Data Mining With R

刘思喆

December 9, 2008

R 对数据的前期处理

R 对数据的前期处理

1. R 对数据库的支持
2. 描述性分析与绘图
   - 单维变量的展示
   - 多维变量的展示

R 对数据的前期处理

1. R 对数据库的支持
2. 描述性分析与绘图
   - 单维变量的展示
   - 多维变量的展示
3. 数据的预处理过程

R 的扩展算法

4 同 Data Mining 相关的 packages

R 的扩展算法

R 的扩展算法

R 的扩展算法

4 同 Data Mining 相关的 packages

5 关联规则

6 分类方法

7 聚类 (cluster)
- 层次聚类
- 分割聚类

R 的扩展算法

4 同 Data Mining 相关的 packages

5 关联规则

6 分类方法

7 聚类 (cluster)
- 层次聚类
- 分割聚类

8 随机森林

R 的扩展算法

4 同 Data Mining 相关的 packages

5 关联规则

6 分类方法

7 聚类 (cluster)
- 层次聚类
- 分割聚类

8 随机森林

9 神经网络

1. R 应该是所有数据分析软件里方法 (函数) 最多的，截至 2008 年 11 月 11 日，cran 上共提供了 1624 个包，涵盖了贝叶斯推断、分类方法、计量经济学、生态学、金融学、遗传学、机器学习、稳健统计、空间统计、生存分析、时间序列等多个方面。
2. 广泛的数据接口。比如 R-base 可以良好的接入 CSV(Comma Separated Values) 数据，或者通过其他包来扩展，直接读入 SPSS、SAS、Minitab、Stata、Excel 等文件，或者直接读取 MySQL、SQL Server、DB2、Oracle 等数据库。
3. 强大的绘图功能。R 提供了为 “高水平”(High level)、“低水平”(Low level) 和 “交互式”(Interactive) 三种绘图命令，而且很容易生成 ps、pdf、png、jpeg、bmp、gif、SVG 甚至 LaTeX或 HTML 输出。
4. 最重要的一点：R is free (in both senses).

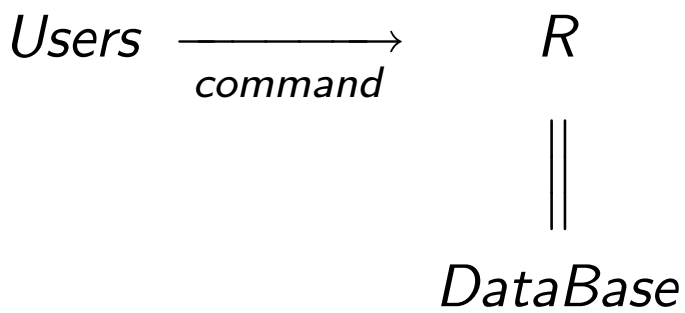
Part I

R 对数据的前期处理

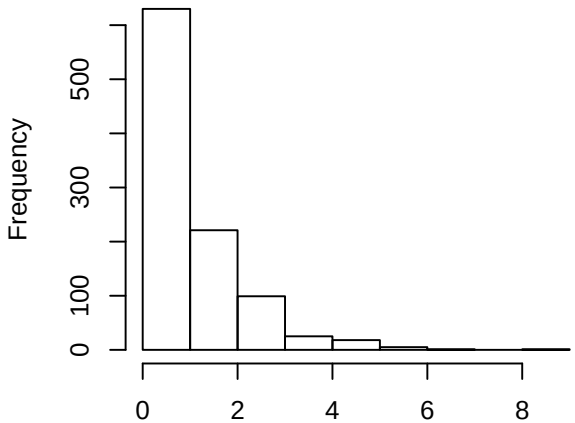
CRAN 上提供了三种 R 同数据库之间的接口:

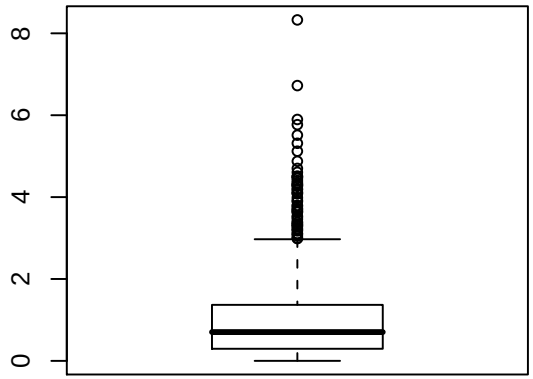
1. **DBI** (DataBase Interface) 包并存的实施模式;
2. Michael Lapsley 负责维护的 **RODBC** 包提供了 R 同支持 ODBC 数据库之间的接口
3. 通过 JDBC(Java DataBase Connectivity) 的方式。

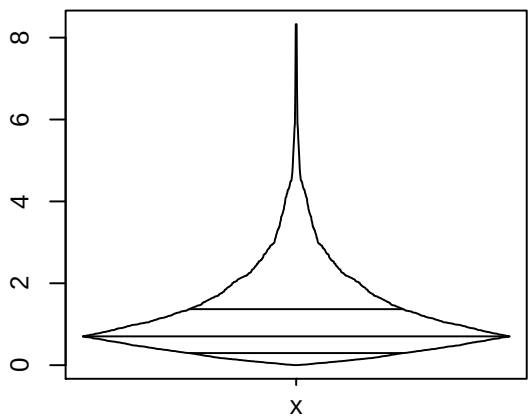
三者的逻辑图

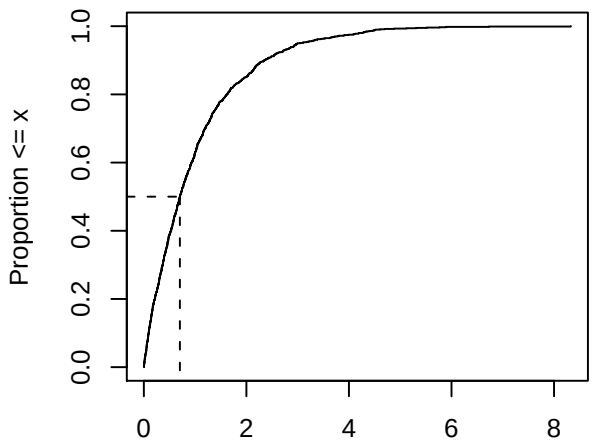


计算均值 (*mean*)、中位数 (*median*)、标准差 (*sd*)、方差 (*var*)、分位数 (*quantile*) 的函数。更一般地，*summary* 函数提供了了解数据的直接途径：


Frequency
0
100
300
500
0
2
4
6
8


0
2
4
6
8


0
2
4
6
8
x


Proportion <= x
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8

传统的可视化技术

- 展示聚集程度和频度的条形图 (*barplot*)；
- 揭示变量分布的直方图 (*hist*)；
- 演示顺序数据趋势的曲线图 (*plot*)；
- 占总体百分比的饼图 (*pie*)；
- 揭示二元变量的关系的散点图 (*plot*)。

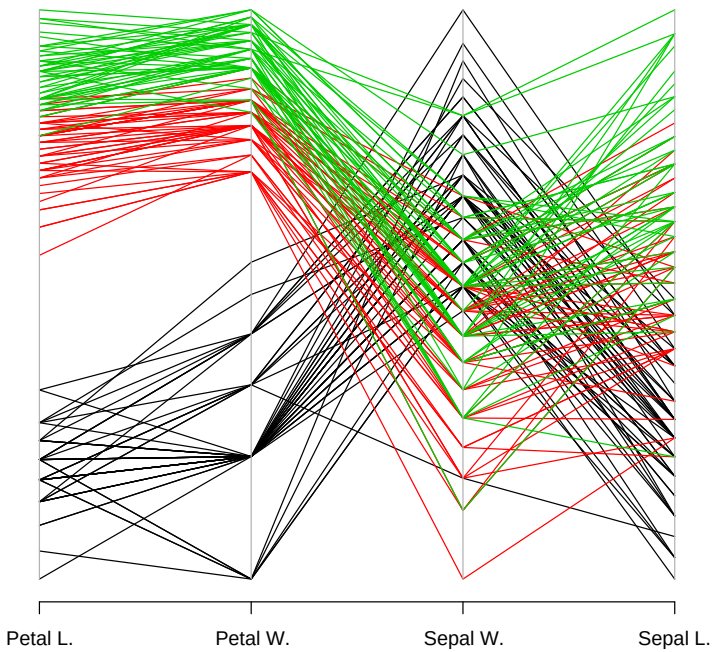
现代数据可视化技术

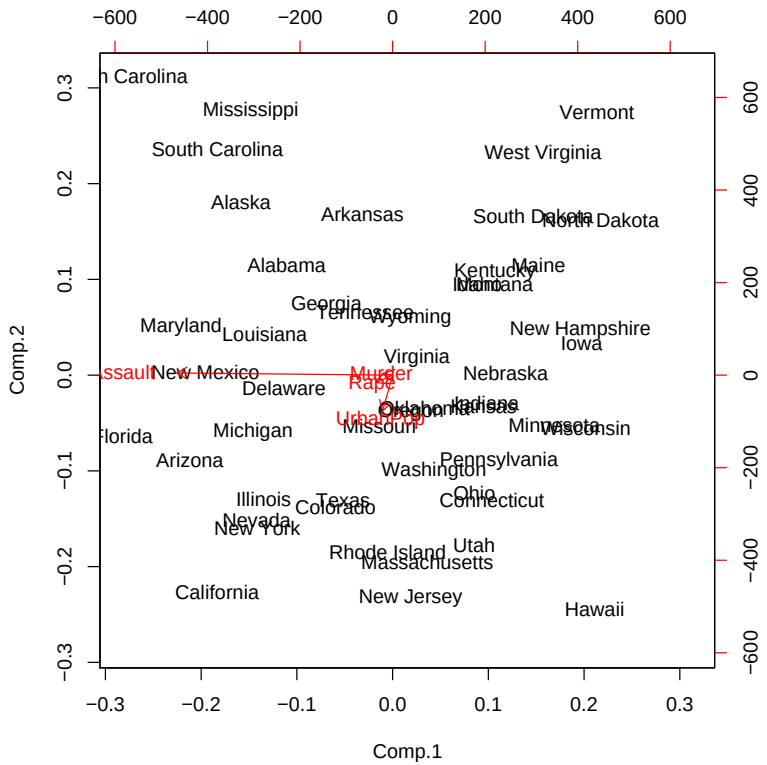
更适合于高维数据的展示，和 R 语言比较相关的技术有

几何投影： 将高维数据通过几何投影，展示在低维度平面；

图像演示： 将单个高纬度数据影射到一个图像上。

几何投影


Petal L.
Petal W.
Sepal W.
Sepal L.

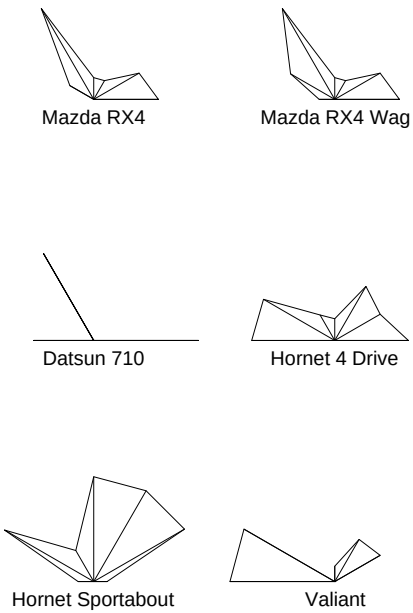

−600
−400
−200
0
200
400
600
Carolina
Mississippi
Vermont
South Carolina
West Virginia
Alaska
Arkansas
South Dakota
North Dakota
Alabama
Kentucky
Maine
Georgia
Tennessee
Wyoming
Maryland
Louisiana
New Hampshire
Iowa
Virginia
New Mexico
Murder
Nebraska
Rape
Delaware
Michigan
Missouri
Florida
Arizona
Pennsylvania
Washington
Illinois
Texas
Ohio
Colorado
Connecticut
Nevada
New York
Rhode Island
Utah
Massachusetts
California
New Jersey
Hawaii
Comp.2
Comp.1
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
600
400
200
0
−200
−400
−600

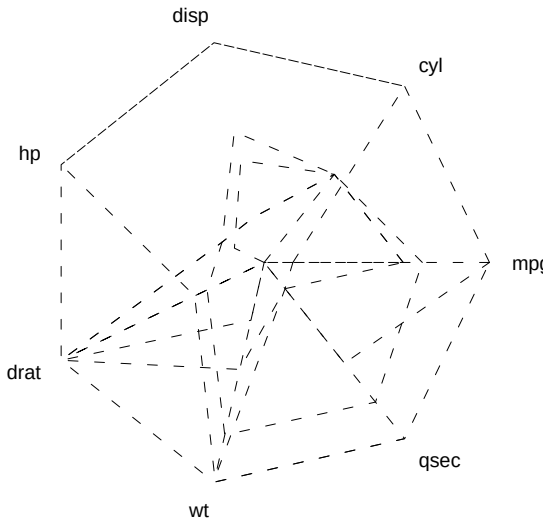
Andrews 调和曲线图

思想是将多维变量 $X = (x_1, x_2, \cdots, x_n)$ 投影到二维平面的一条曲线 $f(t), -\pi \leq t \leq \pi$ 上。

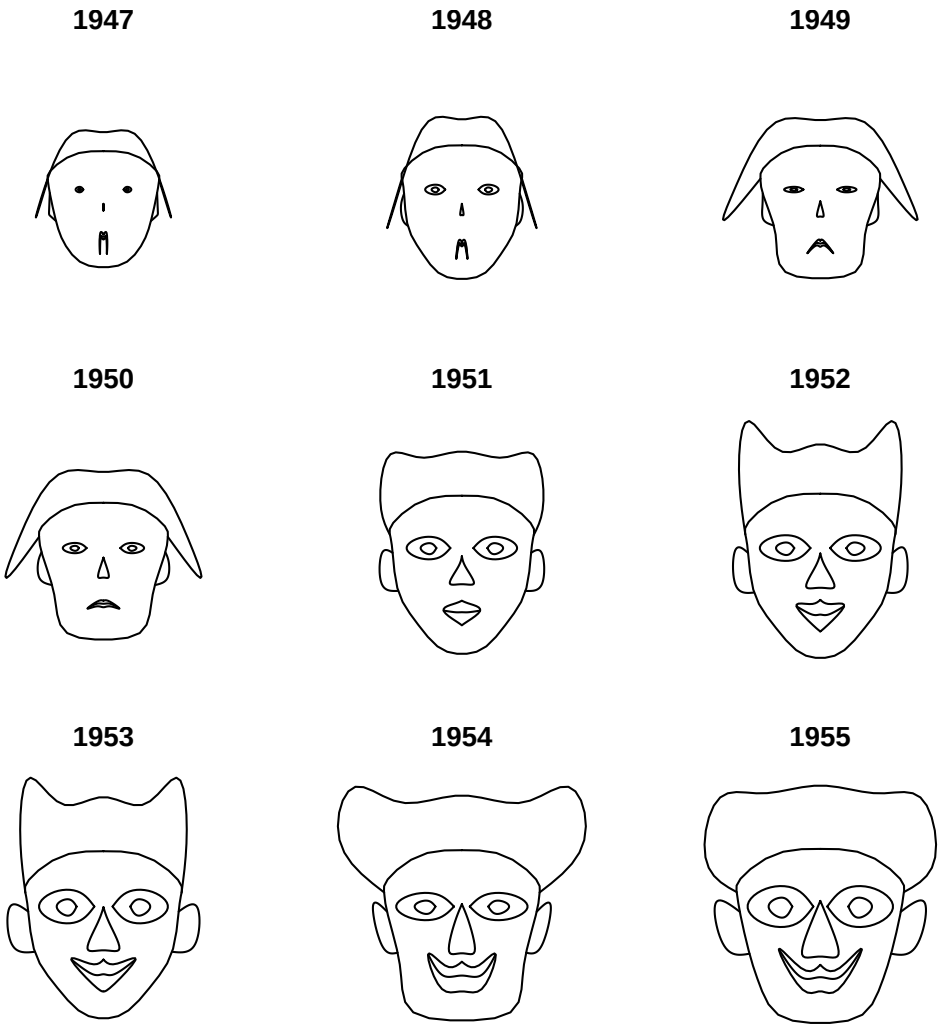
$$f(t) = \frac{x_1}{\sqrt{2}} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \cdots$$

图像演示

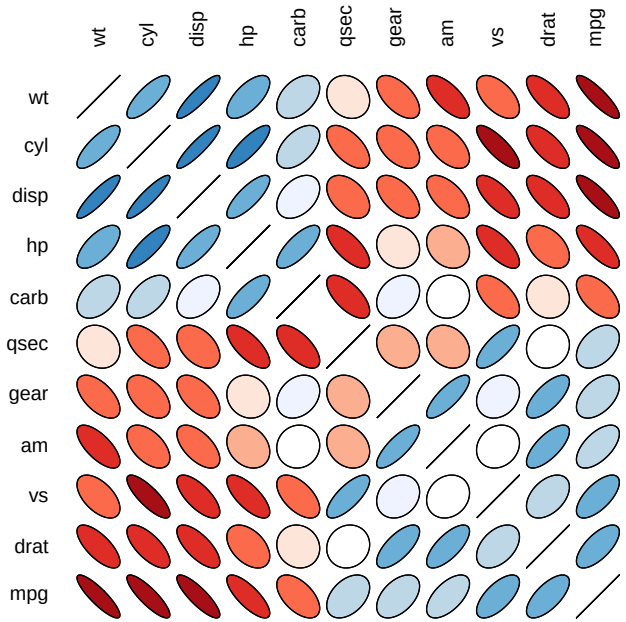

Mazda RX4
Mazda RX4 Wag
Datsun 710
Hornet 4 Drive
Hornet Sportabout
Valiant

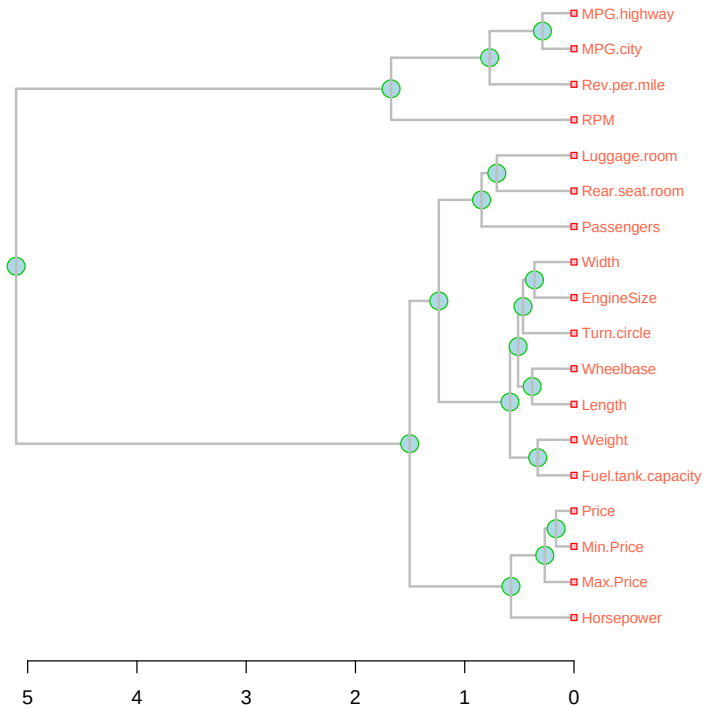

disp
cyl
hp
mpg
drat
qsec
wt

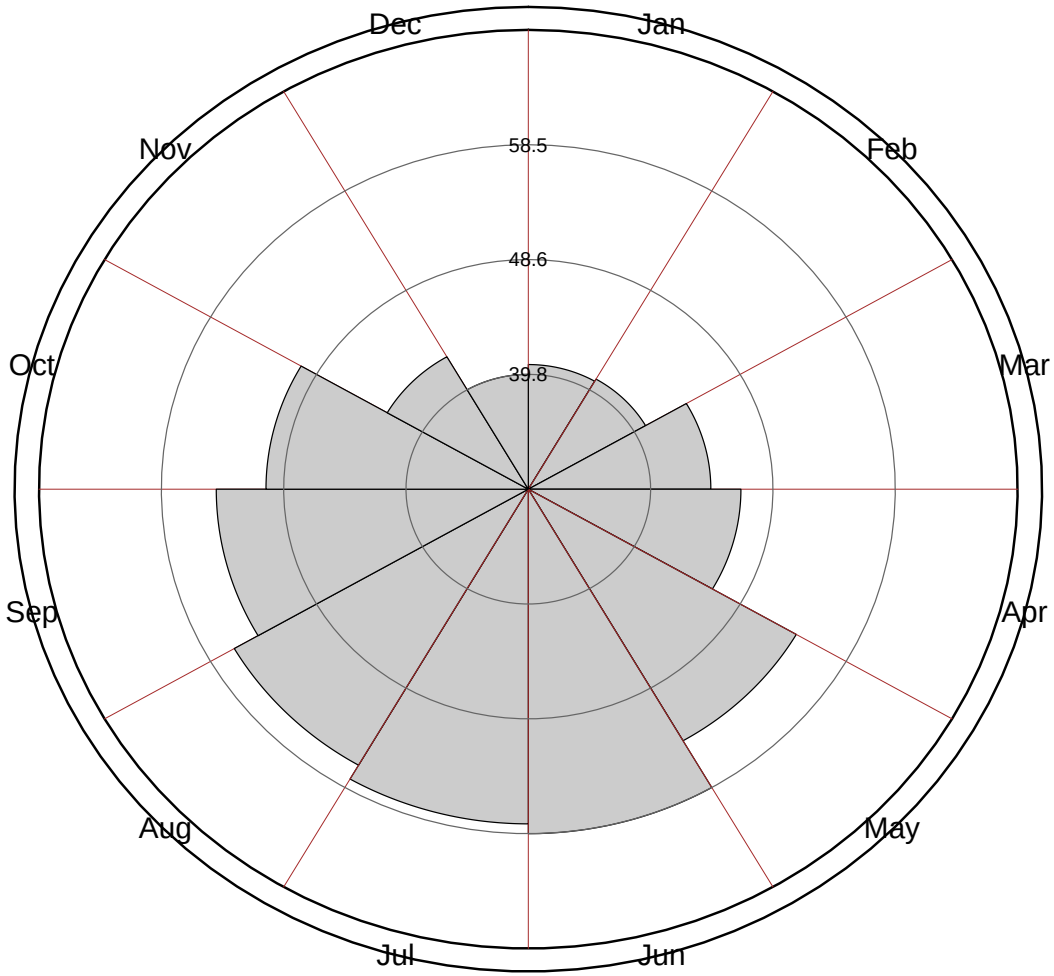

Mazda RX4
Mazda RX4 Wag
Datsun 710
Hornet 4 Drive
Hornet Sportabout
Valiant


1947
1948
1949
1950
1951
1952
1953
1954
1955

相关系数展示


wt
cyl
disp
hp
carb
qsec
gear
am
vs
drat
mpg


MPG.highway
MPG.city
Rev.per.mile
RPM
Luggage.room
Rear.seat.room
Passengers
Width
EngineSize
Turn.circle
Wheelbase
Length
Weight
Fuel.tank.capacity
Price
Min.Price
Max.Price
Horsepower
5
4
3
2
1
0


Dec
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
58.5
48.6
39.8

选取和转化

$$subset(airquality, Temp > 80, select = c(Ozone, Temp))$$

$$transform(airquality, new = -Ozone, Temp = (Temp - 32)/1.8)$$

rescaler 函数

range: scale to [0, 1]

rank: convert values to ranks

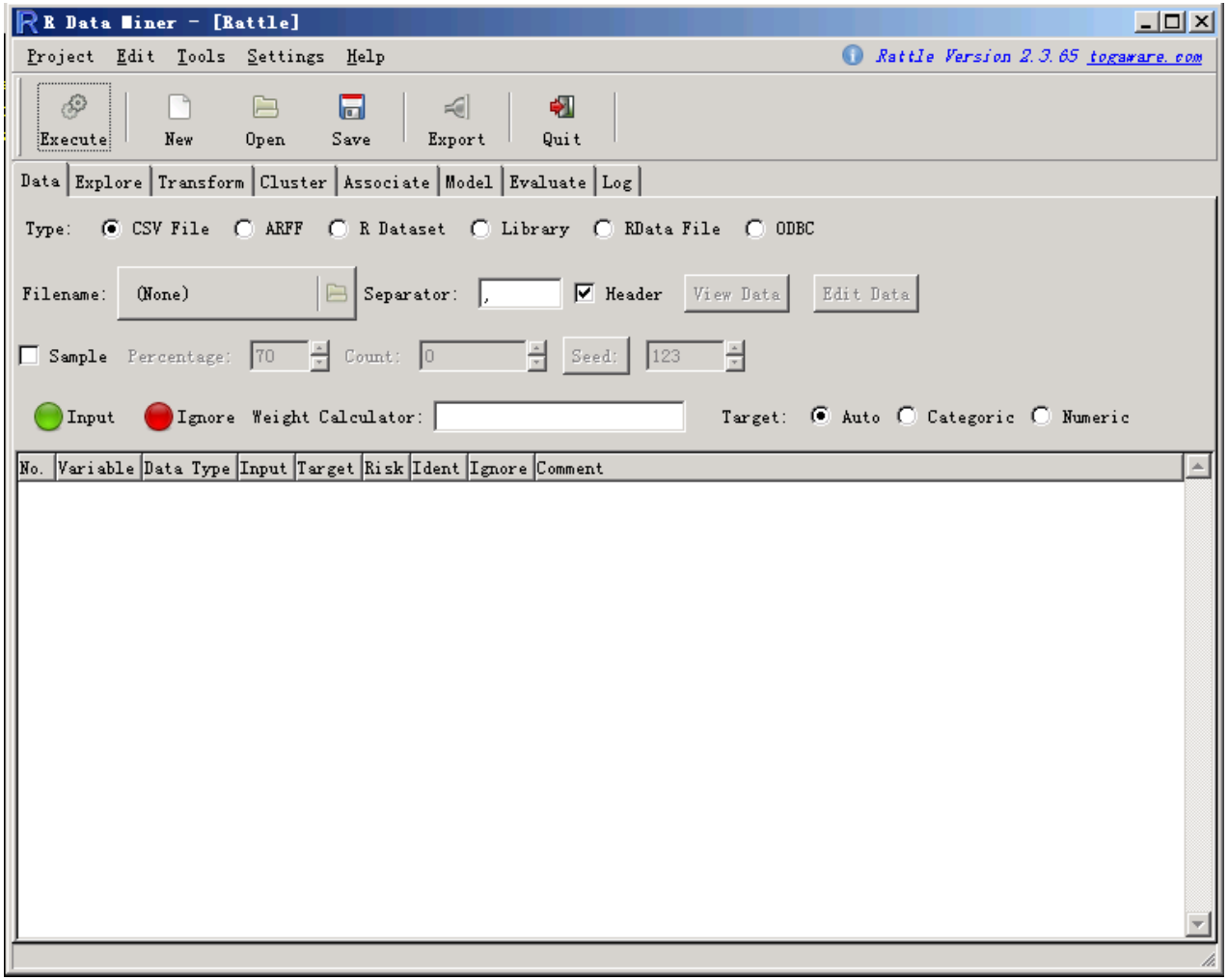
robust: robust version of sd, substract median and divide by median absolute deviation

sd: subtract mean and divide by standard deviation

Part II

R 的扩展算法

Rattle(the R Analytical Tool To Learn Easily)



Rattle 提供了一个基于 Gnome 环境 (需要 RGtk2 包支持) 的数据挖掘界面，通过 Rattle，即使对 R 语言不是很了解的用户也可以通过简单的点击鼠标来读入、转换、探索数据。而且用户可以在它的 log 中了解 Rattle 所使用的 R 语言命令记录。

rattle 的弱点

1. Gnome 环境在 Windows 下需要安装 Gtk+，实践证明，在分析过程中存在不稳定特点。
2. 基于处理大型数据集的应用，R 常采用服务器模式，rattle 的图形界面弱化很多。
3. rattle 提供的算法虽然已达到实际应用水平，但仍然有些包不能调用，略显遗憾。

Weka

Weka 是基于 Java 环境的数据挖掘软件，它的强项主要在分类 (classification) 领域，在这个领域几乎包含了所有的机器学习的方法，而且它还集成了如回归、关联规则、clustering 算法。

RWeka 包提供了 R 同 Weka 之间的接口，通过 **RWeka** 包，R 可以调用所有 Weka 的算法。

	主题	核心函数	对应扩展包
统计模型	主成分分析	princomp	base
	因子分析	factanal	base
	回归模型	lm,nlm,rlm	base,stats,MASS
	Logistic	glm,polr,lrm	base,MASS,Design
	cox 比例模型	coxph,cph	survival,Design
	方差分析	aov,TukeyHSD	stats
	时间序列	ar,arima,garch	stats,tseries

	主题	核心函数	对应扩展包
统计模型	主成分分析	princomp	base
	因子分析	factanal	base
	回归模型	lm,nlm,rlm	base,stats,MASS
	Logistic	glm,polr,lrm	base,MASS,Design
	cox 比例模型	coxph,cph	survival,Design
	方差分析	aov,TukeyHSD	stats
	时间序列	ar,arima,garch	stats,tseries
数据挖掘算法	关联规则	apriori,Tertius	arules,RWeka
	K-methods	kmeans,clara	base,cluster
	朴素贝叶斯	LBR	Rweka
	判别分析	lda,qda,fda	MASS,mda
	决策树	rpart,J48,tree,ctree	rpart,RWeka,tree,party
	随机森林	randomForest	randomForest
	支持向量机	svm,SMO	e1071,RWeka
	层次聚类	hclust,agen	base,cluster
	k 近邻	knn,IBk	class,RWeka
	神经网络	nnet	nnet

关联规则

关联规则 (Association Rules) 是发现交易数据库中不同商品（项）之间的联系，通过这些规则来找出顾客购买行为模式，如购买了某一商品对购买其他商品的影响。发现这样的规则可以应用于商品货架设计、货存安排以及根据购买模式对用户进行分类。

arules 包提供了两种快速的、基于 C 语言实现的挖掘方法 Apriori 和 Eclat。

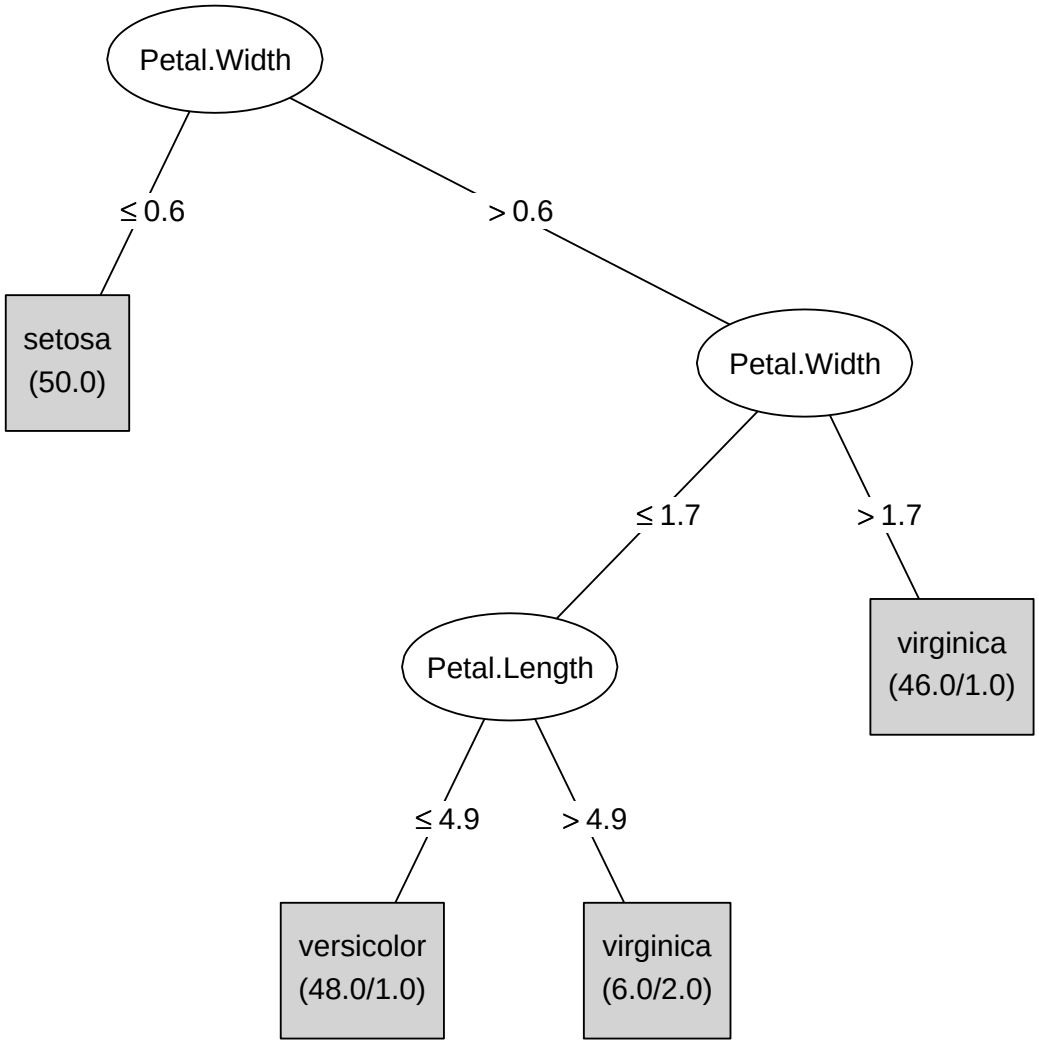
分类 (classification)

分类方法 (classification) 被称为 Tree-based Methods。主要应用有两种情形，一种是当输出 (因变量) 是分类变量 (factor) 时，这时叫做分类树 (Classification trees)，这种情况比较常见；还有一种是当输出变量为连续变量，称为回归树 (Regression trees)。

决策树的优点

- 决策树是非参数的，因此并不需要数据的正态性假设；
- 决策树可以处理不同的数据类型，包括连续型数值变量、分类变量、顺序变量、二元变量。因此可以不需要进行数据变换；
- 在探测变量重要程度、交互效应和离群点方面非常有益，尤其是在模型探索期间；
- 决策树在广泛的领域都有应用。

Clark and Pregibon(1992) 在 S 语言中最早引入 tree-based models，并且可以通过函数 tree 及其支持函数实现，而 Brian Ripley 在 1998 年通过 **tree** 将这个 *tree* 函数 “克隆” 到 R，具有类似功能 (或者更好、更有效) 的包还有 Terry M Therneau and Beth Atkinson 的 **rpart** 包和 Hothorn，Hornik，Zeileis 的 **party**。


Petal.Width
≤ 0.6
> 0.6
setosa
(50.0)
Petal.Width
≤ 1.7
> 1.7
Petal.Length
virginica
(46.0/1.0)
≤ 4.9
> 4.9
versicolor
(48.0/1.0)
virginica
(6.0/2.0)

聚类 (cluster)

聚类 (cluster) 是将数据自动划分为若干个群 (clusters) 的一种算法集。这种算法的结果会使群内的观测变量比较类似，群间的差别比较大。这种方法在不论在统计分析还是机器学习都有很长的历史，它在数据探索的初期意义明显

提供数据的等级划分，群的个数由等级划分过程来决定，所以群的个数可以是从一到样本数的任意值。层次聚类有两种实现方式：

- 凝聚方式：首先每一个单独的样本点都被认为是一个群，然后每个群开始汇聚形成一个更大的群，直到汇聚到一个包含所有样本的最大群。
- 分裂方式：汇聚方式的逆向过程。假设所有的样本点构成一个大群，根据划分方法将其分裂为更小的群，直到每个样本点最后都形成单独的群。

如果事先已经知道群的个数，即需要用户指定群的个数。这种方式开始会有一个初始划分 (随机的)，然后通过叠代，最后达到群内样本差别小，群间差异大的结果。

随机森林

随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。Leo Breiman 和 Adele Cutler 发展出推论出随机森林的算法。而”Random Forests” 是他们的商标。

随机森林的生长方式

1. 假设训练样本的个数为 N，那么采取放回式抽样 (即 bootstrap 抽样) 的抽样集用来生长树；
2. 假设有 M 个变量，对于每一个节点，随机选择 m($m << M$) 个基于此点上的变量。根据这 m 个变量，计算其最佳的分割方式。
3. 每一颗树都尽量生长，而不被修剪。

在 R 中 **randomForest** 包提供了实现随机森林的 *randomForest* 函数。

$$iris.rf < -randomForest(Species \sim ., data = iris,$$
$$importance = T, proximity = T)$$

OOB estimate of error rate: 4.67%

table	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	3
virginica	0	3	47

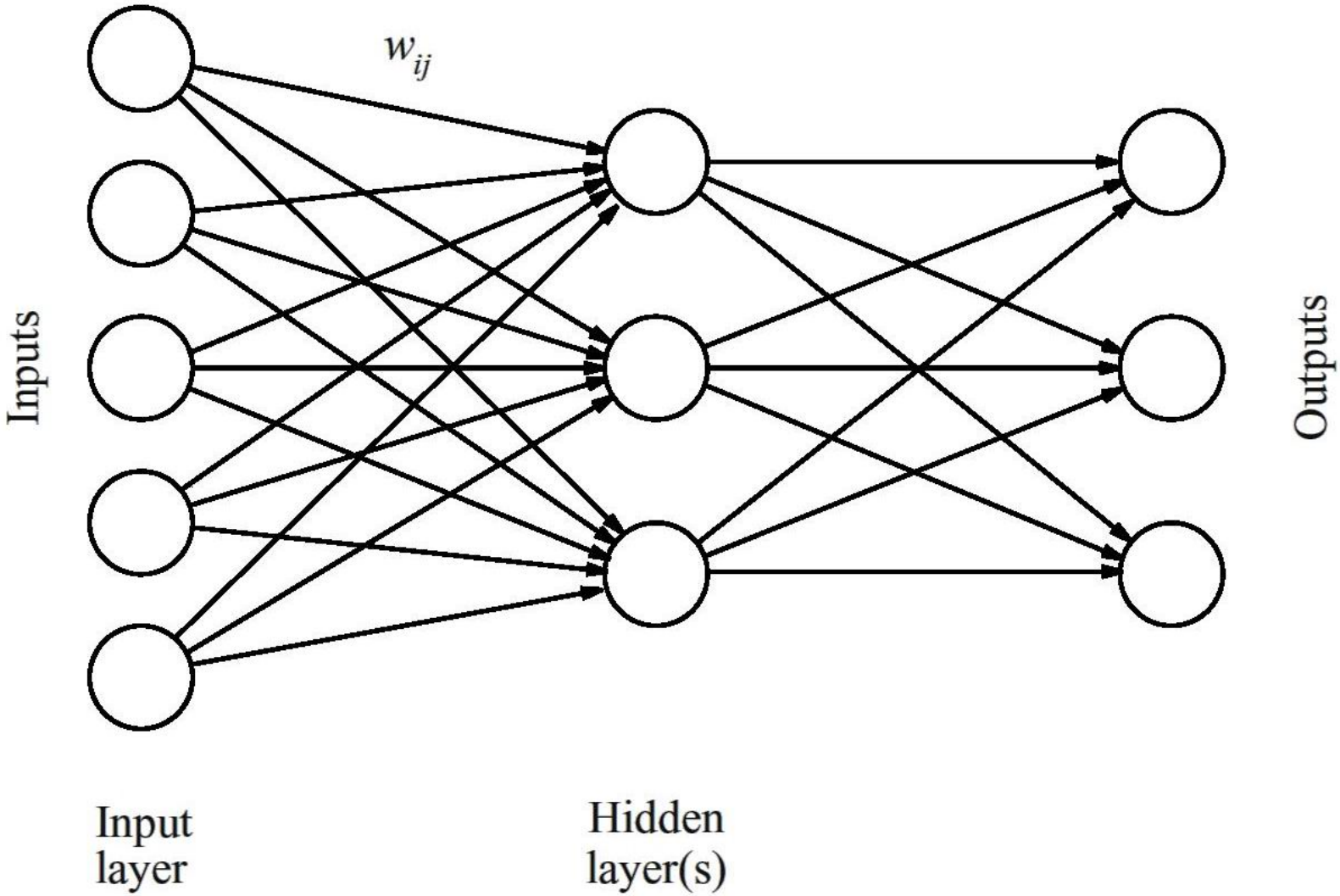
神经网络

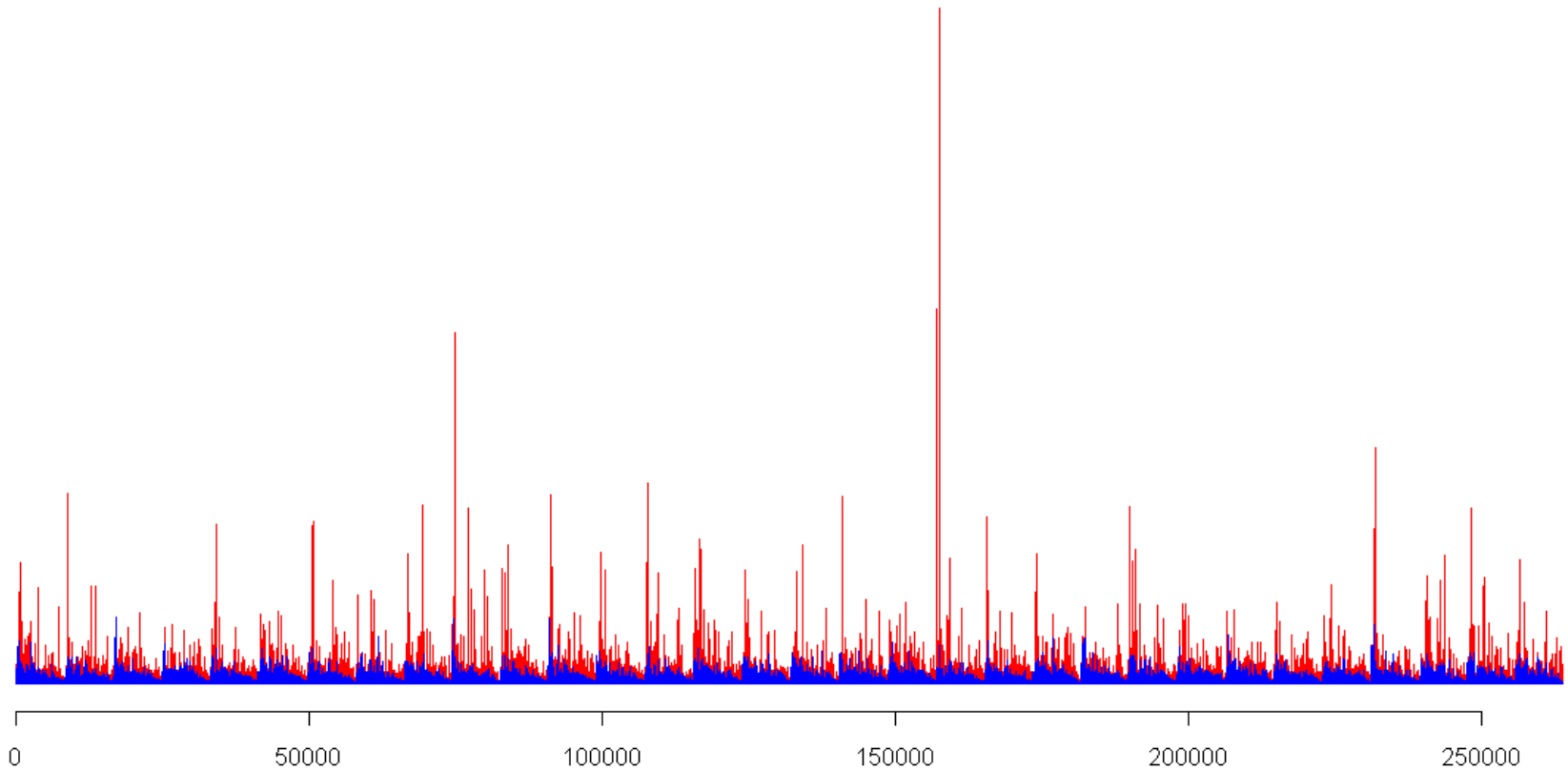
一种常见的多层结构的前馈 (Feedforward) 网络由三部分组成：

输入层 (Input layer) 接受输入信息 (解释变量)。

隐藏层 (Hidden layer) 是输入层和输出层之间众多神经元和链接组成的层面。隐层可以有多层，习惯上一般使用一层。隐层的节点 (神经元) 数目不定，但数目越多神经网络的非线性越显著。

输出层 (Output layer) 输入信息在隐藏层中传输、变换、加权后形成输出结果 (被解释变量)。


w_{ij}
Inputs
Outputs
Input layer
Hidden layer(s)


0
50000
100000
150000
200000
250000

Thanks