RUTGERS

School of Communication and Information
Machine Learning for Data Science
Special Topic Course
Fall 2017

Course Delivery:	On campus
Course Website:	http://chirags.comminfo.rutgers.edu/teaching/2017_fall/ML/

Instructor:	Chirag Shah
Email:	chirags@rutgers.edu
	(24-hour turnaround on email correspondence)
Office Phone:	*848-932-8807*
Office Hours:	Wednesdays 2-3pm

Catalog Description

This course offers students a practical introduction to using Machine Learning algorithms, tools, and techniques for solving problems that fall under the umbrella of Data Science. It is structured around learning concepts from the field of Machine Learning and applying them on data-intensive problems. While the course covers theories of Machine Learning and tools such as R, the focus is on using them for solving data-driven problems. The students will be introduced to several real-life problems that involve analyzing data for prediction, classification, organization, estimations, and pattern recognition.

Course Description

As a constant flux of rapidly growing amounts of data is created and used in industries and research environments, there is an increasing demand for people who are able to pursue data-driven thinking and decision-making using meaningful insights derived from large and diverse data. Many of these situations involving data-driven decision-making require approaches that are designed around finding patterns in large amounts of data, and using such discoveries for solving data problems and making useful predictions. This course offers students a practical introduction to using Machine Learning algorithms, tools, and techniques for solving problems that fall under the umbrella of Data Science. It is structured around learning concepts from the field of Machine Learning and applying them on data-intensive problems. While the course covers theories of Machine Learning and tools such as R, the focus is on using them for solving data-driven problems. The students will be introduced to several real-life problems that involve analyzing data for prediction, classification, organization, estimations, and pattern recognition.

Pre- and Co-requisites

This course requires computational thinking, statistics, and a basic understanding of linear algebra. Specifically, the student will need to have prior experience with programming as well as statistics. Examples of courses that fulfill such requirements are 04:547:202, 16:198:509, and 17:610:562 for programming; 15:291:531, 16:954:581, and 17:610:511 for statistics. An exception may be made for a student who could demonstrate technical readiness through some other method, including a technology course taken elsewhere or industry experience. Consult with the course instructor for further information.

Learning Objectives

By the end of the course, students will be able to:

1. Exhibit familiarity with Machine Learning methods by learning and experiencing essential algorithms and approaches.
2. Use Machine Learning techniques to explore and analyze data, and derive decision-making insights.
3. Identify data-driven analytics problems, and design solutions and applications to solve them using Machine Learning techniques.

Instructional Method

The course will be taught as a mixture of lecture, discussion and in-class lab, in an effort to provide an accelerated path to experience. The lectures will incorporate theories from Machine Learning, and practices from Data Science. The discussions will involve thinking about ways to apply Machine Learning techniques to real-life problems. The in-class lab will engage the students in writing code to implement Machine Learning algorithms for solving data problems. With each class, an assignment will be given that will typically extend or reinforce the concepts learned during the class.

Major Readings

There is no textbook for this course. The instructor will provide a companion ebook for free. There are several online articles that the students will be recommended to read. These links will be provided through the course Website. For those who lack proper background in statistics or programming are recommended to look up the following books.

- Machine Learning with R (2/e) by Brett Lantz.
  https://smile.amazon.com/Machine-Learning-Second-Brett-Lantz/dp/1784393908/
  This book provides a good introduction to both Machine Learning and R. It can be a nice reference book for most of this course, but do not count on it to take you further than the very basics.
- R for Data Science by Hadley Wickham and Garrett Grolemund
  https://smile.amazon.com/Data-Science-Transform-Visualize-Model/dp/1491910399/
  This is a very comprehensive book for R. We don't need most of what's in this book for this course, but if you are planning on continuing with R and Data Science, this could be a good reference book to have.

- Practical Statistics for Data Scientists: 50 Essential Concepts by Peter Bruce and Andrew Bruce
  https://smile.amazon.com/Practical-Statistics-Data-Scientists-Essential/dp/1491952962/
  This is a good book to have for basic to intermediate statistics, especially relating to Data Science. It also covers some aspects of learning (classification, regression, clustering).

Online Course Shell/Site

The course will have a Website at http://chirags.comminfo.rutgers.edu/teaching/2017_fall/ML/. Use the Website to learn about the most updated schedule, reading materials, and assignments. The actual assignments will be available from and submitted through Sakai, unless instructed otherwise.

If you need help with accessing or using Sakai site for this course, please contact Sakai helpdesk: 848-445-8721 (Mon-Fri 8am-6pm) or at sakai@rutgers.edu

Methods of Assessment and Grading

The content of this course is best understood by assimilating the lectures, by readings, by analyzing examples and by practice. The assessment for this course is based on a series of assignments that match the real-world process and on class participation. Assignments are of two types: smaller exercises and a multi-part course project. Descriptions of the assignments are available on the course Website. There will also be exercises that are not graded – in all cases, you will later use the same techniques/methods as a part of your project. Class participation includes participation in discussions; reading descriptions. Course grades are calculated based on cumulative and weighted points from all the assignments, mid-term-exam, final project, and class participation. Following list shows the mapping of % points to letter grade, along with an interpretation of the grades.

- A (91-100%): Outstanding and excellent work of the highest standard, mastery of the topic, evidence of clear thinking, good writing, work submitted on time, well organized and polished.
- B+ (85-90%:) Very good work, substantially better than the minimum standard, very good knowledge of the topic; error free.
- B (80-84%): Good work, better than the minimum standard, good knowledge of the topic.
- C+ (74-79%): Minimum standard work, adequate knowledge of the topic.
- C (70-73%): Work barely meeting the minimum standard, barely adequate knowledge of the topic; errors.
- D (65-69%): Writing not up to standard, disorganized, many errors
- F (< 65%): Unacceptable, inadequate work
- T: Temporary.

The final grade will be weighted based on the following: Assignments: 45%, Mid-term: 20%, Final project: 30%, Class participation: 5%.

Key Assignments

This is a practice-oriented course. That means most of the assignments (homework, in-class, exams) will be based on tackling real-life problems and applying skills learned in the class. The classes and assignments directly relate to the learning objectives (LO). Specifically,

- LO-1 is associated with units 1-6 and 9-13, during which the students will learn about various Machine Learning techniques and how they can be used for solving Data Science problems.
- The weekly assignments will address LO-2.
- LO-3 will be met through the mid-term and the final projects.

Weekly assignments (45%): The course will have weekly assignments – given with each class. The assignment will typically be an extension of what is covered during that week. In other words, a typical assignment will ask to take what was taught and practiced during the class and take it a few steps further. One can expect to spend roughly 5-8 hours a week to work on an assignment.

Mid-term project (20%): The mid-term exam will start in the class on October 25th. It will be open-book exam, which means you can use any and all resources you like, including online. If you have to miss that particular class, contact the instructor to find an alternative time and place for you to take the mid-term exam.

Structure:
This will be a problem-solving assignment. You will be asked to solve the given set of data problems using Machine Learning techniques covered so far.

You can use existing pieces of code under the following conditions:
(1) There should be at least 30% new code in your project;
(2) Any code from somewhere (including your own work) has to be attributed properly.

Time: You will be given 2 days starting the class.

Grading: The exam (project) will be worth 100 points and will be graded according to the following rubric.

Final project (30%): The final project will be done in individual (or team). After mid-term exam, you will be given time to find topic of interest with writing a brief proposal, then you will do presentation of the project with written report.

Possible topics/ideas:

- Opinion mining using social media data
- Author identification using text
- Text categorization
- Movie revenue prediction

- Predicting the winners of Oscars for each category
- Predicting the winners of Golden Globe awards for each category
- Evaluating the chances of various NBA teams for winning a bracket
- Predicting election results
- Weather forecast
- Creating a rank-list for Man of the Year (Time magazine)
- Stock market (closing prices of each index/stock on x/xx)

Structure:
This is a problem-solving assignment. You will be asked to choose one of the suggested problems or find your own problem with available data, and solve the problem using the Machine Learning techniques covered in the semester. In addition to the learned tools and methods, you can employ any external resources if they are needed to solve problem(s).

You can use existing pieces of code under the following conditions:
(1) There should be at least 30% new code in your project;
(2) Any code from somewhere (including your own work) has to be attributed properly.

Time: You will be given approximately 4 weeks.

Grading: The project will be worth 100 points and will be graded according to the following rubric.

- Practicality of the topic to the phenomena in society, technology, and business area: 20
- Comprehensiveness of data processing and computation: 20
- Suitability of algorithms and methods: 20
- Proper error-checking in the code: 10
- Internal documentation (e.g., comments): 10
- External documentation (report): 10
- Class presentation: 10

Organization of the Course and Course Calendar

This course is about solving problems that fall under the larger umbrella of Data Science using Machine Learning tools and techniques. Conceptually, it is divided up in the following five broad categories. With each category, corresponding units from the schedule table are also listed.

1. Predicting from existing data
A large part of Data Science deals with using the data we have to make predictions about the future or unseen data. We will do this using concepts such as regression and gradient descent

(Unit 1). This is also a good starting point from traditional statistics or introductory data analytics course, where concepts such as correlation and regression are covered.

2. Classifying the data
This category of problems deals with figuring out how to put data points into appropriate classes. To do this, we need to learn the structure and the nature of the data. We will do this using different algorithms such as logistic regression (Unit 2), softmax regression (Unit 3), kNN (Unit 3), as well as decision trees and random forests (Unit 4). When we are dealing with data in high dimensions, techniques such as SVM (Unit 10) and dimensionality reduction (Unit 11) will come in handy.

3. Organizing and explaining the data
Often, it's not clear how the data is or can be classified. But we still want to understand the underlying structure of the data. We can do this using clustering or density estimation (Unit 6).

4. Creating estimations using the data
Many real-life applications require us to deal with hidden or missing data. Sometimes the data is not all static and new data may be coming sporadically. To deal with such situations, we need to estimate, evaluate how well we are doing, and repeat this process. Bayes' theorem (Unit 5) and Expectation-Maximization (EM) (Unit 9) are very popular techniques to do this.

5. Finding patterns from the data
Humans are quite effective in finding patterns in the world around them. Can we build machines that could do this at least reasonably well? What better way than to try to mimic how a human brain works? This is where we study reinforcement learning (Unit 11) and neural networks (Unit 12).

#	Day	Learning Objectives	Topics	Readings	Assignments
1.	09/06	● Explain data-driven decision-making ● Discuss learning methods for problem-solving	● Introduction ● Basics of differential equations ● Regression ● Gradient Descent	● Machine Learning: Two-page intro, Primer ● R Primer ● Intro to calculus: Differentiation, Partial derivatives ● Gradient descent and linear regression	Assignment-1
2.	09/13	● Apply probability theory to classification problems	● Basics of probability theory ● Logistic regression	● Introduction to probability ● Introduction to logistic regression	Assignment-2

3.	09/20	● Construct generalized classification approach using linear algebra	● Basics of linear algebra ● Generalized Linear Model (GLM) ● Classification with softmax regression ● Classification with kNN	● Linear algebra review ● Softmax regression	Assignment-3
4.	09/27	● Create human-readable decision tree using classification ● Solve classification problems without overfitting the data	● Decision trees ● Random forests	● Tutorial on decision trees and random forests	Assignment-4
5.	10/04	● Use conditional probability to classification problems	● Basics of conditional probability ● Bayesian decision theory	● Introduction to conditional probability ● Short explanation of Bayes' theorem	Assignment-5
6.	10/11	● Demonstrate how data can be organized in clusters ● Quantify likelihood of data discovery using density estimation	● Clustering ● Introduction to density estimation	● Introduction to clustering with k-means	Assignment-6
7.	10/18	● Review machine learning techniques for solving data-driven problems ● Solve data-extensive problems using machine learning techniques	*Review and practice*		Assignment-7
8.	10/25	● Review machine learning techniques for solving data-driven problems ● Solve data-	*Mid-term exam*		--

		extensive problems using machine learning techniques			
--	11/01	*No class – instructor away for ASIST 2017 Conference*			
9.	11/08	● Apply a method of discovering patterns with hidden or missing data	● Expectation-Maximization (EM)	● Introduction to EM	Assignment-8
10.	11/15	● Perform classification with high dimensional data	● Support Vector Machine (SVM)	● Beginner's guide to SVM ● SVM in R	Assignment-9
--	11/22	*No class – Friday schedule*			
11.	11/29	● Identify prominent features in high dimensional data ● Explain reinforcement learning	● Dimensionality reduction ● Introduction to reinforcement learning	● Tutorial for reinforcement learning	Assignment-10
12.	12/06	● Experiment with human-like learning using machines	● Neural networks	● Basic neural network theory	Assignment-11
13.	12/13	● Recognize other potentials of machine learning in Data Science	● Introduction to advance topics (e.g., Deep Learning, HMM)	● Primer on deep learning	--
14.	12/20	● Demonstrate the use of machine learning in real-life data-driven problem-solving	*Project presentations and class wrap-up*		Final project

Late Submission Policy

Unless otherwise noted, all written assignments, group projects, etc., are due at the time and date listed in the syllabus. If you experience an unavoidable personal situation that prevents you from completing work on time, please inform the instructor prior to the date the work is due. Late work will result in points taken off, a lowering of the assignment grade, and/or an "F," depending on the assignment.

Attendance and Participation Policy

It is University policy (University Regulation on Attendance, Book 2, 2.47B, formerly 60.14f) to excuse without penalty students who are absent from class because of religious observance, and to allow the make-up of work missed because of such absence. Examinations and special required out-of-class activities shall ordinarily not be scheduled on those days when religiously observant students refrain from participating in secular activities. Absences for reasons of religious obligation shall not be counted for purposes of reporting. Students are advised to provide timely notification to instructors about necessary absences for religious observances and are responsible for making up the work or exams according to an agreed-upon schedule.

Students are expected to attend all classes. If you expect to miss one or two classes, please use the University absence reporting Website https://sims.rutgers.edu/ssra/ to indicate the date and reason for your absence. An email is automatically sent to the instructor. Note that if you must miss classes for longer than one week, you should contact a dean of students to help verify your circumstances. Also note that class participation accounts for 5% of the final grade (see the grading policy above). You are responsible for obtaining any material that might have been distributed in class the day when you were absent.

Communication

For emails, Rutgers accounts preferred. Always include your name (esp. if emailing from non-Rutgers account) and always include the course number in subject line. If you don't, your email most likely will not be read. This course uses Sakai, primarily for submitting assignments and posting grades. Speaking of communication, please turn off or silent your cellphones and anything that can spontaneously make noise before entering the class.

Library Resources

Rutgers University Libraries offer numerous resources to assist students. Librarians can help guide you through research and reference tools. A series of LibGuides are available to get you started. Here are some of the LibGuides you may find useful:

Introduction to Rutgers University Libraries
(http://libguides.rutgers.edu/intro)
Congressional Research
(http://libguides.rutgers.edu/congress)
Communication Studies
(http://libguides.rutgers.edu/cat.php?cid=25866)
Government Information
(http://libguides.rutgers.edu/cat.php?cid=25881)
Journalism and Media Studies
(http://libguides.rutgers.edu/cat.php?cid=34201)
Law
(http://libguides.rutgers.edu/cat.php?cid=25854)
Library and Information Science

(http://libguides.rutgers.edu/cat.php?cid=25870)
Political Science
(http://libguides.rutgers.edu/cat.php?cid=25871)

Academic Integrity

The consequences of scholastic dishonesty are very serious. Rutgers' academic integrity policy is at http://academicintegrity.rutgers.edu/. Multimedia presentations about academic integrity may be found at http://www.scc.rutgers.edu/douglass/sal/plagiarism/intro.html and http://wps.prenhall.com/hss_understand_plagiarism_1/0,6622,427064-,00.html

Academic integrity means, among other things:

- Develop and write all of your own assignments.
- Show in detail where the materials you use in your papers come from. Create citations whether you are paraphrasing authors or quoting them directly. Be sure always to show source and page number within the assignment and include a bibliography in the back.
- Do not look over at the exams of others or use electronic equipment such as cell phones or MP3 players during exams.
- Do not fabricate information or citations in your work.
- Do not facilitate academic dishonesty for another student by allowing your own work to be submitted by others.

If you are doubtful about any issue related to plagiarism or scholastic dishonesty, please discuss it with the instructor.

Students with Disabilities

(For undergraduates) Students with documented disabilities who need accommodations in this class must do so through the Rutgers Disabilities Services Office. See http://disabilityservices.rutgers.edu/ for details. You can also speak with a SC&I adviser about your disability by visiting the Office of Student Services in the SC&I Building, Room 214 or calling them at 848-932-7500 (dial 2 as your menu choice).

(For graduate students) Students with documented disabilities who wish accommodations in this class must do so through the Rutgers Disabilities Services Office. See http://disabilityservices.rutgers.edu/ for details. SC&I Assistant Dean Kevin Ewell (kevin.ewell@rutgers.edu) will coordinate your services locally. Student who develop disabling medical problems or other issues during the semester that affect your ability to complete coursework should request advising from Lilia Pavlovsky (pavlovsk@comminfo.rutgers.edu) or SC&I Assistant Dean Kevin Ewell (kevin.ewell@rutgers.edu).

How to Succeed in this Course

- Successful students will attend class regularly. If you know you must miss a class, please contact the instructor in advance, either by phone or email. You can obtain assignments

or notes from a fellow classmate or from the instructor. In the case of a prolonged absence from class, you should schedule an appointment with the instructor so we can discuss the course material and concepts that you missed.

- Successful students will pay close attention to the course goals and objectives, because they will help you master the course material. If you have any questions about any of the objectives, please ask the instructor. Questions are encouraged during class for clarification. Remember that you're probably not the only one in the class with the same question. If you have questions about material from previous classes, please email me prior to the next class session, and the instructor will address your question at the beginning of the class session, prior to any exams.
- Successful students will talk to their classmates about the course material. You will find that they can help you understand many complex issues.
- Successful students will come prepared to the class with assigned readings for that class. This will help you comprehend the material for that class better. Regular assignments will also be given at the end of each class. Doing these assignments and turning them on time (typically before the next class), will help you obtain higher-order learning goals for this course.

Professionalism

- Access the class material promptly and on time.
- Respect yourself, classmates, and the instructor.
- Participate in class discussions.
- Display preparedness for class through completing reading assignments.
- Present content knowledgeably with supported reasoning.

Biographical Information about the Instructor

Chirag Shah is an Associate Professor in both the School of Communication & Information (SC&I) and the Department of Computer Science at Rutgers University. His research interests include information seeking/retrieval in social and collaborative contexts. Dr. Shah received a PhD in information science from the University of North Carolina (UNC) at Chapel Hill. He directs the InfoSeeking Lab at Rutgers where he investigates issues related to information seeking, interactive information retrieval, and social media, supported by grants from National Science Foundation (NSF), Institute of Museum and Library Services (IMLS), Google, and Yahoo! He also serves as a consultant to the United Nations Data Analytics on various Data Science projects involving social and political issues, peacekeeping, climate change, and energy.

Weather and Other Emergencies

The university rarely cancels classes for inclement weather. To check if classes are cancelled, visit http://campusstatus.rutgers.edu/. You can also try to call 732-932-7799. During severe weather conditions, announcements are made over the following radio stations: WCTC (1450AM), WMGQ (98.3FM), WRSU (88.7FM), WMCA (570AM), WOR (710AM), WCBS (880AM), WABC (770AM), WBGO (83.3FM), WHWH (1350AM), WPST (97.5FM), WJLK (131OFM), WMTR (1250AM).

Other Information to Keep in Mind

Students are expected to take the initiative to become aware of university policies and services that will help them succeed in their academic work. You are responsible for following the guidelines specified in the university's academic integrity policy, procuring information literacy skills needed to succeed in academics, seeking advisement when needed, and taking advantage of support services.

Students seeking help with the content of this course should contact the instructor either during office hours, or make a separate appointment. Students seeking help with the scheduling of classes or registration should contact the SC&I Student Services Office in Room 214 of the SC&I Building. Check here for contact information: http://comminfo.rutgers.edu/student-services/contact-us.html.

A great deal of information is available on the SC&I Website, including course descriptions and details about all degree programs: http://comminfo.rutgers.edu.

Rutgers has Learning Centers on each campus where any student can obtain tutoring and other help; for information, check http://lrc.rutgers.edu/ Rutgers also has a Writing Program where students can obtain help with writing skills and assignments: http://plangere.rutgers.edu/index.html

SC&I IT Services offers help with a variety of technology problems. They are located in the SC&I Building in Room 120 (first floor); 848-932-5555; help@comminfo.rutgers.edu .