北京数博思达信息科技有限公司
Email: contact@supstat.com
http://supstat.com

R 语言培训课程基础模块

我们在 R 语言方面有很强的优势，且今年将引进美国著名 R 语言数据分析公司 Revolution Analytics 和 RStudio 的成熟课程体系，以下是今年我们将推出的部分课程模块：

培训模块	主要内容
R 语言简介	R 语言编程基础、数据处理、作图和数据分析
用 R 语言做数据可视化	学习用 R 的几种作图包：基础包、lattice、ggplot2，以及中级的包：iPlots、googleVis 和 RGGobi
R 软件及应用统计方法	语言入门、用 R 编程、用 R 绘图。基础统计、经典回归分析、现代回归和分类、列联表分析、关联规则分析、广义线性模型、多层模型、通过顾客满意度模型引入路径模型及其 PLS 偏最小二乘方法和 SEM 模型、时间序列简介及状态空间模型、多元分析
R 语言数据挖掘介绍	聚焦于 R 在数据挖掘领域的应用，内容包括以 caret 包为框架，介绍分类回归算法和聚类算法。
R 语言高级编程	R 是种面向对象的编程语言，本课程强调良好的编程实践，设计和开发清晰、简洁、有效的代码。内容涉及 S3、S4 对象，函数式编程，以及 R 包的编写。
R 与自动化报告	使用 R 里的 knitr 包，实现可重复研究，自动化报告
R 与最优化	介绍数学规划以及最优化的几个经典理论及其模型，并用 lpSolve、TSP、Rsolnp、Rglpk 等包分析实际的工程数据，帮助企业更好的进行最优化管理
R 与 web 服务	把 R 的应用程序部署在服务器，用客户端通过 web 服务的 API 连接服务器 Rserver/opencpu
R 在量化投资中的应用	介绍量化投资的理论及其模型，并运用 R 高效率编程实现。
使用 R 和 RHadoop 做大数据分析	在 ubuntu 系统下配置 RHadoop 的相关知识以及利用 RHadoop 完成数据分析工作
利用 R 语言进行文本挖掘	介绍 HMM 模型原理及其应用，以及利用 Rwordseg、 tm、rmmseg4j 包进行文本挖掘，
R 语言与数据库管理	使用 R 里面 mysql 包，实现 R 语言与数据库的对接，并进行相应数据操作以及进一步的数据挖掘
R 语言与高性能计算	介绍在 R 中如何进行内存管理，提高计算效率，并分别并行计算给出常见 R 包在工作实践中的应用，实现高性能统计分析。

除以上课程外，我们可能还会根据参加者的反馈与需求增加入门、进阶课程，或按模块、行业划分的课程。欢迎来电咨询，或联系 training@supstat.com

北京数博思达信息科技有限公司
Email: contact@supstat.com
http://supstat.com

R 语言专题培训

主题一：R 语言金融数据可视化与金融建模

本课程主要介绍应用于量化金融方面的 R 包——quantmod，顾名思义，quantmod 就是提供给宽客们建模所用的工具。Quantmod 具有强大的数据抓取，绘画专业行情分析图表以及各种技术指标计算等功能，常常只要几行函数就能完成从数据获取和处理到画图的复杂功能，其工作效率之高让行家里手都觉得瞠目结舌。学习者可以通过学习本课程，了解金融数据分析的基本问题和目标，步入这个金领云集的高收入领域，也为后续的金融建模课程奠下基础，成为变身宽客的敲门砖。炒股爱好者学习 quantmod，可以应用于股票数据处理，帮助快速判断行情状态，以及提出自己独到的模型，无论您身为何种角色，都可以从中获益。

课程大纲

1. 数据获取

1.1 获取网络开放数据
1.1.1 股票数据
- 日交易数据
- 股息数据
- 财务报表数据

1.1.2 重金属价格数据
1.1.3 汇率市场数据
1.2 获取数据库数据
1.2.1 小型数据文件的读取：
- 文本
- .csv
- excel

1.2.2 MySQL、redis 等数据库的读取
- MySQL
- redis

2. 基本数据操作：清洗、转换和抽取

2.1 数据判断
- is 族函数
- has 族函数

2.2 数据提取
- 提取列数据
- 提取子集

2.3 数据转换
- 滞后
- 收益率计算
- 日期转换

2.4 其他操作
- 峰谷查询
- 阈值过滤

3. 金融数据可视化

3.1 基本绘图函数
- chartSeries 详解
- rechart

3.2 基本图形
- 蜡烛图
- 柱状图
- 线图

3.3 技术分析图
3.4 图形修饰、缩放和存储
- 图形修饰
- 图形缩放
- 图形存储

4. 金融建模

4.1 建模准备
4.2 设定模型
- 线型模型
- 支持向量机
- 随机森林

4.3 估计模型参数
4.4 模型的诊断、应用与解释

5. 未来的改进措施

5.1 现有缺点
5.2 改进与发展

北京数博思达信息科技有限公司
Email: contact@supstat.com
http://supstat.com

主题二：R 语言基础与数据挖掘

第一天：R 语言基础部分

1.1 基础语法入门 9:00-11:30	* 什么是 R * 如何学习 R * 如何得到帮助，相关资源和书籍 * RStudio，扩展包，工作空间 * 数据对象 * 向量化操作 * 函数和控制语句
摘要：讲解 R 的基本操作知识，了解 R 的特点、资源获取方式，并掌握基本的程序编写	
案例和练习：使用 R 语言完成若干欧拉项目(Euler Project)	
1.2 数据获取 13:00-14:00	* 本地文档的读写 * 连接数据库 * Web 数据抓取 * API 数据源 * 其它数据资源
摘要：R 语言如何从各种方式读取数据，通过基本的 WEB 知识进行网页抓取，连接数据库，通过 sql 语句调用数据，从本地读取 excel 等各种文件数据。	
案例和练习：使用网页抓取和 API 获得豆瓣网站上的数据。	
1.3 数据整理 14:00-15:30	* 数据变换 * 数据重塑 * 拆分合并 * 汇总数据 * 字符串操作 * 日期操作
摘要：讲解 R 语言如何操作数据，能对数据进行自由的操作转换，特别是对于字符串和日期的操作处理。	
案例和练习：分析 NBA 数据，判断金融市场中的一月效应，找出Q Q群中的话唠	
1.4 数据可视化 15:30-17:00	* R 语言中的可视化函数 * 单变量的特征 * 比例的构成 * 多变量的关系 * 展现时间的变化 * 地理信息
摘要：讲解 R 语言中基础绘图函数和高级绘图包的使用，理解可视化探索的各种方法	
案例和练习：使用图形来对之前的 NBA 数据进行分析	

北京数博思达信息科技有限公司
Email: contact@supstat.com
http://supstat.com

第二：R 数据挖掘之一

2.1 数据挖掘工业流程 9:00-10:00	
摘要：讲解数据挖掘的主要任务，各挖掘方法的简要介绍和应用场景，典型的工作流步骤，以及 R 和数据挖掘之间的关系。	
2.2 先修统计知识 10:00-11:30	* 描述统计 * 推断统计 * 统计模拟计算
摘要：讲解必要的统计学知识，了解各种概率分布、统计描述函数，以及学习假设检验的原理和应用	
案例和练习：报摊进货决策	
2.3 挖掘预处理 13:00-14:00	* 数据的类型 * 数据的可能问题 * 数据预处理的方法
摘要：讲解分析前的数据处理工作，包括如何识别数据中可能的问题，如数据缺失和数据噪声，如何解决这些问题。	
2.4 线性回归 14:00-16:00	* 简单线性回归 * 回归诊断 * 多元回归 * 非参数回归
摘要：讲解最经典的回归分析方法，即对连续数值进行预测，学习其基本原理和前提假设，以及如何在 R 中实施回归。	
案例和练习：葡萄酒的定价分析	
2.5 logistic 回归 16:00-17:00	* logstic 回归原理 * 实现原理 * 输出解释
摘要：讲解 logistic 回归，即对离散的类别数据进行预测，了解系数解释和模型的意义	
案例和练习：是否能升入大学的数据预测	

第三天：R 数据挖掘之二

3.1 **模型评价 9:00-10:30**	* 回归模型评价 * 分类模型评价 * underfit 和 overfit
摘要：讲解如何对分类模型和回归模型进行评价，学习过度拟合与拟合不足的区别，并如何规避这些陷阱	
3.2 K 近邻方法 10:30-11:30	* 最近邻分类器原理 * 各种距离的定义 * 使用 KNN 对乳腺癌数据分析
摘要：学习 KNN 分类方法的原理和实现过程，它的优点缺点，以及如何在 R 中的函数应用	
3.3 决策树方法和集成学习 13:00-14:00	* 理解决策树算法 * 决策树和回归的结合 * 随机森林 * 提升算法
摘要：学习决策树分类器，以及在其基础上的装袋算法、提升算法、随机森林	
3.4 K 均值聚类 14:00-15:00	* K-means 聚类 * 改进算法 * K 的选择
摘要：学习用 kmeans 来进行无监督学习，了解其原理和改进方法，并了掌握如何选择关键参数	
案例和练习：对青少年群体进行分群	
3.4 推荐算法 15:00-17:00	* 协同过滤算法简介 * 奇异值分解
摘要：介绍基本的协同过滤方法，并使用奇异值分解来避免高维问题	

第四天：文本挖掘与大数据

4.1 文本挖掘入门 09:00-10:30	* 文本挖掘的基础术语 * 朴素贝叶斯方法 * 文本挖掘案例：垃圾邮件过滤、红楼梦文本分析、航空公司满意度分析
摘要：介绍基于 R 环境的文本挖掘入门，如何处理将文本这种非结构化数据转为结构化数据，并使用 bayes 进行分析。	
4.2 突破 R 内存瓶颈的若干技术 10:30 - 11:30	* memoise 包：使用本地缓存，提升 CPU 性能 * 利用 bigmemory、ff 将硬盘空间用于数据处理 * R 的并行计算：parallel、foreach+doMC 等包
摘要：在单机情况下如果突破内存限制针对大数据进行建模，介绍本地缓存、扩展存储、多核以及 GPU 运算	
4.3 rhadoop 的使用 13:00 - 16:00	* rhdfs 安装与使用 * rmr2 安装与使用 * rhbase 安装与使用 * rmr 的 mapreduce 案例 * logistic 回归基于 rhadoop 的实现
摘要：介绍 R 与 hadoop 环境结合使用，如何在 mapreduce 框架下使用 R，因时间限制，仅以 logistic 回归在 rhadoop 下使用的案例	
4.4 RevoSacleR 功能介绍 16:00 - 17:00	* RevoSacleR 数据挖掘算法介绍 * 案例分析：航班延误数据的分析
摘要：RevoSacleR 是商业版 R——Revolution R Enterprise 的大数据模块，用这一模块，可以非常方便地对大数据建模，可以用在多核、多进程场景，换到 hadoop 场景下仅需要修改环境配置语句，而不需要自己动手实现 mapreduce 框架的代码	

主题三：R 语言和 Hadoop 系统架构在大数据分析中的应用

培训目的和特色

1. 第一天的基础课程不单讲 R，也讲解与大数据应用相关的 Linux 和 Java 的基础知识
2. 课程由在业界有多年应用经验的讲师提供，确保培训不是纸上谈兵；
3. 将 R 语言的教学与工程实践相结合，可以让初学者更快应用到实际工作中；

课程大纲

第一天：基础知识	
1. linux 基础	● Ubuntu 的环境 ● 基本命令行操作
2. JAVA 基础	● JAVA 基本概念 ● JAVA 基本核心包和语法 ● JAVA 开发工具
3. R 基础	● R 基本概念 ● R 核心函数和语法
4． 安装系统配置基本环境	● 系统要求：ubuntu 12.04 LTS 64 位 server 版 ● 如果使用 window 的朋友 ,安装虚拟机 VirtualBox，在虚拟机中安装 ubuntu

第二天：RHadoop 基础	
1. Hadoop 环境搭建	● 环境规划 ● 软件安装 ● 环境测试
2. R 与 Hadoop 集成	● 安装 R ● 安装 rhdfs ● 安装 rmr
3. RHadoop 操作入门	● HDFS 基本操作 ● MapReduce 原理及案例 ● Streaming 案例

第三天：RHadoop 高级分析	
1. 线性回归	● 线性回归基于 MapReduce 的拆分原理 ● 线性回归基于 RHadoop 的实现 ● 案例：预测汽车的销售量
2. logistic 回归	● logistic 回归基于 MapReduce 的拆分原理 ● logistic 回归基于 RHadoop 的实现 ● 案例：评估电信客户流失的概率
3. 朴素贝叶斯分类	● 朴素贝叶斯分类基于 MapReduce 的拆分原理 ● 朴素贝叶斯分类基于 RHadoop 的实现 ● 案例：将网页文本分类
4. Kmeans 聚类	● Kmeans 聚类基于 MapReduce 的拆分原理 ● Kmeans 聚类基于 RHadoop 的实现 ● 案例：细分银行的潜在客户
5. 协同过滤	● 协同过滤基于 MapReduce 的拆分原理 ● 协同过滤基于 RHadoop 的实现 ● 案例：基于历史评分推荐图书

大数据技术课程

学习目标

快速入门大数据、数据科学领域技术，可以自己完成动手实践

课程大纲

第一天课程：

1. 大数据生态圈：hadoop 生态圈目前现状介绍，hadoop 生态中各个流派的业务应用场景
2. Hadoop 集群部署管理：集群的部署安装过程以及操作
3. Hadoop 基础：HDFS、MapReduce 原理介绍以及具体应用方式
4. 数据采集技术：原理介绍以及 Flume 的具体应用方式

第二天课程：

1. 列式存储：HBase、Cassandra 原理介绍以及应用
2. 数据分析：原理介绍以及 Hive 的具体应用方式
3. 文档数据库 MongoDB、内存数据库 Redis 原理以及具体应用方式
4. 数据挖掘、机器学习算法介绍及应用

学习方式

1. 在学员学习完课后，讲师在线答疑两周时间，保证课程教授内容已经完全消化；
2. 上课采用“理论+实践”方式，需要学员自带电脑，准备好上课环境（电脑最低配置 2 核 CPU、4G 内存、200G 硬盘）；

往期培训现场图片


讲师团队主要成员介绍

刘思喆

现就职于京东商城网站智能和商业化部推荐团队，主要负责用户行为，商品特征建模等内容。8 年来，一直追求为服务企业提供高效、完备的数据解决方案，尤其在统计分析、预测分析、数据可视化、机器学习、文本挖掘、社交网络等领域。在加入京东商城前，供职于亚信联创 BOC、神州数码思特奇 DSS，主要为电信运营商提供数据挖掘及业务咨询等顾问服务。10 年 R 语言使用经验，R 语言企业级应用的践行者，中国 R 语言会议、数据科学沙龙联合发起人，中国最大的统计社区-统计之都常务理事，06 年至今一直担任 R 语言版版主。

刘思喆 2005 年毕业于中国人民大学统计学院，《153 分钟学会 R》的作者，《R in a nutshell》译者。

肖凯

肖凯是 SupStat 上海地区的负责人，精通 R 语言环境下的统计建模和数据挖掘，熟悉 ggplot2 和 D3 进行数据可视化展现，曾在第五届 R 语言大会上做过专题演讲。曾在长江水利委员会网络与信息中心水利发展研究所工作，利用决策树模型研究河流水质的影响因素；利用随机森林算法建立需水预测模型。后就职于上海万达信息股份有限公司，负责公共卫生领域的数据分析和挖掘工作。主要工作有：建立灰色模型预测人口死亡率指标；利用集成学习算法和健康档案数据建立预测模型,计算糖尿病发病概率,筛查高危人群。

肖凯也是《数据科学中的 R 语言》的作者（西安交通大学出版社，出版中）

向磊

中国首个开源 Hadoop 部署管理平台 EasyHadoop 的作者，首个开源 Hive 数据仓库可视化管理查询平台 phpHiveAdmin 作者、Hadoop 资深讲师，曾任中国联通，深交所大数据内部培训专家，在大数据领域有多年的技术研发和项目实施经验。

曾任暴风影音数据部门 Hadoop 项目组负责人及架构师，为暴风影音提供大数据的系统平台支持及技术研发工作。2012 年凭借开源项目 EasyHadoop 和 phpHiveAdmin 获得首届阿里云开发者大赛二等奖。并获得 2012 年 51CTO 中国十大杰出 IT 博客。目前组建北京数衡科技有限公司并任技术总负责人，致力于大数据技术的推广与 Hadoop 生态系统产品的落地，为企业提供高效、稳定的大数据平台，提供数据挖掘及数据可视化等一体化解决方案。

马延辉

目前就职于暴风影音，担任公司 HBase 业务集群负责人。
从事搜索、大数据行业 4 年开发经验，6 年 java 开发经验，先后在淘宝、Answers.com 从事垂直搜索、大数据分析和挖掘等方向 的研发。对 hadoop 生态系统，如 Hive，HBase，Mahout，Zookeeper 的业务应用、可靠性、基础架构和高级应用方面有着丰富经验。

擅长领域和课程：HBase 运维、性能调优和应用开发、Hive 使用和性能调优、Java MapReduce 研发。

做过的开源项目：hbase secondary index、Ella（HBase Monitor）、Hive-orc-mr

史东杰

目前就职于暴风影音，phphbaseadmin 开源项目作者，hadoop 平台工程师,曾就职于高德、趣游等公司。

对 hadoop 运维、hive 数据分析、hbase 应用开发等方面有着丰富经验。擅长领域有 hive 使用和调优，hadoop 监控等。做过的项目有游戏数据分析平台、ambari 二次开发、phphbaseadmin。

赵修湘

目前就职于缔元信，擅长数据挖掘，曾就职于暴风影音。项目经历：

- 一览视频推荐系统，使用算法：Fptree、 ItemBaseCF
- 用户 demographic 识别，使用算法：bayesian classification
- 亿赞普网站分类系统， 使用算法：svm
- 文本分类广告点击分析，使用算法：asocciation rules
- 中科院虚拟经济和数据科学研究中心的软件可信性评估，使用算法：文本分类