02-620: Machine Learning for Scientists

Seyoung Kim

Spring 2020

1. Course Information

1.1 Vital information

Course		**Lecture**	
	Time	MWF 10:30-11:20 AM	
	Location	GHC 4215	
		Recitation	
	Time	TBD	
	Location	TBD	
Instructors		**Seyoung Kim**	
	E-mail	sssykim@cs.cmu.edu	
	Website	cs.cmu.edu/~sssykim/	
	Office	GHC 7721	
	Office Hour	M 11:30AM -12:30 PM	
TAs		**Cathy Su**	**Yutong Qiu**
	E-mail	qsu1@andrew.cmu.edu	yutongq@andrew.cmu.edu
	Office Hour	TBA	TBA
	Location	TBA	TBA

1.2 Course description

With advances in scientific instruments and high-throughput technology, scientific discoveries are increasingly made from analyzing large-scale data generated from experiments or collected from observational studies. Machine learning methods that have been widely used to extract complex patterns from large speech, text, and image data are now being routinely applied to answer scientific questions. The course will select scientific questions that arise in genomics, populaton genetics, and medicine and discuss how to address these questions using machine learning techniques. It will cover disease-related genetic variant discovery with regression methods; clinical decision making for patients with classification methods; pathway discovery with clustering algorithms; learning gene regulatory networks with probabilistic graphical models; genome sequence analysis with hidden Markov models; making functional annotations of genomes with deep learning methods; and selecting an appropriate machine learning technique for the given scientific problem using learning theories. This course is intended for graduate students interested in learning

machine learning methods for scientific data analysis and modeling. Programming skills and basic knowledge of linear algebra, probability, statistics are assumed. Homework assignments will consist of written problems and analyses of genetic and genomic data drawn from the literature in biology. The course grade will be computed as the result of homework assignments, midterm tests, and class participation.

1.3 Pre-requisites

Prerequisites: 02-680 or an equivalent class.

1.4 Course Details

Canvas Homepage. The course homepage will be hosted on Canvas. Canvas will be used for attendance and as a central repository for grades. You should be automatically enrolled at `https://canvas.cmu.edu/courses/13407`.

Discussion Forum. An online forum is provided on Piazza as an area for discussion and questions. The forum will be moderated by the course staff who will respond to questions, but students are encouraged to help each other via discussion. However, assignment specifics should not be discussed — any hints will be provided by the teaching staff. You can find the class on Piazza at `https://piazza.com/class/k4agcvccufz4j6`.

1.5 Textbooks

Reading material will be drawn from the following textbooks. All textbooks listed below are freely available online as pdf files.

Machine Learning: a Probabilistic Perspective by Kevin Murphy [M]

The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, and Jerome Friedman [H]

Additional text book:

Machine Learning by Tom Mitchell

2. Curriculum

2.1 Key dates

Feb 21 Friday: Midterm I

March 6-15: Spring break, no class

April 3 Friday: Midterm II

April 17: Spring carnival, no class

Final week: Final exam

2.2 Tentative course schedule

As with course topics, the lecture-by-lecture below is tentative and subject to change.

COURSE SCHEDULE			
Week	**Topic**	**Lecture**	**Reading**
Week 1	Introduction to ML	Supervised/unsupervised learning MLE, MAP estimation	M Ch 1.1-1.3, 2.2.3, 3.1-3.4, 4.6.2.2
Week 2	Regression	Linear regression, regularization [*Finding genetic variants that control gene expression levels*]	M Ch 7.1-7.3.1, 7.5.1, 7.5.4, 13.3.1, 13.3.3-13.3.4
Week 3	Classification	K-nearest neighbor method Naive Bayes classifier, logistic regression Generative and discriminative classifier [*Genetic variants that affect disease state*]	H Ch 13.3; M Ch 1.4.2, 3.5.1-3.5.3, 8.1-8.3, 8.6
Week 4		Decision tree, random forest [*Cancer classification based on expression*]	M Ch 16.1-16.2
Week 5		SVM [*Cancer classification based on expression*]	M Ch 14.5
Week 6	Clustering	Hierarchical clustering, K-means [*Gene module/pathway discovery*]	M Ch 25.1, 25.5; H Ch 13.1-13.2
Week 7	Probabilistic graphical models	Bayesian networks: models, learning, and inference [*Learning gene networks*]	M Ch 10
Week 8		Dynamic Bayesian networks, Gaussian graphical models [*Modeling longitudinal data*]	H Ch 17.1-17.3
Week 9	Bayesian learning	Bayesian estimation, MCMC [*Modeling uncertainty in gene networks*]	H Ch 8.6; M Ch 15.1-15.2
Week 10	Model selection	Bias-variance trade-off, VC dimension [*Which model is the best?*]	H Ch 7.1-7.10
Week 11	Neural networks	Models, backpropagation algorithm [*Predicting sequence specificity in functional genomics*]	M Ch 16.5
Week 12	Dimensionality reduction	SVD, PCA [*Population structure discovery in genome sequences*]	M Ch 11.2.1-11.2.3
Week 13	Latent variable models	Mixture models, EM algorithm [*Gene module/pathway discovery with uncertainty*]	M Ch 11.4.1-11.4.2, 11.4.7, 11.5
Week 14		Hidden Markov model, forward-backward algorithm, Baum-Welch algorithm [*Gene discovery from genome sequences, modeling recombination*]	M Ch 17.1.1-17.2.2, 17.3-17.5
Week 15	Applications	Genomic data analysis and ML	

3. Coursework

Coursework will consist of the following components. **You will have two late days you can use to receive full credits for any late homework submission during the semester.**

Homework assignments. (45% of grade) Written homework assignments will test your knowledge of the material covered in class.

Attendance and participation (10% of grade) Attendance will be taken, and we will have occasional in-class exercises that serve to reinforce the concepts we have covered. These exercises will not be graded, but participation will be expected in order to receive a complete grade for that day. You are allowed three "dropped" attendance grades without penalty. These can be used for any purpose.

Examinations. (45% of grade) The exams will test your knowledge of the material from the class. The two midterms will be held in class, and the final exam will be held during the university's scheduled time. The exam dates are:

- Midterm 1 (15% of grade): Feb 21 in class
- Midterm 2 (15% of grade): April 3 in class
- Final exam (15% of grade): Time and location TBD (will be posted when set by university)

The midterms will not be cumulative: midterm 2 will cover material encountered after midterm 1. The final exam will cover the material from the entire semester. You are allowed to bring one letter-sized note to each exam, for which you will receive 5 points. The note should be hand-written.

4. Collaboration Policy and Academic Integrity

All class work should be done independently unless explicitly indicated on the assignment handout. You may *discuss* homework problems with classmates, but must write your solution by yourself. If you do discuss assignments with other classmates, you must supply their names at the top of your homework. No excuses will be accepted for copying others' work, and violations will be dealt with harshly. (Getting a bad grade is much preferable to cheating.)

The university's policy on academic integrity can be found at the following link: `http://www.cmu.edu/academic-integrity/`. In part, it reads, "Unauthorized assistance refers to the use of sources of support that have not been specifically authorized in this policy statement or by the course instructor(s) in the completion of academic work to be graded. Such sources of support may include but are not limited to advice or help provided by another individual, published or unpublished written sources, and electronic sources." You should be familiar with the policy in its entirety. **The default penalty for any academic integrity violation is failure of the course.**

In particular: use of a previous semester's answer keys or online solutions for graded work is absolutely forbidden. Any use of such material will be dealt with as an academic integrity violation.

5. Other policies

Classroom etiquette: To minimize disruptions and in consideration of your classmates, we ask that you please arrive on time and do not leave early. If you must do either, please do so quietly. **The use of phones or other electronic devices during class is forbidden and will result in a zero discussion grade for the day (counts as missed class).**

Excused absences: Students claiming an excused absence for an in-class exam must supply documentation (such as a doctor's note) justifying the absence. Absences for religious observances must be submitted by email to the instructor during the first two weeks of the semester. Note that job or internship interviews are not a justification for an excused absence.

Other: The following policies of 15-110 also apply to this class. This text is mostly quoted from the 15-110 website (with some modifications):

- **I must be out of town for a university related event (e.g. member of a team). What should I do about my assignments?**

  If you have an official excuse we will make special arrangements for you to submit the assignment, please contact the instructors.

- **I am out of town attending a family/important event. How can I submit my assignments due for the week?**

  The assignment must be submitted online before the due date.

- **I missed the in-class exam because I fell sick. What should I do?** You must immediately seek medical treatment and receive an official medical excuse. You must also contact the instructors prior to the exam or as soon as possible. If you can produce documentation we can make arrangements to give you a makeup test. Otherwise, we will be unable to make any exceptions.

- **I am failing the course. Is there any extra work I can do to get a passing grade?** Unfortunately, we cannot make exceptions. The best way to avoid this situation is to talk to one of the instructors as soon as possible to find out what you need to do. Do not wait until the last few weeks of classes to discuss your performance.

6. Accommodations for Students with Disabilities

If you have a disability and have an accommodations letter from the Disability Resources office, we encourage you to discuss your accommodations and needs with us as early in the semester as possible. We will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, we encourage you to contact them at `access@andrew.cmu.edu`.

7. Provost's Statement on Student Well-Being

Take care of yourself. Do your best to maintain a healthy lifestyle this semester by eating well, exercising, avoiding drugs and alcohol, getting enough sleep and taking some time to relax. This will help you achieve your goals and cope with stress.

All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at `http://www.cmu.edu/counseling/`. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night:

CaPS: 412-268-2922

Re:solve Crisis Network: 888-796-8226

If the situation is life threatening, call the police:

On campus: CMU Police: 412-268-2323

Off campus: 911

If you have questions about this or your coursework, please let us know.