I320D: Topics in Human Centered Data Science: Applied Machine Learning with Python

Monday/Wednesday: 3:30PM-5:00PM, PAR 103

Instructor: Dr. Abhijit Mishra (he/his)
Email: abhijitmishra@utexas.edu

TA: TBA
Email: TBA

Office Hours
Abhijit Mishra: Thursday 11:30AM-1:00PM via zoom (https://utexas.zoom.us/j/8979599959)
TA: TBA

Canvas: https://utexas.instructure.com/courses/1352220

Communication and Asking for Help
Please ask all questions that are applicable to the entire class on Canvas, so that others may benefit from the discussion. Only use email for questions unique to individual circumstances; in those cases, please address all questions to both abhijitmishra@utexas.edu and <TA>

Course Description

This course will cover relevant fundamental concepts in machine learning (ML) and how they are used to solve real-world problems. Students will learn the theory behind a variety of machine learning tools and practice applying the tools to real-world data such as *numerical data*, *textual data (natural language processing)*, and *visual data (computer vision)*. Each class is divided into two segments: **(a) Theory and Methods**, a concise description of an ML concept, and **(b) Lab Tutorial**, a hands-on session on applying the theory just discussed to a real-world task on publicly available data. We will use Python for programming.

Intention and Objectives

The course can be a starting step to building a machine learning / data science career-profile. Our intention is to:

- Provide a bird's eye view of the field of ML and enable students to make informed decisions while choosing from different career options, i.e., working on an ML-based cutting-edge product, joining the industry in an ML-centric role, or pursuing a master's or a doctoral study specializing in ML or data science.
- Provide motivation and preparedness for advanced courses in AI / ML offered in iSchool (or other departments), such as Theoretical and Foundations of Machine Learning, Deep Learning, Natural Language Processing, and Computer Vision

By the end of the course, the goals for the students are to:

- Develop a sense of where to apply machine learning and where not to, and which ML algorithm to use
- Understand the process of garnering and preprocessing a variety of "big" real-world data, to be used to train ML systems

- Characterize the process to train machine learning algorithms and evaluate their performance
- Develop programming skills to code in Python and use modern ML and scientific computing libraries like SciPy and scikit-learn
- Propose a novel product/research-focused idea (this will be an iterative process), design and execute experiments, and present the findings and demos to a suitable audience (in this case, the class).

Prerequisites

[1] **Programming in Python** (i.e., Programming for Informatics - I304) : The proposed ML is applied in nature and there is a lab session in each class where students will code in Python. While the instructor will provide handouts for python basics, there is no way a student without any knowledge in programming will be able to pick up and fully participate in classes. Hence, programming for informatics (or equivalent programming course) is a necessary prerequisite.

[2] **I310D - Introduction to Human Centered Data Science:** Students are expected to have been exposed to harnessing and processing data, probability and statistics and linear algebra. I310D provides a suitable background and hence a preferred prerequisite. Alternatively, students may opt for one or more of the following courses (or courses that are similar in nature):

SDS 321 - Introduction to Probability and Statistics
SDS 323 - Statistical Learning and Inference
CS-329E: Elements of Data Analytics

Instruction Modality

Class meetings will be in person, with some exceptions and dependent on the state of the COVID-19 pandemic. If we are unable to meet in person, classes will be held virtually via Zoom. Classes will be a mixture of lecture and hands-on sessions.

Accommodations for Students with Disabilities

The university is committed to creating an accessible and inclusive learning environment consistent with university policy and federal and state law. Please let me know if you experience any barriers to learning so I can work with you to ensure you have equal opportunity to participate fully in this course. If you are a student with a disability, or think you may have a disability, and need accommodations please contact Services for Students with Disabilities (SSD). Please refer to SSD's website for contact and more information: http://diversity.utexas.edu/disability/. If you are already registered with SSD, please deliver your Accommodation Letter to me as early as possible in the semester so we can discuss your approved accommodations and needs in this course.

Required Materials

There is no required textbook for this course; all assigned readings will be available online at no cost. Reading materials/resources will be added to canvas for each module.

Required Devices

This course requires students to bring their laptop computers, although it is device agnostic (PC and Mac preferable but do let me know beforehand if you are working with any customized hardware+ OS , something like Raspberry PI board + Linux) . Students will be required to install Python, SQL and Jupyter notebooks. For resource heavy exercises, we may use Google Colaboratory.

Class Participation

Students are expected to attend every class and actively engage themselves in class discussions and complete the lab tutorial at the end of every session. They may polish and submit the tutorial by 11:59PM on the class day.

Assignments and Course Project

The class format is split between reading and coding assignments for the first half of the semester followed by a project the second half of the semester.

1. Assignments

SIX assignments will be given in the first half of the semester. Each assignment will have: (a) a theoretical question based on weekly assigned readings and (b) a coding exercise similar to the lab tutorials. Assignments are intended to bring conceptual clarity, stimulate algorithmic thinking and emulate practical ML implementation scenarios. Moreover, students will be encouraged to reuse the code from the coding assignments in their course projects.

2. Course Project

The goal of the course project is to promote effective planning, execution, and communication of an ML-centric product/research idea. Assignments related to the course project will be related to (a) Project Planning (b) Gathering Resources (c) Experiment Design and Execution, and (d) Preparing presentation, report, and demo. Students will be required to present before the class.

Late Work and Missed Work

In an effort to accommodate any unexpected personal events, I have enacted a grace policy of two days for this course. You do not have to utilize this policy, but if you find yourself struggling with unexpected personal events, I encourage you to email me as soon as possible (in advance of the due date) to notify me that you are using our grace policy. You may either have a two-day grace period for one assignment, or you may have 2 one-day extensions for two different assignments. The only absences that will be considered excused are for religious holidays or extenuating circumstances due to an emergency. If you plan to miss class due to observance of a religious holiday, please let us know at least two weeks in advance. You will not be penalized for this absence, although you will still be responsible for any work you will miss on that day if applicable. In the event of an unexcused absence, we do not guarantee the opportunity to make up missed in-class work, but one may be granted. Check with us for details or arrangements.

Grading Policies

Course grades will be made up of the following components. Final letter grades will be awarded according to the grade cutoffs below, including pluses and minuses.

Grade Component	Percentage
Attendance, Participation in class and Lab Completion	20%
Eight Assignments (due Friday each week)	40%
Final Project	40%%

Grade Breaks

Grade	Cutoff
A	94%
A-	90%
B+	87%
B	84%
B-	80%
C+	77%
C	74%
C-	70%
D+	67%
D-	60%
F	< 60%

Course Outline

Week 1. Introduction

Lecture: Introduction and Motivation, Definition of ML Types of learners, Data Formats

Lab: Basics of Python and using Jupyter, Statistics in Python, Loading, processing and visualising different data-formats

Assignment 0: Pre-course survey

Week 2. Feature Extraction and Representation

Lecture: Representing data - Splitting and Extracting Features

Lab: Extracting, representing and visualizing features in weather domain

Assignment 1: Feature Extraction and Visualization

Week 3: Supervised Learning - Regression

Lecture: What is supervised learning, Linear Regression, Gradient Descent

Lab: Training and evaluating linear regression for prediction in weather domain, analysing feature importance

Assignment 2: Linear Regression theory + playing with gradient descent code
Project: Group formation

Week 4: Overfitting and Regularization

Lecture: Overfitting and under-fitting, Regularization strategies, Cross validation

Lab: Exploring regularization techniques and K-fold cross validation

Project: Project pre-proposal

Week 5: Supervised Learning - Classification 1

Theory: Classification, Intro to Decision Trees, Bagging and Boosting, Random Forests

Lab: Training decision tree and random forests on weather data, analyzing feature importance

Assignment 3: Decision Tree theory + Application (Code) on a Kaaggle dataset

Week 6: Supervised Learning - Classification 2

Theory: Logistic Regression, Support Vector Machines, Kernel trick

Lab: Training Logistic Regression and SVMs using weather data

Assignment 3: LR theory + SVM Application (Code) on a Kaaggle dataset with kernel methods

Week 7: Neural Networks

Lecture: Perceptrons and Neural Networks, Back propagation

Lab: Show how a perceptron works in code, Implementing stacked NN layers using PyTorch

Project: Project proposal and planning

Week 8: Image as Data

Lecture: Image as Data, Computer Vision basics, Convolutional Neural Networks

Lab: CNNs using PyTorch, Use Cloud Infrastructure with GPU for training an Object classifier - Face Identification

Assignment 4: Deep learning theory + Image segmentation exercise with python

Week 9: Text as Data

Lecture: Text as Data, NLP basics, Recurrent Neural Networks

Lab: RNNs using PyTorch, Use Cloud Infrastructure (Google Colaboratory) with GPU for training an NLP system - Sentiment Analysis

Assignment 5: RNN theory + News Classification using LSTMs

Week 10: SPRING BREAK

Week 11: Unsupervised Learning

Lecture: Clustering basics, K-means and outlier detection

Lab: K-Means Clustering walkthrough on student data

Assignment 6: K-means theory + Anomaly detection exercise in Python

Week 12: Semi-supervised Machine Learning and Transfer Learning

Lecture: Semi-supervised learning and transfer basics, Auto-encoders, Representation Learning for text and images

Lab: Word and sentence vector extraction, Image Representation using PyTorch

Assignment 7: K-means theory + Anomaly detection exercise in Python

Week 13: Evaluation of ML Models

Lecture: Precision-Recall, Confusion Matrix, RoC Curve, Course Project Foundations

Lab: Precision, Recall, F-score and RoC curves, Ablation Studies

Project: 1-page interim progress on projects

Week 14: ML Operations

Lecture: ML Ops basics, Model and experiment logging, Serving models on cloud and on-devices

Lab: Deploying ML models and designing a web-based demo using Flask

Assignment 8: Model serving using flask (only coding)

Week 15: Privacy and Ethics in ML

Lecture: Privacy and Consent, Differential Privacy, Bias and Fairness

Lab: Project status, blockers and other issues

Week 16: Project Presentations and Demos

Week 17: NO CLASS (Project report submission by May 1)

Academic Integrity

Students who violate University rules on academic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since such dishonesty harms the individual, all students, and the integrity of the University, policies on academic dishonesty will be strictly enforced. For further information, please visit the Student Conduct and Academic Integrity website at http://deanofstudents.utexas.edu/conduct.

Confidentiality of Class Recordings

In the event that class should be recorded, class recordings are reserved only for students in this class for educational purposes and are protected under FERPA. The recordings should not be shared outside the class in any form. Violation of this restriction by a student could lead to Student Misconduct proceedings.

Religious Holy Days

By UT Austin policy, you must notify me of your pending absence as far in advance as possible of 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, you will be given an opportunity to complete the missed work within a reasonable time after the absence.

Names and Pronouns

Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, culture, religion, politics, sexual orientation, gender, gender variance, and nationalities. I will gladly honor your request to address you by your chosen name and by the gender pronouns you use. Class rosters are provided to the instructor with the student's chosen (not legal) name, if you have provided one. If you wish to provide or update a chosen name, that can be done easily at this page, and you can add your pronouns to Canvas.

Basic Needs Security

Any student who faces challenges securing their food or housing and believes this may affect their performance in the course is urged to contact the Dean of Students for support. UT maintains the UT Outpost which is a free on-campus food pantry and career closet.

Mental Health Support

I urge students who are struggling for any reason and who believe that it might impact their performance in the course to reach out to me if they feel comfortable. This will allow me to provide any resources or accommodations that I can. If immediate mental health assistance is needed, call the Counseling and Mental Health Center (CMHC) at 512-471-3515 or you may also contact Bryce Moffett, LCSW (iSchool CARE counselor) at 512-232-2983. Outside CMHC business hours (8a.m.-5p.m., Monday-Friday), contact the CMHC 24/7 Crisis Line at 512-471-2255.

Land Acknowledgement

I would like to acknowledge that we are meeting on the Indigenous lands of Turtle Island, the ancestral name for what now is called North America. Moreover, I would like to acknowledge the Alabama-Coushatta, Caddo, Carrizo/Comecrudo, Coahuiltecan, Comanche, Kickapoo, Lipan Apache, Tonkawa and Ysleta Del Sur Pueblo, and all the American Indian and Indigenous Peoples and communities who have been or have become a part of these lands and territories in Texas.

Title IX Reporting

Title IX is a federal law that protects against sex and gender-based discrimination, sexual harassment, sexual assault, unprofessional or inappropriate conduct of a sexual nature, dating/domestic violence and stalking at federally funded educational institutions. UT Austin is committed to fostering a learning and working environment free from discrimination in all its forms. When unprofessional or inappropriate conduct of a sexual nature occurs in our community, the university can:

1. Intervene to prevent harmful behavior from continuing or escalating.
2. Provide support and remedies to students and employees who have experienced harm or have become involved in a Title IX investigation.
3. Investigate and discipline violations of the university's relevant policies.

Beginning January 1, 2020, Texas Senate Bill 212 requires all employees of Texas universities, including faculty, report any information to the Title IX Office regarding sexual harassment, sexual assault, dating violence and stalking that is disclosed to them. Texas law requires that all employees who witness or receive any information of this type (including, but not limited to, writing assignments, class discussions, or one-on-one conversations) must be reported. **I am a Responsible Employee and must report any Title IX related incidents** that are disclosed in writing, discussion, or one-on-one. Before talking with me, or with any faculty or staff member about a Title IX related incident, be sure to ask whether they are a responsible employee. If you would like to speak with someone who can provide support or remedies without making an official report to the university, please email advocate@austin.utexas.edu. For more information about reporting options and resources, visit http://www.titleix.utexas.edu/ , contact the Title IX Office via email at titleix@austin.utexas.edu, or call 512-471-0419.

Although graduate teaching and research assistants are not subject to Texas Senate Bill 212, they are still mandatory reporters under Federal Title IX laws and are required to report a wide range of behaviors we refer to as unprofessional or inappropriate conduct of a sexual nature, including the types of conduct covered under Texas Senate Bill 212. The Title IX office has developed supportive ways to respond to a survivor and compiled campus resources to support survivors.