Instructor Name and Contact Information:

Aidong Zhang, PhD

aidong@virginia.edu
Office Hours: Forum questions, and appointment only as needed

Subject Area and Catalog Number: SYS 6018

Year and Term: Spring 2020

Class Title: Data Mining
Level (Graduate or Undergraduate): Graduate

Credit Type: Graded

Class Description:
Data mining describes approaches to turning data into information. Rather than the more typical deductive strategy of building models using known principles, data mining uses inductive approaches to discover the appropriate models. These models describe a relationship between a system's response and a set of factors or predictor variables. Data mining in this context provides a formal basis for machine learning and knowledge discovery. This course investigates the construction of empirical models from data mining for systems with both discrete and continuous valued responses. It covers both regression and classification and explores both practical and theoretical aspects of data mining

Required Text: James, Witten, Hastie, and Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer, 2013. A free downloadable pdf version is available on the website. (https://www-bcf.usc.edu/~gareth/ISL/)

Learning Outcomes:
Upon successful completion of this course, you will be able to:
1. Build classification and regression models for a given data set using R statistical software.

2. Explain the statistical theory used in data mining that affects how each type of model makes predictions.

3. For a given data set and model, determine the optimal algorithmic parameters to customize the results of the model based on practical goals.

4. Evaluate the performance of a model in terms of various factors such as accuracy, computational cost, interpretability, and practical requirements.

5. Determine the most appropriate algorithm for a given data set based on the needs of the user.

6. Use visualization techniques to help users understand and interpret the data mining results.

Assessment Components:

Assignment	Percent of Grade	Due
Guided Question Set Participation	10	Modules 4, 7, and 11
Homework Question Sets	10	Modules 1-11
Exam 1	15	Module 3
Project 1	15	Module 7
Exam 2	25	Module 9

Delivery Mode Expectations:

This course is delivered in 12 learning modules through Collab. Each course module consists of an online lesson and a live lesson.

Meetings (video and teleconference) for live lessons will be held on Tuesdays from 8:15 to 9:15pm Eastern Time via Zoom. Students are expected to actively participate in the live class discussions by asking questions and contributing to the discussion.

Students are expected to come to class prepared to participate in each week's live activity. Learners are expected to complete the asynchronous learning activities prior to the synchronous session.

The synchronous session will be used for a combination of walkthroughs for using R, small-group discussion of one of the assigned textbook exercises, and/or a Q&A period on the remaining assigned textbook exercises.

After live meetings, students are expected to complete the remaining textbook exercises. Students may work together on these, and are encouraged to ask and answer questions on the discussion board, but each student's answers must be in their own words.

Required Technical Resources and Technical Components:

The statistical software R will be used in this course and is available for free download.

Class Instruction and Activities

The primary purpose of this course is to enable students to build and optimize classification and regression models and use these models to gain information from data. This course will enable students to use data mining methods to solve prediction problems in practical research, giving them hands-on experience with real data sets.

While the focus is on intuitive understanding and practical implementation in the R software, students will also learn the underlying theory behind many of the methods. By the end of the course, students will be able to select regression and classification methods for a given prediction or inference problem, will be able to implement, test, and optimize the methods, and explain the underlying theory.

Evaluation Standards and Assessments

Quizzes (40%)

The module quizzes have a combination of multiple-choice and true/false questions, and must be completed and submitted before the live session for the module.

Textbook Exercises (20%)

The textbook exercises will generally be assigned weekly. The textbook exercises consist of a mixture of conceptual and applied problems, which give students the opportunity to apply statistical learning methods to real-world problems. Students are expected to attempt the textbook exercises before the live session, but these exercises are due after the live session. This gives students the opportunity to ask questions about the textbook exercises during the Q&A period of the live session. The instructor will choose one of the textbook exercises for students to discuss in small groups during the live session.

Students are encouraged to collaborate on the textbook exercises, and in this class you are encouraged to share R-code for assignments in the discussion forum, but each student must prepare their solutions individually. Working together is a great way to learn the material but copying others' work and submitting it as your own is in direct violation of the Honor Code and will be treated as such.

Disaster Relief Project Part 1 and Part 2 (30%)

In this project, students will use classification methods covered in this course to solve a real historical data-mining problem, locating displaced persons living in makeshift shelters following the destruction of the earthquake in Haiti in 2010.

Students will use data from the actual collect over Haiti. The goal is to test each of the algorithms studied in this course on data from the Haiti imagery and determine which works most accurately and in a timely way to try to locate as many of the displaced persons in imagery so they can be provided food and water in time.

Students will document the accuracy and runtime for classification for each algorithm in a Powerpoint file. In Module 6, students will submit the Powerpoint file with some of the algorithm results filled out. In Module 12, at the end of the course, students will submit the completed file, which will include the conclusions from their research and recommendations for what algorithms should be used.

Engagement (Live Session Attendance and Participation (10%)

Your engagement grade is based on your attendance at the live sessions. Students are expected to come to the live lesson on time and prepared to fully participate in the discussion.

Course Topics

Module	Name	Assignments
1	Introduction to Data Mining	Module 1 textbook exercises
2	Linear Regression in Data Mining	Module 2 textbook exercises
3	Classification with LR, LDA and QDA	Module 3 textbook exercises
4	Resampling and Validation	Module 4 textbook exercises
5	Linear Model Selection and Regularization	Module 5 textbook exercises
6	Nonlinear Regression	Module 6 textbook exercises Disaster Relief Project Part 1
7	Tree-Based Models--CART	Module 7 textbook exercises
8	Tree-Based Models--Ensemble Methods	Module 8 textbook exercises
9	Support Vector Machines	Module 9 textbook exercises
10	Unsupervised Learning and Visualization	Module 10 textbook exercises
11	Implementation and Feature Engineering	
12	The Cutting Edge	Disaster Relief Project Part 2

Live Session Schedule:

Live sessions will be held on Tuesday evenings from 8:15-9:15pm EDT. Below is a list of specific dates.

Date	Time
1/14/20	No Live Session

1/21/20	8:15-9:15pm EDT
1/28/20	8:15-9:15pm EDT
2/4/20	8:15-9:15pm EDT
2/11/20	8:15-9:15pm EDT
2/18/20	8:15-9:15pm EDT
2/25/20	8:15-9:15pm EDT
3/3/20	8:15-9:15pm EDT
3/10/20	Spring Break - No Live Session
3/17/20	8:15-9:15pm EDT
3/24/20	8:15-9:15pm EDT
3/31/20	8:15-9:15pm EDT
4/7/20	8:15-9:15pm EDT
4/14/20	No Live Session
4/21/20	8:15-9:15pm EDT
4/28/20	Exams

Communication & Student Response Time:

Discussion forums are set up so students are encouraged to pose questions on the discussion forums. Classmates are encouraged to help each other.

Emails can be sent to the instructor. Be sure to have our class name and number labeled clearly in the subject heading and allow for 1-2 business days for a response.

Electronic Submission of Assignments

All assignments must be submitted electronically through Collab by the specified due dates and times. It is crucial to complete all assigned work - failure to do so will likely result in failing the class.

DSI Grading Policies

The standing of a graduate student in each course is indicated by one of the following grades: A+, A, A-; B+, B, B-; C+, C, C-; D+, D, D-; F. B- is the lowest satisfactory grade for graduate credit.

Attendance

Students are expected to attend all class sessions. Instructors establish attendance and participation requirements for each of their courses. Class requirements, regardless of delivery mode, are not waived due to a student's absence from class. Instructors will require students to make up any missed coursework and may deny credit to any student whose absences are excessive. Instructors must keep an attendance record for each student enrolled in the course to document attendance and participation in the class.

University Email Policies

Students are expected to check their official UVa email addresses on a frequent and consistent basis to remain informed of University communications, as certain communications may be time sensitive. Students who fail to check their email on a regular basis are responsible for any resulting consequences.

Mid-Term and End-of-Class Evaluations

Students may be expected to participate in an online mid-term evaluation. Students are expected to complete the online end-of-class evaluation. As the semester comes to a close, students will receive an email with instructions for completing this. Student feedback will be very valuable to the school, the instructor, and future students. We ask that all students please complete these evaluations in a timely manner. Please be assured that the information you submit online will be anonymous and kept confidential.

Academic Integrity

The Data Science Institute relies upon and cherishes its community of trust. We firmly endorse, uphold, and embrace the University's Honor principle that students will not lie, cheat, or steal, nor shall they tolerate those who do. We recognize that even one honor infraction can destroy an exemplary reputation that has taken years to build. Acting in a manner consistent with the principles of honor will benefit every member of the community both while enrolled in the Data Science Institute and in the future. Students are expected to be familiar with the university honor code, including the section on academic fraud (http://www.student.virginia.edu/~honor/proc/fraud.htm).

Each assignment will describe allowed collaborations, and deviations from these will be considered Honor violations. If you have questions on what is allowable, ask! Unless otherwise noted, exams and individual assignments will be considered pledged that you have neither given nor received help. (Among other things, this means that you are not allowed to describe problems on an exam to a student who has not taken it yet. You are not allowed to show exam papers to another student or view another student's exam papers while working on an exam.) Sending, receiving or otherwise copying electronic files that are part of course assignments are not allowed collaborations (except for those explicitly allowed in assignment instructions). Assignments or exams where honor infractions or prohibited collaborations occur will receive a zero grade for that entire assignment or exam. Such infractions will also be submitted to the Honor Committee if that is appropriate. Students who have had prohibited collaborations may not be allowed to work with partners on remaining homework assignments.

Special Needs

It is my goal to create a learning experience that is as accessible as possible. If you anticipate any issues related to the format, materials, or requirements of this course, please meet with me outside of class so we can explore potential options. Students with disabilities may also wish to work with the Student Disability Access Center to discuss a range of options to removing barriers in this course, including official accommodations. Please visit their website for information on this process and to apply for services online: https://www.studenthealth.virginia.edu/sdac.

If you have already been approved for accommodations through SDAC, please send me your accommodation letter and meet with me so we can develop an implementation plan together.

Technical Specifications: Computer Hardware

- Operating system: Microsoft Windows 8.1 (64-bit) or Mac OS X 10.10
- Minimum hard drive free space: 100GB, SSD recommended

- Minimum processor speed: Intel 4th Gen Core i5 or faster
- Minimum RAM: 8GB

Technical Support Contacts

Login/Password: https://in.virginia.edu/helpdesk

UVaCollab: collab-support@virginia.edu