CS 4641\7641 A: Machine Learning (Spring 2021)

Course Information

- **Lecture**: Tuesdays and Thursdays, 9:30am-10:40am
- **Q/A for Lecture**: Tuesdays and Thursdays, 5:00pm-6:10pm - Offline Questions
- **Location**: BlueJeans
- **Piazza**: http://piazza.com/gatech/spring2021/mlspring2021aasynch/home

Instructor:
Mahdi Roozbahani
(mahdir@gatech.edu)

Head TA:
Ruijia Wang
(rwang@gatech.edu)

Head TA:
Kevin Tynes
(kdtynes@gatech.edu)

TA:
Shalini Chaudhuri
(shalini.chaudhuri@gatech.edu)

TA:
Shreeshaa Kulkarni
(shrek@gatech.edu)

TA:
Jayanta Bhowmick
(jbhowmick3@gatech.edu)

TA:
Huili Huang
(hhuang413@gatech.edu)

TA:
Shubhangi Upasani
(shubhangi.upasani@gatech.edu)

TA:
Danrong Zhang
(dzhang373@gatech.edu)

TA:
Kunal Chawla
(kunalchawla@gatech.edu)

TA:
Yuzi Hu
(yhu495@gatech.edu)

TA:
Harrison “Clay” Tyler
(htyler7@gatech.edu)

TA:
Rohit Das
(rohdas@gatech.edu)

TA:
Sidharta Vadaparty (Sid)
(svadaparty3@gatech.edu)

TA:
Pallavi Misra
(pmisra9@gatech.edu)

TA:
Ruslina Utomo (Rusty)
(rutomo6@gatech.edu)

TA:
Bharat Mamidibathula
(bsm7@gatech.edu)

TA:
Daniel Smith
(dsmith466@gatech.edu)

TA:
Dheeraj Eidnani
(deidnani@gatech.edu)

Course Overview

This course introduces techniques in machine learning with an emphasis on *algorithms and their applications to real-world data*. We will investigate the following question: how to computationally extract useful knowledge from data for decision making and task support? We will focus on machine learning methods, which are organized into three parts:

1. **Basic math for data science and machine learning**
    - Linear algebra
    - Probability and statistics
    - Information theory
    - Optimization
2. **Unsupervised machine learning for data exploration**
    - Clustering analysis
    - Dimensionality reduction
    - Kernel density estimation

3. **Supervised learning for predictive data analysis**
    - Tree-based models
    - Support vector machines
    - Linear classification and regression
    - Neural networks

Prerequisites for this course include (1) basic knowledge of probability, statistics, and linear algebra; (2) Basic programming experience in Python.

In addition to the technical content, this class includes the following learning objectives:

- Structuring a task into a machine learning work flow
- Collaborating effectively on team projects in a remote environment
- Conducting peer evaluation in a constructive format
- Communicating technical content in a concise and effective manner

Schedule

For all dates used in this course, their times are 23:59 Anywhere on Earth (11:59 pm AoE). For example, a due date of "January 8" is the same as "January 8, 23:59pm AoE". Convert the times to your local times using a Time Zone Converter.

Date	Class	Assignments	Project	Quizzes	Readings
Thu, Jan 14	*Course overview; Refer to Lecture1 video on Canvas Refer to Lecture1-QA video on Canvas	Piazza signup			GT Honor Code
Mon, Jan 18					
Tue Jan 19	*Data analysis toolbox; *Broadcasting toolbox; *Modified in lecture-Data analysis toolbox; *Modified in lecture-Broadcasting toolbox; *QA notes; Project information; Refer to Lecture2 video on Canvas Refer to Lecture2-QA video on Canvas				Undergrad Canvas Access for previous ML projects; Grad Canvas Access for previous ML projects; Heilmeier catechism; Visual Information Theory by Chris Olah; GitHub Pages; YAML Configuration; NumPy Tutorial; Matplotlib Tutorial; Project Examples; seaborn: statistical data visualization; Overleaf for GT students;
Wed, Jan 20				Q0: Introduction,pre-test	
Thu, Jan 21	*Foundations: Linear algebra; *Class notes; *QA notes; Refer to Lecture3 video on Canvas Refer to Lecture3-QA video on Canvas				Correlation vs Covariance Linear Algebra Review by Zico Kolter
Mon, Jan 25				Q1: Data analysis toolbox and Linear algebra	
Tue, Jan 26	*Foundations: Probability and statistics; *Class Notes; *QA Notes; Refer to Lecture4 video on Canvas Refer to Lecture4-QA video on Canvas	A1 out			Probability Theory Review by Andrew Moore; The Differences Between Data, Information and Knowledge;

Wed, Jan 27		Project Seminar 1 by Ruijia and Kevin [Undergrad Canvas Access] [Grad Canvas Access]		
Thu, Jan 28	*Foundations: Information Theory; *Class Notes; *QA Notes; Refer to Lecture5 video on Canvas Refer to Lecture5-QA video on Canvas			The Differences Between Data, Information and Knowledge; More about Cross Entropy and KLD; Cross Entropy as loss function;
Mon, Feb 1			Q2: Probability and Information Theory	
Tue, Feb 2	*Foundations: Optimization; *QA Notes; Refer to Lecture6 video on Canvas Refer to Lecture6-QA video on Canvas			KKT for inequality constrained optimization; Why Cross Entropy over MSE for Classification;
Wed, Feb 3		Project seminar 2 by Rohit and Dheeraj [Undergrad Canvas Access] [Grad Canvas Access]		
Thu, Feb 4	*Data analysis toolbox - part 2; Data analysis toolbox QA Notes; Refer to Lecture7 video on Canvas Refer to Lecture7-QA video on Canvas	Project team composition due		NumPy Tutorial; Matplotlib Tutorial;
Mon, Feb 8			Q3: Optimization and Data Analysis toolbox	
Tue, Feb 9	*Clustering Analysis and K-Means; *Class Notes; *QA Notes; Refer to Lecture8 video on Canvas Refer to Lecture8-QA video on Canvas			Curse of dimensionality (Euclidean space example); Jupyter Notbook (Kmeans and DBSCAN);
Wed, Feb 10		Project seminar 3 by Shreeshaa and Rusty [Undergrad Canvas Access] [Grad Canvas Access]		
Thu, Feb 11	*Gaussian Mixture Model; *Class Notes; *QA Notes; Refer to Lecture9-part1 video on Canvas Refer to Lecture9-part1-QA video on Canvas			GitHub Student Application;
Mon, Feb 15			Q4: Clustering Analysis, Kmeans and GMM	
Tue, Feb 16	*Gaussian Mixture Model; *Class Notes; *QA Notes; Refer to Lecture9-part2 video on Canvas Refer to Lecture9-	A1 due | A2 out		

	part2-QA video on Canvas			
Web, Feb 17		Project seminar 4 by Shalini [Undergrad Canvas Access] [Grad Canvas Access]		
Thu, Feb 18	*Hierarchical Clustering; *Hierarchical Class Notes; *QA Notes; Refer to Lecture10 video on Canvas Refer to Lecture10-QA video on Canvas			Understanding the concept of Hierarchical clustering Technique; Dendrogram Visualization ;
Mon, Feb 22			Q5: GMM and Hierarchical	
Tue, Feb 23	*Density-Based Clustering; *Class Notes; *QA Notes; Refer to Lecture11 video on Canvas Refer to Lecture11-QA video on Canvas			Jupyter Notbook (Kmeans and DBSCAN);
Wed, Feb 24		Project proposal forum to discuss your project topic with the instructional Team before proposal submission [Undergrad Canvas Access] [Grad Canvas Access]		
Thu, Feb 25	*Evaluation of Clustering Algorithms; *Class Notes; *QA Notes (no notes for this lecture); Refer to Lecture12 video on Canvas Refer to Lecture12-QA video on Canvas			
Mon, Mar 1			Q6: DBSCAN and Cluster evaluation	
Tue, Mar 2	*Density Estimation *Class Notes; *Class QA; Refer to Lecture13 video on Canvas Refer to Lecture13-QA video on Canvas			KDE interactive visualization ; KDE sampling ; KDE SKLearn and sampling ; Jupyter Notebook Kernel Density Example;
Wed, Mar 3		Project proposal due Project seminar 5 by Daniel [Undergrad Canvas Access] [Grad Canvas Access]		
Thu, Mar 4	*Dimensionality reduction; *Class Notes; *QA Notes; Refer to Lecture14 video on Canvas Refer to Lecture14-QA video on Canvas			Image reconstruction using PCA ; Feature extraction using PCA ; PCA for images ; PCA as linear combination of features ; PCA and Linear Discriminant Analysis ;
Mon, Mar 8			Q7: Density estimation and dimensionality reduction	
	*Linear Regression;			

Tue, Mar 9	*Class Notes; *QA Notes; Refer to Lecture15 video on Canvas Refer to Lecture15-QA video on Canvas	A2 due | A3 out		Simple Linear Regression in Matrix Format; Adding Noise to Regression Predictors;
Wed, Mar 10				
Thu, Mar 11	*Linear Regression [Contd]; *Regularization and Linear Regression; *Class Notes; *QA Notes (No QA Notes); Refer to Lecture16 video on Canvas Refer to Lecture16-QA video on Canvas (No QA Video)			
Mon, Mar 15				
Tue, Mar 16	No assignments or classes or assessments			
Wed, Mar 17				
Thu, Mar 18	*Regularization and Linear Regression [Contd]; *QA Notes; Refer to Lecture17 video on Canvas Refer to Lecture17-QA video on Canvas		Q8: Linear Regression	
Fri, Mar 19		Project seminar 6 by James [Undergrad Canvas Access] [Grad Canvas Access]		
Mon, Mar 22			Q9: Regularization and Linear Regression	
Tue, Mar 23	*Naïve Bayes and Logistic Regression; *Class Notes; *QA Notes; Refer to Lecture18 video on Canvas Refer to Lecture18-QA video on Canvas			
Wed, Mar 24				
Thu, Mar 25	*Naïve Bayes and Logistic Regression; *QA Notes; Refer to Lecture19 video on Canvas Refer to Lecture19-QA video on Canvas			
Mon, Mar 29			Q10: Naive Bayes and Logistic Regression	
	*Neural Networks(Forward pass and Back propagation);			NN Playground ; Interactive NN

Tue, Mar 30	*Class Notes; *QA Notes [No note, watch the QA video]; Refer to Lecture20 video on Canvas Refer to Lecture20-QA video on Canvas			initialization ; The role of a hidden layer; Back propagation numerical example; More detailed introduction;
Wed, Mar 31		Project seminar 7 by Kunal and Yuzi [Undergrad Canvas Access] [Grad Canvas Access]		
Thu, Apr 1	*Neural Networks and Deep learning (CNN); *Class Notes; *QA Notes [No note, watch the QA video]; Refer to Lecture21 video on Canvas Refer to Lecture21-QA video on Canvas	A3 due | A4 out		CNN Live Demo; A guide to an efficient way to build CNN and optimize its hyper-parameters; Back Propagation in CNN; Transfer learning in CNN; Project Scoring Guidance;
Mon, Apr 5			Q11: NN and CNN	
Tue, Apr 6	*Decision Tree; *Decision Tree Class Notes; *QA Notes [No note, watch the QA video]; Refer to Lecture22 video on Canvas Refer to Lecture22-QA video on Canvas			
Wed, Apr 7				
Thu, Apr 8	*Decision Tree; *Decision Tree Class Notes; *Ensemble Learning and Random Forest; *Ensemble Learning Class Notes; *Ensemble Learning QA [No note, watch the QA video]; Refer to Lecture23 video on Canvas Refer to Lecture23-QA video on Canvas	Project midpoint report		
Mon, Apr 12			Q12: Decision Tree, Ensemble Learning and Random Forest	
Tue, Apr 13	*Support Vector Machine; *Class Notes; *QA Notes [No note, watch the QA video]; Refer to Lecture24 video on Canvas Refer to Lecture24-QA video on Canvas			KKT and SVM;
Wed,				

Apr 14			
Thu, Apr 15	*Support Vector Machine Class Notes; *Kernel Method \ SVM; *Kernel Method \ SVM Class Notes; *QA Notes [No note, watch the QA video]; Refer to Lecture25 video on Canvas Refer to Lecture25-QA video on Canvas		
Mon, Apr 19			Q13: SVM
Tue, Apr 20	*Kernel Method \ SVM Class Notes: *QA Notes [No note, watch the QA video]; Refer to Lecture26 video on Canvas Refer to Lecture26-QA video on Canvas		
Wed, Apr 21			
Thu, Apr 22	*Ethics in ML; Class Notes [No note, watch the QA video]; Refer to Lecture27 vsideo on Canvas Refer to Lecture27-QA video on Canvas	A4 due	
Mon, Apr 26			Q14: Kernel Method
Tue, Apr 27	*ML Course review; Refer to Lecture28 video on Canvas Refer to Lecture28-QA video on Canvas	Final project due [Undergrad Canvas Access for Projects] [Grad Canvas Access for Projects]	

Course policies

- **Attendance:** Our class will be offered in an Asynch mode. We will not meet in person and every lecture will be live-streamed and recorded. The recordings will be made available to all students after class time. My on-campus class was mandatory and it helped me and my students A LOT to work with each other for a better environment to facilitate learning. The class is in Asynch mode. Therefore, attendance is not required for this class, but I highly recommend students to come to the live-streaming lectures. Trust me it will be fun and you will give me a lot of energy to teach better. The fact that you need to listen to the lectures without fast-forwarding me can help you to learn the materials much better and you will have the chance to ask questions if you are confused anywhere in the lectures.
- **Class deliverables:** All class deliverables will be handled via Gradescope excep quizzes which will be on Canvas. The time span offered to complete the course objectives is plentiful and deadlines will not be extended under any circumstances. To ensure the class is fair for all students, you will receive zero credit for work submitted after the deadline. Regrade requests should be submitted directly on Gradescope within one week of grade publication. Should you find yourself in an impasse with the TA responsible for your grading, feel free to contact the head TA or course instructor.
- **Piazza:**
  - Piazza will be the main and only place for the course discussions and announcements. If you have questions, please ask it on Piazza first because 1) other students may have the same question; 2) you will get help much faster.
  - If it’s something you do not like to discuss publicly on Piazza, you can use private messaging on Piazza.
  - Anytime you want to send a private message to just me on Piazza, please make sure to add our HEAD TAs too in case I may miss your message.
  - **Piazza GOOD questions**

        - I don't understand this part of the lecture, can you explain it to me?
        - This **certain** part of the hw is not clear to me, would it be possible to explain that more?
        - I have a question about the project ...
        - I found an issue on the website, hw or the lectures, can you clarify ...
        - Any feedback, suggestions, ... would be greatly appreciated.
        - Usually, most of the questions are good.
    - **Piazza BAD questions**
        - Can you debug my code? [our team will not do that. You need to be specific about your question]
        - Can you find where the problem is in my code?
- **Exceptional circumstances:** Any request for exceptions to these policies should be made in advance when at all possible. Requests should be due to incapacitating illness, personal emergencies, or similarly serious events. Your request should be accompanied by a supporting letter issued by the Dean of Students.

Diversity and inclusion

Just as machine learning algorithms cannot accomplish complex tasks if trained on datasets of limited variability, our course cannot be successful without appreciating the diversity of our students. In this class we aim to create an environment where all voices are valued, respecting the diversity of gender, sexuality, age, socioeconomic status, ability, ethnicity, race, and culture. We always welcome suggestions that can help us achieve this goal. Additionally, if any of our class scheduled activities conflicts with religious events, please inform the instruction team so that we can make appropriate arrangements for you.

Students with disabilities: your access to this course is extremely important to us. The institute has policies regarding disability accommodation, which are administered through the Office of Disability Services: http://disabilityservices.gatech.edu. Please request your accommodation letter as early in the semester as possible, so that we have adequate time to arrange your approved academic accommodations.

Office hours and questions

Office hours will start on the second week of clases. Please follow the instruction on this Excel Sheet [Undergrad] [Grad] to signup for a 10-minutes slot with one of the TAs. If you require more than ten minutes, please advise the TAs. They'll return to your BlueJeans meeting once they have completed their appointments with other students. You just need to add your name, question of interest and your BlueJeans meeting link. Please do not change the other part of the Excel Sheet [Undergrad] [Grad]. The TA meetings are designed to be one-on-one. Please do not join another student's BlueJeans meeting. The sole exception to this policy being discussions about the project, in which your fellow team members can also join. In addition to the one-on-one meetings open office hours with the instructor will be held weekly where you can ask general questions about the topics covered in class. In-person office hours are only available by appointment and will likely be held outdoors, in line with the aforementioned Georgia Tech's and CDC guidelines with respect to preventing the spread of the coronavirus.

404 Not Found

---

nginx/1.18.0 (Ubuntu)

Grading

- **Assignments (50%)**
  - There will be four assignments. Each one is designed to improve and test your understanding of the materials. Assignments will have both programming and written analysis components.
  - You will need to submit all your assignments using Gradescope. Instructions on how to submit your code and written portions will follow with every assignment. **Handwritten solutions WILL NOT BE ACCEPTED** and you will not receive credit for a handwritten submission.
  - You are required to use Markdown, Latex (watch the tutorial created by our own team [Undergrad Access] [Grad Access] and OverLeaf Latex Example in the Video), or a word processing software to generate your solutions to the written questions. Because **handwritten solutions WILL NOT BE ACCEPTED.**
  - All assignments follow the "no-late" policy. Assignments received after the due date and time will receive **zero credit.**
  - All students are expected to follow the Georgia Tech Academic Honor Code.
  - You can easily export your Jupyter Notebook to a Python file and import that to your desired python IDE to debug your code for assignments.
  - You are **NOT allowed to share any assignment codes or answers with other students**. Piazza is the best place to have discussion regarding assignments and course topics. Discussions are just for the better understanding of questions and should not directly answer the questions.
  - We have 4 big assignments in total. Visit this course's Canvas site for the assignment documents. See the schedule table above for deliverable due dates. (Topics are subject to change)
    - [12.5%] HW1: Linear Algebra, Probability and Statistics, Maximum Likelihood Estimation, Optimization, Information Theory
    - [12.5%] HW2: KMeans, Expectation Maximization, Gaussian Mixture Model, Clustering Evaluation
    - HW3: Singular Value Decomposition, Principal Component Analysis, Linear Regression, Rgularization Naive Bayes
    - [12.5%] HW4: Decision Trees, Random Forest, Support Vector Machine, Neural Networks, CNN

- **Project (30%)**
  - **Proposal (5%)**
    - A project proposal should be written on your GitHub page. It is also a good starter to come up with the first draft of your project.
    - You need to provide us the link to your GitHub page. Make sure your GitHub repository is private.
    - It should be less than 500 words single spaced. References are not the part of the word count.
    - A project proposal should include:
      - Introduction/Background
      - Problem definition
      - Methods
      - Potential results
      - Discussion
      - At least three references (preferably peer reviewed). You need to properly cite the references on your proposal.
    - A checkpoint to make sure you are working on a proper machine learning related project.
    - Your group needs to submit a presentation of your proposal. Please provide us a public link which includes a **3 minutes** recorded video. I found that OBS Studio and GT subscribed Kaltura are good tools to record your screen. Please make your visuals are clearly visible in your video presentation.
  - **Midterm report (10%)**
    - A checkpoint to make sure that you have had major progress in your project. You will add information to your project Proposal and turn it into your midterm report.
    - You need to provide us the link to your GitHub page. Make sure your GitHub repository is private.
    - The midterm report does not have a word count limitation.
    - A project midterm report is quite similar to your proposal with the exception of having actual results instead of potential ones:
      - Introduction/Background
      - Problem definition
      - Data Collection
      - Methods
      - Results
      - Discussion
    - We need to see where you obtain your data and how you have done your data cleaning. Make sure to talk about the features and different feature selection approaches.
    - You do not submit any video for the midterm report.
  - **Final report (15%)**
    - You need to provide us the link to your GitHub page. Make sure your GitHub repository is private.
    - A final report should include:
      - Introduction/Background
      - Problem definition
      - Data Collection
      - Methods

        - Results
        - Discussion
        - Conclusions
    - Your group needs to submit a presentation of your final report. Please provide us a public link which includes a **6 to 9 minutes** recorded video. I found that OBS Studio and GT subscribed Kaltura are good tools to record your screen. Please make your visuals are clearly visible in your video presentation.

- **General project guidance**
    - Your project will be graded based on the following criteria:
      Was the motivation clear?
        - What is the problem?
        - Why is it important and why we should care?

      Were the dataset and approach used effectively?
        - How did you get your dataset?
        - What are its characteristics (e.g. number of features, # of records, temporal or not, etc.)
        - Why do you think your approach can effectively solve your problem?
        - What is new in your approach?

      Were the experiments, results, and conclusion satisfactory?
        - How did you evaluate your approach?
        - What are the results?
        - How do you compare your method to other methods?

      How was the presentation in general?
        - Finished on time?
        - Effective visualizations? (Are they relevant? Do they help you better understand the project's approaches and ideas?)
        - Use of text (Succinct or verbose?)
    - **Undergrad students** can ONLY team up with Undergrad Students, and **Grad students** can ONLY team up with Grad students. If you are in a **Grad students** team, you are required to have both unsupervised and supervised learning in your project. I highly recommend **Undergrad students** to use both unsupervised and supervised learning in your project. However, if you were to pick one, please go with supervised learning.
    - In order for you to obtain hands-on experience applying the topics covered in this course, you are expected to complete a term project utilizing real-world data. The project will encompass both unsupervised and supervised learning.
    - Each project needs to be completed in a team of five people (you will be forming your team on your own. In case you cannot find a team, we will randomly assign you a team). Team members need to clearly claim their contributions in the project report. Once your teams have been formed and you have selected a topic, you will be assigned a mentor, who will provide you with general guidance on your project. It is important to note that your team will lead the project effort: obtaining the data, researching data-driven approaches to accomplish your project goal and coordinate your own activities. The role of the mentor is solely to advise you, should you find yourself stuck and unable to make progress. We also accept a team of four, if you really cannot find the fifth team member.
    - You will create a **GitHub page** page for your project, which you will use to publish your main deliverables. There will be three deliverables published to your GitHub: a proposal, a midterm checkpoint, and a final report.
    - **Seminars:** To help you conduct your project successfully, We have project seminars where one or two TAs will present their research or industrial projects. Doing so, you will gain a good sense of what it is being done in both Academia and Industry. Besides that, students can ask general questions about their class project and how to improve that in each seminar. Seminars will be streamed online and recorded and they will be published on the course website. Similar to the class lectures, Please ensure that you join to these seminars and get yourself familiar with the practical and real-world application of ML. They will be held on Wednesdays from 11:00 am to 12:00 pm. Timing may change for some of the seminars. We will have Piazza post for each seminar, and its exact time. All the information will be updated for each seminar on the course webiste with a BJ Event link.
    - Google colaboratory allows free access to run your Jupyter Notebook. I strongly suggest you use it for your project, especially for teams that are going to employ Deep Learning. Don't forget to take advatage of Google Cloud Platform and AWS Educate as well.

- **Quizzes (14%)**
    - There will be 14 quizzes throughout the semester.
    - We will consider your top 9 quizzes' scores. Each quiz will have 1.55% of your final score.
    - All quizzes are mandatory to be taken even if they do not count toward your final grade.
    - The topic of each quiz will coincide roughly with the content covered in class on that week.
    - Quizzes will have a duration of **seven-minutes for Undergrad** students and **six-minutes for Grad students**. Each quiz will have five multiple choice questions. They will be available from 12:00 am AOE Monday until 11:59 pm AOE Monday.
    - Quizzes measure your understanding of the topics and they will be more conceptual questions.
    - Quizzes' answers will be released on Saturdays. Please do not ask any questions about a quiz that you just take on Piazza before we release the answers.
    - Quizzes questions are selected randomly from our question bank, which means that students will not receive

the same questions for their quiz.

- **Class participation (6%)**
  - Piazza has statistics which give us many measurements regarding how much a student has been involved on Piazza's activities such as viewing posts, answering questions, asking questions and so on. We use this to account for your Class Participation score. At the end of the semester, we will define a minimum and maximum number of involvement considering all the students and your grade will be defined based on that.

- **Bonus points (up to 8%)**
  - **About bonus points:** Bonus points will be counted to always be beneficial for your final grade. More information on bonus points for assignments will be provided as the semester progresses. If it becomes necessary to curve grades, bonus points will be applied after curving, not before.
  - **Undergrad and grad:** You can obtain up to 5% bonus points would be answering the challenging questions we may have in some of the hws.
  - **Undergrad:** You will notice that we have bonus points for all the hws, where grad students are required to answer those questions, but it will be optional for undergrad students. You will receive up to 3%, if you answer those questions. Note that these are different than the challenging questions. Challenging questions are bonus for both grad and undergrad.
  - **How does it work?** For example, hw 1 may have 30 bonus points, hw 2 may have 20 bonus points and so on. If you receive all the bonus points for all your hws, we will add 5% to your final grade. If you are an undergrad and you answer all the challenging and Grad students questions, you will receive 8%.

COVID-19 Policy

The fall semester 2020 is especially challenging due to the Covid-19 pandemic and a growing awareness of racial inequities. The following information relates to specific services and guidelines for courses during this semester. The most up-to-date information on Covid-19 is on the TECH Moving Forward website and in the Academic Restart Frequently Asked Questions.

Expectations and Guidelines

Each of us has a responsibility to ourselves and our fellow Yellow Jackets to be mindful of our shared commitment.

- We are all required to wear a face covering while inside any campus facilities/buildings, including during in-person classes, and to adhere to social distancing of at least 6 feet. If an individual forgets to bring a face covering to class or into any indoor space, there will be a clearly marked supply of these in each building. If a student fails to follow Georgia Tech's policies on social distancing and face coverings, they will initially be reminded of the policy and if necessary, asked to leave the class, meeting, or space. If they still fail to follow the policy, they may be referred to the Office of the Dean of Students. Information on the Institute's policy on face coverings.
- Students are expected to sit in assigned seats and to come to class only on days that are assigned to them.
- Papers, projects, tests, homework, and other assignments will only be accepted in electronic form unless the assignment is a physical artifact.

Additional information is available in the Student Guidebook.

Instructor Illness or Exposure to Covid-19

During the fall 2020 semester, some faculty members may be required to quarantine due to exposure or isolate due to a Covid-19 diagnosis. Some disruption to classes or services is inevitable, but Georgia Tech is making every effort to ensure continuity of operations. As is the case in any semester, faculty may cancel a class if they have an illness or emergency situation and cover any missed material at their own discretion. If an instructor needs to cancel a class, they should notify students as early as possible.

Faculty who are staying home due to symptoms should monitor their health closely and consult with their school chair to determine if remote instruction or substitute instruction is most appropriate for the course. If they need to cancel a class repeatedly, a backup will be supplied in the form of a temporary substitute instructor or asynchronous work. No course will be canceled after the first class has occurred.

If you have not tested positive but are ill or have been exposed to someone who is ill, please follow the Covid-19 Exposure Decision Tree for reporting your illness.

Student Illness or Exposure to Covid-19

During the semester, you may be required to quarantine or self-isolate to avoid the risk of infection to others. Quarantine is the separation of those who have been exposed to someone with Covid-19 but who are not ill; isolation is the separation of those who have tested positive for Covid-19 or been diagnosed with Covid-19 by symptoms.

If you have not tested positive but are ill or have been exposed to someone who is ill, please follow the Covid-19 Exposure Decision Tree for reporting your illness.

During the quarantine or isolation period you may feel completely well, ill but able to work as usual, or too ill to work until you recover.

Remote courses and remote class sessions during hybrid courses. Unless you are too ill to work, you should be able to complete your remote work while in quarantine or isolation.

If you are ill and unable to do course work this will be treated similarly to any student illness. The Dean of Students will have been contacted when you report your positive test or are told that it is necessary to quarantine and will notify your instructor that you may be unable to attend class events or finish your work as the result of a health issue. Your instructor will not be told the reason. We have asked all faculty to be lenient and understanding when setting work deadlines or expecting students to finish work, and so you should be able to catch up with any work that you miss while in quarantine or isolation. Your instructor may make available any video recordings of classes or slides that have been used while you are absent, and may prepare some complementary asynchronous assignments that compensate for your inability to participate in class sessions. Ask your instructor for the details.

CARE Center, Counseling Center, Stamps Health Services, and the Student Center

These uncertain times can be difficult, and many students may need help in dealing with stress and mental health. The **CARE Center** and the **Counseling Center**, and **Stamps Health Services** will offer both in-person and virtual appointments. Face-to-face appointments will require wearing a face covering and social distancing, with exceptions for medical examinations. Student Center services and operations are available on the **Student Center** website. For more information on these and other student services, contact the Vice President and Dean of Students or the **Division of Student Life**.

Accommodations for Students at Higher Risk for Severe Illness with Covid-19

Students may request an accommodation through the Office of Disability Services (ODS) due to 1) presence of a condition as defined by the Americans with Disabilities Act (ADA), or 2) identification as an individual of higher risk for Covid-19, as defined by the Centers for Disease Control (CDC). Registering with ODS is a 3-step process that includes completing an application, uploading documentation related to the accommodation request, and scheduling an appointment for an “intake meeting” (either in person or via phone or video conference) with a disability coordinator.

If you have been approved by ODS for an accommodation, I will work closely with you to understand your needs and make a good faith effort to investigate whether or not requested accommodations are possible for this course. If the accommodation request results in a fundamental alteration of the stated learning outcome of this course, ODS, academic advisors, and the school offering the course will work with you to find a suitable alternative that as far as possible preserves your progress toward graduation.

Resources

No textbook will be required for this course, however you are strongly encouraged to complete the readings indicated for each class. You may also find the following books very helpful:

- Learning from data, by Yaser S. Abu-Mostafa
- Pattern recognition and machine learning, by Christopher Bishop
- Machine learning, by Tom Mitchell
- Data Mining: Concepts and Techniques, by Jiawei Han, Micheline Kamber, and Jian Pei
- The Elements of Statistical Learning, by Trevor Hastie, Robert Tibshirani, and Jerome Friedman
- Deep Learning, by Ian Goodfellow, Yoshua Bengio, and Aaron Courville

Other resources, such as machine learning toolboxes and datasets, will be provided throughout the course.

Dataset Ideas (may need API, or scraping) - Thanks to Polo and everyone who contributed with suggestions to these datasets

- Google Dataset Search
- Google public datasets.
- Kaggle public datasets
- Awesome Public Datasets.
- NYC Taxi data for 2013 Trip Data (11.0GB). 2013 Fare Data (7.7GB). Visualization for a days trip.
- Large datasets publicly available.
- Georgia Tech's campus data (has APIs): bus info, directory, building, T-square, room reservation, building facilities usage (e.g., electricity, lights, A/C, etc.), Oscar/course info/registration, etc.
- Yahoo WebScope
- Data.gov: U.S. Government's open data
- IPEDS data: Postsecondary education data from National Centre for Education Statistics
- Bureau of Labor Statistics data
- Uber data: Anonymized data from over 2 billion trips
- Freebase
- Yelp
- Microsoft Academic Graph
- Numerous APIs from Google (e.g., Maps, Freebase, YouTube, etc.)
- Zillow: real estate listing site

- Numerous graph datasets (large and small): SNAP, Konect
- Movies data: IMDB
- List of lists of datasets for recommendations.
- Million song dataset by Echo Nest.
  It contains not only the basic information of songs (artist, genre, year, length etc), but also some musical features(like tempo, pitch, key, brightness).
- Dataset about soccer games, players, clubs.
  No API, but easy to scrape.
  For a soccer player: transfer history, performance, nationality, birth date, etc.
  For a soccer club: performance, squad, etc.
- The Free 'Big Data' Sources Everyone Should Know
- Quandl - a dataset search engine for time-series data.
- UCI also has a collection of links to various datasets sorted for various tasks (Classification, Regression, etc)
- Amazon AWS Public Data Sets
- KDD Cup: annual competition in data mining, like Kaggle
- Academic domain: Microsoft Academic Search, DBLP
- Retrosheet: MLB statistics (Game/Play logs)
- Classification datasets
- Various geophysical datasets for the oceans (magnetism, gravity, seismology, etc).
- Social trends
- Beer data Website offline :(. Older version at web.archive.org
- Academic torrents (terabytes)
- Article Search API from the New York Times (all the way back to 1851!)
- Civil Engineering Dataset
- (Kayak: flight, hotel, car, etc.)
- Data Science Initiative - Microsoft Research has various datasets and access to tools that can aid in data science research

×

Cite

Copy Download