CSCI 8000 (CSCI – 51343): Advanced Topics in Data Mining

Fall 2021

1   Course Information

- **Instructor**: Dr. Ninghao Liu (Email: ninghao.liu@uga.edu)

- **Format**: Face-to-face

- **Time&Location**:

  – Tuesday and Thursday: 3:55pm - 5:10pm (Boyd Grad Rsch Ctr 306)
  – Wednesday: 4:10pm - 5:00pm (Boyd Grad Rsch Ctr 208)

- **Office Hours**:

  – Thursday 11:00 - 12:00 AM, or by an email appointment
  – Location: 616 Boyd Graduate Studies Research Center

- Piazza sign-up link: `piazza.com/uga/fall2021/csci8000`

2   Course Description

The goal of this course is to familiarize students with several advanced topics in data mining, including anomaly detection, text mining, network analysis, and recommender systems. Students will have the opportunity to learn the recent advances not only in research theories but also in various industrial applications.

This is a graduate-level and research-oriented course. Students are expected to have a prior background in data mining or machine learning. In this course, the instructor will review the basic concepts of data mining, with brief introduction about the advanced topics as preliminaries. Then, students will in turn present research papers from a reading list. In addition, students will work on a course project on data mining related methodology or applications.

- **Recommended Prerequisites**: CSCI 4380/6380: Data Mining

- There are no final exams.

3   Course Topics

- Anomaly Detection

  – Fundamental Methods:
    1. Statistical methods
    2. Proximity-based methods
    3. Clustering-based methods
    4. Classification-based methods

- PCA, Autoencoders for Outlier Detection

- Text Representation Learning and Text Mining

  - Text Retrieval
    1. TF-IDF
    2. Vector space representation
  - Embedding methods
    1. SVD
    2. Word2Vec, EMLO
    3. BERT

- Network Analysis

  - Link analysis, Page-Rank
  - Embedding methods
    1. Deepwalk, LINE
    2. Graph Neural Networks

- Recommender Systems

  - Classical models
  - Deep models for recommender system

4 Course Materials

- **Textbook**: No book is required for this course, but the following books are recommended:

  1. Jiawei Han, Micheline Kamber, Jian Pei. "Data Mining: Concepts and Techniques" (3rd Ed.) (ISBN-10: 9780123814791)
  2. Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze. "Introduction to Information Retrieval" (ISBN-10: 0521865719)

- **Reading Materials**: Research papers relevant to the topics covered in this course will be introduced in the course.

  - "LOF: identifying density-based local outliers." SIGMOD. 2000.
  - "Support vector method for novelty detection." NIPS. 1999.
  - "Isolation forest." ICDM. 2008.
  - "Anomaly detection with robust deep autoencoders." KDD. 2017.
  - "Efficient estimation of word representations in vector space." 2013.
  - "A simple but tough-to-beat baseline for sentence embeddings." ICLR. 2016.
  - "Attention Is All You Need." NeurIPS. 2017.
  - "Deepwalk: Online learning of social representations." KDD. 2014.
  - "Line: Large-scale information network embedding." WWW. 2015.
  - "Semi-supervised classification with graph convolutional networks." ICLR. 2017
  - "Item-based collaborative filtering recommendation algorithms." WWW. 2001.
  - "Item2vec: neural item embedding for collaborative filtering." 2016.
  - "Real-time personalization using embeddings for search ranking at airbnb." KDD. 2018.
  - (P1) "Few-shot Network Anomaly Detection via Cross-network Meta-learning." WWW. 2021.

- (P2) "Language Models as Knowledge Bases?" EMNLP. 2019.
- (P3) "OpenWGL: Open-World Graph Learning." ICDM. 2020.
- (P4) "Disentangling User Interest and Conformity for Recommendation with Causal Embedding." WWW. 2021.

5 Grading Policy

- **Homework assignment:** 35%

  - Assignment 1: 15%
  - Assignment 2: 20%

- **Paper review:** 15%

- **Course project:** 50%

  - Proposal presentation: 15%
  - Final presentation: 10%
  - Final report: 25%
  - No regular final exams

- **Extra Credit - Advanced Paper Presentation:** 5%

6 Final Letter Grades

[90, 100] : A	[87, 90) : A-	[84, 87) : B+	[80, 84) : B	[77, 80) : B-
[74, 77) : C+	[70, 74) : C	[67, 70) : C-	[60, 67) : D	[0, 60) : F

7 Late Submission and Regrading Policy

For homework assignments, 20% is deducted for each late day for up to 48 hours (including weekends) after which submissions are not accepted. Late presentation materials and project reports will not be accepted.

You may request a re-grade of any graded item any time within 7 days of receiving the grade on eLC. For grades posted on or after reading day, students must send a regrade request within 2 days after the grade was posted on eLC. To make a regrade request, you should send a regrading request from your UGA email account. If a rubric is posted for an assignment, then the regrade request must include which parts of the grading rubric were incorrectly graded. Regrade requests may result in a lower grade.

8 Communication

Important announcements will be made through eLC. Lecture notes/slides and grades will also be posted to that site. You need to **turn on email notifications related to Announcements and Grades**. To do so, log into eLC, click on your name/profile picture on the upper right hand corner, click "Notifications", scroll down and check the appropriate boxes. **Any course materials downloaded from eLC must not be shared with anyone, posted on any websites, or used for commercial purposes**.

When emailing the instructor, please include a [CSCI8000] tag in the subject line. Please include your full name in the email. Students must email the instructor using their official UGA addresses (yourEmail@uga.edu); the instructor will not respond to emails from other domains. Please allow 24-hours for a response on a weekday, and 48-hours for a response on the weekend or holiday/break.

9 Course Project Description

The course project consists of a proposal presentation, a final presentation, and a final report. **Two students** will form a group for each project.

9.1 Proposal Presentation

- Both team members should present their slides.
- Each presentation is: a talk (28 minutes) + QA (7 minutes) = 35 minutes.
- Slides should be emailed to the instructor by midnight (11:59 pm) before the scheduled lecture.
- Content to be covered:
  - Background
  - Related work (could be the papers in paper reviews)
  - The problem to be tackled in the project, e.g.,
    * **New Algorithms/Models**: Extend the state-of-the-art methods, derive solutions, and validate it on benchmark datasets
    * **New Applications**: Identify a new application of a model, especially interdisciplinary research
  - Tentative experimental settings: Benchmark datasets, baseline methods, evaluation metrics.

9.2 Final Presentation

- Scheduled in the final exam week (Tuesday, Dec. 14, 3:30 - 6:30 pm).
- Both team members should present their slides.
- Each presentation is talk (15 minutes) + QA (3 minutes) = 18 minutes.
- Slides should be emailed to the instructor by the presentation start time.
- Content to be covered:
  - The formal problem definition
  - The details of your solution
  - Experimental results on benchmark datasets, when compared with baseline methods (usually ≥ 2), under the evaluation metrics
  - Conclusion

9.3 Final Project Report

- Latex template: `https://www.overleaf.com/gallery/tagged/acm-official#.WOuOk2e1taQ`. Select "ACM Conference Proceedings" as the template. Change "documentclass[manuscript,screen,review]{acmart}" to "**documentclass[sigconf]{acmart}**"
- Length: ≥ 2.5 pages + unlimited references
- Format:
  - Introduction
  - Problem Definition
  - The Propose Method

- Experiments
- Conclusion
- References

10 Paper Review Details

The goal of paper review is to make students be more familiar with the topic of the course project.

- Students will be required to read **5 papers** and write a review for each of them.
- The paper report should be prepared using using a word processing system and be roughly **1.5 pages** in length (single-spaced, at most 11 point font, about **800 words**). The detailed format:
  - Paper Bibliography Information: title, leading author, where and when published.
  - Problem Statement: In your own words, please identify the problem addressed in the paper and why the problem is significant.
  - Proposed Solution: Describe the solution method proposed by the author(s) or the analysis completed in the paper.
  - Results/Findings: Summarize the major results/findings/conclusions of the paper.
  - Contribution to the Project: Why you choose this paper.
- **Where can you choose the papers from?**
  - ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)
  - The Web Conference (WWW)
  - The Conference on Information and Knowledge Management (CIKM)
  - International Conference on Web Search and Data Mining (WSDM)
  - ACM Special Interest Group on Information Retrieval (SIGIR)
  - Annual Meeting of the Association for Computational Linguistics (ACL)
  - The Conference on Empirical Methods in Natural Language Processing (EMNLP)
  - Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)
  - Conference on Neural Information Processing Systems (NeurIPS)
  - International Conference on Machine Learning (ICML)
  - The International Conference on Learning Representations (ICLR)
- Please inform the instructor and get approval if your really want to include papers beyond the venues listed above.
- For each student, at most 1 paper could be selected from the "Reading Materials".
- **At least 3** of the reviewed papers should be **officially** published in **2019, 2020, 2021, or later**.
- Paper reviews are done individually (students in the same group should not submit reviews of the same paper).
- Must be chosen from **full paper**.

11 Tentative Timeline

- Group registration due: Tuesday, Aug. 27, 11:59 pm
- Paper reviews due:
  - Paper 1&2 reviews: Thursday, Sept. 16, 11:59 pm
  - Paper 3&4&5 reviews: Thursday, Oct. 21, 11:59 pm
- Proposal presentation: Oct. 26 ∼ Oct. 28, Nov. 2 ∼ Nov. 4
- Final presentation: Tuesday, Dec. 14, 3:30 - 6:30 pm
- Final report due: Tuesday, Dec. 14, 11:59 pm

12 Academic Honesty

UGA Student Honor Code: "I will be academically honest in all of my academic work and will not tolerate academic dishonesty of others."

You must abide by the academic honesty policy of the University of Georgia. The penalties for academic dishonesty are severe and ignorance is not an acceptable defense. For more detail, please see the UGA policy at: `https://honesty.uga.edu/Academic-Honesty-Policy/`. Questions related to course assignments and the academic honesty policy should be directed to the instructor.

13 Accommodations for Disabilities

Students with a disability or health-related issue who need a class accommodation should make an appointment to speak with the instructor as soon as possible. Paperwork from the Disability Resource Center (`https://drc.uga.edu/`) will be required.

14 Mental Health and Wellness Resources

- If you or someone you know needs assistance, you are encouraged to contact Student Care and Outreach in the Division of Student Affairs at 706-542-7774 or visit `https://sco.uga.edu/`. They will help you navigate any difficult circumstances you may be facing by connecting you with the appropriate resources or services.

- UGA has several resources for a student seeking mental health services (`https://www.uhs.uga.edu/bewelluga/bewelluga`) or crisis support (`https://www.uhs.uga.edu/info/emergencies`).

- If you need help managing stress anxiety, relationships, etc., please visit BeWellUGA (`https://www.uhs.uga.edu/bewelluga/bewelluga`) for a list of FREE workshops, classes, mentoring, and health coaching led by licensed clinicians and health educators in the University Health Center.

- Additional resources can be accessed through the UGA App.

15 Coronavirus Information for Students

- **Face Covering.** Following guidance from the University System of Georgia, face coverings are recommended for all individuals while inside campus facilities.

- **Guidelines for COVID-19 Quarantine Period.** Students who are fully vaccinated do not need to quarantine upon exposure unless they have symptoms of COVID-19 themselves. All others should follow the Georgia Department of Public Health (DPH) recommendations:

  Students who are not fully vaccinated and have been directly exposed to COVID-19 but are not showing symptoms should self-quarantine for 10 days. Those quarantining for 10 days must have been symptom-free throughout the monitoring period and continue self-monitoring for COVID-19 symptoms for a total of 14 days. You should **report the need to quarantine on DawgCheck** (`https://dawgcheck.uga.edu/`), and communicate directly with your faculty to coordinate your coursework while in quarantine. If you need additional help, reach out to Student Care and Outreach (sco@uga.edu) for assistance.

Students, faculty and staff who have been in close contact with someone who has COVID-19 are no longer required to quarantine if they have been fully vaccinated against the disease and show no symptoms.

- **How can I obtain the COVID-19 vaccine?** University Health Center is scheduling appointments for students through the UHC Patient Portal (`https://patientportal.uhs.uga.edu/login_dualauthentication.aspx`). Learn more here – `https://www.uhs.uga.edu/healthtopics/covid-vaccine`.

  The Georgia Department of Health, pharmacy chains and local providers also offer the COVID-19 vaccine at no cost to you. To find a COVID-19 vaccination location near you, please go to: `https://georgia.gov/covid-vaccine`.

  In addition, the University System of Georgia has made COVID-19 vaccines available at 15 campuses statewide and you can locate one here: `https://www.usg.edu/vaccination`.

- **What do I do if I have COVID-19 symptoms?** Students showing COVID-19 symptoms should self-isolate and schedule an appointment with the University Health Center by calling 706-542-1162 (Monday-Friday, 8 a.m.-5p.m.). Please DO NOT walk-in. For emergencies and after-hours care, see, https://www.uhs.uga.edu/info/emergencies.

- **What do I do if I test positive for COVID-19?** If you test positive for COVID-19 at any time, you are required to **report it through the DawgCheck Test Reporting Survey**. We encourage you to stay at home if you become ill or until you have excluded COVID-19 as the cause of your symptoms. UGA adheres to current Georgia Department of Public Health (DPH) quarantine and isolation guidance and requires that it be followed. Follow the instructions provided to you when you report your positive test result in DawgCheck.

Disclaimer: *The course syllabus is a general plan for the course; deviations announced to the class by the instructor may be necessary.*