CSE6250 Syllabus (O01/OAN) Fall2020

Big Data Healthcare

Instructor Information

Instructor	Email	Office Hours & Location
Jimeng Sun	jsun@cc.gatech.edu	Schedule by Email
Teaching Assistant(s)	**Email**	**Office Hours & Location**
Ming Liu	mliu302@gatech.edu	TBD

General Information

Description

Data science plays an important role in many industries. In facing massive amount of heterogeneous data, scalable machine learning and data mining algorithms and systems become extremely important for data scientists. The growth of volume, complexity and speed in data drives the need for scalable data analytic algorithms and systems. In this course, we study such algorithms and systems in the context of healthcare applications.

In healthcare, large amounts of heterogeneous medical data have become available in various healthcare organizations (payers, providers, pharmaceuticals). This data could be an enabling resource for deriving insights for improving care delivery and reducing waste. The enormity and complexity of these data-sets present great challenges in analyses and subsequent applications to a practical clinical environment. In this course, we introduce the characteristics of medical data and associated data mining challenges on dealing with such data. We cover various algorithms and systems for big data analytics. We focus on studying those big data techniques in the context of concrete healthcare analytic applications such as *predictive modeling*, *computational phenotyping* and *patient similarity*. We also study big data analytic technology:

1. Scalable machine learning algorithms such as online learning and fast similarity search;
2. Big data analytic system such as Hadoop family (Hive, Pig, HBase), Spark and Graph DB

Pre- &/or Co-Requisites

1. Good machine learning and data mining concepts such as classification and clustering;
2. Proficient programming and system skills in Scala , Python and Java;
3. Proficient knowledge and experience in dealing with data and understand the ETL process(recommended skills include SQL, NoSQL such as MongoDB)
4. Minimum grade of C for MATH 3215 or MATH 3225 or ECE 3077 or ISYE 2027. Two of the following:

- CX 4240. Introduction to Computing for Data Analysis
- CS 4400 - Introduction to Database Systems
- CX 4242. Data and Visual Analytics

Course Requirements & Grading

Grading Scheme

- **50% Homework** 5 homework 10% each
- **25% Project**
    - 3% proposal
    - 7% paper draft
    - 5% final presentation
    - 10% final paper
    - Note: Penalty will be reflected due to inactive team project participations
- **20% Final exam**
- **5% Participation**
    - Piazza activities(contributions to Piazza: questions answers and threads created)

Grading Scale

Your final grade will be assigned as a letter grade according to the following scale:

A	90-100%
B	80-89%
C	70-79%
D	60-69%
F	0-59%

Prof. Sun will determine if we can curve at the end of semester. Please see http://registrar.gatech.edu/info/grading-system for more information about the grading system at Georgia Tech.]

Course Materials Other Classroom Management Tools

Course website: http://sunlab.org/teaching/cse6250/fall2020/

Piazza (O01/OAN) piazza.com/gatech/fall2020/cse6250bdh

Gatech Canvas: used for Assignment/Project submission

Course Schedule

Please email Ming (mliu302@gatech.edu) if you have further question on the schedule.

Week #	Dates	Video lessons	Lab	Deliverable Due
1	Aug 17-21	[1. Intro to Big Data Analytics], [2. Course Overview]		
2	Aug 24-28	[3. Predictive Modeling]	[Hadoop & HDFS Basics]	HW1 Due (Oct 30)
3	Aug 31- Sep 4	[4.MapReduce]& [HBase]	[Hadoop Pig & Hive]	
4	Sep 7-11	[5.Classification evaluation metrics], [6.Classification ensemble methods]		HW2 Due (Sep 13)
5	Sep 14-18	[7. Phenotyping], [8. Clustering]	[Scala Basic], [Spark Basic], [Spark SQL]	
6	Sep 21-25	[9. Spark]	[Spark Application] & [Spark MLlib]	HW3 Due & Project Group Formation & Project Requirements Release (proposal/ draft/final) (Sep 27)
7	Sep 28- Oct 2	[10. Medical ontology]	[NLP Lab]	
8	Oct 5-9	[11. Graph analysis]	[Spark GraphX]	Project Proposal Due (Oct 11)
9	Oct 12-16	[12. Dimensionality Reduction], [13. Patient similairty], [14. CNN]	[Deep Learning Lab]	HW4 Due (Oct 18)
10	Oct 19-23	[15. DNN], [16. RNN]		
11	Oct 26-30	Project Discussion		HW5 Due (Nov 1)
12	Nov 2-6	Project Discussion		
13	Nov 9-13	Project Discussion		Project Draft Due (Nov 8)
14	Nov 16-20	Project Discussion		
15	Nov 23-27	Project Discussion		Final Exam (Dec 1)
16	Noc 30-Dec 4	Project Submission		Final Project Due (code + presentation + final paper) (Dec 6)

Academic Integrity

Georgia Tech aims to cultivate a community based on trust, academic integrity, and honor. Students are expected to act according to the highest ethical standards. For information on Georgia Tech's Academic Honor Code, please visit http://www.catalog.gatech.edu/policies/honor-code/ or http://www.catalog.gatech.edu/rules/18/.

Any student suspected of cheating or plagiarizing on a quiz, exam, or assignment will be reported to the Office of Student Integrity, who will investigate the incident and identify the appropriate penalty for violations.

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.

ASSIGNMENT TURN-IN

Assignments will be released and turned in via the **Canvas** Platform during the week in which they occur.

ATTENDANCE AND/OR PARTICIPATION

Participation is important for this course and learning in general. We use posting on piazza (piazza discussion links are listed on the home page) are as the proxy to measure the participation level.

COLLABORATION & GROUP WORK

Homework assignments are strictly individual efforts, while final projects can be done by small groups (3 people or less) or individuals. You can discuss high level concepts regarding to lectures or homework on the piazza but you shouldn't share your own (or others') solution and code with other students (either on piazza or through other means).

EXTENSIONS, LATE ASSIGNMENTS, & RE-SCHEDULED/MISSED EXAMS

Each student is allowed 2 days of late submission in total, to be used for homework only. You can use these 2 days for any assignment (not project). Once you have used up your late days, late assignments will be penalized at a rate of 10% per day. Assignments more than 5 days late will not be accepted.

STUDENT USE OF MOBILE DEVICES IN THE CLASSROOM

Not applicable

STUDENT-FACULTY EXPECTATIONS

At Georgia Tech we believe that it is important to continually strive for an atmosphere of mutual respect, acknowledgement, and responsibility between faculty members and the student body. See http://www.catalog.gatech.edu/rules/22/ for an articulation of some basic expectations – that you can have of me, and that I have of you. In the end, simple respect for knowledge, hard work, and cordial interactions will help build the environment we seek. Therefore, I encourage you to remain committed to the ideals of Georgia Tech, while in this class.