 章你真正需要了解的 经典统计学知识221 起源221 统计学的定义 223 从统计学角度解释 226 示例：公式测验vs 图解法 229 控制实验vs 观察研究 230 实验设计232 前景 234 (选学)贝叶斯统计—— 另一种观点 235 用频率论来解释概率235 用贝叶斯方法来理解概率 236 贝叶斯数据分析： 一个实际有效的例子238 贝叶斯推理：总结与讨论 241 工作坊：关于R 243 扩展阅读249 第11 章插叙：数学大搜捕——大脚怪和最小二乘等253 如何平均均值 253 辛普森(悖论 254 标准差 256 如何计算258 (选学)应该选择哪一个259 (选学)标准误差 259 最小二乘 260 统计参数估计 261 函数逼近263 扩展阅读 264 第Ⅲ部分计算：数据挖掘 …… 附录A科学计算与数据分析的编程环境435 附录B应用：微积分447 附录C 使用数据485 索引499

章节摘录

版权页： 插图： 我很喜欢假色图，因为它既能保留定量信息，又能表示大量信息资料。
然而，假色图的准确性主要取决于调色板的质量。
映射过程，就是将数值与颜色联系在一起的过程。
让我们快速回顾一下颜色和计算机图形的相关知识。
计算机图形中的颜色常常是由一组数据指定的，而这组数据则由红、绿、蓝三种基本色彩元素强度来表示。
虽然RGB三基色技术在技术层面上看似很好，但它并不是特别直观。
相反，我们倾向于从颜色的色调、饱和度和明暗度来考虑颜色表示问题（例如，亮度或颜色的浅淡）。
一般来说，色调包含彩虹的所有颜色（从红到黄、绿、蓝、紫）。
奇怪的是，色谱似乎绕了一圈又回到本身，就像紫最终又平滑地渐变为红。
（这种情况是因为彩虹中色谱是按各种色彩的主要电磁频率的顺序来排列的。
对于紫／品红来说，它们没有主要的频率，而紫色是一种由低频率的红色和高频率的蓝色混合而成的色调。
）大多数计算机图形程序用色调—饱和度—明暗度（HSV）三基色技术来生成彩色图形。
我们很难找到一个可靠的调色板设计方案。
更不幸的是，人们的权宜心理和常识似乎常常导致设计出来的调色板非常糟糕。
这里有一些想法和建议供大家参考。
保持简单 简单地使用红、白、蓝的调色板往往能产生非常好的效果。
对于连续的颜色变化，可以使用蓝—白—红调色板，而对于分割任务，可以使用一个白—蓝—红—白的调色板，分割线上使用蓝—红进行过渡。
分割任务和平滑性改变的区别 分割任务（例如，找到超过一定阈值的所有点，找出过零数据的分布情况）要求在区域两边的临界线上都使用亮丽的色彩过渡，而一个数据集的平滑变化则要求用连续的颜色渐变来表示。
当然，可以在单个调色板中既使用颜色渐变，又使用强烈的对比色。
保持直观上的有序性 在调色板中，可以将低值映射为冷色、高值映射为暖色，让人从直观上感觉井然有序。
类似的例子包括简单的蓝—红调色盘和“强烈、对比色系”（黑—红—黄—白——稍后将讨论为什么不建议使用“强烈的对比色”）。
其他能让人感觉尽然有序的调色方案是“改进的彩虹”（包括蓝—青—绿—黄—橙—红—紫）和与地形图（蓝—青—绿—棕色—黄褐色—白）相似的“地理系列”。

媒体关注与评论

“一本通俗易懂的参考书，有助于理解如何征服海量数据。”
——Austin King，Mozilla资深Web开发人员
“造就数据科学家的必读工具书。”
——Michael E. Driscoll，Dataspora的CEO兼创始人

<<数据之魅>>

编辑推荐

《数据之魅:基于开源工具的数据分析》结构合理，通俗易懂，适合数据分析爱好者和从业者阅读，也适合以科学计算为工具的科研人员参考。
同时，《数据之魅:基于开源工具的数据分析》还适用于计算机科学、数学、工程技术和其他相关专业本科或研究生的数据分析课程，是一本不错的参考书。

名人推荐

“ Google，Facebook，Amazon和Netflix，更别说华尔街和制造业、零售业到保健行业的企业，他们的成功越来越得益于选择正确的工具从海量数据中抽取和挖掘出有意义、有价值的信息。
现在，‘ 数据科学家 ’是硅谷最抢手的人物。
”——Tim O'Reilly“ 一本通俗易懂的参考书，有助于理解如何征服海量数据。
”——Allstin King. Mozilla资深Web开发人员“ 造就数据科学家的必读工具书。
”——Michael E.Driscoll. Dataspora的CE0兼创始人

版权说明