MSGC 310 Statistical Models for Business Analytics (Intro to Machine Learning)
Course Syllabus
Fall 2019
Version 1 (last edited 8/26/19)

Instructor: Jonathan Hersh, Ph.D
Assistant Professor, Economics and Management Science
Argyros School of Business

Teaching Assistants: Sam Webster (swebster@chapman.edu)
Muhammad Karkoutli (karko100@mail.chapman.edu)

Location: Tuesdays and Thursdays, 11:30am – 12:45pm (Section 2)
Tuesdays and Thursdays, 4:00pm – 5:15pm (Section 3)

Office: BK 307G

Office Hours: Thursdays, 2:00pm – 4:00pm and by appointment

TA Office Hours: Thursdays, 5:00pm – 7:00pm, Beckman 303
Fridays, 10am – 12:00pm, Beckman 303

Contact Information: Email: hersh@chapman.edu
Website: jonathan-hersh.com

Course Description

This is a course in how to use data and analytical models to learn actionable information for businesses. Data science and machine learning are all the rage – and these skills are being highly compensated by firms – but in this class we will learn actual technical skills so that you can have a foundation in this growing area. We will cover basic statistics and machine learning models, and learn how to implement them in the R programming language.

This will be a class where coding is required. If you have no coding experience – not to worry. This class is designed to teach you to code in R. R is a powerful language, and with user written extensions called packages, the capabilities are continually expanding. It also has a very active and friendly user community. To make the language more accessible you may find it useful to install RStudio as a front end GUI (or graphical user interface). Did I mention both R and RStudio are free? You can download R at the following link: http://cran.stat.ucla.edu/ and Rstudio here: https://www.rstudio.com/products/rstudio/download/

Course Learning Outcomes

After successfully completing this course you will be able to:

- Compute and interpret descriptive statistics in R using the tidyverse
- Perform basic data manipulation in R using dplyr
- Demonstrate working knowledge of basic matrix algebra
- Build, analyze and interpret linear regression models in R

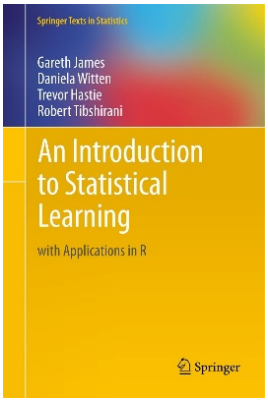
- Build, analyze and interpret binary prediction models such as logistic regression in R
- Articulate the importance of and implement resampling techniques such as bootstrapping and cross-validation
- Build, analyze and interpret regularized regression models (Lasso, Ridge and Elastic-net) for model selection in R
- Build, analyze and interpret tree-based methods (regression trees, random forests, extreme gradient boosted trees)
- Execute and understand simple unsupervised learning techniques (Principle component analysis and k-means clustering)

Course Prerequisites

At least one of the following:
MGSC 209-Introduction to Business Statistics
MATH 203-Introduction to Statistics
PSY 203-Statistics for the Behavioral Sciences
BUS 603-Business Statistics
BUS 609-Business Analytics

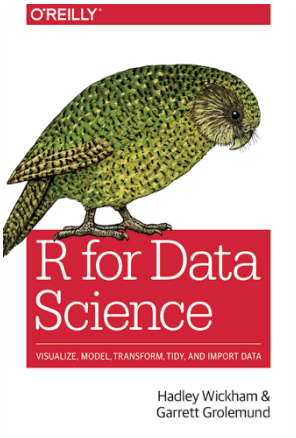
Readings

Textbooks:



James G., Witten D., Hastie T., and Tibshirani, R. (2013). An introduction to statistical learning. New York, NY: *Springer Science and Business Media*. 978-1-4614-7137-0.

This is a wonderful introductory machine learning text. It will cover the details of the statistical models and give some coding examples. It is available as a free download using the link here: http://www-bcf.usc.edu/~gareth/ISL/



Grolemund, Garrett, and Wickham, Hadley (2017). R for Data Science, second edition.

While the above book is excellent in describing the algorithms and presenting their statistical properties, it is not the best for learning how to code in R. That is the purpose of this book. Even if you know R, or have used it, we're going to be learning some advanced features that make coding in R fun, easy, and very powerful. This book is also freely available here https://r4ds.had.co.nz/

All books will be placed on one hour reserve at the Leatherby Libraries reserve desk.

Software
If you want to use your personal laptop, please install the following software:

1. R Studio v.1.2.1335
    - https://www.rstudio.com/products/rstudio/download/#download
2. R 3.6.1
    - Windows: https://cran.r-project.org/bin/windows/base/
    - Mac: https://cran.r-project.org/bin/macosx/
3. Miktex:
    - http://miktex.org/download

Even if you have already installed R/R Studio please ensure you have these versions installed!

You may also use the free, cloud based RStudio server at https://rstudio.cloud/

Software issues is not sufficient excuse for late/unfinished problem sets.

Course Format
The course will consist of a mix of lectures and interactive coding exercises that we will do together. You are welcome to use the computers in the lab or bring your own laptop.

TAs will hold office hours in the days before the problem set is due. Please attend these sessions!

Evaluation
Course Term Project (35%): **In self-assigned groups of two to three you will choose a real-world setting in which to apply some of the techniques learned in this class.** You may self-select both your groups and the topic of study. However, I may suggest some datasets to work with to get you started. Appropriate topics include, "What are the characteristics of movies that do well at the box office?"; "Assessing demand for food trucks based on neighborhood characteristics?".

There will be two components to this term project 1) a presentation, and 2) a brief report. Additional details on the term project will be provided in a second, separate document presented later in the term.

"Midterm" Exam (25%): **A single, comprehensive modeling skills exam will account for 25% of your grade. The exam is cumulative in that you may be required to demonstrate all skills learned up to the time of the exam.**

In the exam, students will be responsible for an analysis of a management problem requiring modeling to provide insight into the problem. **The exam will be in-class, spanning two days, and will be open book, notes, Internet, etc. The only restriction being you may not share answers or code among students, and you may not copy and paste any answers found anywhere else. Students will need to demonstrate use of the R software to complete the exam.** Additional information on the exam will be made available prior to the exam.

Problem sets (20%): To learn how to code, one must spend time coding. To aid in your learning process, we will have near weekly problem sets, where we will apply skills learned in the class.

Problem sets are due at 11:59pm on Friday and must be submitted via Blackboard. If a particular problem set requires compilation using RMarkdown and a compiled file is not submitted, I will penalize the grade on this problem set by 50%. I will drop the single lowest homework score from your final evaluation. Late problem sets will be penalized 50% for each day late.

Students may and are encouraged to work on homework assignments together. Assignments may be submitted in groups of up to two. Assignments that do not represent the student's own work will be awarded a score of 0. This includes copying and pasting other students' code. You are encouraged to learn from each other, however, you must write your own code. I may use plagiarism detection software to detect whether students work is indeed their own.

Quizzes (10%): **Weekly quizzes will ensure students are keeping up with the material presented in the class.** These quizzes are meant to be brief and will take approximately 10-15 minutes at the beginning of the class period. The lowest quiz score will be dropped.

Quizzes will be held every Tuesdays. Students will be responsible for providing definitions for terminology, describing concepts, reading computer results, answering basic analysis questions and, perhaps very straight-forward computations. These quizzes will generally be in multiple choice or short answer format. The quizzes will be closed book and notes.

Participation (10%): Participation counts for the remainder of your grade**. To receive full points for participation, you should come to class, and be engaged with the material.** Students who are continually distracted by matters outside the classroom, including texting and email, will receive no points for participation. **If you come to class and are disengaged, your grade will reflect it. On many occasions we will work on exercises together in-class, and you will be required to upload your code, which will factor into your participation grade.**

Exam Dates

Midterm exam: on or about October 29th and October 31st (in class)

Grade Disputes

All student evaluation on problem sets and quizzes will be made available to students within one week of submission. Students are permitted to dispute a score on an assignment they have received. **Students are expected to dispute the evaluated score within the one week period of the availability of the evaluation. Failure to dispute a grade within this period indicates that the student has accepted the mark.**

First round disputes should be made in writing via email or in other written form to me, e.g. comments written on the returned homework assignment. Disputes should include a convincing evidence as to why a particular mark should be changed. Convincing evidence should be logical and concise. If the dispute is not resolved to the student's satisfaction, the student may request an in-person meeting to further discuss the issue.

Final Course Evaluation

Final course letter grades will be assigned on the following scale:

A	95-100 %	C	73-76.9%
A-	90-94.9%	C-	70-72.9%
B+	87-89.9%	D+	67-69.9%
B	83-86.9%	D	63-66.9%
B-	80-82.9%	D-	60-62.9%
C+	77-79.9%	No Pass	<60%

Course Resources

Website: I will use Blackboard to distribute course materials.

Office hours: In general, I encourage you to come visit me during office hours. Note: in the event that I am traveling and need to cancel office hours, I will reschedule accordingly.

Email: I will do my best to respond in a reasonable time. If I have not replied in 24 hours, please email again as it might have slipped my notice.

Slack: Slack is an "asynchronous communication" platform that is very popular in the tech world. You can use this workspace to ask me questions about problem sets, to ask students to help you with coding bugs, or to communicate with your group members. You do not need to use Slack, but you may find it helpful.

To sign up for our Slack workspace use the following link: https://tinyurl.com/310slack19

TAs: The TAs are available to answer your coding questions! Email them first as I might not respond as timely as they can.

Poll Everywhere: We will on occasion use Poll Everywhere software. My polls can be reached via PollEv.com/hersh

Technology Policy

My technology policy can be summarized as follows: *I ask that you use technology thoughtfully in the classroom*. Technology is wonderful, but the careless use of it can distract others and prevent you from being fully engaged with the material. Please completely silence your phones – *note, vibration mode is not silent* – and resist being distracted by texting, or viewing Facebook or other distracting website during class. If I see that you are using technology for any purposes other than for course material, you will receive no points for class participation on that day.

Classroom Etiquette

Classroom is a sacred space. We should all strive to respect that space. This course will be hard at times, and that can be frustrating. I believe students learn best when they feel comfortable and can struggle with the material without fear of criticism from either the instructor or fellow students. Civility of peers

and your instructor is expected at all times, both in the class and outside of it. Please come speak to me if you feel this is not being met.

Chapman University Academic Integrity Policy

I take academic integrity very seriously. I am obligated to report any evidence of violations of Chapman's policy on academic integrity to the dean's office.

"Chapman University is a community of scholars that emphasizes the mutual responsibility of all members to seek knowledge honestly and in good faith. Students are responsible for doing their own work, and academic dishonesty of any kind will be subject to sanction by the instructor and referral to the university's Academic Integrity Committee, which may impose additional sanctions up to and including dismissal. (See the Undergraduate Catalog for the full policy.)
"Academic dishonesty can take a number of forms. It includes, but is not limited to, cheating on a test or examination, claiming the work of another as your own, plagiarizing any paper, research project or assignment (to appropriate for use as one's own passages or ideas from another), or falsely submitting material to fulfill course requirements."

Chapman Policy on Students with Disabilities

In compliance with ADA guidelines, students who have any condition, either permanent or temporary, that might affect their ability to perform in this class are encouraged to inform the instructor at the beginning of the term. You also may contact the Office of Disability Services at www.chapman.edu/disabilities. The University, through the Disability Services Office, will work with the faculty member who is asked to provide the accommodations for a student in determining what accommodations are suitable based on the documentation and the individual student needs. The granting of any accommodation will not be retroactive and cannot jeopardize the academic standards or integrity of the course.

Important Addresses and Telephone Numbers

Disabilities Services:	Tutoring, Learning, and Testing Center:
410 N. Glassell	Cecil B. DeMille Hall 130
Phone: (714) 997-6778	Phone: (714) 997-6828

Chapman's Diversity Policy

Chapman University is committed to ensuring equality and valuing diversity. Students and professors are reminded to show respect at all times as outlined in Chapman's Harassment and Discrimination Policy. Please see the full description of this policy at http://www.chapman.edu/faculty-staff/human-resources/eoo.aspx. Any violations of this policy should be discussed with the professor, the dean of students and/or otherwise reported in accordance with this policy.

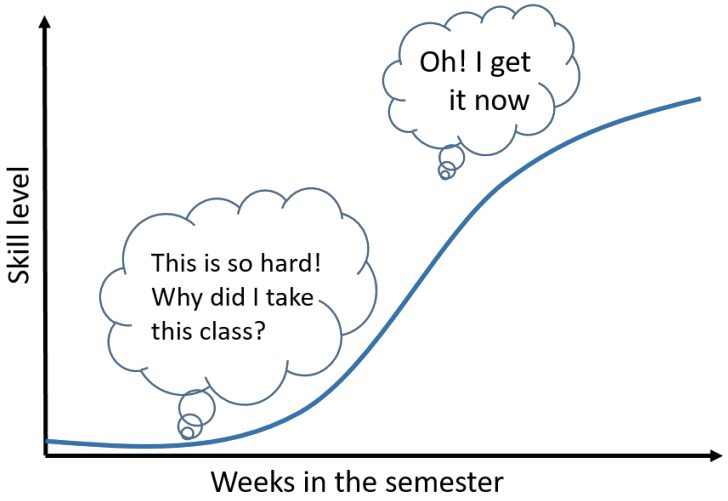
FAQS:

I'm worried I don't have enough coding or math background to be successful in this course. Should I still take the course?

Yes! This course is designed such that anyone with the prerequisites can be successful and can still get a good grade. The more math, statistics, and programming experience you have, the easier you might find this course. However, we all have to start somewhere, and this course is designed to get you started and dangerous regardless for where you are at.

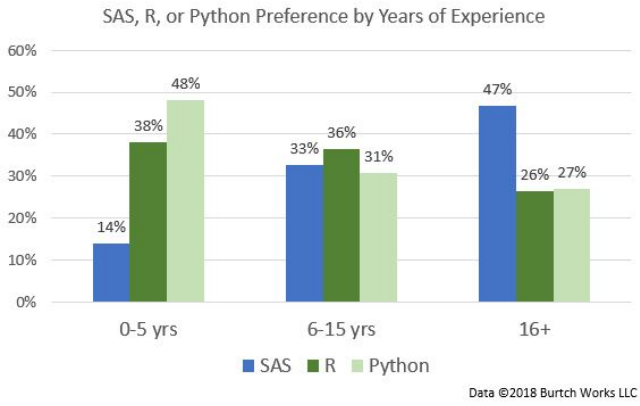
That's good…but I'm worried about my grade. I want to take this course, but I also don't want my GPA to get crushed.

I understand and sympathize with the pressures of keeping your grades up. Many people in this class have a learning curve that looks like the image to the right. Initially things are very difficult, and this period can last for a month or two. All of a sudden, things 'click', and the student makes rapid progress. Note that the grading structure is very generous if this is your trajectory. I drop two of your lowest problem sets and quizzes. And the largest component of your grade is the final project, which will take place towards the end of the semester. So if much of this is new and you are worried about your grade, this class is structured so that you can be successful.



Why are we learning R?

R is one of the most popular tools for data science (see bottom left). It has a robust, active user community that is surprisingly friendly and open to newbies. It is also one of the most well compensated skills to learn (see bottom right).



AVERAGE SALARY FOR High Paying Skills and Experience

SKILL	2013	YR/YR CHANGE
R	$ 115,531	n/a
NoSQL	$ 114,796	1.6%
MapReduce	$ 114,396	n/a
PMBok	$ 112,382	1.3%
Cassandra	$ 112,382	n/a
Omnigraffle	$ 111,039	0.3%
Pig	$ 109,561	n/a
SOA (Service Oriented Architecture)	$ 108,997	-0.5%
Hadoop	$ 108,669	-5.6%
Mongo DB	$ 107,825	-0.4%

I'm much better at Stata/Python/SAS/Excel? Can I use these instead?

It's wonderful to know multiple coding languages. Certain languages have their strengths. Python is excellent deep learning and data munging. R is great for implementing the latest statistical package. SAS is good for…paying a lot for a license. With the exception of SAS I regularly use all of these tools and think they have their place. However, because of coordination costs, we need to use one language together. R is great for beginners and experts alike, so we will use it in this course. However, I certainly encourage you to learn as many languages as you can, including Stata/Python/Excel and even SAS.

Argh! I'm stuck with my code and quite frustrated. What should I do?

Your first resource should be your textbooks. Read the topic again, this time more slowly. Sometimes it just takes two or three readings (or four or five) to really understand a topic. Your second resource should be your fellow students. Have friends in your class that you can ask questions. Ask a question to the classes' Slack channel. Often there are very small mistakes that can be frustrating and take a while to debug. If you ask them, they might know and be able to spot the error. Your third resource should be the TAs. You can email them anytime and ask questions, and please attend their office sessions. Your fourth resource should be online resources and documentation. Sites like www.stackoverflow.com, www.statmethods.net, and others are a wealth of resources. Don't forget you can always read the package/function help by typing `help(`*`package or function name`*`)` into the console. Your fifth resource is me. Please send me an email when you have exhausted all the above possibilities.

I wish there were an easier way, but this is just the process of learning how to code. Even I get stuck for hours on some small part of my code. My advice is to start your problem sets early and have smart people you can ask when you get stuck.

To be good at this, you just have to be gritty:

https://www.ted.com/talks/angela_lee_duckworth_grit_the_power_of_passion_and_perseverance?language=en

Date	Topic	Reading	Assignment (Problem Sets Due Friday)
Tue, Aug 27	Intro/Welcome		
Thu, Aug 29	Data Visualizations and Rmarkdown	RFDS: Chapters 1-3	
Tue, Sep 3	Basic Data Transformations and Exploratory Data Analysis	RFDS: Chapters 4-7	Quiz 1
Thu, Sep 5	Bias-Variance Tradeoff	ISLR: Chapters 1-2	Problem set 1
Tue, Sep 10	Linear Regression 1: Regression Review	ISLR: Chapter 3	Quiz 2
Thu, Sep 12	Linear Regression 2: Model Building and Extensions	ISLR: Chapter 3	Problem set 2
Tue, Sep 17	Linear Regression 3: Model Building and Extensions	ISLR: Chapter 3	Quiz 3
Thu, Sep 19	Classification 1 : Logistic Regression	ISLR: Chapter 4	Problem set 3
Tue, Sep 24	Classification 1 : Logistic Regression	ISLR: Chapter 4	Quiz 4
Thu, Sep 26	Classification 2: Binary Model Diagnostics	ISLR: Chapter 4	Problem set 4
Tue, Oct 1	Classification 2: Binary Model Diagnostics	ISLR: Chapter 4	Quiz 5
Thu, Oct 3	Resampling Methods: Bootstrap and Cross-Validation	ISLR: Chapter 5	Problem set 5
Tue, Oct 8	Linear Selection and Regularization 1: Lasso	ISLR: Chapter 6	Quiz 6
Thu, Oct 10	Linear Selection and Regularization 2: Ridge	ISLR: Chapter 6	Problem set 6
Tue, Oct 15	Linear Selection and Regularization 3: ElasticNet	ISLR: Chapter 6	Quiz 7
Thu, Oct 17	Tree Based Methods: Decision Trees	ISLR: Chapter 8	Problem set 7
Tue, Oct 22	Tree Based Methods: Bagging	ISLR: Chapter 8	Quiz 8
Thu, Oct 24	Tree Based Methods: Random Forests	ISLR: Chapter 8	Problem set 8
Tue, Oct 29	In Class Exam Day 1		
Thu, Oct 31	In Class Exam Day 2		
Tue, Nov 5	Tree Based Methods: Boosting	ISLR: Chapter 8	
Thu, Nov 7	Unsupervised Learning: Clustering	ISLR: Chapter 10	
Tue, Nov 12	Unsupervised Learning: PCA	ISLR: Chapter 10	Quiz 9
Thu, Nov 14	Tidy Data, Merging Datasets, Strings and Factors	RFDS: Chapter: 12-16	Problem set 9
Tue, Nov 19	Caret Package	Barter, Rebecca. "A basic tutorial of caret"	Quiz 10
Thu, Nov 21	Project Presentations		
Tue, Nov 26	Thanksgiving Recess (no class)		
Thu, Nov 28	Thanksgiving Recess (no class)		
Tue, Dec 3	Project Presentations		
Thu, Dec 5	Project Presentations		
Finals Week	Final projects due December 14th at midnight		Final project due

[This page intentionally left blank]

A note on cheating:

- ***Copying and pasting code from other students in your class is cheating***
- ***Copying and pasting code from solutions you found online is cheating***
- ***Copying and pasting code from a previous course is cheating***
- ***Posting exam/problem set solutions to any online repository or website is cheating***

Copying and pasting is not acceptable in this class. You will fail this class if I find evidence you are cheating.

I completely encourage you to help and work with each other. Just don't allow students to copy and paste your answers.

I understand that copying and pasting code from any source is cheating. I agree to do honest work in this class.

_____

(Sign Name)

_____

(Print Name)