R语言培训
——R的应用

李舰

Email: lijian.pku@gmail.com

Homepage: www.leejian.name

第三届中国R语言会议（上海分会）

2010 年 11 月

目 录

目 录

目 录

目 录

- 如果您对R的需求属于以下一条或几条，可以开始午睡了zzzzzz
  - 使用R内置的标准统计函数进行简单的数据分析
  - 用R当做图工具
  - 用R当科学计算器
  - ……
- 如果您对R的需求属于以下一条（没有或几条！），您应该不在这里。。。
  - 简单便捷的数据分析（不如SPSS）
  - 便捷的数据整理（不如Excel）
  - 轻松实现高性能的统计分析（不如SAS）
  - 只追求计算能力（不如C和Fortran）
  - 开发商业应用系统（不如JAVA）
  - 开源灵活通用的语言（不如Python）
  - 数学领域通用便捷的工具（不如Matlab）
  - ……

- 如果您平时的工作或研究需要大量的
  - 数据分析和统计建模
- 同时，对以下要求也比较高
  - 统计图形
  - 数据处理
  - 矩阵运算
  - 易于学习
  - 自由免费
  - 灵活的编程
  - 轻量化的平台
  - 可扩展的能力
  - 强大社区的支持
  - ……
- 那么，R将会是最好的选择

灵活的语言

- S语言是一种用于数据分析和图形展示的高级语言
  - 为数据而生的程序设计语言
  - 设计之初就开始追求的可扩展性
- 最早的R也参照了很多Scheme语言的特性
  - Scheme是LISP的一个方言
  - Scheme是一种链表处理语言，擅长符号计算
  - Scheme可以像操作数据一样操作函数

开放的平台

- R能很方便地和其他软件系统整合，从而扩展应用
  - R与底层运算，调用C或者Fortran
  - R与其他语言的调用，rjava、rJython、rcom
  - R与数据库，rodbc、DBI、RODM
  - R与Office，statconnDCOM
  - R与网络信息，RCurl、XML
  - R与大内存处理，bigmemory、ff
  - R与并行计算，snow、rmpi、foreach
  - ……

强大的社区支持

- R的贡献者们分布在全球各地，活跃在各种应用领域，很多专业领域都能找到适用的R项目或者R包
  - 统计动画，animation
  - 矩阵可视化，corrplot
  - 生物数据分析，Bioconductor
  - 最优化方法，glpk、lpSolve
  - 贝叶斯统计，MCMCpack、WinBUGS
  - 抽样，sampling、survey
  - 水文模拟，TOPMODEL、RHydro
  - 随机微分方程，sde
  - 心理学，AnalyzeFMRI、fmri、psychometric
  - 决策树，rpart
  - 图形界面，gWidgets
  - 文本挖掘，tm、tau
  - 混合效应模型，nlme
  - ……

矩阵乘法的应用

- 对如下矩阵，求出每一行最大数所属的类别，用1,2,3,4表示

```
m.data
    class1    class2    class3     class4
0.3909334 0.8309291 0.6547570 0.52187303
0.7665842 0.4181528 0.7140503 0.70818227
0.2036516 0.4182180 0.6572924 0.02469917
```

- 最后得到类别向量$c(2,1,3)$

示例

- apply(abs(m.data==apply(m.data,1,max))%*%diag(1:4),1,sum)

奇异值分解的应用

- 对于 $m \times n$ 矩阵 A,存在 m 阶正交阵 u 和 n 阶正交阵 v，使得 $A = u \times d \times v'$，其中 d 是对角阵，是为奇异值分解
  - u 和 v 是 A 的奇异向量
  - d 是 A 的奇异值
- 在实际应用中，奇异值分解可以产生很好的效果
  - 奇异值分解可以快速地帮助文本分类
  - 对于大型的稀疏矩阵，通过奇异值进行化简，可以节省存储，提高运算能力

- 下例中的每一行表示一个词在不同文章中出现的次数，每一列表示一篇文章包含某个词的数目

svd.data

words	a1	a2	a3	a4	a5	ba	b2	b3	b4
人类	1	0	0	1	0	0	0	0	0
接口	1	0	1	0	0	0	0	0	0
计算机	1	1	0	0	0	0	0	0	0
用户	0	1	1	0	1	0	0	0	0
系统	0	1	1	2	0	0	0	0	0
响应	0	1	0	0	1	0	0	0	0
时间	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
调查	0	1	0	0	0	0	0	0	1
树	0	0	0	0	0	1	1	1	0
图	0	0	0	0	0	0	1	1	1
子式	0	0	0	0	0	0	0	1	1

- 可以对该稀疏矩阵进行奇异值分解

- svd.res = svd(svd.data[,2:10])

```
>
> round(t(svd.res$u[,1:2]),2)
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12]
[1,] -0.22 -0.20 -0.24 -0.40 -0.64 -0.27 -0.27 -0.30 -0.21 -0.01 -0.04 -0.03
[2,] -0.11 -0.07  0.04  0.06 -0.17  0.11  0.11 -0.14  0.27  0.49  0.62  0.45
>
> round(svd.res$v[1:2,],2)
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
[1,] -0.20 -0.06  0.11 -0.95  0.05 -0.08 -0.18  0.01  0.06
[2,] -0.61  0.17 -0.50 -0.03 -0.21 -0.26  0.43 -0.05 -0.24
>
> round(svd.res$d,2)
[1] 3.34 2.54 2.35 1.64 1.50 1.31 0.85 0.56 0.36
>
```

少用显式循环，多用apply系列函数

- apply(array, margin, function, ...)
  - 将一个函数作用于一个数组的某个维度，并返回一个降维后的数组
  - 通过 margin 指定作用于哪个维度
- lapply(list, function, ...)
  - 将函数作用于列表或向量的每个元素，并返回一个相同长度的列表
  - 可以代替循环
- sapply(list, function, ...)
  - 将函数作用于列表或向量的每个元素，并返回一个向量、矩阵或者列表
  - 如果simplify设为F，则等价于lapply
  - 默认simplify为T，会根据函数值自动判断返回的数据类型，如果是纯量则返回向量，如果是等长的量则返回矩阵，否则返回列表

- tapply(array, indicies, function, ..., simplify)
  - 将一个函数作用于一个不规则数组（ragged array）的每个单元
  - 对某个数组按照位置分好组，然后对每组进行函数运算
  - 可以按照因子进行分组
- sweep(array, margin, stats, function, ...)
  - 将运算符作用于数组的某个维度，使之和某个减维后的数组运算

R中的高性能运算

- 内存管理
  - 内存上限和垃圾清理
- 函数库的优化
  - BLAS和LAPACK
- 大内存的处理
  - 数据库
  - bigmemory
  - MapReduce
- 并行计算
  - 显式并行
  - 隐式并行

两道关于概率的应用题

- 2009年6月12日，湖北武汉5141名困难家庭市民参与一个经适房小区公开摇号，结果中签的124名市民当中有6人的购房资格证明的编号是连号。请计算该事件发生的概率。
  - 媒体报道概率为千万亿分之一
- 2009年7月29日，湖北老河口市第二期经济适用住房把摇号结果发在了网上，很快被网民发现在1138户具有购房资格的申请者中，抽中了514户购房者，其中有14户资格证编号相连。请计算该事件发生的概率。
  - 统计系副主任估算出来的百分之一左右

- 精确求解的方法
  - 将1138个球投入514个框内，求每个框中最多球数为14个且至少有一个框中正好有14个球的概率
  - 计算最多数目为k个时，确定某个框中正好有14个，全排列后减去n-14个球投入513个框的数目
  - 递归求解
- 蒙特卡罗方法
  - 也称随机模拟或者统计模拟方法
  - 利用随机变量模拟现实的过程
  - 经过多次模拟，统计数值特征
- 该题蒙特卡罗方法的详细步骤可以参考：http://taiyun.cos.name/tag/%e6%a6%82%e7%8e%87/

遗传算法的思路

- 依据基因建模
  - 将解的数据结构尽量设为离散数据，0-1型最好
- 构造初始种群
  - 针对优化问题，在整个解空间，通过随机数的方式产生满足约束条件的初始解
- 制定遗传规则
  - 繁殖资格的规则，轮盘法
  - 基因交换的规则
  - 自然选择的规则
  - 变异的规则
- 迭代运算
  - 每次迭代模拟一个世代，使得种群（可行解）不断进化，直到搜寻到满意解或到达指定的上限

线性规划示例

- 假设工厂生产A、B、C，三种商品，A和C的利润都为1元，B的利润为9元，生产A、B、C耗费甲类原材料的数目分别为1、2、3，耗费乙类原材料的数目分别为3、2、2，甲类原材料总量为9，乙类原材料总量为15，如何安排生产使得工厂利润最高？
- 优化模型如下

  Set up problem: maximize

  $x_1 + 9x_2 + x_3$ subject to

  $x_1 + 2x_2 + 3x_3 \leq 9$

  $3x_1 + 2x_2 + 2x_3 \leq 15$

用R求解

- Rglpk
- lpSolve
  - library(lpSolve)
  - f.obj $<-$ c(1, 9, 3)
  - f.con $<-$ matrix (c(1, 2, 3, 3, 2, 2), nrow=2, byrow=TRUE)
  - f.dir $<-$ c("$\leq$", "$\leq$")
  - f.rhs $<-$ c(9, 15)
  - lp ("max", f.obj, f.con, f.dir, f.rhs)

R的文本处理

- grep(pattern, x, ignore.case = FALSE, perl = FALSE, value = FALSE, fixed = FALSE, useBytes = FALSE, invert = FALSE)
  - 通过正则表达式搜索文本，并把匹配的行打印出来
  - pattern表示正则表达式
  - x表示要查找的向量
  - perl表示是否使用Perl的正则表达式规则
- sub(pattern, replacement, x, ignore.case = FALSE, perl = FALSE, fixed = FALSE, useBytes = FALSE)
  - 将找到的子串替换成replacement

正则表达式

- 正则表达式元字符
  - "^" 匹配一个字符串的开始
  - "$" 匹配一个字符串的结尾
  - "." 匹配除了换行符以外的任一字符
  - "*" 表示将其前的字符进行0个或多个的匹配
  - "?" 匹配0或1个正好在它之前的那个字符
  - "+" 匹配1或多个正好在它之前的那个字符。
  - "|" 表示逻辑的或
- 贪婪匹配和懒惰匹配
  - 默认情况下是匹配尽可能多的字符，是为贪婪匹配
  - 如果要进行懒惰匹配，也就是匹配最短的字串，需要在后面加个"?"

文本挖掘技术

- 中文分词
  - 大学/生活/像/白纸
  - 大学生/活像/白纸
- 文档模型
  - 布尔模型
  - 向量空间模型
  - 概率模型
- 文本检索
  - 全文检索，TR(Text Retrieval)
  - 检索的度量：查准率(precision)、查全率(recall)、F-度量(F-measure)
  - 检索工具：Smart系统、Lucene系统、Okapi系统、Lemur Toolkit系统

- 文本自动分类
  - 分类体系
  - 训练集
  - 分类模型
- 话题检测跟踪
  - Topic Detection and Tracking (TDT)
  - 话题检测，将新闻分为话题类簇
  - 话题跟踪，监控新闻报道信息流以便发现与某一已知话题有关的新报道
- 文本过滤
  - 信息过滤(IF)，从动态的信息流中将满足用户兴趣的信息挑选出来
  - 关注用户建模

- 关联分析
  - 类似于DM中的关联规则
- 文档自动摘要
  - 利用计算机自动地从原始文档中提取全面准确地反映该文档中心内容的简单连贯的短文
  - 摘要方法：位置法、提示字串法、频率统计法、文章框架法、仿人算法
  - 评价方式：利用文档摘要代替原文档执行某个文档相关的应用（检索、分类等）
- 信息抽取
  - Information Extraction
  - 实体提取和关系提取
  - 应用：情报系统，从竞争对手的网站上提取关键数据

文本分类步骤示例

- 中文分词
  - 将网络信息处理为标准化文本，进行分词操作
- 文档建模
  - 将分词后的文章转化成向量模型（Term Vector）
  - 计算文章d中每个词w在t时刻取词的权数：

$$weight_t(d,w) = \frac{tf(d,w)log((W_t+1)/(wf_t(w)+0.5))}{\sqrt{\sum_{w^1 \in d}(tf(d,w^1)log((W_t+1)/(wf_t(w^1)+0.5)))^2}}$$

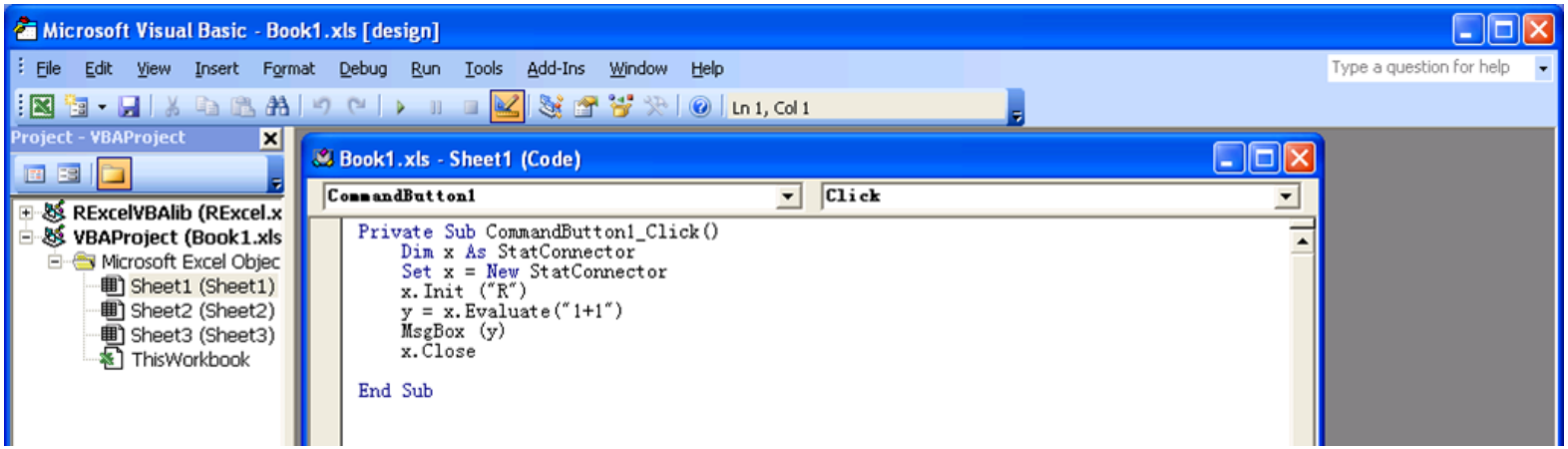
- 按照话题聚类
  - 将t时刻采集的文本向量按相似度（Similarity）聚类
  - 使用余弦夹角的方式计算相似度
- 判别分析
  - 将类簇归入某个已知的类别
- 评价和检验
  - 计算precision、recall、F-measure
- 更新训练集

R与JAVA的整合

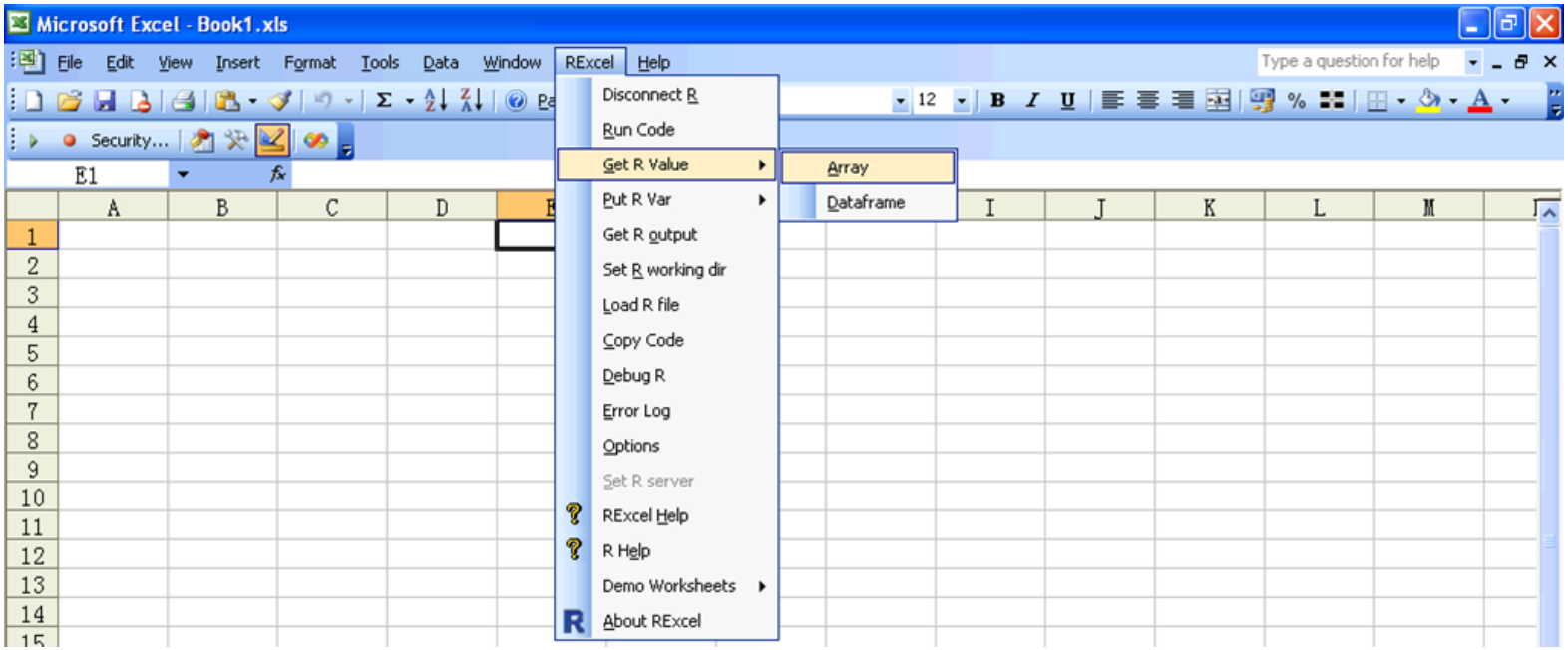
- Rserve
  - http://www.rforge.net/Rserve/
- rJava
  - http://www.rforge.net/rJava/

R与Office的整合

- http://sunsite.univie.ac.at/rcom/
- R与VBA



- RExcel


Microsoft Excel - Book1.xls
File Edit View Insert Format Tools Data Window REXcel Help
Type a question for help
Security...
E1
Disconnect R
Run Code
Get R Value
Put R Var
Get R output
Set R working dir
Load R file
Copy Code
Debug R
Error Log
Options
Set R server
REXcel Help
R Help
Demo Worksheets
About RExcel
Array
Dataframe

R与数据库

- DBI
  - RSQLite
- rodbc

Thank you!

Email: lijian.pku@gmail.com
Homepage: www.leejian.name