RUTGERS
Rutgers Business School
Newark and New Brunswick

Business Analytics and Information Tech
COURSE NUMBER: 33:136:494
COURSE TITLE: Data Mining and Business Intelligence

COURSE DESCRIPTION

This course presents computing tools and concepts for all stages of dealing with the modern data deluge---statistical computing at the center, but also the essential surrounding tasks, including data organization, presentation of results and the user interface. This approach is needed to deal with the challenges posed by modern technology, challenges that are also opportunities for better use of data. The size and complexity of data sources has increased enormously, while the importance of learning from the data has been recognized as never before. New modes of computing such as large-scale parallelism and cloud computing can help, but require new approaches to programming. But the key challenge is to use our own time effectively by choosing the best programming approach for each stage of a project.

The course also covers linear and polynomial regression, logistic regression and linear discriminant analysis; cross-validation and the bootstrap, model selection and regularization methods (ridge and lasso); nonlinear models, splines and generalized additive models; tree-based methods, random forests and boosting; support-vector machines; Some unsupervised learning: principal components and clustering (k-means and hierarchical). Computing is done in R, through tutorial sessions and homework assignments.

We present a range of computing paradigms and corresponding languages, each designed for ease of use but also providing a rich set of tools. We use the R language and the thousands of packages written for it for core statistical computing.

This course also presents concepts and techniques as related big data analytics. Big Data Analytics with R and Hadoop exposes students to the paradigm of Mining of Massive data Sets.

COURSE MATERIALS

Recommended Textbooks:

1. Software for Data Analysis by John Chambers, Springer 2008 (PDF Downloadable from Rutgers Library for free).

2. An Introduction to Statistical Learning with Applications in R by Gareth James, Daniell Witten, Trevor Hastie, Robert Tibshirani: Springer

3. Big Data Analytics with R and Hadoop by Vignesh Prajapati: PackT publishing Opensource (http://www.packtpub.com/big-data-analytics-with-r-and-hadoop/book)

4. Mining of Massive Datasets by Anand Rajaraman, Jure Leskovec, Jeffrey D. Ullman: downloadable free from The Stanford University Infolab (http://infolab.stanford.edu/~ullman/mmds/book.pdf)

Reference Textbooks:

ggplot2: Elegant Graphics for Data Analysis, Hadley Wickam, Springer, 2009.

Learning Python, Marj Lutz, O'Reilly.

Parallel R, O'Reilly.

Advanced R Development (forthcoming) by Hadley Wickam. See Advanced R Wiki

Visualizing Data. Ben Fry. O'Reilly.

Prerequisites:
No previous knowledge of programming languages is required. However those of you that are familiar with some other language, particularly C or a C derivative, will have an easier ride in the first few weeks. You need to have access to a personal computer (Windows, Mac or Unix will all work.) You need to be able to download and install software on this machine. You also need to have access to the internet.

CLASS ORGANIZATION & ADMINISTRATION

Attention:
This course is fundamentally different from other courses you have ever taken or will take in this program. It is not about learning a few formulas, principles, definitions, and applying them using the inventory of skills you have already acquired in your previous education. This course is about expanding exactly this inventory of skills that forms the underlying basis of your education to a totally new area, and develop a way of thinking that is unlike those you are employing in other coursework. Programming is not easy for those who have no prior experience with it, yet it becomes easy as you practice. Programming projects and homework are the heart and soul of this course. You have to do them in order to learn. Therefore, you may very well need to spend more time working on this course than on any other, practicing how to write programs. This is the only way you can acquire a skill essentially different from others that you already have.

Attendance:
Regular attendance is compulsory. You are not allowed to check your emails, access Web sites not related to the course or work on something that is beyond the scope of this course during the class time.

Assignments:
You may have discussions with your class members, but you have to submit your own work. Please be sure to keep a copy of the assignment by yourself in case that there is any problem with your hand-in/online submission or you have to use it later this semester. Assignments have to be submitted **before** the beginning of the class on the specified due day. **No late submissions will be accepted.**

Exams: There will be **no make-up exams.** You are required to present a written proof for situations such as going on to an emergency room due to unexpected and serious illness. Chatting during the exam is **not** allowed. **Email communication during the exam will be considered cheating. No** collaboration between class members will be allowed during any exam. There will be **no** extra-credit project.

Collaboration and Cheating: Collaboration of any kind is **strictly forbidden** on all exams, and quizzes. Any violations that I detect will be formally prosecuted. Students should familiarize themselves with the RBS honor code pledge, "I pledge, on my honor, that I have neither received nor given any unauthorized assistance on this examination (assignment)." See http://academicintegrity.rutgers.edu/academic-integrity-at-rutgers for more information.

FINAL GRADE ASSIGMENT

In-class work, Assignments		10%
Exam I		20%
Exam II	20%	
Project		15%
Final		35%

I reserve the right to make changes to the grade calculation scheme.

Business Analytics and Information Tech **(33:136:494)**

COURSE SCHEDULE

Week of	**Week**	**Topic**
01/20	1	Introduction to Data Mining and Business Intelligence Functional programming and R; objects in R Introduction to Statistical Learning: chapter 1&2
01/27	2	Dataframes in R R packages Design, checks, publishing Introduction to Statistical Learning: chapter 3
02/03	3	S4 Classes and Methods. Introduction to Statistical Learning: chapter 4
02/10	4	OOP computing model in R, Reference Classes and other languages Introduction to Statistical Learning: chapter 5
02/17	5	Databases, SQL, ODBC, drivers and interfaces from R (DBI) XML, Xschema, XSL Introduction to Statistical Learning: chapter 6
02/24	6	Intersystem interfaces: R and C,C++, Python, Java, etc Spreadsheet model of Computing, interface to R Introduction to Statistical Learning: chapter 7
03/03	7	Exam I
03/10	8	Data Visualization: R graphics, ggplot2, graphs Introduction to Statistical Learning: chapter 8
03/17	9	SPRING BREAK
03/24	10	Debugging R, interactively in R and at the C level

		Introduction to Statistical Learning: chapter 9
03/31	11	Large computations and large data; vectorizing; measuring efficiency Cluster Computing, MPI, R facilities, CUDA examples if time permits Introduction to Statistical Learning: chapter 10
04/07	12	Cluster Computing, MPI, R facilities, CUDA examples if time permits
04/14	13	Exam II
04/21	14	Map-reduce computations, Hadoop
04/28	15	Web based interfaces, libraries, publishing. Examples Review
05/08	16	Final Exam Period 05/08 to 05/14

Scholastic Dishonesty Policy

The University defines academic dishonesty as cheating, plagiarism, unauthorized collaboration, falsifying academic records, and any act designed to avoid participating honestly in the learning process. Scholastic dishonesty also includes, but not limited to, providing false or misleading information to receive a postponement or an extension on assignments, and submission of essentially the same written assignment for two different courses without the permission of faculty members. The purpose of assignments is to provide individual feedback as well to get you thinking. Interaction for the purpose of understanding a problem is not considered cheating and will be encouraged. However, the actual solution to problems *must* be one's own.