CSCI 4360/6360

Data Science II

Mondays, 3:00 - 3:50, Forest Resources-4 0517

Tuesdays & Thursdays, 2:20 - 3:35, Physics 0303

Dr. Shannon Quinn

Email: `squinn@cs.uga.edu`
Website: `https://eds-uga.github.io/csci-x360-fa21`
Office Location: Boyd GSRC 638A
Office Hours: `Office Hours` channel on Discord

The course syllabus is a general plan; when (not if) deviations arise, they will be announced.

Course Description: This course introduces the students to advanced analytics techniques in data science, including computer vision, semi-supervised learning, spectral analytics, randomized algorithms, and deep learning.

This course aims to provide students with deep knowledge of sophisticated data science techniques for making sense of data across domains. Students are instructed how to process high-dimensional spatiotemporal data, use evolutionary models or hybrid techniques such as semi-supervised learning, and design novel metrics to measure data. Furthermore, students are given the opportunity to explore the latest tools in the PyData scientific computing stack, as well as tools in other domains (e.g., Julia). The course is appropriate both for students preparing for research in Data Mining and Machine Learning, as well as Bioinformatics, Data Science, and students who want to apply Data Mining techniques to solve problems in their fields of study.

It is assumed that students taking this course have thorough knowledge of basic machine learning concepts (classification, clustering, regression). Very little time will be spent introducing the fundamental machine learning concepts covered in CSCI 3360 Data Science I.

It is NOT assumed that students taking this course will have extensive prior programming knowledge, but it WILL be expected that students can pick this up quickly following one or two class meetings covering the Python programming language. Additionally, the first few meetings will also entail a thorough review of fundamental linear algebra concepts.

Prerequisites: Any one or more of the following:

- CSCI 3360 Data Science I (undergraduates)
- CSCI 6380 Data Mining (graduates)
- CSCI 6850 Biomedical Image Analysis (graduates)

It is highly, highly, **HIGHLY recommended** that you take one of the above prerequisites OR have prior knowledge of and experience with machine learning / linear algebra / probability and statistics techniques before you take this course.

Credit Hours: 4

No required textbooks! Recommended texts include:

1. Trevor Hastie, Robert Tibshirani, Jerome Friedman. *Elements of Statistical Learning* (2^{nd} ed., 2016). ISBN-13: 978-0387848570.

2. Trevor Hastie, Robert Tibshirani, Martin Wainwright. *Statistical Learning with Sparsity* (1^{st} ed., 2015). ISBN-13: 978-1498712163.

3. Christopher Bishop. *Pattern Recognition and Machine Learning* (1^{st} ed., 2006). ISBN-13: 978-0387310732.

4. Kevin Murphy. *Machine Learning: A Probabilistic Perspective* (1^{st} ed., 2012). ISBN-13: 978-0262018029.

5. Stephen Boyd and Lieven Vandenberghe. *Convex Optimization* (1^{st} ed., 2004). ISBN-13: 978-0521833783.

6. Richard Szeliski. *Computer Vision: Algorithms and Applications* (2011^{th} ed., 2010). ISBN-13: 978-1848829343.

7. Andrew Blake and Michael Isard. *Active Contours* (1^{st} ed., 1998). ISBN-13: 978-1447115571.

8. Daphne Koller and Nir Friedman. *Probabilistic Graphical Models* (1^{st} ed., 2009). ISBN-13: 978-0262013192.

9. Michael W. Mahoney. *Lecture Notes on Spectral Graph Methods* (arXiv, 1608.04845, 2016).

10. Eric D. Kolaczyk. *Statistical Analysis of Network Data* ((2009^{th} ed., 2009). ISBN-13: 978-0387881454.

11. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning* (1^{st} ed., 2016). ISBN-13: 978-0262035613.

12. Tom Mitchell. *Machine Learning* (1^{st} ed., 1997). ISBN-13: 978-0071154673.

Topical Course Outline

1. Lightning review of basic linear algebra, programming concepts in Python, Data Science I fundamentals
2. Dense motion analysis and autoregressive models
3. Object segmentation and tracking using linear dynamical systems
4. Graphs
5. Spectral clustering, semi-supervised learning on graphs
6. Metric learning
7. Kernel and Sparse PCA, dictionary learning, kernel methods
8. Fundamentals of deep neural networks
9. Backpropagation, convolutional and recurrent networks, autoencoders

10. Deep generative models
11. Ethical data science
12. Current data science tools and frameworks

Grade Distribution:

Homeworks	45%
Workshops	15%
Midterm exam	40%
Final project	40%
Total	**140%**

There are five homework assignments; only **three** are required. Students can **choose** whether to take the midterm exam **or** participate in the final [team] project, which will result in a final grade tabulation of 100%. Students in 4360 are required to organize and present **one workshop** during the semester; students in 6360 are required to organize and present **two workshops** during the semester.

Course Policies

- **Announcements**
  - I use either Slack or Discord for disseminating information and making course announcements outside of in-person class meetings. Signing on is **required** for the class, as is checking for new announcements at least once every 24 hours. Beyond that is entirely at your discretion.
  - While you can and are more than welcome to directly DM or email me with any questions you have, I **strongly recommend** you post your questions in the `#questions` forum of the Slack/Discord chat. In over 5 years of offering this course, I can count on one hand total the number of times I've been asked questions by students whose answers pertained only to that student. Far, far more often, I get 5-10 DMs asking me the exact same question. If you're struggling with something, chances are someone else is, too!
- **Attendance**
  - **There is no attendance requirement.** Until the pandemic eases, and/or mask/vaccine mandates are put in place, you are welcome to attend class or not at your discretion, no permission required. In general this is also the case: you are all adults and do not need my permission to attend class or not, for any reason.
  - To reiterate: don't *ask* if you can miss a day of lecture for a doctor's appointment / family event / nap, just *inform* me that you'll be missing lecture on a given day (don't even need to explain why) so that I don't go looking for you.
  - **If you are experiencing symptoms, do NOT attend class.** Please notify me that you will be out of class for a bit so we can coordinate getting you the needed materials.
  - All class meetings will be broadcast over Discord. Office hours will not be recorded, but lectures will, and they will be posted as soon as the videos are rendered.
- **Homework**

- Homework assignments are due by 11:59:59pm on the noted date. Assignments submitted after this deadline will lose 25/100 points for every subsequent 24 hour-period they are late. **This takes effect as of 12:00:00am after the deadline**.
- Homework assignments are to be done **individually**. You are more than welcome to collaborate on them, but please make a note of who you collaborated with on your submission.

- **Final projects**
  - The final project is team-based, ideally on teams of 2 or 3. As such, only one set of deliverables is required per team.
  - The final project has several components. The first is the **proposal**, a 1-2 page document outlining the major elements of your project, including but not limited to: the title, who your teammates are, the overall goals of your project, and how you will clearly delineate success from failure.
  - The second are a pair of **updates** spaced out by a few weeks. These updates are 1-page documents that identify progress on each of your objectives since the last update/proposal, any roadblocks you have run into, how you worked around / are planning to work around them, and any deviations you are taking from your original proposal.
  - The third is a **presentation**, given during the final week of class. This is a 30-minute talk (25-minute presentation on your overall project, with 5 minutes for Q&A) in which all or some of your teammates can participate. This should be a deep and thorough dive into the background, motivation, goals, findings, and conclusions of your project and will constitute the majority of the grade on the final project. Grades will be determined based on completeness of your findings, degree to which you achieved your initial objectives, and your ability to adjust to unexpected obstacles / changes.
  - The fourth is a **peer review**, to be completed by one of the other project teams. Ahead of presentations, each project team will be randomly assigned another project team to review. These reviews will completely anonymous, and will provide constructive feedback on the strengths and weaknesses of the projects based on their presentations.

Academic Honesty

As a University of Georgia student, you have agreed to abide by the University's academic honesty policy, "A Culture of Honesty," and the Student Honor Code. All academic work must meet the standards described in "A Culture of Honesty" found at: `https://ovpi.uga.edu/academic-honesty/academic-honesty-policy`. Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation. Questions related to course assignments and the academic honesty policy should be directed to the instructor.

- Read "A Culture of Honesty," the UGA academic honesty policies and CS Academic Integrity policies.
- You must not allow others to copy or look at your work.
- You must not give/share your lab/project assignment work to a fellow student.
- Copying significant portions of code from a fellow student or any other source (including internet) is plagiarism and will be dealt with as such.

- Seriously though: I've caught a few students who tried to use previous years' solutions. It's harder than you think to pass off code written by others as your own, especially with machine learning on my side. You can use blocks of code here and there **provided you cite where you got them**, but please assemble the full solution yourselves.
- If in doubt, **cite**.
- If you have questions about an assignment or if you run into problems, **ask me**.
- This course has no exams or quizzes, only projects, so please don't ask about extra credit. There is none.
- All of your coursework must meet the aforementioned policies and rules. Students that violate any of these rules or the UGA Academic Honesty policies will be liable to a penalty. The instructor will strictly enforce Academic Honesty policies and report any violation of the aforementioned policies and rules.

****Highly Tentative** Course Outline**
Subject to change.

Week	Content
Week 1	• Course introduction; Python crash course • *Workshop 0*
Week 2	• Linear algebra review • *Workshop 1; Homework 1 out*
Week 3	• Dense Motion Analysis • *Workshop 2*
Week 4	• Linear dynamical systems • *Workshop 3 ; Homework 1 due*
Week 5	• Ethics in Data Science, Technical debt in machine learning • *Workshop 4; Homework 2 out*
Week 6	• Graphs • *Workshop 5*
Week 7	• Spectral Clustering • *Workshop 6; Homework 2 due*
Week 8	• Semi-supervised learning on graphs • *Midterm Exam; Homework 3 out*
Week 9	• Metric learning • *Workshop 7; Final project proposals*
Week 10	• Evolutionary Computing • *Workshop 8; Homework 3 due; Final project assigned*
Week 11	• Kernel and Sparse PCA, Dictionary learning • *Workshop 9; Final project update #1; Homework 4 out*
Week 12	• Deep neural networks, backpropagation • *Workshop 10*
Week 13	• Convolutional and recurrent neural networks • *Workshop 11; Final project update #2; Homework 4 due*
Week 14	• Autoencoders and deep generative models • *Workshop 12; Homework 5 out*
Week 15	• Course Wrap-Up • Final project talks
Week 16	• Final project talks • *Homework 5 due*