SYS 4582/6018: Data Mining

Spring 2019

Schedule (./index.html)	Syllabus (./syllabus.html)	Homework (./homework.html)

Course Info

Course Info	
Class Time:	M,W 9:30am-10:45am
Class Location:	Mechanical Engr Bldg 339
Instructor:	Dr. Michael D. Porter
Office:	102F Olsson Hall
Email:	mdp2u {at} virginia.edu
Office Hours:	Wed 2:30-3:45p (and by appt.)

TA:	Tim Eddy
Office:	TBD
Email:	tle7pa {at} virginia.edu
Office Hours:	Fri 12:00-2pm (and by appt.)

Course Prerequisites:

Students taking this course should have prior knowledge in linear regression analysis (e.g., SYS 4021, SYS 6021, STAT 5120), statistical inference (e.g., APMA 3120), and linear algebra (e.g., APMA 3080). Students should also have a basic working knowledge in a scientific programming language (e.g., R, Python, Matlab). All course examples will be in R.

Course Description:

Data mining is a process that transforms raw data into generalizable knowledge concerning relationships among variables. This course describes and investigates a variety of data mining methods. We consider supervised and unsupervised methods and show their applicability to a range of problems.

Student Learning Objectives:

Students will learn how and when to use common data mining methods, understand their comparative strengths and weaknesses, and how to critically evaluate their performance. Students completing this course should be able to: (i) construct and apply novel data mining methods for predictive modeling, (ii) use unsupervised learning methods to find structure in data, (iii) incorporate appropriate techniques to detect patterns and anomalies in complex data, and (iv) properly select, tune, and assess models.

Required Textbooks:

1. **An Introduction to Statistical Learning** by James, Witten, Hastie and Tibshirani.
    - An electronic version of this book is freely available at http://www-bcf.usc.edu/~gareth/ISL/ (http://www-bcf.usc.edu/~gareth/ISL/). This text has labs that show the R code for many of the methods we will cover. This book is also good for a less technical description of common statistical learning methods.
2. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd Edition)** by Hastie, Tibshirani, and Friedman.
    - An electronic version of this book is freely available at http://www-stat.stanford.edu/~tibs/ElemStatLearn/ (http://www-stat.stanford.edu/~tibs/ElemStatLearn/). We will only cover some parts of this text.
3. **Mining of Massive Datasets** by Jure Leskovec, Anand Rajaraman, Jeff Ullman
    - An electronic version of this book is freely available at http://www.mmds.org/ (http://www.mmds.org/). We will only cover some parts of this text.
4. **Introduction to Data Mining (Second Edition)** by Tan, Steinbach, Karpatne, and Kumar.
    - An electronic version of select sections of this book is freely available at https://www-users.cs.umn.edu/~kumar001/dmbook/index.php (https://www-users.cs.umn.edu/~kumar001/dmbook/index.php). We will only use the free sections.

Other Course Materials:

- This course requires the use of the following statistical and typesetting software:
    - R (http://cran.us.r-project.org (http://cran.us.r-project.org)) is a free command-line based statistical language
    - RStudio is a free IDE for R (http://www.rstudio.com/ide (http://www.rstudio.com/ide))
    - LaTeX is a free typesetting system for producing technical documents (e.g., journal articles and presentations).
- Other course material and reading assignment will come from instructor notes and recent journal articles.
- The textbook **Applied Predictive Modeling** by Kuhn and Johnson [webpage] (https://link.springer.com/book/10.1007%2F978-1-4614-6849-3) contains some useful information. Unfortunately, we do not have free access at UVA.

Course Assessment:

- The course grade will be based on ten homework assignments (40%), two midterm exams (each worth 15%), a final project (25%), and class participation/labs/attendance (5%).
- A+: >98%, A: 92-97%, A-: 90-91%, B+: 88-89%, B: 82-87%, B-: 80-81%, etc.
- There is no grade "curving" in this course. However, the lowest homework grade will be replaced by the highest homework grade.
    - Because of this policy, there will be no make-up homework, exams, or quizzes!
    - Note: There will be no "extra credit" assignments; spend your time on the assigned work.
- The exam dates are posted in the Class Schedule (on the course website). Note these now so there are no conflicts.
- All assignment submissions will be made through Collab. Late submissions (even 1 second late) will not be accepted and will receive a score of zero. The time stamps produced by Collab will be the authoritative reference for all such decisions. If you have special circumstances (e.g., a documented physical condition) that prevent you from adhering to the posted deadlines, please inform me at least 1 week in advance of the deadline so that I can make arrangement to accommodate you.
- Homework solutions will consist of a pdf document and code.
    - My recommendation is to use RMarkdown (http://www.stat.cmu.edu/~cshalizi/rmarkdown/) which will produce the pdf and contain the code.
    - Overleaf is another option for those unfamiliar with latex.
    - All code must be easy to follow (e.g., by good commenting)
- Pre-class assignments (reading and coding) will prepare you for the lecture. Absences or neglect of the pre-class assignments will result in lowering your participation grade.
- The final project allows you to put into practice what you have been learning and demonstrate your mastery of the course material. You must come up with your own project idea. Ideally your project will be related to your current research, job aspirations, or interests.

Course Outline:

- Unsupervised Learning
    - Association Analysis
    - Clustering
    - Non-parametric Density Estimation
    - Network Analytics and PageRank
    - Anomaly Detection
- Supervised Learning
    - Penalized Regression (e.g., Ridge, Lasso, elastic net)
    - Nonparametric Methods (e.g., basis functions, kernel methods, GAM)
    - Classification and Naive Bayes
    - Trees and Random Forest
    - Ensembles and Boosting

- Resampling (bootstrap, cross-validation, monte carlo)
- Variable Selection and Pre-processing

Academic Calendar:

Important dates for the semester can be found on the academic calendar:
http://www.virginia.edu/registrar/calendar.html (http://www.virginia.edu/registrar/calendar.html)

Policy on Academic Misconduct (Honor Code):

I trust every student in this course to fully comply with all provisions of the University's Honor Code and work together to maintain UVA's Community of Trust (https://honor.virginia.edu/overview). By enrolling in this course, you have agreed to abide by and uphold the Honor System of the University of Virginia, as well as the following policies specific to this course.

- All submitted work must be pledged.
- All work must be completed individually unless specific permissions are given on the assignment.
  - Homework and in-class exercises can be discussed with classmates, but the final write-up, code, and solutions must be your own.
- It is not always easy to tell what qualifies as a violation, so do not be afraid to talk to me about it. Such discussions do not imply guilt of any kind.
- All suspected violations will be forwarded to the Honor Committee, and you may, at my discretion, receive an immediate zero on that assignment regardless of any action taken by the Honor Committee.

Please let me know if you have any questions regarding the course Honor policy. If you believe you may have committed an Honor Offense, you may wish to file a Conscientious Retraction by calling the Honor Offices at (434) 924-7602. For your retraction to be considered valid, it must, among other things, be filed with the Honor Committee before you are aware that the act in question has come under suspicion by anyone. More information can be found at http://honor.virginia.edu (http://honor.virginia.edu). Your Honor representatives can be found at: http://honor.virginia.edu/representatives (http://honor.virginia.edu/representatives).

Disability Statement:

The University of Virginia strives to provide accessibility to all students. If you require an accommodation to fully access this course, please contact the Student Disability Access Center (SDAC) at (434) 243-5180 or sdac@virginia.edu (mailto:sdac@virginia.edu). If you are unsure if you require an accommodation, or to learn more about their services, you may contact the SDAC at the number above or by visiting their website at http://studenthealth.virginia.edu/student-disability-access-center/faculty-staff (http://studenthealth.virginia.edu/student-disability-access-center/faculty-staff).

Your Well Being

The University of Virginia and School of Engineering serve as a safe space for students and aims to promote your well-being. If you are feeling overwhelmed, stressed, or isolated, there are many individuals here who are ready and wanting to help. If you wish, you can make an appointment with me to discuss in private. Alternatively, the Student Health Center offers Counseling and Psychological Services (CAPS) https://www.studenthealth.virginia.edu/caps (https://www.studenthealth.virginia.edu/caps). If you prefer to speak anonymously and confidentially over the phone, call Madison House's HELP Line 24/7 at434-295-8255 https://www.madisonhouse.org/overview-helpline/ (https://www.madisonhouse.org/overview-helpline/).

If you or someone you know is struggling with gender, sexual, or domestic violence, there are many community and University of Virginia resources available. The Office of the Dean of Students (https://odos.virginia.edu/), Sexual Assault Resource Agency (SARA) (http://www.saracville.org/), and UVA Women's Center (http://womenscenter.virginia.edu/) are ready and eager to help. Contact the Director of Sexual and Domestic Violence Services at 434-982-2774.

Religious Accommodations

Students who wish to request academic accommodation for a religious observance should submit their request to me by email as far in advance as possible. If you have questions or concerns about your request, you can contact the University's Office for Equal Opportunity and Civil Rights (EOCR) https://eocr.virginia.edu/accommodations-religious-observance (https://eocr.virginia.edu/accommodations-religious-observance). Accommodations do not relieve you of the responsibility for completion of any part of the coursework you miss as the result of a religious observance.