Advanced Methods for Data Analytics

DA-350

Spring 2019
M – W- F 9:00am – 10:20am

Burton Morgan 219

Course Description

This course is designed to develop students' understanding of the cutting edge methods and algorithms of data analytics and how they can be used to answer questions about real-world problems. These methods, and the underlying models, can be used to learn from existing data to make predictions about new data. The course will examine both supervised and unsupervised methods and will include topics such as clustering, classification, and network analysis.

Instructor

Anthony Bonifonte
Office: Burton Morgan 410
Office Hours: Mon, Tu 1:30-4:00, Wed 1:30 – 3:00
If you are unable to attend these hours, you are encouraged to email questions or schedule an appointment.
Email: bonifontea@denison.edu
TA: Anamay Agnihotri, agniho_a1@denison.edu
TA office hours: TBA

Course Goals

At the end of the course, students should be able to:

- Apply advanced data analytics methods to solve real world problems
- Generate predictions for classification and regression problems using cutting edge machine learning techniques
- Evaluate the performance of predictions and choose between potential models
- Prescribe actions to take by solving optimization problems with data
- Handle missing data with both simple and sophisticated techniques
- Describe structures in data using unsupervised algorithms
- Reduce dimensionality of data for easier storage, computation, and analysis
- Understand and implement analytics algorithms from primary sources and documentation
- Estimate algorithm run times and efficiently handle large computations

"The world is one big data problem."

– Andrew McAfee

Course Logistics

Prerequisites: CS 181
and DA301 (co-requisite)

Textbook: An Introduction to Statistical Learning by James et. al.
Free pdf: http://www-bcf.usc.edu/~gareth/ISL/
Printed copy: https://link.springer.com/book/10.1007/978-1-4614-7138-7

"A breakthrough in machine learning would be worth ten Microsofts."

– Bill Gates

Expectations

Academic Honesty: Academic honesty, the cornerstone of teaching and learning, lays the foundation for lifelong integrity. Academic dishonesty is intellectual theft. It includes, but is not limited to, providing or receiving assistance in a manner not authorized by the instructor in the creation of work to be submitted for evaluation. This standard applies to all work ranging from homework assignments to major exams. I will assume that you are familiar with the Code of Academic Integrity. To learn more about it, please go to https://denison.edu/academics/curriculum/integrity

Class communication: All electronic communication will be through announcements using NoteBowl and delivered to your Denison email. You are responsible for checking these messages periodically to stay informed of important dates and potential changes to the syllabus.

I am pleased to reply to questions via email at bonifontea@denison.edu. I check my email frequently, but I reserve the right to a 48-hour response period. This means questions immediately before an assignment due date may not receive a timely response. Please send all emails through your Denison account

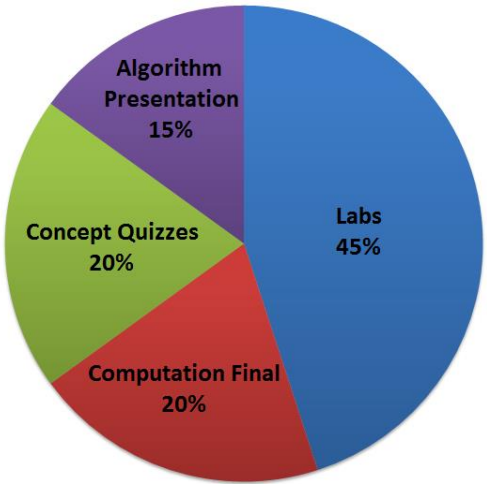
Disability Accomodations: Any student who feels he or she may need an accommodation based on the impact of a disability should contact me privately as soon as possible to discuss his or her specific needs. I rely on the Academic Resource Center (ARC) in 020 Higley to verify the need for reasonable accommodations based on the documentation on file in that office.

Technology Policy: Please be respectful with your use of laptops and technology in class. I request you only use them for class related purposes, as I and others may find them distracting. Cell phones should be kept silent and away, and you can expect the same from me.

Software: All topics in this course will be demonstrated using R. For solving optimization problems we will use Gurobi, the industry standard and most powerful optimization solver available. This software has an R plugin and we will call it from there. You can request a free license at: https://user.gurobi.com/download/licenses/free-academic

The lab computers also have this software downloaded.

Assignments and Grading



Final Course Grade:

A+:	98%	A:	92%	A-:	90%
B+:	88%	B:	82%	B-:	80%
C+:	78%	C:	72%	C-:	70%
D+:	68%	D:	62%	D-:	60%

Labs 45%

Fridays will be reserved for discussing and working on lab assignments. Lab assignments will be mostly solving applied real-world data problems in a variety of disciplines with methods from the course. Some lab problems will use simulated data to more clearly illustrate theoretical properties and examine differences between methods.

Some labs will be submitted individually, some will allow groups of 2 to work together. Discussing with fellow students is highly encouraged, although code sharing on individual assignments is not. Labs will due at the start of class the following Friday after assigned. Labs will involve computational problems and writing paragraphs interpreting the results and putting them in context.

Concept Quizzes 20%

The first 15 minutes of class each Monday will be short quizzes examining your understanding of the previous week's content. Questions will be mostly short answer and will examine your understanding of the context, usage, and theoretical basis of the methods and algorithms discussed. To succeed on these quizzes you will need to keep up with class notes and read the corresponding sections of the text.

Algorithm Presentation 15%

The fields of data analytics and machine learning are evolving at a lightning pace. You will need to learn and implement new methods many times in your career. There exist hundreds of machine learning algorithms, each with pros, cons, and peculiarities in usage and implementation, but we only have time to discuss a few.

For this assignment, you will work with two partners of your choice and pick a method from a pre-selected list. You will give a 15 minutes class presentation explaining the high level algorithm idea, a little of the theory motivating the method, and show the performance of the method on a test problem compared to methods we discussed in class. To understand and implement the method, you will need to read primary sources and the code documentation.

We will have one presentation at the start of class each Wednesday starting mid-February and continuing as long as necessary. Presentation order will be choosable based on a random assignment in class.

Computation Final 20%

A cumulative take home final assessment will test your ability to use the algorithms from throughout the semester. You will be challenged with a real-world problem to decide what types of methods to employ, how to choose between them, describing the data, making predictions, and estimating the accuracy of your predictions. You will be permitted to use any outside resources including the textbook, prepared R scripts, google searches, and existing posts on coding advice websites such as Stack Overflow.

The assessment will be distributed on the last day of class, Monday May 6, and will be due at the latest Friday May 10 at midnight (earlier submissions are acceptable). With adequate preparation, it should take no longer than 3 hours of work.

Artificial Intelligence, deep learning, machine learning—whatever you're doing if you don't understand it—learn it. Because otherwise you're going to be a dinosaur within 3 years.

— Mark Cuban

Course Schedule

Dates	Topics	Textbook Reading
1/21 – 1/25	Review : Cross Validation	2.1, 2.2, 5.1
1/29 – 2/01	Unsupervised learning: Principle Component Analysis	6.3, 10.2
2/04 – 2/08	Unsupervised learning: Clustering	10.3
2/11 – 2/15	Supervised learning – k-Nearest Neighbors	2.2.3
2/18 – 2/22	Supervised learning – two class classification	4.3
2/25 – 3/01	Supervised learning – multiclass classification	4.4
3/04 – 3/08	Supervised learning – tree methods	8.1, 8.2
3/11 – 3/15	Supervised learning – neural networks	Supplemental reading
3/18 – 3/22	Spring Break	
3/25 – 3/29	Missing Data Methods and Multiple Imputation	Supplemental reading
4/01 – 4/05	Optimization	Supplemental reading
4/08 – 4/12	Optimization Continued	Supplemental reading
4/15 – 4/19	Reinforcement learning : Multi Armed Bandit Model	Supplemental reading
4/29 – 5/03	Other topics of interest	
5/06	Wrap-Up and Conclusion	

"Doctors can be replaced by software – 80% of them can. I'd much rather have a good machine learning system diagnose my disease than the median or average doctor."

-Vinod Khosla