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

Semester: Fall 2024
Time / Venue: Monday/Wednesday: 8:00AM-9:30AM, **UTC1.118**

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

TA:Shaunak Pusalkar (he / him)
Email: shaunakp@utexas.edu

Canvas: https://utexas.instructure.com/courses/1395041

Office Hours
Abhijit Mishra: Monday 12:00PM-2:00PM and Wednesday 12:00PM-2:00PM (https://utexas.zoom.us/j/8979599959) or by appointments.

TA Office Hours: TBD

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 shaunakp@utexas.edu.

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. **Slides and lecture notes will be provided one week in advance.**

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 (https://colab.research.google.com/)

Class Participation (Attendance through quiz and tutorials)

Students are expected to attend every class, actively participate in discussions, and complete the lab tutorial at the end of each session. Tutorials can be polished and submitted by 11:59 PM on the same day.

While attendance will not be explicitly recorded, I may occasionally administer quizzes during the theory period. Students present in class will be provided with an access code and 10 minutes to complete the quiz. In-class participants will have multiple opportunities to submit the quiz and earn full points. Those absent “without an excused absence” will receive a ZERO for the quiz. **Given that all attendees have the potential to earn full points, quiz dates will not be announced in advance.**

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 take home assignments, each carrying 10 points assignments will be given. Each assignment is a Python based coding exercise 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.

All take-home assignments will be posted on Mondays (after lectures) and will be due on Wednesdays in the following week (10 days of turnaround time).

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
Quiz in class and Lab Completion (i.e., Graded hands-ons/Labs)	30%
Assignments	40%
Final Project	30%

Grade Breaks

Grade	Cutoff
A	94%
A-	90%
B+	87%
B	84%
B-	80%
C+	77%
C	74%
C-	70%
D+	67%
D	64%
D-	60%
F	< 60%

Course Outline

WEEK 1. Introduction (Aug 26- Aug28)

Lecture 1: Introduction and Motivation, Definition of ML Types of learners, Data Formats
Lecture 2: Feature Engineering and representation basics
Readings (For week 1 and 2):

1. Introduction to ML Lecture notes: https://sebastianraschka.com/pdf/lecture-notes/stat451fs20/01-ml-overview__notes.pdf
2. Feature Engineering Explained https://builtin.com/articles/feature-engineering#
3. Fundamental Techniques of Feature Engineering for Machine Learning: https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114

4. Exploratory Data Analysis: Comparing the Five Most Popular EDA tools: https://towardsdatascience.com/comparing-five-most-popular-eda-tools-dccdef05aa4c
5. **[optional]** Zytek, A., Arnaldo, I., Liu, D., Berti-Equille, L., & Veeramachaneni, K. (2022). The Need for Interpretable Features: Motivation and Taxonomy. arXiv preprint arXiv:2202.11748.

For Lab / Tutorial:

1. Introduction to Python: https://www.geeksforgeeks.org/data-science-tutorial/#pyt
2. [Virtual environment + Jupiter] https://towardsdatascience.com/python-virtual-environments-jupyter-notebook-bb5820d11da8
3. Markdown in Jupyter Notebook: https://www.geeksforgeeks.org/markdown-cell-in-jupyter-notebook/

Assignment (not graded): Pre-course survey

WEEK 2. Feature Engineering and Representation Cont… (Sept 4)

Lecture: Representing data - Splitting, Defining and Extracting Features
Tutorial: Exploratory Data Analysis, Selecting, extracting and visualizing features
Readings:
Refer to week 1

Assignment 1: Feature Engineering

WEEK 3: Supervised Learning - Regression (Sept 9-Sept 11)

Lecture: Supervised learning overview, Linear Regression, Least Square Method, Gradient Descent
Tutorial: Training and evaluating linear regression for prediction with IMDB dataset, Analyzing feature importance using R1 scores and coefficients
Readings:

[1] Linear Regression Theory: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/
[2] Gradient Descent: https://www.analyticsvidhya.com/blog/2021/03/understanding-gradient-descent-algorithm/
[3] Implementation using scikit learn: https://realpython.com/linear-regression-in-python/

Assignment 2: Linear Regression on Boston Housing Dataset

WEEK 4: Overfitting and Regularization (Sept 16- Sept 18)

Lecture: Overfitting and under-fitting, Regularization strategies, Cross validation
Tutorial: Exploring regularization techniques and K-fold cross validation

Readings:
[1] What is overfitting? https://www.ibm.com/topics/overfitting

[2] Intuitions on L1 and L2 Regularization: https://towardsdatascience.com/intuitions-on-l1-and-l2-regularisation-235f2db4c261
[3] Ryan P. Adams. Overfitting and Regularization https://www.cs.princeton.edu/courses/archive/fall18/cos324/files/regularization.pdf

WEEK 5: Supervised Learning - Classification 1 (Sept 23-Sept 25)

Lecture: Logistic Regression, Support Vector Machines, Kernel trick
Tutorial: Training Logistic Regression and SVMs using diabetes dataset
Readings:
[1] The Ultimate Guide to Logistic Regression for Machine Learning: https://www.keboola.com/blog/logistic-regression-machine-learning

[2] The last blog you would need on SVM: https://medium.com/analytics-vidhya/the-last-blog-you-will-read-on-svm-8598cf3d0603

Formal but optional reading:
[1] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer. (Chapter 6 and 7).

Assignment 3: Logistic Regression and SVMs

WEEK 6: Supervised Learning - Classification 2 (Sept 30 - Oct 2)

Lecture: Classification, Intro to Decision Trees, Bagging and Boosting, Random Forests
Tutorial: Training decision tree and random forests on diabetes data, analyzing feature importance
Readings:
Decision Trees:
[1] Brandon Rohrer (2018) How decision trees work (https://e2eml.school/how_decision_trees_work.html)

[2] Anshul Saint (2021) An Introduction to Random Forest Algorithm for beginners https://www.analyticsvidhya.com/blog/2021/10/an-introduction-to-random-forest-algorithm-for-beginners/

Ensemble Methods:
[1] Pavan Vadapalli (2022) Bagging vs Boosting in Machine Learning: Difference Between Bagging and Boosting https://www.upgrad.com/blog/bagging-vs-boosting/

Formal but optional reading:
[1] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer. (Chapter 14, Section 14.3 and 14.4).

Assignment 4: Decision Tree and Ensemble Methods

WEEK 7: Neural Networks (Oct 7 - Oct 9)

Lecture: Perceptrons and Neural Networks, Back propagation
Tutorial: Show how a perceptron works in code, Implementing stacked NN layers using TensorFlow (keras)

Readings:
1. Quick introduction to neural networks (https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks/)
2. Neural Network with Python (https://victorzhou.com/blog/intro-to-neural-networks/)

Optional Reading:

3. Introduction - The Perceptron (https://web.mit.edu/course/other/i2course/www/vision_and_learning/perceptron_notes.pdf)
4. A Neural Network in 11 lines of Python Part 1 (http://iamtrask.github.io/2015/07/12/basic-python-network/)
5. Yann LeCun, Yoshua Bengio, Geoffrey Hinton, Deep Learning (https://www.cs.toronto.edu/~hinton/absps/NatureDeepReview.pdf) Nature 521, no. 7553 (2015): 436-444. doi:10.1038/nature14539

Project: Group Formation (not-graded)

WEEK 8: Image as Data (Oct 14- Oct 16)

Lecture: Image as Data, Computer Vision basics, Convolutional Neural Networks
Tutorial: CNNs using TensorFlow (keras), Object classifier - Handwriting Recognition
Readings:
[1] Convolutional Neural Networks for Dummies
https://towardsai.net/p/deep-learning/convolutional-neural-networks-for-dummies

[2] O'Shea, K., & Nash, R. (2015). An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458

Project: Project proposal and planning document

WEEK 9: Text as Data (Oct 21 - Oct 23)

Lecture: Text as Data, NLP basics, Recurrent Neural Networks
Tutorial: Text classification for sentiment analysis using TensorFlow (keras)

Readings:
[1] Vectorization Techniques in NLP https://neptune.ai/blog/vectorization-techniques-in-nlp-guide

[2] What are recurrent neural networks? https://www.ibm.com/topics/recurrent-neural-networks

[3] Recurrent Neural Networks https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks

[4] [Optional] Medsker, L. R., & Jain, L. C. (2001). Recurrent neural networks. Design and Applications, 5, 64-67.

WEEK 10: Unsupervised Learning (Oct 28 - Oct 30)

Lecture: Clustering basics, K-means and outlier detection, Evaluating clusters (distortion and silhouette)

Tutorial: K-Means Clustering walkthrough

Readings:
[1] K-means Clustering - A theoretical foundation https://medium.com/@akshay.sinha/k-means-clustering-8ef58ca0d024

[2] K-means clustering blog: https://neptune.ai/blog/k-means-clustering

[3] What is hierarchical clustering: https://www.displayr.com/what-is-hierarchical-clustering/

Optional but formal readings:

[1] Blömer, J., Lammersen, C., Schmidt, M., & Sohler, C. (2016). Theoretical analysis of the k-means algorithm–a survey. Algorithm Engineering: Selected Results and Surveys, 81-116.

[2] Bishop, C. M., & Nasrabadi, N. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: springer. (Chapter 9 Section 1)

Assignment 5: Unsupervised Learning

WEEK 11: Semi-supervised Machine Learning and Transfer Learning (Nov 4-Nov 6)

Lecture: Semi-supervised learning and transfer basics, Auto-encoders, Representation Learning for text and images
Tutorial: Extracting representations from images and text from pre-trained models
Readings:
[1] Semi-Supervised Learning, Explained with Examples https://www.altexsoft.com/blog/semi-supervised-learning/

[2] Transfer Learning for Machine Learning https://www.seldon.io/transfer-learning

[3] Transfer Learning Guide: A Practical Tutorial With Examples for Images and Text in Keras https://neptune.ai/blog/transfer-learning-guide-examples-for-images-and-text-in-keras

Project: One page interim progress on projects

WEEK 12: Time Series Data Analysis (Nov 11 - Nov 13)

Lecture: Lecture: Time Series Data, Trend, Seasonality, Feature Engineering for Time Series Data, Autoregressive integrated moving average (ARIMA) model
Tutorial: ARIMA based forecasting
Readings:
[1] Introduction to Time Series Data Analysis: https://www.quantstart.com/articles/Beginners-Guide-to-Time-Series-Analysis/
[2] Time series data analysis with Python: https://www.machinelearningplus.com/time-series/time-series-analysis-python/

Assignment 6:Time series forecasting

<u>WEEK 13: Feature Selection, Ranking and Evaluation of ML Models (Nov 18-Nov 20)</u>

Lecture+Lab1: Feature Selection methods, Feature Ranking, Dimensionality Reduction.
Lecture+Lab2: Precision-Recall, Confusion Matrix, RoC Curve

Readings:
[1] Feature Selection Techniques in Machine Learning (Updated 2023) https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/

[2] t-SNE Clearly Explained https://towardsdatascience.com/t-sne-clearly-explained-d84c537f53a

[3] Visualize & Interpret PCA Results via Biplot https://statisticsglobe.com/biplot-pca-explained

[4] Neural Network Embeddings Explained https://towardsdatascience.com/neural-network-embeddings-explained-4d028e6f0526

[5] Recall, Precision, F1, ROC, AUC, and everything https://medium.com/swlh/recall-precision-f1-roc-auc-and-everything-542aedf322b9

WEEK 14: FALL BREAK -NO CLASSES HELD (Nov 25- Nov 30)

WEEK 15: Human Centered ML & Course Wrapup + Project Presentation 1 (Dec 02 - Dec 04)

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

Readings:
1. Human-Centered Machine Learning https://medium.com/google-design/human-centered-machine-learning-a770d10562cd

2. What is Differential Privacy: definition, mechanisms, and examples https://www.statice.ai/post/what-is-differential-privacy-definition-mechanisms-examples

Project: Final Presentation 1 (12/04/2024)

WEEK16: Project Presentation - 2 (Apr 29)

Project: Final Presentation 2 (12/09/2024)

Week 17: NO CLASS (Project report submission by May 6)

Project: Final Report (Due: 12/12/2024)

Mantra for Student Success : Navigating the Machine Learning Course

- Achieve higher attendance, aiming for 100% to maximize exposure and engagement during lectures and practical exercises.
- Submit practicums and assignments promptly, recognizing that minor errors can be overlooked while focusing on continuous improvement.
- Prioritize transparency by appropriately citing tools, resources, and data sources, showcasing your commitment to ethical and accountable work.
- Approach in-class quizzes with a clear understanding and well-organized thoughts, leveraging your conceptual clarity to excel.
- If programming presents challenges, embrace deliberate practice to strengthen your skills and confidently navigate technical aspects.
- Embrace iteration as you prepare presentations, ensuring impactful task demonstrations, comprehensive analyses, and well-structured reports.
- Recognize that success in the DL course is a result of these concerted efforts, culminating in your growth as a proficient and accomplished DL practitioner.

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.

AI Tools Usage Policy:

The utilization of AI-powered tools, including platforms like ChatGPT, Google Gemini, Meta LLaMa, DALL-E, or ANY other small/large language/image/audio/video generative models, to create content such as text, code, images, multimedia, or any related materials intended for assignments, quizzes, or projects that contribute directly to the evaluation of grades within this course is **strictly proscribed**. Exceptions to this rule apply only if the incorporation of such systems aligns with the specified objectives of the assignment or project. Breaching this policy may result in the initiation of proceedings related to student misconduct.

Should there be any suspicion surrounding the content submitted by a student, suggesting the involvement of an AI tool, I retain the authority to request clarification from the student. This clarification may be sought through email communication or arranged verbal discussions in the form of one-on-one meetings. In the event of any inconsistencies between the provided explanations and the submitted solutions, I reserve the right to instigate misconduct proceedings against the concerned student. Upon enrolling in this course, students inherently express their agreement to adhere to this policy as well as any forthcoming policies described below.

Course Material Sharing Policy

Unauthorized sharing or distribution of lecture notes, slides, or examination questions is strictly prohibited without prior permission from the instructors. Failure to adhere to this policy may result in the initiation of legal actions. 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 these restrictions 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.