ONLINE MASTER OF SCIENCE IN ANALYTICS

ISYE/CSE 6740 – COMPUTATIONAL DATA ANALYSIS / MACHINE LEARNING I

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

PROFESSOR: Yao Xie; yao.xie@isye.gatech.edu

TEACHING ASSISTANTS:
- (HEAD TA) MOYI GUO, MOYI@GATECH.EDU
- (HEAD TA) HELEN LE LU, HELENLULE@GATECH.EDU
- ALFIE PENGFEI CHEN, PFCHEN@GATECH.EDU
- RUYI DING, DINGRUYI1996@GATECH.EDU
- NEVIN THOMAS, NEVINTHOMAS@GATECH.EDU
- ANKUR BHARDWAJ, ABHARDWAJ75@GATECH.EDU
- BENJAMIN BRAY, BENRBRAY@GATECH.EDU
- SIAWPENG ER, SER8@GATECH.EDU

PREREQUISITES

- Undergraduate level probability, linear algebra, and statistics.
- Any programming language of your choice – MATLAB, R, Python, etc.

COURSE DESCRIPTION

Machine learning is a field of computer science that gives computers the ability to learn without being explicitly programmed. The course is designed to answer the most fundamental questions about machine learning: What are the most important methods to know about, and why? How can we answer the questions such as “is this method better than that one' using asymptotic theory”? How can we answer the question 'is this method better than that one' for a specific dataset of interest? What can we say about the errors our method will make on future data? What's the “right” objective function? What does it mean to be statistically rigorous?

This course is designed to give graduate students a thorough grounding in the methods, theory, mathematics and algorithms needed to do research and applications in machine learning. The course covers topics from machine learning, classical statistics, and data mining. Students entering the class with a pre-existing working knowledge of probability, statistics and algorithms will be at an advantage, but the class has been designed so that anyone with a strong numerate background can catch up and fully participate. Some experience with coding is expected (at a language of your choice, e.g., MATLAB or Python.)

For detailed course topics, please see the tentative course schedule.

LEARNING OBJECTIVES

After taking this course, students should be able to:

- Gain thorough understanding in the methods, theory, mathematics and algorithms needed to do research and applications in machine learning.
- Implementing and use machine learning algorithms.
- Gain experience with analyzing real data.

TEXTBOOKS/READINGS

- **Textbook**: The course material will be based on lectures slides provided in the course.
- **Recommended References**:

- (PRML) Pattern recognition and Machine Learning, Christopher M. Bishop.
- (ESL) The elements of Statistical Learning: Data Mining, Inference, and Predictions, 2nd edition, Trevor Hastie, Robert Tibshirani, and Jerome Friedman.
- (FML) Foundations of Machine Learning, 2nd edition. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar.

COURSE SCHEDULE
Please see Canvas for the course schedule.

COMMUNICATION
Instructor Communication: All communication from your instructor will take place in Canvas. You are expected to check Canvas every day for important course-related information. However, by following the instructions provided in the course, you can also ensure that you do not miss important instructions, announcements, etc. by adjusting your account settings to receive important information directly to your email account or cell phone. For more details, log into Canvas, enter the course, and see the section entitled "Before You Begin: Instructions for Getting Started."

Content Questions and Help: Because questions can often be addressed for the good of the group, please do not email your questions directly to the instructor. Instead, course and content questions will be addressed on Piazza. Feel free to set your post to private to ask questions about your grade or other issues unique to you. For more information, log into Canvas, enter the course, and see the section entitled "Before You Begin: Instructions for Getting Started."

Office Hours. Live office hours will be conducted every week via Blue Jeans. These sessions will be both an opportunity for the instructor to discuss course logistics and content but also an opportunity for you to ask questions. While it is strongly suggested that you participate in these meetings, all sessions will be recorded and archived if you are unable to attend or wish to reference them later. For the meeting schedule, links, and archives, please see the section in your Canvas course entitled "Weekly Videoconferences."

STUDENT EFFORT
Students are expected to devote 8-10 hours per week to complete the course requirements. This guideline encompasses all class activities, including reading the textbook and supplementary resources, watching lesson videos, participating in office hours and forum discussions, completing homework assignments, and studying for exams. Of course, students can spend as much time as necessary, but it is important to be careful not to fall behind.

Assignment Distribution and Grading Scale

Assignment	Release Date	Due Date	Weight
HW1	Jan 9	Jan 19	(total of 7 HWs) 60%
HW2	Jan 20	Feb 2	
HW3	Feb 3	Feb 16	
HW4	Feb 17	March 1	
HW5	March 2	March 15	
HW6	March 23	April 5	
HW 7	April 6	April 19	
Final Exam	April 20	April 30	40%

GRADING

Grades will be assigned on the following basis:

Homework:	60%
Final Exam:	40%

Important: Make sure the scores in Canvas are consistent with what you got. We will not make any change in grading for works older than 2 weeks.

The following grading scale, with scores rounded to the nearest whole number, will be used in the course:

- 90-100%: A
- 80-89%: B
- 70-79%: C
- 60-69%: D
- below 60%: F

LESSONS

Video lessons for this course will be housed on edX. For more details on creating and linking your edX account, log into the Canvas, enter the course, and see the section entitled "Before You Begin: Instructions for Getting Started."

HOMEWORK

Homework should be submitted in Canvas **by 11:59 pm EST on the date it is due**. No submission will be accepted through email. We strongly encourage the use of LaTeX for your submission. Assignments will include both exercises and computer problems; the computer problems will ask you to carry out statistical analysis using computer statistical software. **Keep in mind that you should not hand in raw computer output. Conclusions and interpretation of results are more important than good printouts. Compute output with proper explanation will not receive full grades.**

*** (important) Homework code submission requirement**

For homework questions involving programming, please submit all your files using a single ZIP file (other format not acceptable), name the file as: 'lastname_firstname_HWx.zip', any file format or file name not following the requirement will have 10 percent points reduced from that submission. When you submit code, make sure all the code files and the data files are included in a separated folder, with name: 'lastname_firstname_hwx_code.zip'. TA will not be allowed to modify your code, so if your code have path issue, it will be considered as "not executable".

You can work together with other students on homework, as long as you write-up and turn in your own solutions. You are also allowed (and encouraged) to ask me questions, although you should try to think about the problems before asking. Request for re-grading the Homework/Exams/Quizzes should be made within a week of returning Homework/Exams/Quizzes. Any kind of academic misconduct is subject to F grade as well as reporting to the Dean of students.

Moreover, we have the following accommodation policies to help with emergent situations. (1) You can have ONE CHANCE to ask for a one-week homework extension. (2) If you have already used the one-chance one-week extension, and if you submit the homework late: one day late the

grade will be discount to 75% of your total, two days late the grade will discount to 50% of your total, three days late the grade will discount to 25% of your total. Past three days, your homework will not be accepted.

EXAMS

The following exams will be administered in this course:

- Final Exam Date: **Monday April 20 - Thursday April 30**
  - The final exam is comprehensive.

Final Exam must be taken between Monday April 20, and Thursday April 30. No late submission for Final Exam is accepted. There is NO GRACE PERIOD OR MAKE-UPS for the exam. If the Final Exam is failed to be submitted, the final grade will be "incomplete," and you may choose to take the final exam in the future and finish the course.

Exams in this course will be open book. You are expected to finish the exam independently. If there is any conflict of time, please let us know beforehand. There are no make-ups.

The regrading requests for both the homework and final exam have to be submitted within 2 weeks from the grades are released.

PLAGIARISM

Plagiarism is considered a serious offense. You are not allowed to copy and paste or submit materials created or published by others, as if you created the materials. All materials submitted and posted must be your own original work.

STUDENT HONOR CODE

You are responsible for completing your own work.

All OMS Analytics degree students are expected and required to abide by the *letter* and the *spirit* of the Georgia Tech Honor Code. The teaching assistants and I will also abide by these honor codes. I am very serious about this expectation because ethical behavior is extremely important in all facets of life. To review the Georgia Tech Honor Code, please visit http://osi.gatech.edu/content/honor-code . Any OMS Analytics degree student suspected of behavior in violation of the Georgia Tech Honor Code will be referred to Georgia Tech's Office of Student Integrity.

ACCOMMODATIONS FOR STUDENTS WITH DISABILITIES

If you are a student with learning needs that require special accommodation, contact the Office of Disability Services at (404) 894-2563 or http://disabilityservices.gatech.edu/, as soon as possible, to make an appointment to discuss your special needs and to obtain an accommodations letter. Please also e-mail me as soon as possible in order to set up a time to discuss your learning needs.

Course Schedule

All assignments and exams are due by 11:59 pm EST on the date listed below.

Week/Dates	Module/Topic	Weekly Overview	Deliverables
Week 1 Jan 6 - 10	Introduction and Overview	Overview of the topics and scope of the class	
Week 2 Jan 13 - 17	Clustering and k-means	We will introduce a building block of a fundamental problem in unsupervised learning (clustering): K-means	Homework 1, release Jan 9, Due Jan 19
Week 3 Jan 20 - 24	Spectral Clustering	We will discuss another type of clustering algorithm: spectral clustering, which is different from k-means since it is based on geometry (connectivity) of data	
Week 4 Jan 27 - 31	Dimensionality Reduction and PCA	We will present linear dimensionality reduction technique called PCA	Homework 2, release Jan 20, Due Feb 2
Week 5 Feb 3 - 7	Nonlinear Dimensionality Reduction	We will study non-linear dimensionality reduction techniques	
Week 6 Feb 10 - 14	Density Estimation	Discuss basic density estimation method, which captures the distributional information of the data	Homework 3, release Feb 3, Due Feb 16
Week 7 Feb 17 - 21	Gaussian Mixture Model and EM Algorithm	We will present a popular type of model for densities called Gaussian mixture models and discuss how to fit such models	
Week 8 Feb 24 - 28	Basic of Optimization Theory	We will introduction the essentials of optimization theory which is a foundation of developing machine learning algorithms	Homework 4, release Feb 17, Due March 1
Week 9 March 2 - 6	Classification Naïve Bayes and Logistic Regression	Introduce classification problem and two basic methods for classification	
Week 10 March 9 - 13	Support Vector Machine (SVM)	Introduce SVM classifier	Homework 5, release March 2, Due March 15
Spring Break March 16 - 20			
Week 11 March 23 - 27	Neural Networks	Understand basic neural networks	
Week 12 March 30 – April 3	Boosting Algorithms and AdaBoost	Introduce basic boosting algorithms and AdaBoost	Homework 6, release March 23, Due April 5
Week 13 April 6 - 10	Random Forest	Introduce tree-based methods for regression and classification, and random forest	
Week 14 April 13 - 17	Bias-Variance Tradeoff and Cross-Validation	Introduce principle of bias-variance tradeoff and how it is used to cross-validation for model selection and parameter tuning	Homework 7, release April 6, Due April 19
Week 15 April 20 - 30	Final Exam Week	Summary of class	**Final Exam – must be taken between Monday April 20, and Thursday April 30**