

机器学习 第1周

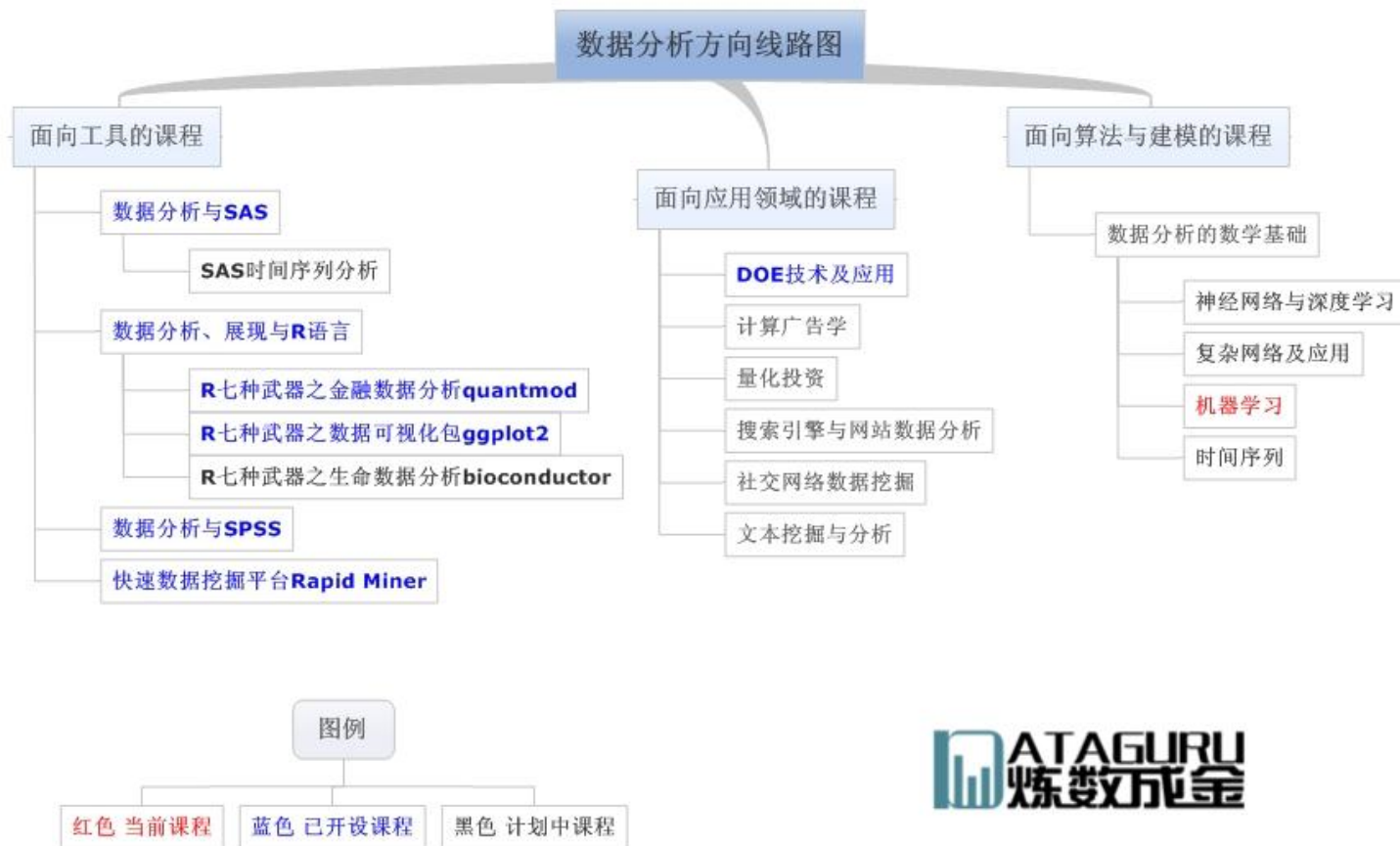
【声明】 本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

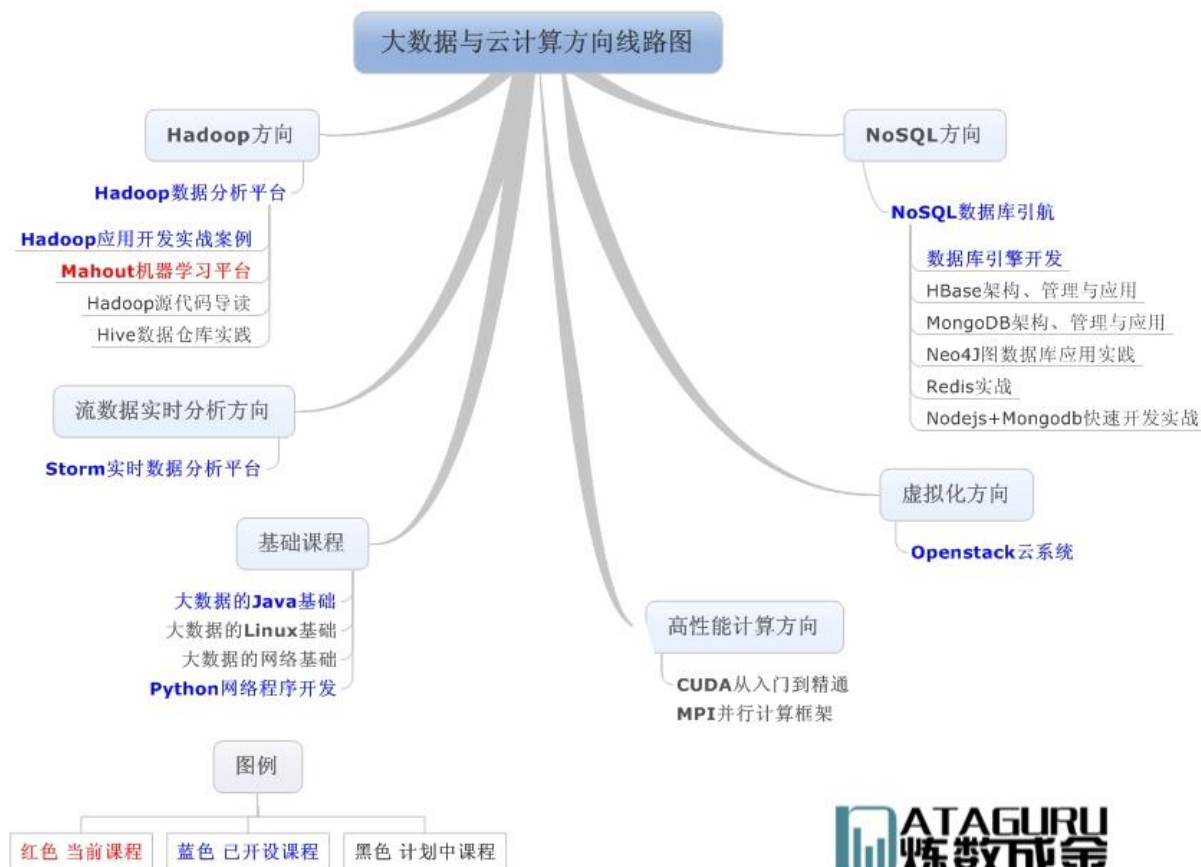
<http://edu.dataguru.cn>

- 机器学习算法为主的课程，结合软件的使用和部分案例
- 课程需要有一定的数学素养，数学是能表达量化关系和算法的唯一语言
- 将机器学习算法用于大数据挖掘，是本课程的主要目标，但也会讲述其它的机器学习覆盖领域
- 参考书大多艰涩，言简意赅，但通过精读即使能掌握部分也能有很大收获。希望学习者可以坚持
- 怎样把复杂的算法用浅显易懂的方式和例子，向非专业人士清晰表达，是本课程的最大挑战
- 课程内容可能会根据授课情况作出调整
- 课程周期视内容难度每1-2周1次授课
- 请大家重视交流，不要留下知识盲点

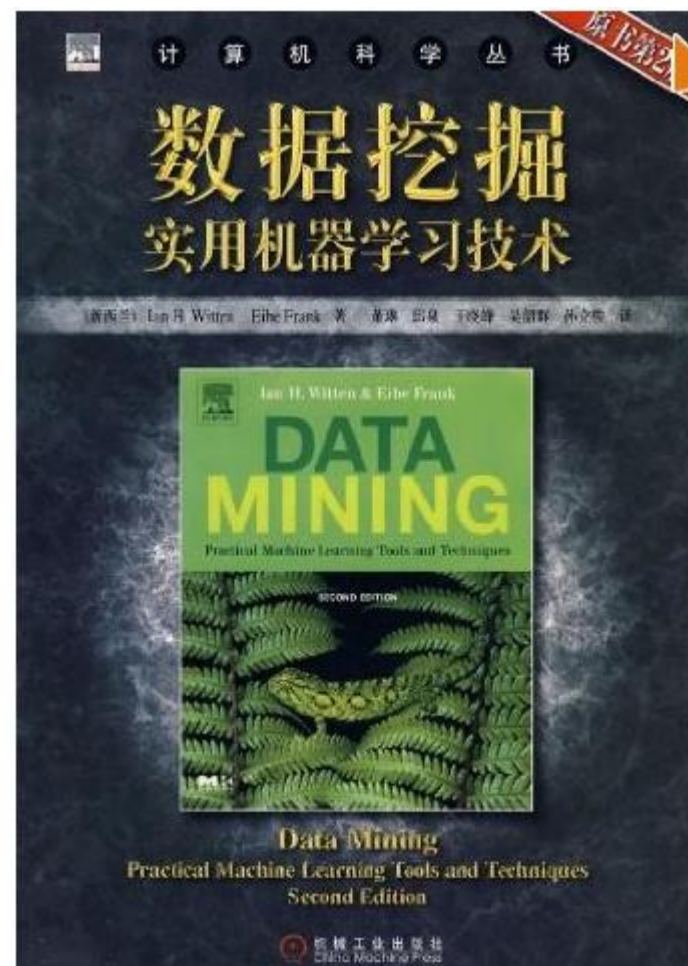
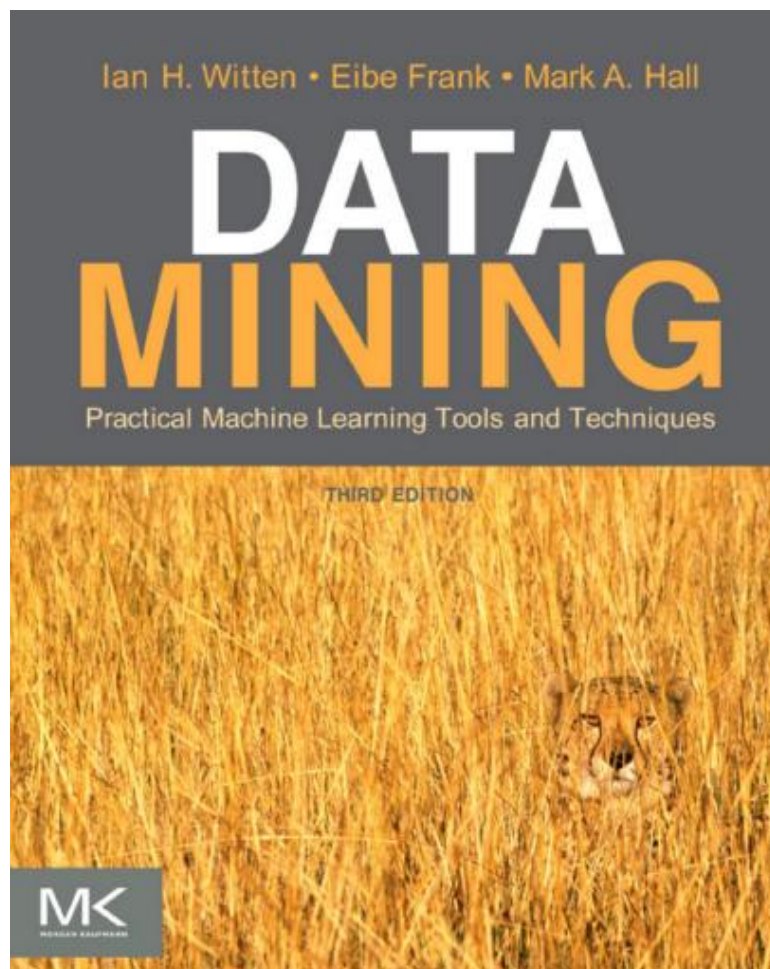
在炼数成金课程线路图中的位置

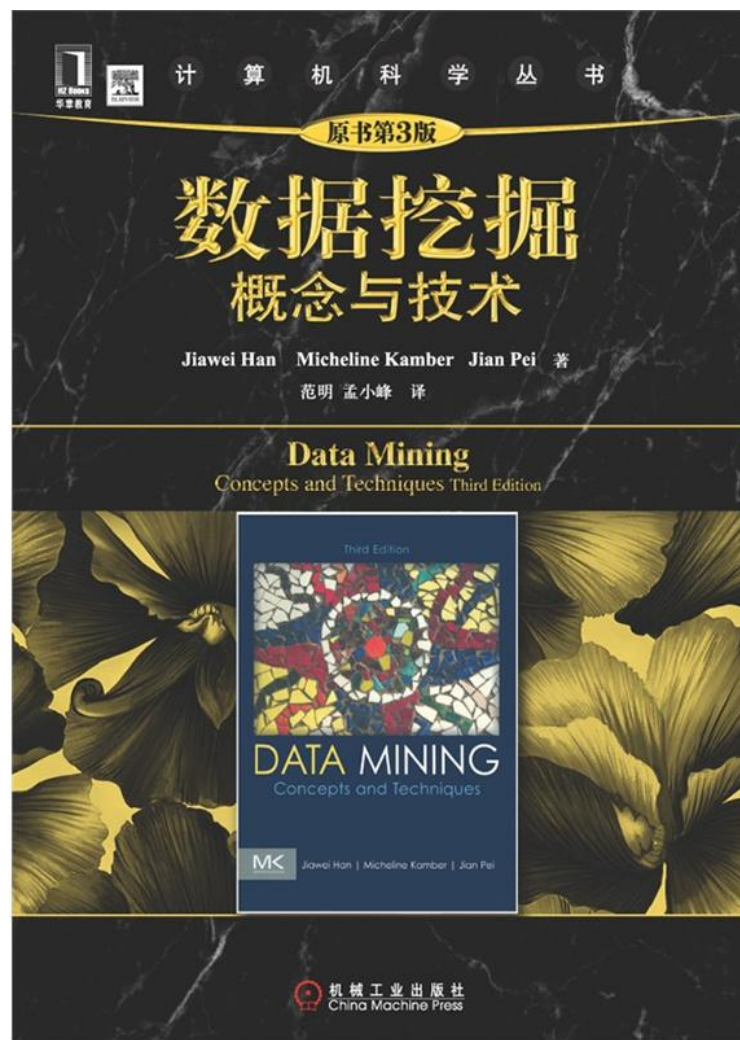


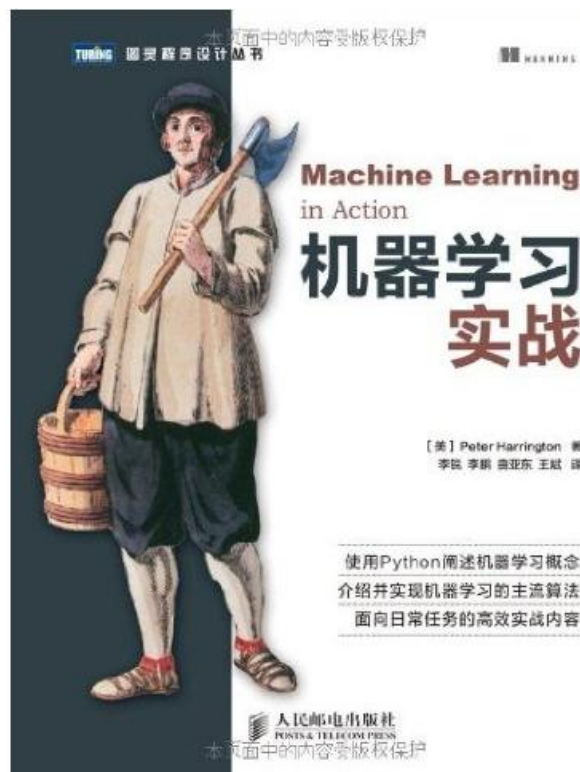
在炼数成金课程线路图中的位置

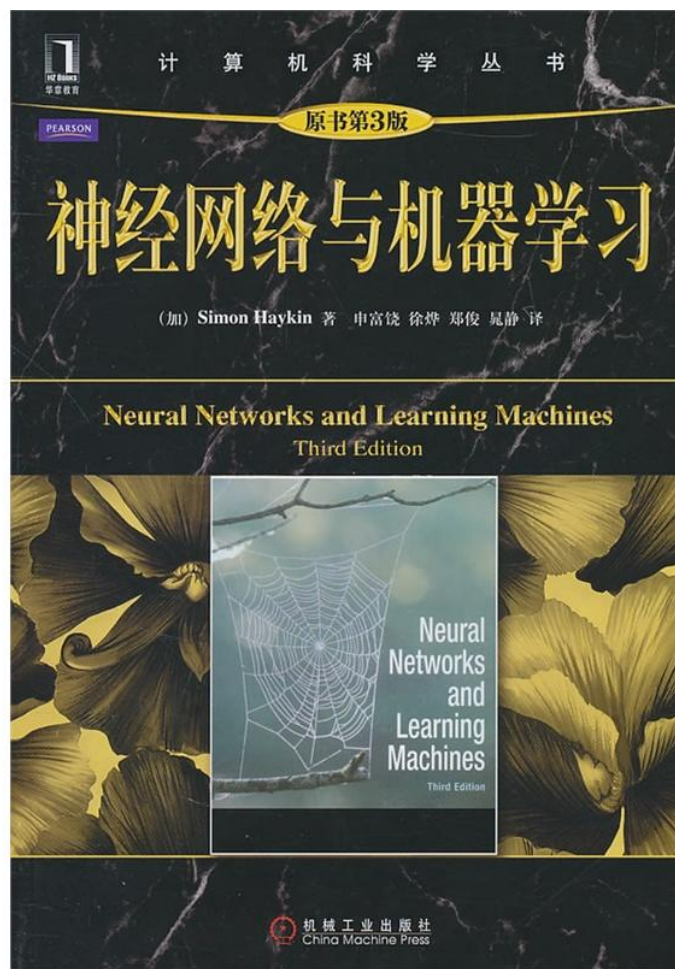


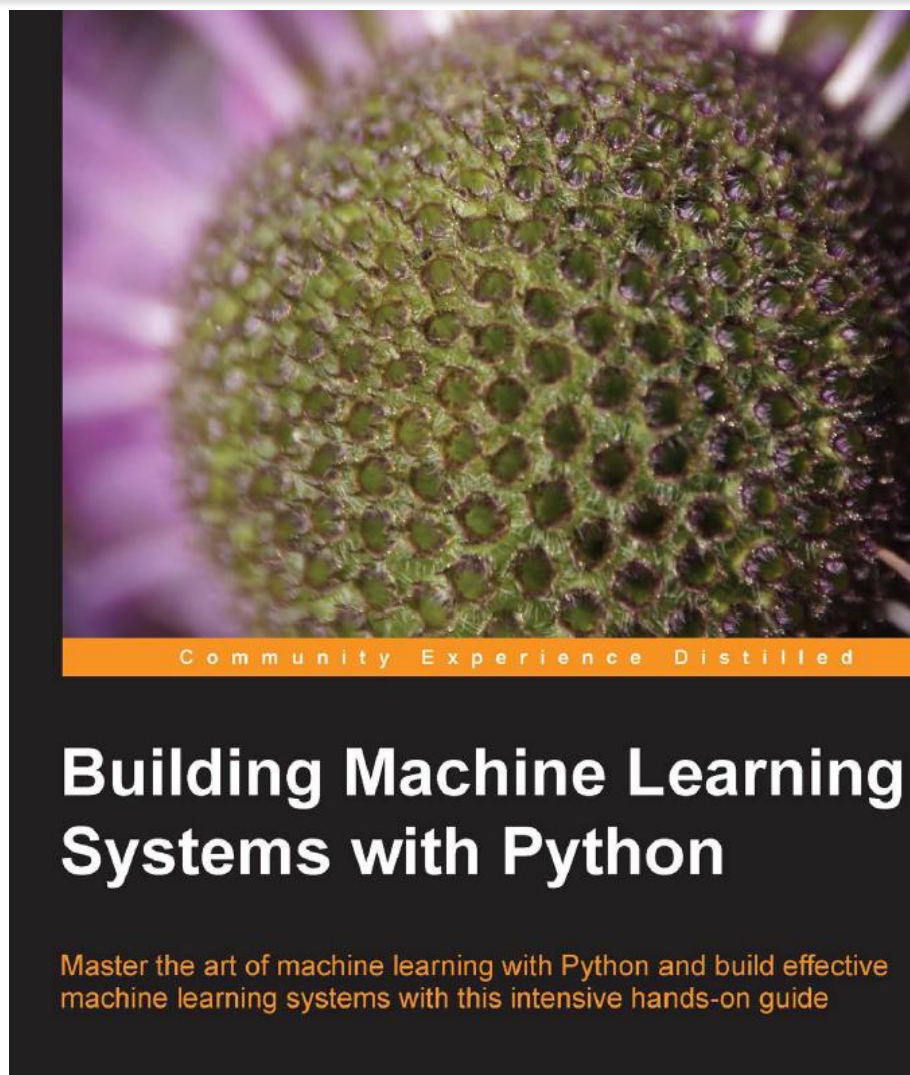
- 熟悉课程里所介绍的各种算法的细节
- 懂得如何使用这些算法去解决实际场景问题
- 熟悉了解常用的机器学习和数据挖掘软件
- 育成目标：数据分析师，算法设计师，具备算法设计能力的高层次程序员











- 机器学习是指是一门多领域交叉学科。专门研究计算机或其它软硬件设备怎样模拟或实现人类的学习行为，以获取新的知识或技能，重新组织已有的知识结构使之不断改善自身的性能。
- 应用机器学习技术到产品中，给用户带来“机器具备人类般高智能”的震撼性体验。
- 人力成本又越来越高，机器学习能降低企业成本，提高投入产出比。
- 第二次机器革命——以具备人类智能为核心价值的机器占主导地位（第一次机器革命——动力系统革命），对国家软实力具有重要作用。
- 机器学习是人工智能研究的核心内容。它的应用已遍及人工智能的各个分支，如专家系统、自动推理、自然语言理解、模式识别、计算机视觉、智能机器人等领域。
- 机器学习在数据挖掘里被大量使用，其技术内涵几乎通用，可以看作同一座山峰在不同视角下的侧影。

- 数据分析和数据挖掘：机器学习实现一套工具、方法或程式，从现实世界的海量数据里提炼出有价值的知识，规则和模式。并把该提炼成果应用到前台系统，辅助业务的进行，使其达到更好的效果，例如推荐，辅助决策（沙盘推演，博弈，预测结果），精准辨别，参与服务等，使到业务能产生更大的效益
- 图像和语音识别：语音输入，OCR，手写输入，通讯监控，车牌识别，指纹识别，虹膜识别，脸像识别
- 智慧机器，机器人：生产线机器人，人机对话，电脑博弈

- 当当网的图书推荐
- 汽车之家同类汽车推荐
- 淘宝的同类商品推荐
- 新浪的视频推荐
- 百度知道的问题推荐
- 社交推荐
- 职位推荐

推荐系统：京东商城

item.jd.com/11224757.html

全部商品分类

首页

服装城

京东超市

团购

夺宝岛

闪购

图书 > 计算机与互联网 > 编程语言与程序设计

电子书 | 音像 | 在线读书 | 团购 | 版权补贴 | 图书榜 | 新书榜 | 特价 | 预售 | 所有图书分类

购买此书的读者还购买了



R语言实战

¥56.20 (7.2折)



图灵程序设计丛书：统计思维：程序员数学之概率统计

¥23.20 (8折)



华章科技：R语言编程艺术

¥47.90 (7折)



R语言经典实例

¥59.30 (7.6折)



Hadoop技术内幕：深入解析MapReduce架构

¥48.10 (7折)



R和Ruby数据分析之旅

¥36.80 (8.2折)



Hadoop技术内幕：深入解析Hadoop Common和

¥66.80 (7.6折)



图灵程序设计丛书：代码的未来

¥65.10 (8.3折)



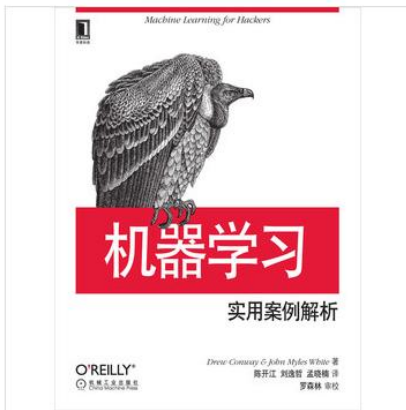
O'Reilly：Hadoop权威指南（第2版）

¥63.80 (7.2折)



利用Python进行数据分析

¥61.40 (6.9折)



分享到：

企业批量购书通道

Click here for international delivery

机器学习：实用案例解析

（机器学习和数据挖掘领域的经典图书，基础理论与实践完美的结合）

定 价：¥69.00

京 东 价：¥50.80 (7.4折) (降价通知)

商品评分：★★★★★(已有210人评价)

配 送 至：北京朝阳区管庄

有货，11:00前完成下单，预计今日（03月22日）送达

作 者：Drew Conway, John Myles White 著 陈开江, 刘逸哲, 孟晓楠 译

出 版 社：机械工业出版社

出版时间：2013-04-01

I S B N：9787111417316

所属分类：图书 > 计算机与互联网 > 编程语言与程序设计
TOP177

购买数量：-

1

+

加入购物车

+ 加关注

51个卖家在售 ¥48.3 起

51个卖家在售

查看全部

恒久图书专营店 ¥48.30

永腾宇辉图书专营店 ¥48.30

阳光图书专营店 ¥48.30

最佳组合



机器学习：实用案例解析

+



图灵程序设计丛书：机器学习实战
557条 (96%好评)
¥48.70 (4折)

+



机器学习/计算机科学丛书
811条 (97%好评)
¥34.20 (4折)

+



图灵程序设计丛书：统计思维：程序员数学
173条 (91%好评)
¥32.20 (4折)

+



计算机科学丛书：数据挖掘与R语言
221条 (97%好评)
¥34.00 (4折)

购买最佳组合

京东价：¥50.80
参考价：¥69.00

购买组合

调查问卷

返回

DATAGURU专业数据分析社区

、氯等常用消毒药都很敏感。

6、若有发热及**呼吸道**症状，应戴上口罩，尽快就诊，并切记告诉医生发病前有无外游或与禽类接触史。

7、一旦患病，应在医生指导下治疗和用药，多休息、多饮水，注意**个人卫生**。

评论(9)

 787


 34




sunny闪电雷霆 | 二级 采纳率50%

擅长：暂未定制


其他类似问题

H7N9禽流感有哪些症状 [百度经验]  23 2013-04-09

h7n9禽流感早期症状是什么样的?  217 2013-04-16

H7N9禽流感的症状是什么?  75 2013-04-17

H7N9禽流感症状是什么?  31 2013-04-03

h7n9禽流感什么症状?  14 2013-04-20

[更多关于H7N9的问题>>](#)

问题分类

手机提问 **NEW**

电脑/网络 >

硬件 常见软件 互联网

生活 >

服装/首饰 美容/塑身 购物

医疗健康 >

内科 妇产科 人体常识

体育/运动 >

足球 篮球 健身

电子数码 >

手机/通讯 照相机/摄像机

商业/理财 >

股票 财务税务 创业投资

教育/科学 >

理工学科 外语学习

社会民生 >

法律 求职就业 时事政治

文化/艺术 >

等待您来回答

更多提问 >

我关注的关键词	我关注的分类	为我推荐的问题
春暖花开....性吧		0回答
10 铁观音的茶叶梗子能泡茶喝吗？对身体好吗？		0回答
穿越火线获得英雄武器黑龙的办法了！！！！ [已失效]		0回答
5 给一个可以测定输入的float类型数据小数位数的多少的...		0回答
在常州市老人机哪卖得好？		0回答
跪求小漠国服第一系列泽拉斯三分钟的时候背景音乐		0回答
手拿包什么牌子好呢？请问		0回答
100 品牌折扣店		0回答
想参加云南14年法检考试，但基础有些差，想报个培训班，...		0回答

贝叶斯分类：判定垃圾邮件

收取 发送 撰写 回复 全部回复 转发 删除 邮件提醒 地址簿 远程管理 中转站

Foxmail

huangzh@139.com

收件箱

反垃圾邮件设置

常规 规则过滤 贝叶斯过滤 黑名单 白名单

在学习邮件前需要整理您的邮件夹，以避免把垃圾邮件作为非垃圾邮件学习或把非垃圾邮件作为垃圾邮件学习。

☐ 使用贝叶斯概率模型判定接收的邮件是否垃圾邮件(U)

已学习信息

非垃圾邮件:	2620	垃圾邮件:	4104
非垃圾词:	1106333	垃圾词:	786541
更新时间:	2014-1-16 下午 11:43:34		

学习(L)... 高级(A)...

过滤强度

移动下面的标记设定过滤的强度。

低 中 高

过滤强度设定越高，邮件被判定为垃圾邮件的可能性越高

☒ 自动删除垃圾邮件箱中以下天数之前的旧邮件

30 天之前

☐ "设定为非垃圾邮件"时不显示提示窗(D)

导入... 导出...

确定 取消

发件人 主题 日期

qingbianji88	来自qingbianji88的邮件	2013年12月12日
	这儿有件事要说，最近请要关注一下	2013年12月12日
	自然会议安排	2013年12月11日
	全国1800家分店,星级优眠大床房77元即可入...	2013年12月11日
	老师，您好	2013年12月7日
	韩编辑	2013年12月6日
	论文翻译: stswzh@mail.sysu.edu.cn	2013年12月6日
	2013 研究生优秀论文展示-Emerald	2013年12月5日
	Reference Form Submitted to UBC Graduate Stu...	2013年12月5日
	特色专业建设项目研讨会	2013年12月5日
	可以 799390	2013年12月4日
	三亚旅游国际学术会议邀请函	2013年12月1日
	您有1篇论文成果。确认成果，提高工作效率和...	2013年11月29日

收获明显吗？
发，货比三家，价格战满天飞！
“询盘质量低”，“成交价格低”，“客户忠

解决这件难题唯有：主动出击，抢先同行联系客户，实现一对一交流！双喜外贸客户搜索与开发系统帮助您主动式24小时搜遍你们产品行业的上万上游目标客户资源，模拟手工一对一智能群发，24小时让目标客户知道贵司及产品。具有搜索速度快，搜索质量高，信息准确率高，开发信到达率高，投入成本低等特点。让你一天联系100个客户变为一天联系上万个高质量目标潜在客户。询盘订单不断！！避开

- 分词
- 贝叶斯公式与贝叶斯分类器

若 B_1, B_2, \dots 为一系列互不相容的事件，且

$$\bigcup_{i=1}^{\infty} B_i = \Omega, \quad P(B_i) > 0, i = 1, 2, \dots$$

则对任一事件 A ，有

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{k=1}^{\infty} P(B_k)P(A|B_k)}, \quad i = 1, 2, \dots$$



- 自动化门户系统（百度新闻，谷歌新闻等）
- 搜索引擎根据用户标签类型推送不同类别的搜索结果



评论自动分析

酒店详情

酒店点评 (3027)

立即预订



luya****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

价格便宜 性价比高 交通便捷 靠近市区 服务不错。[详情]

豪华房

有用(0)



luya****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

价格公道 性价比高 交通便捷 酒店餐厅很好吃 服务也很到位。[详情]

高级房

有用(0)



luya****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

五星级酒店而言 价格便宜 性价比高 交通便捷 服务到位。[详情]

豪华房

有用(0)



1100****
2013-12-23

总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

价格合理, 出行方便[详情]

高级房

有用(0)

酒店回复: 2013-12-24

尊敬的顾客您好, 感谢您入住上海明悦大酒店并对我们酒店做出的肯定, 期待您的下次光临!



300720****
2013-12-23

总评: 3.8 卫生: 5 服务: 5 设施: 3 位置: 2

在携程订购的话给的房间都是最小的。别的还行[详情]

高级单人房

有用(0)

来自: 手机用户

酒店回复: 2013-12-24

尊敬的顾客您好, 感谢您入住上海明悦大酒店并对我们酒店做出的肯定, 期待您的下次光临!



109216****
2013-12-23

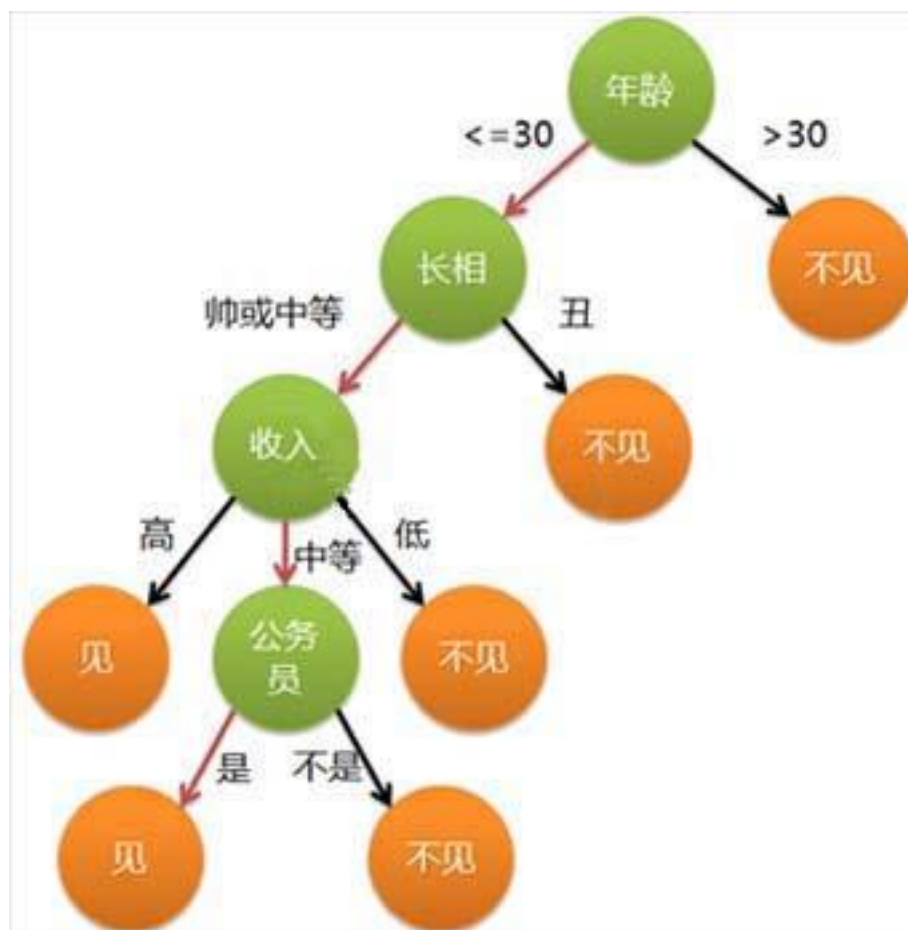
总评: 5.0 卫生: 5 服务: 5 设施: 5 位置: 5

还不错。[详情]

高级单人房

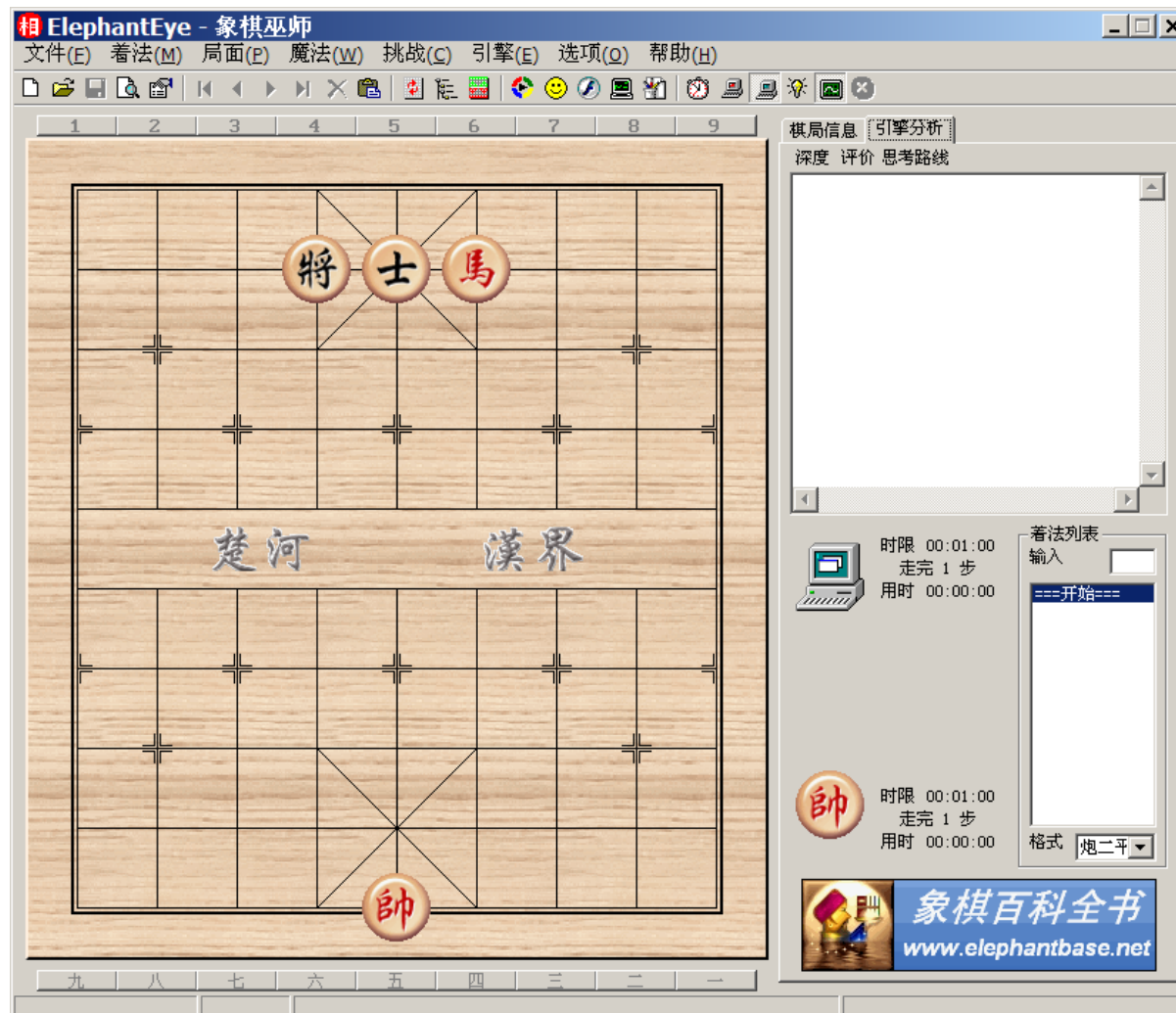
有用(0)

- 给出样本集，学习后输出的产物是一颗决策树



智能博弈：中国象棋云构想

- 局面标准化
- 局面评估函数
- 棋谱学习



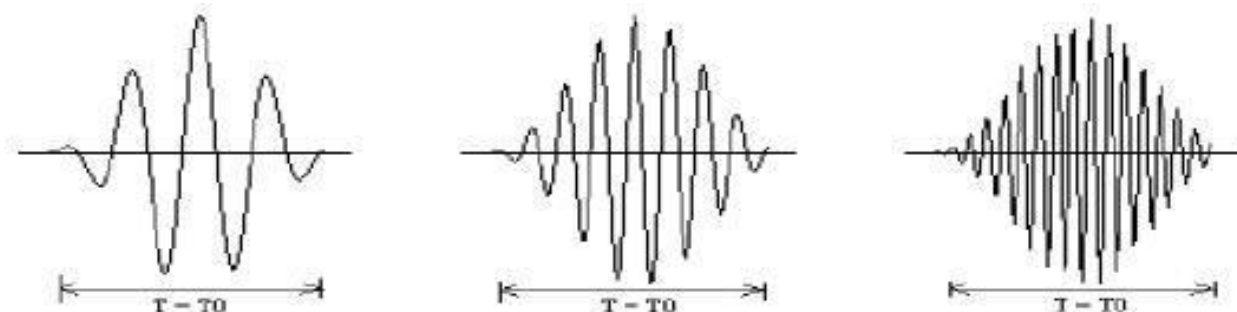


语音识别

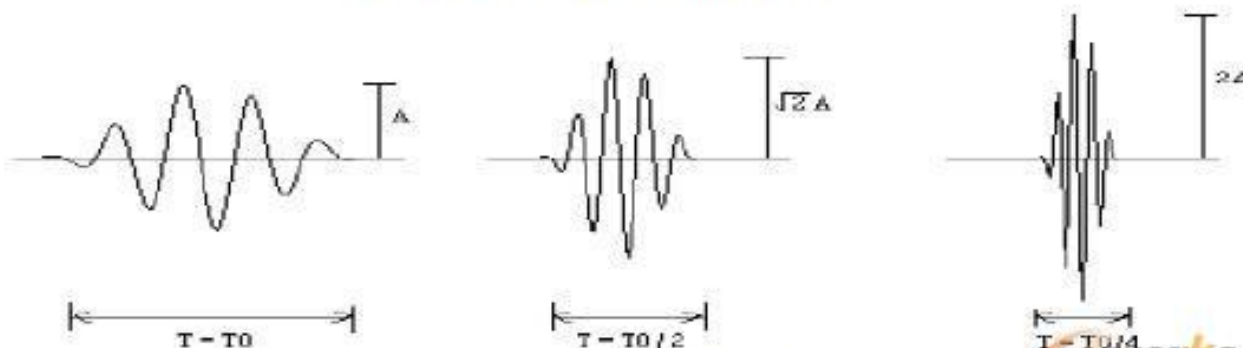
- 语音输入
- 规范化语音：滴滴打车
- 语音属主鉴别



- 指纹、虹膜纹识别
- 脸像识别
- 车牌识别
- 动态图像识别
- **小波分析**



B: 短时傅里叶变换基函数示意图



C: 小波变换基函数示意图

C-works

- R
- Weka
- Matlab
- Python
- 参考：<http://blog.csdn.net/hzxhan/article/details/8548801>

■ R的源起

R是S语言的一种实现。S语言是由 AT&T贝尔实验室开发的一种用来进行数据探索、统计分析、作图的解释型语言。最初S语言的实现版本主要是S-PLUS。S-PLUS是一个商业 软件，它基于S语言，并由MathSoft公司的统计科学部进一步完善。后来Auckland大学的Robert Gentleman 和 Ross Ihaka 及其他志愿人员开发了一个R系统。R的使用与S-PLUS有很多类似之处，两个软件有一定的兼容性。



■ R is free

R是用于统计分析、绘图的语言和操作环境。R是属于GNU系统的一个自由、免费、源代码开放的软件，它是一个用于统计计算和统计制图的优秀工具。

R是一套完整的数据处理、计算和制图软件系统。其功能包括：数据存储和处理系统；数组运算工具（其向量、矩阵运算方面功能尤其强大）；完整连贯的统计分析工具；优秀的统计制图功能；简便而强大的编程语言：可操纵数据的输入和输出，可实现分支、循环，用户可自定义功能。

R是一个免费的自由软件，它有UNIX、LINUX、MacOS和WINDOWS版本，都是可以免费下载和使用的,在那儿可以下载到R的安装程序、各种外挂程序和文档。在R的安装程序中只包含了8个基础模块，其他外在模块可以通过CRAN获得。

R官方网站地址：<http://www.r-project.org>

■ R的特点

1. 有效的数据处理和保存机制。
2. 拥有一整套数组和矩阵的操作运算符。
3. 一系列连贯而又完整的数据分析中间工具。
4. 图形统计可以对数据直接进行分析和显示，可用于多种图形设备。
5. 一种相当完善、简洁和高效的程序设计语言。它包括条件语句、循环语句、用户自定义的递归函数以及输入输出接口。
6. R语言是彻底面向对象的统计编程语言。
7. R语言和其它编程语言、数据库之间有很好的接口。
8. R语言是自由软件，可以放心大胆地使用，但其功能却不比任何其它同类软件差。
9. R语言具有丰富的网上资源



- 商业版本的R

Revolution R (官网 : <http://www.revolutionanalytics.com/>) , 老板是spss的发明者
很多大型厂商也在开始推出自己的R或兼容R的产品 , 例如Oracle、IBM、Sybase

R的CRAN Task View



← → ↺



CRAN

[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R

[R Homepage](#)
[The R Journal](#)

Software

[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

Documentation

[Manuals](#)
[FAQs](#)
[Contributed](#)

CRAN Task View: Machine Learning & Statistical Learning

Maintainer: Torsten Hothorn

Contact: Torsten.Hothorn@R-project.org

Version: 2014-03-07

Several add-on packages implement ideas and methods developed at the borderline between computer science and statistics - this field of research is usually referred to as machine learning. The packages can be roughly structured into the following topics:

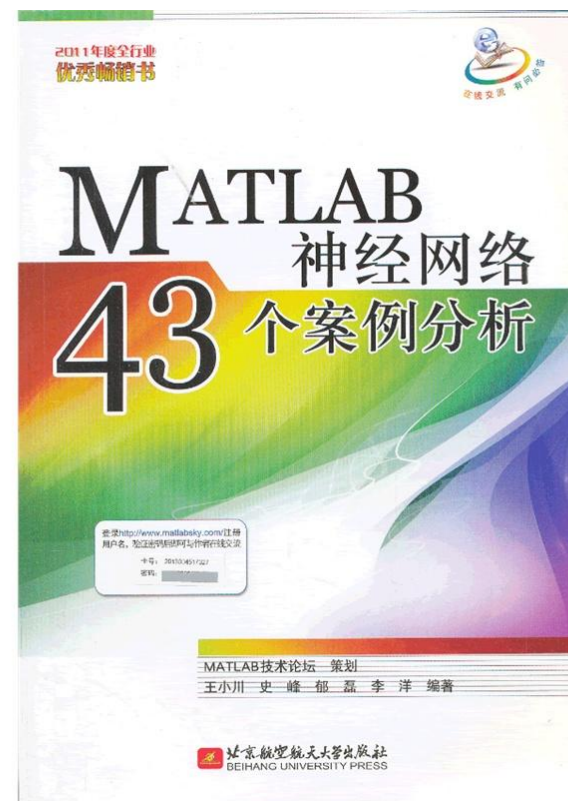
- **Neural Networks**: Single-hidden-layer neural network are implemented in package [nnet](#) (shipped with base R). Package [RSNNs](#) offers an interface to the Stuttgart Neural Network Simulator (SNNS).
- **Recursive Partitioning**: Tree-structured models for regression, classification and survival analysis, following the ideas in the CART book, are implemented in [rpart](#) (shipped with base R) and [tree](#). Package [rpart](#) is recommended for computing CART-like trees. A rich toolbox of partitioning algorithms is available in [Weka](#), package [RWeka](#) provides an interface to this implementation, including the J4.8-variant of C4.5 and M5. The [Cubist](#) package fits rule-based models (similar to trees) with linear regression models in the terminal leaves, instance-based corrections and boosting. The [C50](#) package can fit C5.0 classification trees, rule-based models, and boosted versions of these.
Two recursive partitioning algorithms with unbiased variable selection and statistical stopping criterion are implemented in package [party](#). Function `ctree()` is based on non-parametrical conditional inference procedures for testing independence between response and each input variable whereas `mob()` can be used to partition parametric models. Extensible tools for visualizing binary trees and node distributions of the response are available in package [party](#) as well.
An adaptation of [rpart](#) for multivariate responses is available in package [mvpart](#). For problems with binary input variables the package [LogicReg](#) implements logic regression. Graphical tools for the visualization of trees are available in package [maptree](#).
Trees for modelling longitudinal data by means of random effects is offered by package [REEMtree](#). Partitioning of mixture models is performed by [RPMML](#).
Computational infrastructure for representing trees and unified methods for prediction and visualization is implemented in [partykit](#). This infrastructure is used by package [evtree](#) to implement evolutionary learning of globally optimal trees. Oblique trees are available in package [oblique.tree](#).
- **Random Forests**: The reference implementation of the random forest algorithm for regression and classification is available in package [randomForest](#). Package [ipred](#) has bagging for regression, classification and survival analysis as well as bundling, a combination of multiple models via ensemble learning. In addition, a random forest variant for response variables measured at arbitrary scales based on conditional inference trees is implemented in package [party.randomSurvivalForest](#) offers a random forest algorithm for censored data. Quantile regression forests [quantregForest](#) allow to regress quantiles of a numeric response on exploratory variables via a random forest approach. The [varSelRF](#) and [Boruta](#) packages focus on variable selection by means for random forest algorithms. For large data sets, package [bigrf](#) computes random forests in parallel and uses large memory objects to store the data.
- **Regularized and Shrinkage Methods**: Regression models with some constraint on the parameter estimates can be fitted with the [lasso2](#) and [lars](#) packages. Lasso with simultaneous updates for groups of parameters (groupwise lasso) is available in package [grplasso](#); the [grpreg](#) package implements a number of other group penalization models, such as group MCP and group SCAD. The L1 regularization path for generalized linear models and Cox models can be obtained from functions available in package [glmnet](#), the entire lasso or elastic-net regularization path (also in [elasticnet](#)) for linear regression, logistic and multinomial regression models can be obtained from package [glmnet](#). The [penalized](#) package provides an alternative implementation of lasso (L1) and ridge (L2) penalized regression models (both GLM and Cox models). Package [RXshrink](#) can be used to identify and display TRACES for a specified shrinkage path and to determine the appropriate extent of shrinkage. Semiparametric additive hazards models under lasso penalties are offered by package [ahaz](#). A generalisation of the Lasso shrinkage technique for linear regression is called relaxed lasso and is available in package [relaxo](#). Fisher's LDA projection with an optional LASSO penalty to produce sparse solutions is implemented in package [penalizedLDA](#). The shrunken centroids classifier and utilities for gene expression analyses are implemented in package [pamr](#). An implementation of multivariate adaptive regression splines is available in package [earth](#). Variable selection through clone selection in SVMs in penalized models (SCAD or L1 penalties) is implemented in package [penalizedSVM](#). Various forms of penalized discriminant analysis are implemented in packages [hda](#), [rda](#), [sda](#), and [SDDA](#). Package [Liblinear](#) offers an interface to the LIBLINEAR library. The [ncvreg](#) package fits linear and logistic regression models under the the SCAD and MCP regression penalties using a coordinate descent algorithm. High-throughput ridge regression (i.e., penalization with many predictor variables) and heteroskedastic effects models are the focus of the [bigRR](#)

DATAGURU专业数据分析社区

- Guido van Rossumzai 1989年创立了Python
- I wrote python!
- Python语言的特点
- NumPy
- SciPy <http://scipy.org/install.html>
- Matplotlib <http://matplotlib.org/>



- MATLAB=matrix+laboratory，是由美国mathworks公司发布的主要面对科学计算、可视化以及交互式程序设计的高科技计算环境。
- MATLAB和Mathematica、Maple并称为三大数学软件。它在数学类科技应用软件中在数值计算方面首屈一指。MATLAB可以进行矩阵运算、绘制函数和数据、实现算法、创建用户界面、连接其他编程语言的程序等，主要应用于工程计算、控制设计、信号处理与通讯、图像处理、信号检测、金融建模设计与分析等领域。
- 具有功能完备强大的神经网络包



www.mathworks.cn/products/matlab/whatsnew.html?s_tid=main_release_ML_rp

☆ Google



Accelerating the pace of engineering and science

中国 | 联系我们 | 如何购买

创建帐户 | 登录

产品和服务 解决方案 教育 支持 用户中心 活动 公司

产品和服务 > MATLAB > 新特性

MATLAB

弹出命令历史记录

使用弹出命令历史记录查看，筛选和搜索最近使用的命令。

观看视频

```
clc
plot(f, 2*abs(Y(1:NFFT/2+1)))
Fs = 1000;
T = 1/Fs;
t = (0:L-1)*T;
L = 1000;
clc
plot(f, 2*abs(Y(1:NFFT/2+1)))
fx >> plot(f, 2*abs(Y(1:NFFT/2+1)))
```

马上下载 R2014a

续订软件维护服务

试用软件

在线购买

» See release highlights for all products

- WEKA=Waikato Environment for Knowledge Analysis
- 免费的，非商业化的，基于JAVA环境下开源的机器学习以及数据挖掘软件。
- Weka的主要开发者来自新西兰的Waikato大学。
- 官网：<http://www.cs.waikato.ac.nz/ml/weka/>
- Petaho：<http://community.pentaho.com/projects/data-mining/>



- 回归预测及相应的降维技术：线性回归，Logistic回归，主成分分析，因子分析，岭回归，LASSO
- 分类器：决策树，朴素贝叶斯，贝叶斯信念网络，支持向量机，提升分类器准确率的Adaboost和随机森林算法
- 聚类与孤立点判别
- 人工神经网络

- **Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。**
- **关于逆向收费式网络的详情，请看我们的培训网站 <http://edu.dataguru.cn>**



Thanks

FAQ时间