SI 670 - Applied Machine Learning (Tentative Syllabus)

Fall 2020: Thursday 1:00 PM - 4:00 PM (Zoom Link https://umich.zoom.us/j/93802914291)

Instructor: Grant Schoenebeck

Office: NQ 3341

Office hours: Tuesday 2-3 PM (by appointment) https://umich.zoom.us/j/99488958193
Tuesday 3-4 PM public https://umich.zoom.us/j/98442586030

Course Assistants: Teng Ye (Office hours: Thursday 9:30 - 11:30 am, Zoom link)

Zhuofeng Wu (Office hours: Friday 9:30-11:30 am, Zoom link)

Contact: The course has a discussion board using Piazza set up on Canvas (sign up at: https://piazza.com/umich/fall2020/si670) – all questions regarding course content or that might be of interest to other students, in particular, programming or course logistics questions, should be posted to the Piazza site.

If your question requires confidentiality or is only pertinent to you, please use email.

Staff Email: SI670staff@umich.edu (use unless you need a particular instructor)

schoeneb@umich.edu

tengye@umich.edu

zhuofeng@umich.edu

Course IAs (TBD)

Note: syllabus details subject to change.

Course Description:

In this course, students will learn basic machine learning concepts and methods, along with how to select and apply these methods correctly to solve supervised and unsupervised machine learning problems on real-world datasets. The class focuses more on application than theory and is based on the scikit-learn (for classical ML) and Keras (for Deep ML) libraries in Python.
The course will start with a discussion of how machine learning is different than descriptive statistics, and introduce the scikit learn toolkit through a combination of lectures and labs. The course will introduce supervised approaches for creating predictive models, and students will be able to apply the scikit to learn predictive modelling methods while understanding process issues related to data generalizability (e.g. cross validation, overfitting). Students will also learn effective dimensionality reduction and clustering methods (unsupervised learning) to help find structure in data. The course will end with a look at more advanced techniques, such as building ensembles, deep learning, and practical limitations of predictive models. By the end of this course, students will master basic concepts of supervised (classification) and unsupervised (clustering) techniques, identify which technique they need to apply for a particular dataset and problem, engineer features to meet that need, and correctly evaluate and interpret results from their approach.

Where this course fits in the curriculum

While excellent courses in applied statistics and computer science exist that teach many of the same machine learning concepts, these courses also tend to focus more on theory rather than application, or topics that are not as immediately necessary for successful application to important real-world problems. They also include more advanced mathematical detail, and less of the practical programming focus or diversity of applications that would make this course especially relevant to SI students.

Learning Objectives

***Competency*:**

- Be able to train and apply regression and classification objects (estimators) in scikit-learn.

- Understand how to correctly prepare input data for use, e.g. feature normalization.
- Understand how to evaluate and interpret results from scikit-learn estimators.
- Understand over- and under-fitting and how to detect and prevent these.
- What data leakage is and how to detect it.
- Use model selection methods such as cross-validation to tune the choice of model and key parameters.
- Understand how to select an appropriate machine learning method for a given scenario and dataset.

Literacy:

- Understand the tradeoffs inherent in different machine learning methods: speed, accuracy, complexity of hypothesis space, etc.
- Be able to optimize a classifier for a variety of metrics.
- Be aware of how deep learning methods work and how they are applied.

Awareness:

- Issues of algorithmic bias, transparency, fairness, and other social context in machine learning applications.
- Be aware of different ensemble methods for combining classifiers.

Textbooks and Course Resources

The course will be structured in a way similar to the following two recommended textbooks. They are available for free online to UM students:

Introduction to Machine Learning with Python. A. Mueller and S. Guido. O'Reilly.

http://proquest.safaribooksonline.com/book/programming/machine-learning/9781449369880/firstchapter

Off-campus: http://proquest.safaribooksonline.com.proxy.lib.umich.edu/

The Deep Learning section of the course will be using
Deep Learning with Python, by Francois Chollet Manning.

https://proquest.safaribooksonline.com/book/programming/python/9781617294433

Additional material will be provided in lectures, readings, and other handouts.

Another resource is the Applied Machine Learning Coursera course (Course 3, available for free) as part of the UM Applied Data Science in Python Specialization - which provides videos, background material and examples and can be accessed for free here: https://online.umich.edu/catalog/69/

If you need help getting up to speed on pandas and numpy, you can take Course 1: Introduction to Data Science in Python (also for free)
https://online.umich.edu/catalog/50/

Class Format

The course will combine interactive lectures with in-class lab sessions. **You'll need to bring your laptop to each lab.** Students learn the concepts and techniques in the lecture and practice them by writing programs in the lab portion of the class. In addition to the labs, there will be regular programming assignments and a midterm. Students will use Python with Jupyter / Colab for most of the labs and homework assignments. There will be a few paper discussions. A key part of the class will be a final project.

Course Schedule

Key Dates:

September 3, 2020: First Class.
Oct 15, 2020: Midterm
Nov 5, 2020: Proposal for Final Project Due

Dec ?: Final Project Poster Session
Dec 3: Final Project Deliverables Due, Last Day of Class

Data	Topics

Sept 3	Week 1:
	
	Outline:
	What is ML?
	Key Aspects of Course
	Types of ML
	ML Process
	Data exploration (review)
	ML Process Example
	  K-NN and Accuracy
	  Go Through Process.
	Feature Normalization

Sept 10	Week 2 Outline: Supervised learning concepts. Regression versus Classification k-NN Regression Linear regression, polynomial feature expansion, measuring error: RSS error, k-fold cross validation, Sci-kit learn datasets Overfitting and underfitting
Sept 17	Week 3

	Outline: Logistic regression. Measuring accuracy: ROC, confusion matrix, dealing with categorical and missing data. Regularization: lasso, ridge. Robust regression. Hyper-parameter search.
Sept 24	Week 4 Support vector machines (linear and kernelized): RBF kernels, Multi-class classification, data imputation, data leakage.
Oct 1	Week 5 Decision trees for classification and regression, entropy Boosting, Random forests, gradient boosted decision trees, XGBboost, AdaBoost, feature importance, SVM paper on detecting fraudulent reviews
Oct 8	Week 6 Naive Bayes, pipelines, Unsupervised learning: density estimation.
Oct 15	Week 7 Unsupervised learning: clustering. Agglomerative/tree-based clustering. K-means and variants. Gradient Descent and EM.
Oct 22	Week 8 Unsupervised learning: dimensionality reduction (PCA, multi-dimensional scaling, t-SNE) Evaluation of unsupervised methods. Midterm Examination (tentative).

Oct 29	Week 9: Deep learning 1.Neural networks, Convolutional NN, Embeddings.
Nov 5	Week 10: Deep learning 2. Visualizing ConvNets. Sequence problems: Recurrent NN.
Nov 12	Week 11: Deep learning 3 Generative Adversarial networks (GANs)
Nov 19	Week 12: FAT-ML: bias in training and data collection; implications of privacy
Dec 3	Week 13: Final project presentations (or catch-up). Incentives and Learning, adversarial ML

Class Format

The first half of the class will be typically be lecture-oriented or with interactive activities (e.g. collaborative break-outs). From time to time there may be project-related presentations or activities. The second half will typically be lab-oriented, with interactive problem-solving sessions that are meant to give you an opportunity to get experience actually writing code that uses whatever new tools and concepts have been covered in class that week.

Typically, the instructor and/or GSIs will make slides available online via Canvas before the class.

Homeworks

Homeworks will be due by 11:59pm on the date given in the syllabus and designated in Canvas. There will be a problem set due most weeks, and usually on Monday evening. Please see the late policy section for what happens if you can't make this deadline.

Please make sure you read the section below on Giving/Receiving Help. If you copy someone else's homework solution completely, or almost completely (and/or failing to acknowledge your source), then this will be considered cheating, and I'll refer your case to the academic advising office for disciplinary action.

Please contact the instructor if you have any uncertainty or questions about this policy.

Attendance

Attendance at lectures and participation in labs is necessary for getting the most out of the course.

Readings

There will be readings assigned several weeks to see how the techniques in class are applied by researchers. These readings will be available on Perusall.

Kaggle

We will have one or two Kaggle competitions throughout the term. These allow you to try the techniques we learned in class to a new problem and see how well you can do against others in the course. You are allowed to compete in groups of up to four students.

Final Project

The goal of the final project is to further apply what you've learned in class to real-world datasets. The deliverables for the final project, which will be due at the end of the course, will include an initial proposal, poster presentation, final report, and code/data repository. Details on the final project will be available as we get closer to the project proposal date. You may work individually or in groups of 2-3. However, the scale of the project should reflect the number of participants.

The poster session for the final projects will be on Friday, Dec 6, 2019 tentatively from 11am-2pm.

Grading

The graded work in the course will be weighted as follows to determine a final percentage grade:

Homeworks: 45%

Kaggle: 10%

Midterm: 15%

Final project: 30%

Conversion between percentage grades and letter grades will use the following mapping:

A+ 98; A 93; A- 90; B+ 87; B 83; B- 80; C+ 77; C 73; C- 70; D+ 67; D 63; D- 60; E 50; F 40.

Please email our IAs, TBD, for any question related to the grading of the previous assignments. To avoid our IA being overloaded at the end of the semester, requests for grading corrections must be sent within one week after the grades are released.

Bonus Credits

There may be opportunities for bonus credits during the course.

Late Policy

A **25% penalty** will be assessed for each subsequent 24h period after the deadline that an assignment is late up to 48 hours, after which the assignment will not be accepted. For example, if the due date is 11:59pm on Monday, the possible late penalties would be:

Before 11:59pm Tuesday: 25% deduction

Before 11:59pm Wednesday: 50% deduction

After 11:59pm Thursday: 100% deduction

The deduction will be taken from the maximum allowable points. For example, if you score 90/100 and turn the assignment in 35 minutes late, you will receive a score of 65/100.

We realize that the occasional crisis might screw up your schedule enough to require a bit of extra time in completing a course assignment. Thus, we have instituted the following late policy that gives you a limited number of flexible “late day” credits. These late days are intended to be used in exceptional/emergency circumstances (illness, computer died, power outage), so you should have plenty to get through the term.

You have **two (2)** free late days to use during the course. No fractional late days: they are all or nothing. Each late day used will void one 25% penalty. For example, if an assignment is 2 days late, and you use 1 late days, then it will be assessed a 25% penalty. **No assignment can be turned in more than 2 days late**. Therefore, you cannot use late days on an assignment that is 2 or more days late.

You don't need to explain or get permission to use late days, and we will track them for you. In cases where late days can be assigned in multiple ways (e.g. you have only one late day left but hand in two late assignments) we will always allocate late days in a way that maximizes your grade. Note that resubmissions after the deadline will be counted as late submissions. **Also, late days may not be applied to the final project.**

Exams

There is a midterm exam in this course.

Classroom policy

Students are asked to attend class on time and remain through the entire class. I will expect students to be present and have their video turned on for class. This serves several purposes. First, it allows me to see if students are following. Second, it gives the class a more community feel. Third, it applies a very small amount of peer pressure to pay attention (at least to say seated and tuned in). I understand this may not be possible for everyone. Please let the course staff know in advance if you have a reason for not turning on your video.

Participation Pledge

This course is a lot of fun when the students are engaged. Even in a classroom setting, this is difficult for many students. With remote instruction, there will be many temptations.   In order to aid students (and keep the class exciting), I am offering a participation pledge. During the first week of class, you will be able to sign the pledge if you choose. The pledge will be to fully participate in each class period. This means no looking at your phone, no email, no doing other class assignments, no Facebook, no listening to music, etc. You should behave as if you are actually sitting in a classroom and paying attention.   I know this is hard. It's hard for the instructor! However, I believe that if you take this seriously, you will get a much better experience from the class. Both by learning more, and by having more fun.

You need not sign the pledge, and can still receive full marks in the class. As a small sweetener, I will allow anyone that signs the pledge 2 additional free late days on the problem sets.

Platforms and Technology:

Even offline, this class relies on software. Here is a summary of the different software platforms we will use.

Canvas: Assignments will be posted and turned in on Canvas. Also announcements made (though you should receive email versions of these). Finally, recording of the lectures will be available on canvas.

Zoom   Class will be held on Zoom. Also, virtual office hours.

Piazza Q&A site for questions about class/assignments of a non-personal nature.

Perusall Use this to access readings and comment on, and ask/answer questions about readings. Accessible through Canvas.

Email Direct contact with instructor.

Disconnection If I become disconnected during class: I will try to communicate with the class via email. If I am unable to communicate via email/canvas, please wait 15 minutes (chat amongst yourselves, do other work, etc) before disconnecting.

If you become disconnected during class, please email the instructor. Unfortunately, there is likely little that I can do in the moment.

Academic Integrity and Misconduct

Collaboration

UMSI strongly encourages collaboration while working on some assignments, such as homework problems and interpreting reading assignments as a general practice. Active learning is effective. Collaboration with other students in the course will be especially valuable in summarizing the reading materials and picking out the key concepts. You must, however, write your homework submission on your own, in your own words, before turning it in. If you worked with someone on the homework before writing it, you must list any and all collaborators on your written submission. Each course and each instructor may place restrictions on collaboration for any or all assignments. Read the instructions careful and request clarification about collaboration when in doubt. Collaboration is almost always forbidden for take-home and in class exams.

Giving and Receiving Assistance

We will use a policy used in other SI programming-based courses (e.g. SI106 and SI370 (the following text is taken in modified form from the SI370 F15 syllabus). Learning technical material is often challenging, and a course like this one covers a range of topics and can move quickly. We want you to succeed in the course and we encourage you to get help from anyone you like.

However: In the end, **you** are responsible for learning the material – so you need to make sure that the help you get is focused on gaining knowledge, not just on getting through the assignments. If you rely too much on help so that you fail to master the material, especially the basics earlier in the semester, you will crash and burn later in the course.

The final submission of each homework exercise must be in your own words. If you get help on an assignment, please indicate the nature and amount of help you received. If the assignment involves computer code, add a comment indicating who helped you and how. Any excerpts from the work of others must be clearly identified as such (e.g. quotation with citation, or with comments in the code if it is a code fragment you have borrowed).

If you're a more advanced student and are willing to help other students, please feel free to do so. Just remember that your goal is to help teach the material to the student receiving the help. It is acceptable for this class to ask for and provide help on an assignment via the Q&A site (Piazza), including posting short code fragments (e.g. 2-3 lines). Just don't post complete answers. If it seems like you've posted too much, one of the instructional staff will contact you to let you know, so don't worry about it. When in doubt, err on the side of helping your fellow students. To reiterate, the collaboration policy is as follows. Collaboration in the class is allowed (and even encouraged) for assignments – you can get help from anyone as long as it is clearly acknowledged. Collaboration or outside help is also not allowed on exams or other types of assessments, though for major tests you will be allowed to bring in specific “cheat sheets” with you. Use of solutions from previous semesters is not allowed. The authorship of any assignments must be in your own style.

Plagiarism

All written submissions must be your own, original work. Original work for narrative questions is not mere paraphrasing of someone else's completed answer: you must not share written answers with each other at all. At most, you should be working from notes you took while participating in a study session.

You may incorporate selected excerpts, statements or phrases from publications by other authors, but they must be clearly marked as quotations and must be attributed. If you build on the ideas of prior authors, you must cite their work. You may obtain copy editing assistance, and you may discuss your ideas with others, but all substantive writing and ideas must be your own, or be explicitly attributed to another. See the (Doctoral, MSI, MHI, or BSI) student handbooks available on the UMSI intranet for the definition of plagiarism, resources to help you avoid it, and the consequences for intentional or unintentional plagiarism.

Double Counting Work

If you are taking or have taken another data science-related course this term you may not submit the same work to more than one course (including from a previous course) for your final project. While you might work on a broadly similar problem area, the code and report you turn in must be substantially different and must be unique to this course. Please see the instructor if you’re not sure what this means.

Accommodations for Students with Disabilities

If you think you need an accommodation for a disability, please let me know at your earliest convenience. Some aspects of this course, the assignments, the in-class activities, and the way the course is usually taught may be modified to facilitate your participation and progress. As soon as you make me aware of your needs, we can work with the Services for Students with Disabilities (SSD) office to help us determine appropriate academic accommodations. SSD (734-763-3000; http://ssd.umich.edu) typically recommends accommodations through a Verified Individualized Services and

Accommodations (VISA) form. Any information you provide is private and confidential and will be treated as such.

Student Mental Health and Wellbeing

The University of Michigan is committed to advancing the mental health and wellbeing of its students, while acknowledging that a variety of issues, such as strained relationships, increased anxiety, alcohol/drug problems, and depression, directly impacts students' academic performance. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, contact Counseling and Psychological Services (CAPS) at (734) 764-8312 and https://caps.umich.edu/ during and after hours, on weekends and holidays or through its counselors physically located in schools on both North and Central Campus. You may also consult University Health Service (UHS) at (732) 764-8320 and https://www.uhs.umich.edu/mentalhealthsvcs, or for alcohol or drug concerns, see www.uhs.umich.edu/aodresources. For a more comprehensive listing of the broad range of mental health services available on campus, please visit: http://umich.edu/~mhealth/.