सत्यमेव जयते
Government of India

M S Ramaiah Institute of Technology

Workshop on
Statistical Machine Learning and Game Theory Approaches for Large Scale Data Analysis

9 July 2012 – 14 July 2012

Sponsored by
Mathematical Sciences, Division of Science and Engineering Research Board
Department of Science & Technology
Government of India

Organised by

Department of Computer Science & Engineering
Department of Industrial Engineering & Management
M S Ramaiah Institute of Technology
Vidya Soudha, MSR Nagar, MSRIT Post, Bangalore – 560 054
www.msrit.edu

Organizing Chair
Dr. N V R Naidu
Vice Principal, MSRIT
Professor and Head, Dept. of IEM

Organizing Co-Chair
Dr. R Selvarani
Professor and Head, Dept. of CSE

Coordinator
Dr. Srinivasa K G
Professor, Dept. of CSE

Co-coordinator
Mr. P M Krishna Raj
Assistant Professor, Dept. of ISE

M S Ramiah Institute of Technology

Workshop on
Statistical Machine Learning and Game Theory Approaches for Large Scale Data Analysis

9 July 2012 – 14 July 2012

	9 July 2012	10 July 2012	11 July 2012	12 July 2012	13 July 2012	14 July 2012
9.00 – 9.30	Registration and Breakfast	Breakfast				
9.30 – 11.00	Inauguration	Dr. Arati Deo and Madhusudhan Rao Amazon	Dr. K K Choudary ISI	Dr. K G Srinivasa MSRIT	Ajay Ohri decisionstats.com	Valedictory
11.00 – 11.30	Tea Break					
11.30 – 1.00	Dr. Krishna Kummamuru IBM	Mohan N Dani and Manasa Rao IBM	Dr. K K Choudary ISI	Pramodh Thoughtworks	Ajay Ohri decisionstats.com	
1.00 – 2.00	Lunch Break					
2.00 – 3.30	Dr. N V R Naidu MSRIT	Dr. Ramasuri Narayanam IBM	Srikar. Y. V. Radiant Infosystems	Rohith D Vallam IISc	Swaprava Nath IISc	
3.30 – 3.45	Tea Break					
3.45 – 5. 15	Lab Sessions - Krishna Raj P M					

Distinguished Speakers

Prof. L.M. Patnaik obtained his Ph.D in 1978 in the area of Real-Time Systems, D.Sc. in 1989 in the areas of Computer Systems and Architectures, both from the Indian Institute of Science, Bangalore. During March 2008 – August 2011, he was the Vice Chancellor, Defence Institute of Advanced Technology, Deemed University, Pune. Currently he is an Honorary Professor with the Centre for Electronic Design Technology, Indian Institute of Science, Bangalore. He has published over 635 papers in refereed International Journals and refereed International Conference Proceedings and authored 27 technical reports. He is a co-editor/co-author of twenty one books and authored 12 chapters in other books in the areas of VLSI System Design and Parallel Computing. He has supervised 22 Doctoral theses and over 160 Masters theses in the above areas. As a recognition of his contributions in the areas of Electronics, Informatics, Telematics and Automation, he was awarded the Dr. Vikram Sarabhai Research Award in 1989; the IEEE Computer Society's "1999 Technical Achievement Award" for his contributions in the field of parallel, distributed, and soft computing, and high performance genetic algorithms; the Fourth Sir C V Raman Memorial Lecture Award in 2000; the Pandit Jawaharlal Nehru National Award for Engineering and Technology, in 1999; the Om Prakash Bhasin Award for contributions in the areas of Electronic and Information Technology for the year 2001; the FICCI Award for Innovation in Material Science, Applied Research and Space Science, 2001-2002; the IEEE Computer Society's Meritorious Service Award, 2002; Alumni Award for Excellence in Research for Engineering, the Indian Institute of Science, 2003, Distinguished Engineer Award of The Institution of Engineers(India), 2004; Goyal Prize for Applied Science, 2005; Honorary Fellow, the Indian Society for Technical Education, 2006; Indian Science Congress Association's Srinivasa Ramanujan Birth Centenary Award, 2007-2008. He is Fellow of the IEEE, The Academy of Sciences for the Developing World (TWAS), The Computer Society of India, Indian National Science Academy, Indian Academy of Sciences, National Academy of Sciences, and Indian National Academy of Engineering.

Jai Navlakha received his Ph.D. in Computer Science from Case Western Reserve University in December 1977. Since then, he has been associated with the School of Computer Science (initially, part of the Department of Mathematical Sciences) at Florida International University. He was promoted to the rank of Full Professor in Fall 1987, and served as the Director of the School from 1988 to 1992. Since Fall 1996, he has been the Director of the Center for Computational Research in the School. He has published widely in the areas of software engineering, algorithm analysis, expert systems and neural network applications.

K Rajani Kanth has an extensive experience both in industry and academia. He obtained his Masters and Ph.D from Indian Institute of Science, Bangalore. He served MSRIT as Vice Principal, Principal and Advisor (Academics and Research). He has chaired various high profile committees in the University. An voracious reader and eloquent speaker his area of interest spans many areas across the systems, electronics, computing, pedagogy and education management spectrum.

09 July 2012 (11.30 – 1.00)
Text Mining and Question Answering

IBM built a deep question answering system called Watson which defeated the best players of an American television quiz game show, Jeopardy! The quiz deals with clues from various topics like history, literature, the arts, pop culture, science, sports, geography, wordplay. The show has a unique answer-and-question format in which contestants are presented with clues in the form of answers, and must phrase their responses in question form. Watson system uses a combination of advanced topics from language understanding, machine learning and game theory. In this talk, I review various text mining and natural language processing techniques that are useful in building such a deep question answering system.

Krishna Kummamuru is a software architect in Watson Labs – India, a part of IBM India Software Labs, working on IBM Watson applications to financial sector. Before assuming this role in April 2012, he has been with IBM India Research Lab (IRL) since 1998. During his tenure at IRL, he has worked on building technologies to deliver services in emerging markets based on Spoken Webtechnology; led Research Collaboratory for Service Science at ISB, Hyderabad; led a group on Services Information Management and Analytics working on problems addressing KM issues in IT Service Delivery centers.

He received the B.Sc degree in Physics from Nagarjuna University in 1986. He received the M.E degree and the Ph.D in Electrical Engineering from IISc, Bangalore, India in 1993 and 1999 respectively. He received Alfred Hay Medal for the best graduating student in EE in 1993 from IISc. He has 11 patents granted by USPTO and about 30 publications in refereed conferences and journals (a citation count of about 960 as on April 2012). His research interests include Text & Data mining, Machine learning, User interfaces and Service science.

09 July 2012 (2.00 – 3.30)
Introduction to Statistics and Probablity

N V R Naidu graduated in Mechanical Engineering from Sri Venkateswara University, Tirupathi in the year 1980. He further pursued his M.Tech in Industrial Engineering in the year 1982 and obtained his Ph.D from the same university. Dr. NVR Naidu has been serving the teaching profession with a great devotion for 30 long years. He started his career in 1982 as a Lecturer in the prestigious M.S. Ramaiah Institute of Technology, Bangalore. He is currently the Vice-Principal, Professor & Head of the Department of Industrial Engineering at the same institute. Dr NVR Naidu is a recipient of Dr J Mahajan Award for the year 2008-09 awarded by Indian Institution of Industrial Engineering for his outstanding contribution in the field of education and research and also his name is listed in Marquis Who is Who in the World, USA in the year 2011. Dr. NVR Naidu is well recognized for his research. He has presented and published over 80 research papers in various National and International referred Journals and conferences. Dr.NVR Naidu is the adjudicator for Ph.D thesis for various universities across India. He has produced 3 doctorates and is currently guiding 5 Ph.D scholors in the areas of robust design, design and development of production systems, supply chain network and design of experiments. Dr. NVR Naidu has visited various countries including the USA, Japan and Sri Lanka to collaborate and enhance the Industry-Institution interaction.

10 July 2012 (9.30 – 11.00)
What and How of Machine Learning – Leveraging Machine Learning on Web Data

This talk will introduce Machine Learning concepts and techniques, and discuss various new data mining and machine learning applications being developed for the Internet. Recent advances in cloud computing and big data processing have led to renewed interest in machine learning applications. Machine Learning can not only enhance automation and increase efficiency within various Web related systems and processes, it can also spawn a new range of services and products which would not have been possible otherwise. We will highlight a few real-world examples of such applications and explain the underlying machine learning technology which enables such systems.

Arati has a Ph.D. in Electrical and Computer Engineering (specializing in Robotics) from Rice University, Houston. Arati is currently Senior Manager for Machine Learning at Amazon Bangalore, leading development of applications leveraging machine learning techniques for a variety of business problems within Amazon. Prior to this, Arati led the A9 Bangalore development and operations teams owning multiple product features for Amazon's Ad Technology products. Before joining Amazon in 2005, Arati was Director for Analytics at FICO in San Diego. Arati has experience developing predictive software solutions in various industries including online advertising, credit card fraud, credit risk and healthcare.

Madhu is currently an architect with the Junglee.com team working on making seller onboarding zero-touch. He started at Amazon in 2005 with the Associates/Traffic team where he was responsible for building the Widgets Platform (Large scale Adserver which serves Rich Media Ads for online advertising and affiliates), a contextual recommendation engine, product classifier, and the Publisher Analytics Platform. Madhu graduated with a B.E. in Computer Science from PESIT, worked with ThoughtWorks for 2 years and did a startup on spend-management for a year before joining Amazon. Madhu has expertise in building large scale distributed systems, recommendation and similarities engine.

10 July 2012 (11.30 – 1.00)
The Big Picture of Big Data Analytics

Big Data Analytics deals with problems and solutions emanating from large volumes,varieties and velocity of data. In the present Internet era there is a lot of digitization and data is being exchanged across the world. This leads to peta bytes or exabytes of data exchanges across the Internet. Imagine if we wanted to mine or analyze this volume of data, What technical aids are available and how do we solve it? Big Data is the answer to such large scale analysis and is an information interchange platform.
The session aims at answering questions like What is Big Data?, How do we tackle information interchange in Big Data?, How is analysis of data at rest and In flight analytics done? etc... thereby providing a big picture of Big Data. A case study of social media analytics will be used to help audience appreciate the need and power of Big Data Analytics.

Mohan N Dani is a lead for Infosphere Streams Big Data efforts in IBM-ISL. He has extensive knowledge in implementation of large enterprise solutions and has more than 11.5 Years experience in IT. He has played multiple roles in IBM as a IBM Business Analyst, Development lead and a Solution Architect. His specialization is Streams, He consults for IBM Internal teams for Pre-sales,Post-sales, Delivery Excellence and Thought leadership on Streams/Big Data. He is an MS Caltech and a Banking IT Specialist from Mortgage Bankers Association.

Manasa K. Rao has 3 years of experience in IT and has dealt extensively with Data Quality challenges of Indian Clients. She was part of the cloud computing Data as a Service project and is now working as a Toolkit Lead in the Big Data Streams computing software. She has done her BE from MSRIT (VTU).

10 July 2012 (2.00 – 3.30)
Viral Marketing Through Social Networks

Viral marketing through social networks is a phenomenon of word-of-mouth marketing of news products that exploits the social connections among individuals in an appropriate fashion. This area of research has got significant attention from the research community recently. In this talk, I will introduce the viral marketing problem and then highlight various versions of this problem. I will appropriately discuss the solution techniques to address the different versions of viral marketing problem. I will also present several experimental results to appreciate this notion.

Ramasuri Narayanam is currently a researcher in IBM Research, India. His research interests are social networks and game theory. At IBM, he works on large scale social network data analytics. Prior to joining IBM Research, he obtained both masters degree and Ph.D. degree in Computer Science from Indian Institute of Science (IISc) in 2006 and 2011 respectively. He is a recipient of Microsoft Research Ph.D. Fellowship for the period 2007-2011. A proposal based on his Ph.D. Thesis got an Honorable mention award and a research grant from Yahoo! Key Scientific Challenges Program, 2010.

11 July 2012 (9.30 – 11.00 & 11.30 – 1.00)
Role of Statistical Tests of Hypothesis and Regression Analysis for Handling Large Scale Data Analysis.

Prof.K.K.Chowdhury, presently faculty, SQC & OR unit, Indian Statistical Institute, Bangalore, comes with 30 + years of experience. He comes with Bachelor of Statistics and also, Master of Statistics from Indian Statistical Institute, Kolkata. He is also having Post Graduate Diploma in SQC & OR from Indian Statistical Institute, Kolkata to his credit.

Prof.Chowdhury is engaged in providing Consultancy and Training Services in the field of Quality Management Science to both IT and Non-IT companies in India & abroad, to name a few, Wipro, Ashok Leyland, Saint –Gobain, Grasim Industries, Larsen & Toubro, Reliance Industries, HAL, BHEL, BEL, HMT, etc, Bander Imam Petrochemical Complex & SAIPA Iran, Asian Paints, Johnson & Johnson, Indorama Synthetic, Indonesia; Premisys consulting, Indonesia etc. His consulting assignments are on achieving bottom-line business result improvement through using Six Sigma Programme as well as application of Statistical methods. He conducted several training programs on: Statistical Techniques including Taguchi Methods & Reliability; Six Sigma Green Belt, Black Belt and Master Black Belt programme; for both IT/ITES and non IT/ITES companies

Prof.Chowdhury has many professional laurels to his credit that include: Visiting Faculty of IIM, Indore;Indian institute of Plantation management(IIPM),Bangalore; Head, SQC & OR Division, Indian Statistical Institute during 2008-2010; Proctor for the Certification Programme of ASQ; Contributed more than 20 Technical Papers in National & International Journals e.g. Quality Engineering of ASQ, TQM, UK etc. and also in conference proceeding, Organizing Secretary for the 8th National Convention of National Institution for Quality & Reliability,1998; Audited more than 50 organizations for certification of Quality Management system –QS-9000/ISO – 9000 in India, Indonesia, Malaysia, Philippines, Thailand and Iran on behalf of KEMA, Netherlands.

11 July 2012 (2.00 – 3.30)
Large Volume User, Transaction & Data Management in Government 2 Citizen for Transport Application

We are presenting the implementation and management of User, Transaction and Data Management a large Government 2 Citizen (G2C) Application in Transportation Domain. We will demonstrate the vision and goals of this application and how our solution addressed the requirements and how the solution has been able to provide significant value adds and benefits to different stakeholders of this project. Our Transportation Solutions have assisted in improving the efficiency and effectiveness of our Transportation Clients. They have shown tremendous growth in the passenger volumes, Transaction and Revenue Growth for our clients. More importantly the solutions have made the travel and itinerary management user friendly and convenient for the passengers.

Srikar. Y. V is currently the Delivery Head and manages the delivery function at Radiant Info Systems Ltd, an IT Services firm in Bangalore, India. He's responsible for the delivery management and handles mid-sized development team stationed across multiple locations. He's involved in the project lifecycle from the pre-sales process to implementation and support phase there of. Radiant provides solutions and services in the Online Reservation Space for Transportation Segment, eGovernance Solutions, Smart Card & Biometric Solutions & FMCG segment primarily on Web & eBusiness Technologies. Radiant's transportation solutions are powering the reservation systems of the 7 of the India's largest road transport service providing organizations like KSRTC, TNSTC, GSRTC, OSRTC, PUNBUS, PEPSU, MEGHALAYA etc. Radiant is a leader and pioneer in the Online Reservation Space in India. Srikar has more than 15 years of experience in the Information Technology space and has experience and expertise in the IT Product and Solutions Life cycle management. He is a certified Project Management Professional (PMP) with multiple large end to end product / project development life cycles during my professional career.

12 July 2012 (9.30 to 11.00)
Machine Learning for Large Scale Data Analysis

The session intends to cover:

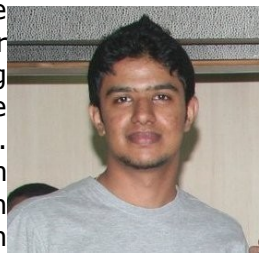
1. Machine Learning: what?
2. Machine Learning Techniques: Clustering, Classification, Recommender examples
3. Supervised and Unsupervised Learning, Reinforcement Learning
4. Trees and Machine Learning
5. Recommender Systems and Machine Learning
6. Applications: Where do you see machine Learning in the real world?

Dr. Srinivasa K G received his PhD in Computer Science and Engineering from Bangalore University in 2007 in the are of Soft Computing for Data Mining Applications. He is now working as a professor in the Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, Bangalore. He is the recipient of All India Council for Technical Education - Career Award for Young Teachers, Indian Society of Technical Education – ISGITS National Award for Best Research Work Done by Young Teachers, Institution of Engineers(India) – IEI Young Engineer Award in Computer Engineering, IMS Singapore – Visiting Scientist Fellowship Award. He has published more than fifty research papers in International Conferences and Journals. He has visited many Universities abroad as a visiting researcher – He has visited University of Oklahoma, USA, Iowa State University, USA, Hong Kong University, Korean University, National University of Singapore are few prominent visits. He has authored two books namely File Structures using C++ by TMH and Soft Computer for Data Mining Applications LNAI Series – Springer. He has been awarded BOYSCAST Fellowship by DST, for conducting collaborative Research in University of Melbourne in the area of Cloud Computing.

12 July 2012 (11.30 to 1.00)
Statistical Machine Learning for Large Scale Data Analysis

This session intends to cover:

1. The paradigm of Machine Learning using statistical methods.
2. Different Statistical machine Learning models and techniques: The math and description of each of these techniques
3. Exploring real-time tools and programming techniques to employ learning in statistical data.
4. What changes when the data is large? An Insight onto Statistical Learning and Large scale data.



Pramod N is working as an application developer in ThoughtWorks Inc. He has completed his engineering degree from Department of Computer Science and engineering, M. S. Ramaiah Institute of Technology. Having worked as a research assistant under Dr Srinivasa K G, he has experience in applying Machine Learning onto various computer science problems. Recent works Include gNIDS- Rule based intrusion detection system employing Genetic Algorithms, Statistical Approach to Network Intrusion detection Using SVM, Hybrid approach to Spoken Language Identification employing Gaussian Mixtures and SVM.

12 July 2012 (2.00 – 3.30)
Game Theory and its applications in Social Network Formation

Game theory is the science of strategy. According to Prof. Roger Myerson, the 2007 Nobel Prize winning economist, game theory may be defined as the study of mathematical models of conflict and cooperation between 'rational', 'intelligent' decision makers. It attempts to determine mathematically and logically the actions that 'players' should take to secure the best outcomes for themselves in a wide array of 'games'. In this talk, we will discuss some of the fundamental concepts of game theory through a number of examples motivated from simple real-world scenarios. Further, we will investigate a particular application of game theory in the context of social network formation.

Rohith received his BE in CSE from Visvesvaraya Technological University, Bangalore in 2002. He has worked for three years in the software industry in the domain of Intelligent Networks. He has completed his MS(Research) in Wireless Networks from Dept of CSE, IIT Madras, India. He is pursuing his PhD at Dept. of CSA, IISc. His current research focus is to apply game theory and mechanism design models in the areas of prediction markets and social networks.

13 July 2012 (9.30 – 11.00 & 11.30 – 1.00)
Big Data Big Analytics

The talk will showcase using open source technologies in statistical computing for big data, namely the R programming language and its use cases in big data analysis. It will review case studies using the Amazon Cloud, custom packages in R for Big Data, tools like Revolution Analytics RevoScaleR package, as well as the newly launched SAP Hana used with R. We will also review Oracle R Enterprise. In addition we will show some case studies using BigML.com (using Clojure) , and approaches using PiCloud. In addition it will showcase some of Google APIs for Big Data Analysis. Lastly we will talk on social media analysis ,national security use cases (I.e. cyber war) and privacy hazards of big data analytics.

Ajay Ohri has been an analytics professional since 2004. He has worked mostly within India and a bit in USA with some very large organizations on leveraging data for business benefits. He is the author of the forthcoming book "R for Business Analytics" (Springer Sept 2012).
For the past five years he has been running his own consulting firm in analytics as well as very focused blog called Decisionstats.com where he has interviewed more than 95 technology leaders.
His research interests are in R and SAS languages with a focus on business analytics. He is a graduate from Delhi College of Engineering and Indian Institute of Management, Lucknow and has also attended graduate courses at University of Tennessee,Knoxville.
In addition to his writing on technology, Ajay has written and co-authored 4 electronic books of poetry and runs a poetry blog.

13 July 2012 (2.00 – 3.30)
Mechanism Design: In Theory and Practice

Ever wondered how Google makes money, or how organizations benefit from outsourcing tasks to experts? Mechanism Design is a tool from Microeconomics that provide solutions to many problems in Internet monetization. In this talk, I am going to give a brief overview of Mechanism Design theory and present some of the challenging problems and their solutions in strategic outsourcing.

Swaprava did his Masters in Telecommunication Engineering in 2008 from Dept. of Electrical Communication Engineering, Indian Institute of Science, Bangalore, where he is currently a PhD Candidate at the Dept. of Computer Science and Automation. His current research interest is in the Game Theoretic questions arising in the area of Internet Economics, Outsourcing, Crowdsourcing, Machine Learning etc. Swaprava's work encompass different areas of strategic task outsourcing. He has completed internships at Xerox Research Centre, Europe (XRCE), in 2010, where he had worked on Incentive Compatible Learning for E-Services, and in EconCS, Harvard University, in 2011, where he has worked with Prof. David C. Parkes, and worked on Economics of Opensource Networks. He is a recipient of the Honorable Mention Award of Yahoo! Key Scientific Challenges Program, 2012. His research is supported by the Tata Consultancy Services PhD Fellowship, 2010. Details about his research and publications are available here: http://swaprava.byethost7.com/

9 - 13 July 2012 (3.45 – 5.15)
Lab Sessions - Large Scale Data Analysis with R

R is a powerful, free software framework which can be used for doing myriad jobs of data mining, statistical analysis and advanced visualisation. It also comes with programming interface and rich set of extendible libraries. Used widely in the research and commercial environment today, R has given rise to a host of tools like Rattle, Gretl which use R engine to do intensive computations in specific domains.

The laboratory sessions spread over five days will cover the following topis
Session 1 : Introduction to GNU Linux and installing R with all options
Session 2 : Basic data handling in R
Session 3 : Advanced data operations and graphics in R
Session 4 : Statistics, Regression and Hypothesis Testing in R
Session 5 : Data Mining Algorithms in R

Krishna Raj obtained his Engineering degree from M S Ramaiah Institute of Technology, Bangalore. He completed his Masters through research working in the area of Free and Open Source Software Engineering. He is pursuing his doctoral studies examining the developmental patterns in Free and Open Source Software. He is currently working as Assistant Professor in the Department of Information Science and Engineering, M S Ramaiah Institute of Technology, Bangalore. He has co-authored a text book on File Structures published by Tata Mc-Graw Hill.
His current interests are Ramayana, Aesthetic Computing and Philosophy of Technology. His technical and general writings can be found at http://krishnarajpm.com

Statistical Machine Learning and Game Theory Approaches for Large Scale Data Analysis

Lab Session I – July 09, 2012

What is R?	R is a free software environment for statistical computing and graphics.
What is Free Software?	Free Software is the one governed by GPL compatible licences.
What is GNU Linux?	GNU Linux is the free operating system environment.
Where can we get R?	http://www.r-project.org/
How to install R?	In UBUNTU – sudo apt-get install r-base In Fedora – yum install R
How to start R?	Type R in command prompt to start interactive prompt
How to install packages?	install.packages("package name")
How to use a library?	library("package name")
How to quit R?	q()
Tools with R Engine	Gretl, Rattle
Alternates for R	WEKA, Orange, RapidMiner

Statistical Machine Learning and Game Theory Approaches for Large Scale Data Analysis

Lab Session II – July 10, 2012

Basic Commands	2+2 , log(10), sqrt(25) ls(), system("cat sample.txt")
Variables	a=10, b=a+10;b
Vectors	A = c(1,3,5,7,9) a = (1:5) a1 = seq(-2,1, by=0.25)
Read data to vector	data.entry(a) a = scan() a= rnorm(10) a = sample(1:6,5,replace=T) RollDie = function(n){ sample(1:6,n,replace=T)}; a =RollDie(5)}
Vector Operations	a, a[1], a[-1] max(a), min(a), length(a) which(a==2) sort(a), diff(a) sum(a> 3), sum(a), mean(a), median(a), var(a), sd(a) sample(a,2) cbind(a,A), rbind(a,A) log(a), log10(a) sqrt(a), exp(a)
Matrices	b = matrix(c(2,5,6,3), nrow=2, ncol=2, byrow=FALSE)
Matrix Operations	Transponse – t(b) Inverse – solve(b) Dimension - dim(b)

Statistical Machine Learning and Game Theory Approaches for Large Scale Data Analysis

Lab Session III – July 11, 2012

See Available Data Sets	Data(), women , ?women, names(women)
Explore Data	summary(women) sd(women) var(women) cor(height, weight)
Use a data for this session	attach(women)
See only a column	height or women$height
Graphs	boxplot(women) hist(height) plot(height, weight) pie(height) X = matrix(rnorm(300),100,3); pairs(X) hist(height, col="green") hist(height, col=heat.colors(length(height))) pareto.chart() plot(ecdf(rnorm(10)) library(scatterplot3d);scatterplot3d(height,weight) pie(table(Species))
Export graph to file	x11(); postscript(file="fig.eps") OR png(file="fig.png"); plot ();graphics.off()
Bivariate Data	b1 = c("a", "b", "a", "b", "b") b2 = c (1,2,3,4,5) table(b1, b2)
Multivariate Data	b3 = c(10:14) b = data.frame(b1, b2, b3)
Import Data	x =read.table(file="sample.txt",header=T) x = read.table(file="sample.csv",header=T, sep=",")
Export Data	x<-matrix(c(1.0, 2.0, 3.0, 4.0, 5.0, 6.0), 2, 3) write (x, file="sample1.txt", ncolumns=3)

Statistical Machine Learning and Game Theory Approaches for Large Scale Data Analysis

Lab Session IV – July 12, 2012

Programming Constructs	if i==10 {b=1} else {b=0} for (i in 1:10){ b[i] = i+100; j[i] = b[i]+1} i=1; while(i<=10) {b[i]=i;i=i+1} ect=function(x) {r = sqrt(x); return(r)} ; s=ect(25); s
Linear Regression	x=c(0,10,20,30,40) y=c(4,22,44,60,82) l=lm(y~x) summary(l) fitted(l) layout(matrix(1:4,2,2));plot(l)
Prediction	x1 = c(5,15,25,35,45) predict(l,data.frame(c = c1), level = 0.9, interval = "confidence") plot(c,s);abline(l)
Hypothesis Testing Chi- Square Test	freq = c(22,21,22,27,22,36) probs = c(1,1,1,1,1,1)/6 chisq.test(freq,p=probs)

Statistical Machine Learning and Game Theory Approaches for Large Scale Data Analysis

Lab Session V – July 13, 2012

Association Rule Mining	titanic<-read.table("Datset.data",header=F) names(titanic)<-c("Class","Sex","Age","Survived") summary(titanic) library(arules) rules<-apriori(titanic) inspect(rules) rules <- apriori(titanic, parameter = list(minlen=2, supp=0.005, conf=0.8), appearance = list(rhs=c("Survived=no", "Survived=yes"), default="lhs"), control = list(verbose=F)) inspect(rules)
K-Means Clustering	iris2 <- iris iris2$Species <- NULL kmeans.result <- kmeans(iris2, 3) kmeans.result kmeans.result["centers"] table(iris$Species, kmeans.result$cluster) plot(iris2[c("Sepal.Length", "Sepal.Width")], col = kmeans.result$cluster)
Hierarchical Clustering	idx <- sample(1:dim(iris)[1], 40) irisSample <- iris[idx,] irisSample$Species <- NULL hc <- hclust(dist(irisSample), method="ave") plot(hc, hang = -1, labels=iris$Species[idx])