INTRODUCTION TO

DATA MINING

CS 363D, SPRING 2019, 51015

TU/TH 12:30-2:00, GDC 1.304

PROFESSOR

Angie Beasley
angie.beasley@utexas.edu
Office Hours: Mon 11:00-12:00
Wed 10:00-11:00
GDC 4.314

TA & PROCTOR

Pratyush Kar
pkar@cs.utexas.edu
Office Hours: Wed 12:00-2:00
GDC 1.302, Desk 4

Mohammad Asif
mohammadasif@utexas.edu
Office Hours: Thurs 2:00-4:00
GDC 1.302, Desk 5

COURSE DESCRIPTION

During "small data" days, data was difficult and tedious to collect. Only the data that was needed to answer a specific question was collected. But we have now entered the world of "big data," where data is constantly collected on everything and everyone. We have an unprecedented amount of data, and classic statistical approaches often do not work on the type of data we have. This is where machine learning comes in...

In this class, you will learn machine learning algorithms to find patterns in large datasets. We will cover classification, clustering, anomaly detection, and association analysis.You will use Python's scikit-learn packages, and Jupyter Notebooks, two industry-standard tools for data mining.

Data mining has already changed the way in which many important decisions are made. In today's data-driven world, it is increasingly critical to understand how these algorithms come to their conclusions and the correct ways to interpret and apply their results.

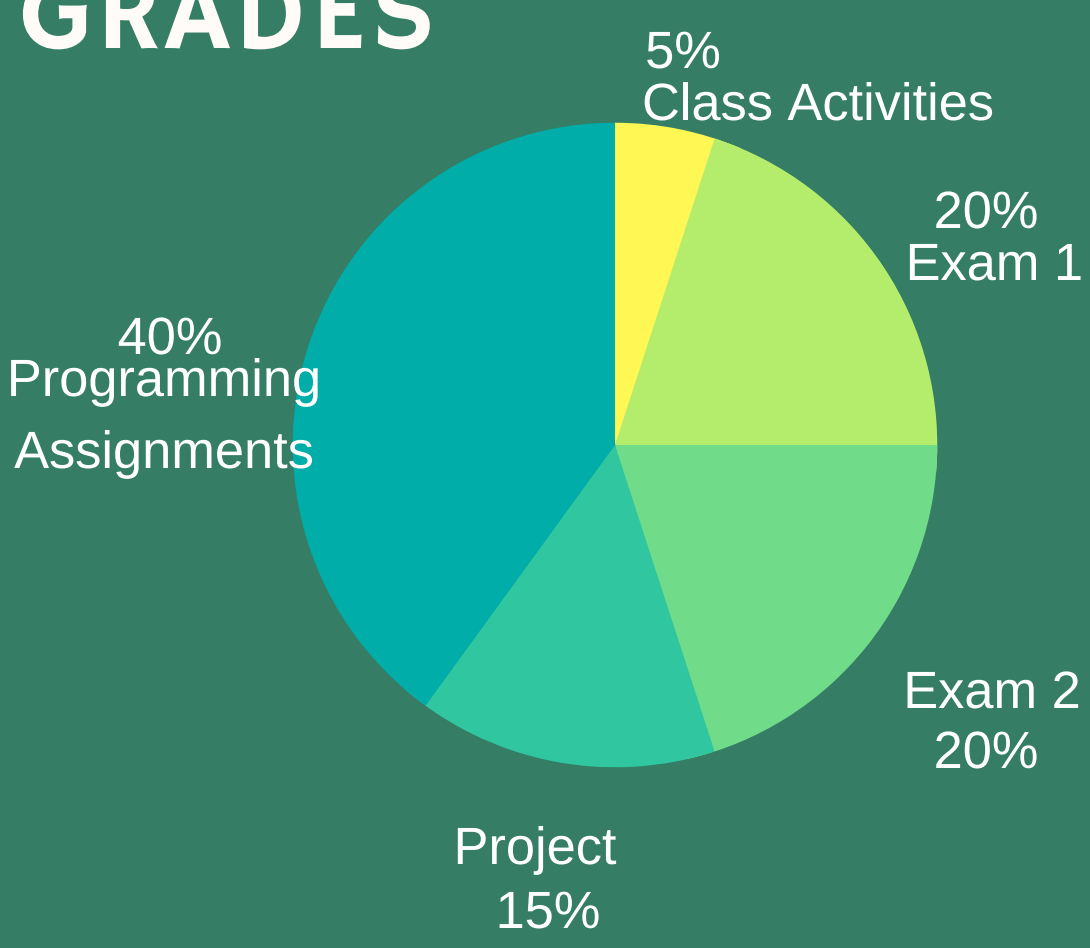
Computer Science 363D and 378 (Topic: Introduction to Data Mining) may not both be counted. Prerequisites: The following coursework with a grade of at least C- : Computer Science 429 (or 310) or 429H (or 310H); Mathematics 362K or Statistics and Data Sciences 321 (or Statistics and Scientific Computation 321); and Mathematics 340L, 341, or Statistics and Data Sciences 329C (or Statistics and Scientific Computation 329C).

TEXTBOOK

Introduction to Data Mining
--> Second Edition <--

by Pang-Ning Tan, Michael Steinbach, Vipin Kumar

GRADES



A	>=94
A-	< 94
B+	< 90
B	< 87
B-	< 84
C+	< 80
C	< 77
C-	< 74
D+	< 70
D	< 67
D-	< 64
F	< 60

All numbers are absolute and will not be rounded up or down at any stage.

PROJECT

This semester, you have an exciting opportunity presented by SparkCognition.

Your project will be to analyze a dataset of your own choosing, using SparkCognition's data analysis tool, Darwin. The results of your analysis will be summarized in both a paper and a short video. You will work in groups of 4 people. This project will make up 15% of your final grade.

In addition to this project allowing you to pursue your own interests (and conduct new research in so doing!), it will also be part of a contest, judged by SparkCognition, in which the first place winners will receive a $10,000 prize! Second place will receive $3000 and third place will receive $2000!

PROJECT SCHEDULE

2/19	Group assignment deadline
3/12	Dataset selection deadline
4/22	Final paper due
5/6-7	Finalists present to SparkCognition

PROGRAMMING ASSIGNMENTS

There will be 6 programming assignments, each equally weighted to total 40% of your final grade.

Programming assignments must be completed using Python 3 and Jupyter Notebooks. Assignments may be worked individually or in pairs. If you work in pairs, you are expected to use the proper pair programming method.

LATE ASSIGNMENTS

You will have 3 late days in 1-day units (that is, 1 minute to 24 hours late = 1 late day) to use throughout the semester. You may divide your late days across the programming assignments in any way you wish. Once you have used all of your late days, late assignments will no longer be accepted.

In the case of pair programming, each member of the pair must have enough late days to cover the late submission. So if the pair submits their code 2 days late, each member must have two late days remaining to use and each member will lose two late days.

To use late days, you only need to submit the assignment. You do not need to email the instructor or the TA, you do not need to indicate that you are using late days. Your late days will be deducted according to when your assignment is submitted. If you submit a late assignment without enough late days to support it, you will receive a zero for that assignment.

Contact me if there are extenuating circumstances.

IN-CLASS ACTIVITIES

Throughout the semester, there will be in-class activities. They will vary in nature and will be random in occurrence. Some will be graded for correctness, and some will only be graded for completion. You may drop your 2 lowest of these and the remaining will make up 5% of your final grade.

COURSE SCHEDULE

Subject to change at instructor's discretion.

1/22	Introduction
1/24	Data Prep & Exploration [Ch 1-2]

CLASSIFICATION

1/29	Feature Engineering
1/31	Decision Trees [Ch 3.1 - 3.3]
2/5	Decision Trees (cont.)
2/7	Overfitting & Cross-Validation [Ch 3.4-3.9]
2/12	Nearest Neighbor [Ch 4.3]
2/14	Naive Bayes [Ch 4.4]
2/19	Evaluating Classifiers [Ch 4.11]
2/21	No class! CNS Career Fair
2/26	Ensemble Methods [Ch 4.10]
2/28	SVMs [Ch 4.9]
3/5	Neural Nets [Ch 4.7]
3/7	Neural Nets (cont.)
3/12	SparkCognition / Exam Review
3/14	EXAM 1
3/19	Spring Break
3/21	Spring Break

CLUSTERING

3/26	Clustering & K-means [Ch 7.1-7.2]
3/28	Density-Based Clustering [Ch 7.4]
4/2	Hierarchical Clustering [Ch 7.3]
4/4	Evaluating Clusters [7.5]
4/9	Anomaly Detection [Ch 9]
4/11	SparkCognition

ASSOCIATION ANALYSIS

4/16	Apriori [Ch 6.1-6.5]
4/18	Scalability Issues
4/23	FP Growth [Ch 6.6]
4/25	Compact Itemsts/Skewed Distributions[6.4,6.9]
4/30	Evaluating Association Patterns [6.7-6.8]
5/2	Sequential Patterns [Ch 7.4]
5/7	Infrequent Patterns [Ch 7.6]
5/9	Exam Review
5/21	EXAM 2, 9:00pm-12:00pm *

* Additional Exam 2 dates/times TBD

ACADEMIC INTEGRITY

Each student in the course is expected to abide by the University of Texas Honor Code:
"As a student of The University of Texas at Austin, I shall abide by the core values of the University and uphold academic integrity."

This means that work you produce on assignments and exams is all your own work, unless it is assigned as group work. I will make it clear for each exam or assignment whether collaboration is allowed or not.

You are responsible for understanding UT's Academic Honesty Policy which can be found here: http://deanofstudents.utexas.edu/sjs/acint_student.php

If you submit code that is not your own, you will be guilty of plagiarism and subject to academic disciplinary action, including failure of the course.

ANONYMOUS FEEDBACK

Anonymous feedback may be provided to the instructor at any time via
Canvas -> Quizzes -> Anonymous Feedback

UNIVERSITY RESOURCES

The Counseling and Mental Health Center (CMHC) provides counseling, psychiatric, consultation, and prevention services: http://cmhc.utexas.edu/

Student Emergency Services
http://deanofstudents.utexas.edu/emergency/

Need help with technology? http://www.utexas.edu/its/

Canvas help is available 24/7 at
https://utexas.instructure.com/courses/633028/pages/student-tutorials

If you have concerns about the safety or behavior of fellow students, TAs or Professors, call BCAL (the Behavior Concerns Advice Line): 512-232-5050. Your call can be anonymous. If something doesn't feel right – it probably isn't. Trust your instincts and share your concerns.

UNIVERSITY POLICIES

RELIGIOUS HOLY DAYS

By UT Austin policy, you must notify me of your pending absence at least fourteen days prior to the date of observance of a religious holy day. If you must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, I will give you an opportunity to complete the missed work within a reasonable time after the absence.

Q DROP POLICY

If you want to drop a class after the 12th class day, you'll need to execute a Q drop before the Q-drop deadline, which typically occurs near the middle of the semester. Under Texas law, you are only allowed six Q drops while you are in college at any public Texas institution.
For more information, see:
http://www.utexas.edu/ugs/csacc/academic/adddrop/qdrop

STUDENT ACCOMODATIONS

Students with a documented disability may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 512-471-6259 (voice) or 1-866-329-3986 (video phone).
http://ddce.utexas.edu/disability/about/

Please request a meeting with me as soon as possible to discuss any accommodations you may need.
Please notify me as soon as possible if the material being presented in class is not accessible to you.
Please notify me as soon as possible if any of the physical space is difficult for you.

EVACUATION INFORMATION

The Office of Campus Safety and Security:
http://www.utexas.edu/safety/
Occupants of buildings on The University of Texas at Austin campus are required to evacuate buildings when an alarm or alert is activated. Exit and assemble outside, unless told otherwise by an official representative. Do not re-enter a building unless given instructions by the Austin Fire Department, The University of Texas at Austin Police Department, or Fire Prevention Services office.
• Familiarize yourself with all exit doors of each classroom and building you may occupy. Remember that the nearest exit door may not be the one you used for entry.
• Students requiring assistance in evacuation shall inform their instructor in writing during the first week of class.
• Information regarding emergency evacuation routes and emergency procedures can be found at:
www.utexas.edu/emergency