DATA 419 — Data Mining (Machine Learning)

Professor: Stephen Davies
Fall semester 2022

Class: TR 2pm in Woodard 132

Final exam: Thursday, December 8th, 3:30–6pm

Office Hours:

Tue	12pm–2pm, **HCC ground floor lobby**
Wed	11am–12:30pm, **James Farmer Hall 221**
Thu	3:30–5pm, **HCC ground floor lobby**

`http://stephendavies.org/data419`

On the short list for "the discovery that has most changed our 21st century world" is the exploding field of Data Mining and its closely-related kin, Machine Learning. I might not need to convince you of this, but in case I do, consider the following applications: speech recognition and autocomplete on smart phones, financial fraud detection, health care cost reduction, product recommendations, Google Translate, criminology, bioinformatics, counterterrorism, intelligent gaming, weather prediction, city planning, supply-chain routing, acceleration of scientific discovery, selection of syndicated content, creation of new types of medications, and self-driving cars. I could easily double, triple, or quadruple the size of that list without even having to think too hard, and you probably could too. Machine Learning is all around us.

And then there's the dark side. It's all over the news: Cambridge Analytica. Deep fakes. The steering of news feeds towards toxic content. The reinforcement of bias in workplace candidate screening. Cathy O'Neill's *Weapons of Math Destruction* chronicles some of the ways Machine Learning has negatively impacted society, even in its relative infancy.

The leaders in this powerful arena are sometimes called "wizards," able to do veritable magic with statistics and computers as their tools. In this course, we'll drink a good, long draught from the cauldron to see what it tastes like.

Course Objectives

- To survey some of the most important applications of Machine Learning (ML), so that you know where the various use cases and technologies "fit" within the

field.

- To tie together the two halves of Data Science. Whereas a course like 420 (Modeling & Simulation) is about creating a model of a data-generating process and seeing whether the data it generates matches the real world, 419 is about analyzing real-world data to infer properties about the process that generated it. It's kind of like 420 in reverse.
- To give you both theoretical understanding and also practical experience working with real ML implementations on real data sets.

Student Learning Outcomes

After completing this course, students will be able to...

- ...identify the different subfields within the field of Machine Learning (ML), and to explain their purposes, assumptions, promises, and limits.
- ...understand – and competently carry out – the various stages in the ML pipeline, from data acquisition all the way to the presentation of results.
- ...both:
  - comprehend the mathematical theory underpinning important ML algorithms, and
  - apply that theory and implement those algorithms concretely in a programming language (Python).
- ...fluently use the key ML-related Python libraries of NumPy, SciPy, matplotlib, and scikit-learn.
- ...explain how data mining (DATA 419) – reasoning from observed data backwards to a plausible model – is the inverse of modeling and simulation (DATA 420) – producing and analyzing the data from a simulated model.

Rules of the game

1. There are absolutely, positively, NO stupid questions!! Your job is not to already know everything before you start the course. Your job is to try hard

to learn, and part of that involves asking questions. I'm a nice guy, and I will not ever belittle you, snub you, or make fun of you; and if anyone else does so I will personally break both of their arms.

2. This class will be interactive. When I point at you in class, say your first name, and be prepared to try and answer questions. (Don't worry if you don't know all the answers.)
3. The book we'll be using (*The Hundred-Page Machine Learning Book*) is mandatory, and you will be required to actually read the sections I assign. However, the quizzes and the final exam will not cover anything from the book that I don't specifically cover in class. Also, the quizzes and the final exam *will* definitely contain lecture material beyond what's in the book.
4. Don't skip class. Just don't. It's bad form. I work hard to prepare for class, to make it compelling and relevant. It hurts my feelings when you don't come. Plus you miss out on important stuff, and you'll end up falling behind if you skip lecture. So come every time. Come happy, fresh, excited, ready to think and to participate.
5. **Absolutely no laptops, cell phones, or other devices during class** I've had students claim that they take notes on their laptop during lecture, but even if it's true, those things are way too big a distraction to you and your fellow students to make it worth it. Just stay tuned in, because I move fast.
6. For any Zoom class that may take place this semester, **you must have your webcam on during the entirety of the lecture.** If you don't have a working Webcam, buy one immediately.

The Honor Code and this course

For this course:

- The Canvas quizzes must be taken alone, in a quiet place, without any form of contact with anyone.
- **You must write all your own Python code in its entirety.** You may not copy any part of anyone else's program. That being said, it is okay to work alongside classmates in a study group, and you are allowed to talk through homework assignments, ask others how they solved a particular problem, and so forth. **If you do this, you must name all the people you worked with on your homework submission. If you fail to give this information clearly on your homework submission, it is an Honor Code violation.**

- For the programming assignments specifically, do **not** copy or share code verbatim with anyone, but you can talk general strategy with classmates, or ask classmates how they solved something. You may **not** discuss any assignment with anyone not currently in the class (which includes people not even at UMW, of course).
- Note that the above statement about study groups does ***not*** include anyone not in the course. You may not discuss the homeworks with former 419 students, other DS minors or CPSC majors, or anyone outside UMW.

Books

- Burkov, A. (2019). *The Hundred-Page Machine Learning Book.* ISBN 978-1-9995795-1-7.
- (Optional) Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd edition). O'Reilly Media. ISBN 978-1-4920326-4-9.

Burkov's as-terse-as-possible text has plenty of admirers, including some well-known ML researchers who praise it as a concise introduction to the entire field. We'll be loosely following its pages, though not in the same order Burkov presents them, as a handy reference to complement what we do in class. This book is **required**. Burkov offers it according to the "read first, buy later" principle he explains at the end of the preface. I leave that to your discretion.

I thought about assigning and requiring Géron's text, but in the end decided against it mostly because (1) we'd only be covering about a third of it, and (2) it's pricey. But I do highly recommend it. The author does a terrific job of explaining difficult concepts and keeping things practical, and I have learned a ton from reading this industry-popular book.

Late policy

No late work will be accepted this semester. Get your stuff in on time!

Grading

Grading this semester will be based on "experience points" (XP). As you complete activities, you will earn XP towards your final total. XP can never be lost, only gained, but you have to earn what you get (*i.e.*, you don't "start off with a 100%" and lose points based on mistakes you make).

There will be opportunities to earn XP throughout the course. Some of these will be spontaneous as the mood strikes me. Some you can earn by completing in-class activities. Some may be in response to impressive things I see you do as the semester progresses. The following opportunities, however, are *guaranteed* to be available to you:

Guaranteed XP opportunities:

Activity	Possible XP
Eight open-book, open-note, timed Canvas quizzes	30 each
Eight homework assignments	40 each
Two responsive readings	50 each
Final exam (comprehensive)	100
Various and sundry others	varies

Grading levels

Here are the levels you may achieve, together with the grade awarded (if any) and the points necessary to reach!

Level	**Total XP**	Semester grade
Gandalf	**700**	A+
Harry Potter	**660**	A
Nynaeve al'Meara	**600**	A–
Merlin	**560**	B+
Dr. Strange	**520**	B
Lady Jessica	**480**	B–
Morgan le Fay	**440**	C+
Kvothe	**400**	C
Professor Albus Dumbledore	**360**	C–
Elminster	**330**	D+
Moiraine Damodred	**300**	D
Melisandre	**270**	
Hermione Granger	**240**	
Raistlin Majere	**210**	
Saruman	**180**	
Egwene al'Vere	**150**	
Thufir Hawat	**120**	
The Scarlet Witch	**100**	
Harry Dresden	**80**	
11	**60**	
The Wicked Witch of the West	**40**	
Willy Wonka	**30**	
Ron Weasley	**20**	
Alex Russo	**2**	
Fruit Pie the Magician	**0**	

Submitting homeworks

Rules for submitting homeworks (whether in Python or in English) will be given when the homework is assigned. For programs, you'll be emailing me your program code as an attachment, and using a specific subject line to distinguish it from my hordes of other email. Meeting the deadline is a matter of sending your email before time expires. For responsive readings, **you must submit a hardcopy of your paper – nothing electronic will serve as an adequate substitute.**

Also, most of my homeworks are due at "midnight." Here's what "midnight" means: if a homework is due "at midnight on Thursday," then it is due after all of Thursday has elapsed, and the clock strikes twelve. (In other words, this is good news: you have all Thursday to work on it.)

Basis for determining mid-semester reports

For midterm progress reports, I look mostly at (a) whether you've been turning assignments in (and preferably on time), and (b) quiz scores. If either or both of these categories are lacking, it's a sign of danger, and I will give you a "U" for your mid-semester grade. Please don't hesitate at all to come talk to me about this so we can figure out how you can do better in the course.

Guidelines for class participation

I believe that students learn best when they participate wholeheartedly in all aspects of the learning process. Hence while your grade will not be partially determined by any "class participation score" *per se*, it is very much to your advantage, and very much recommended, that you join in during class discussions, ask questions, and make comments.

Disabilities

If you have a documented disability, please present me your letter from the Office of Disability Resources and I'll be happy to accommodate you.

Title IX Statement

UMW faculty are committed to supporting students and upholding the University's *Policy on Sexual and Gender Based Harassment and Other Forms of Interpersonal Violence.* Under Title IX and this Policy, discrimination based upon sex or gender is prohibited. If you experience an incident of sex or gender based discrimination, we encourage you to report it. **While you may talk to me, understand that as a "Responsible Employee" of the University, I must report to UMW's Title IX Coordinator what you share.** If you wish to speak to someone confidentially, please contact the confidential resources below. They can connect you with support services and help you explore your options. You may also seek assistance from UMW's Title IX Coordinator; their contact information can be found below. Please visit `http://diversity.umw.edu/title-ix/` to view *UMW's Policy on Sexual and Gender Based Harassment and Other Forms of Interpersonal Violence* and to find further information on support and resources.

Resources
Ruth Davison, Ph.D.
Title IX Coordinator
Lee Hall, Room 401
540-654-5656
`rdavison@umw.edu`

Confidential Resources
On-Campus
Talley Center for Counseling Services
Lee Hall 106, 540-654-1053

Student Health Center
Lee Hall 112, 540-654-1040

Off-Campus
Empowerhouse (24-hr hotline)
540-373-9373

RCASA (24-hr hotline)
540-371-1666

Recording Policy

Classroom activities in this course may be recorded by students enrolled in the course for the personal, educational use of that student or for all students presently enrolled in the class only, and may not be further copied, distributed, published or otherwise used for any other purpose without the express written consent of the course instructor. All students are advised that classroom activities may be taped by students for this purpose. Distribution or sale of class recordings is prohibited without the written permission of the instructor and other students who are recorded. Distribution without permission is a violation of copyright law. This policy is consistent with UMW's *Policy on Recording Class and Distribution of Course Materials.*

Disability resources

The Office of Disability Resources has been designated by the university as the primary office to guide, counsel, and assist students with disabilities. If you receive services through the Office of Disability Resources and require accommodations for this class, please provide me a copy of your accommodation letter via email or during a meeting. I encourage you to follow-up with me about your accommodations and needs within this class. I will hold any information you share with me in the strictest confidence unless you give me permission to do otherwise.

If you have not made contact with the Office of Disability Resources and have reasonable accommodation needs, their office is located in Seacobeck 005, phone number is (540) 654-1266 and email is `odr@umw.edu`. The office will require appropriate documentation of disability.

How to reach me

Come to office hours, see me after class, or e-mail me (`stephen@umw.edu`).

How to reach you

I will be communicating with you outside of class time via e-mail, so make sure to check your UMW e-mail every day! I will also post announcements to the course website, so be sure to subscribe to its RSS feed and check it in your feed reader at least once a day!

Calendar

The official calendar for the course, complete with assignment due dates, tests, *etc.*, will be maintained on the course website at `http://stephendavies.org/data419`. **In any way that the website conflicts with the tentative calendar below, the website is to be considered correct, and the tentative calendar below out of date.**

Week	Topics	Due
1	Machine Learning settings and concepts	HW 1
2	Python's ML libraries	HW 2
3	Feature engineering	
4	Linear Regression models	HW 3
5	Logistic Regression and Linear Discriminant Analysis	HW 4
6	Instance-based learning algorithms: kNN	
7	Instance-based learning algorithms: Support Vector Machines	HW 5
8	Decision Trees	RR 1
9	Random Forests	
10	Model evaluation	HW 6
11	Model tuning: regularization, hyperparameters	
12	Ensemble learning; boosting and bagging	HW 7
13	Consideration of special data types: text mining	RR 2
14	Consideration of special data types: images	
15	Unsupervised learning: clustering algorithms	HW 8

"HW"=Homework, "RR"=Responsive Reading