MIS 382N: Advanced Machine Learning

Course Syllabus

Instuctor: Prof. Joydeep Ghosh, jghosh@utexas.edu

My **office hours** are:
In person (or via Zoom, depending on Covid situation):
Tues 10:30-11:30am, Wed 5-6pm. Location: EER 6.812. My office is in the South tower of the EER building, roughly northeast of 24th and Speedway. Depending on the Covid situation, I may hold some office hours in an outdoor space instead.

By zoom: Tues 6:30-7:30pm. By default, zoom will be open to all drop-ins. First come first served. If you want to discuss something in private, I can put any other student on the call in the "wait room".

My Zoom link is: https://utexas.zoom.us/my/joydeepghosh You need to use your UT ID to access.

TA: Section A (12:30pm), Rishab Khincha, rkhincha@utexas.edu, Office hours: Mon , Fri
TA: Section B (2 pm), Song Wang, song.wang@utexas.edu, Office hours: Wed 10:30AM-11:30AM, Fri 11:00AM-12:00PM. Till Sep 17th, I will hold my office hours on Zoom, my zoom link: https://utexas.zoom.us/j/8115970670.

Course description

In this course we will study a variety of machine learning techniques for predictive analytics, building up from where you were left off in summer. Particular emphasis will be given to (a) a deeper understanding of predictive models, and (b) approaches that are scalable to very large data sets. Many of these capabilities are essential for handing BIG DATA. Connections to relevant business problems shall be made via example studies. We will mostly be using Python (specially Scikit-Learn), and will also get some experience of Tensorflow/PyTorch, mostly through a substantial project. The **central goal** of this course is to convey an understanding of the pros and cons of different predictive modeling techniques, so that you can (i) make an informed decision on what approaches to consider when faced with real-life problems requiring predictive modeling, (ii) apply models properly on real datasets so to make valid conclusions. This goal will be reinforced through both theory and hands-on experience.**Grading information**

(5+10)+15%	Term Project (groups of 3-5): (project proposal outline + presentation) + blog report due end of the semester
30%	5 Asssignments
20%	5 quizzes. Best 4 scores counted
20%	Written Exam in class*

* If we cannot meet safely in class for the written exam, then the exam will be cancelled I'll add two more quizzes (so best 6 of 7 counted), and reassign the 20 points exam credit to quiz, assignments and report (10, 5 and 5) instead.

Dates for quizzes and the exam will be announced on Canvas. There will be no final exam. Quizzes will be held in class and of duration 15 minutes or less. Their objective is to review key concepts introduced in class. I may curve the grade of a quiz if the scores are generally low.

If a quiz is given remotely (via zoom), then it will be open book, open class notes. However you cannot access any other resource (the internet, some other human,…) **as that will be considered cheating.**
You are not allowed to communicate the quiz questions to students in another section as this will be considered as cheating. See note on ACADEMIC DISHONESTY AND POLICIES ON CHEATING below and be sure to understand its implications!

At the end of the course, you will get a score out of 100 based on the percentages stated above. Your final grade will be solely based on this score. The grade is primarily based on the curve, i.e., is relative to how the whole class performs; however entire curve may shift up or down a bit depending on how the class as a whole performs relative to past classes. **Grading is NOT based on absolute thresholds, e.g. 90+ = A etc**.

Textbooks

The material for the lectures is taken from a wide variety of sources, my slides will be available via Canvas. I'll also be placing papers, book chapters, etc, on Canvas as needed.

The **main textbook is**:
(**B**) Christopher M. Bishop (**B**), "Pattern Recognition and Machine Learning", Springer.
See http://research.microsoft.com/en-us/um/people/cmbishop/prml/
A pdf of this book is available under "Resources"

Supplementary references are:

Basic:
1. Max Kuhn and Kjell Johnson (**KJ**), "Applied Predictive Modeling", Springer, 2013
2. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (**JW**), "An Introduction to Statistical Learning with Applications in R", Springer. The authors have kindly provided a free pdf version here, though it may be worth your while to get a hardcopy as well.

Advanced:

1. Trevor Hastie, Robert Tibshirani, and Jerome Friedman (**HTF**), "The Elements of Statistical Learning", Springer. Can get it from Amazon, about $70 but well worth it, or download pdf from http://www-stat.stanford.edu/~tibs/ElemStatLearn/
2. **Math for Machine Learning (2020).**

For help with programming + concepts, "Hands-On Machine Learning with Scikit-Learn and Tensorflow", by A. Geron (O'Reilly, 2nd Ed. 2019)
Deep Learning: Dive into Deep Learning free ebook**:** https://d2l.ai/index.html
"Deep Learning with Python" by Chollet is also pretty good.

The Scikit-Learn website is also reasonably documented.

Course Schedule and Topics

B, HTF, KJ, JW refer to various books referred to above, in case you want to read more. The most important reading sections are bolded below.

Lecture counts are indicative, not binding.

1. Overview and Recap: Types of predictive analytics; Local vs. global models; Role within data mining process, Software demos; Multivariate regression; Tackling Big Data (2 lectures; **KJ: Chapter 1,2; B, Ch 1.2.1 through 1.2.4**; JW Chapters 1,2.1; HTF Ch 1, 2.1-2.6,)
Objective: Recap and revise key concepts from MIS 380, and provide context for this class.

2. Advanced Multivariate regression (partly revision): Basis function expansion; Dealing with large number of features; Ridge, Lasso and Stagewise Approaches; Non-linear methods.
(3 lectures; JW Ch 2.2, **3**, 6.2, **B 3.1, 3.2 (also 1.1, 1.2.6, 1.5.5)**; HTF Ch 2.7, 2.8, 3.1-3.4, 7.1-7.3; KJ: Chapter 4,5,6, 7.1,7.4, 7.5)
Objective: Learn to design, understand and implement predictive models where desired outcome is a numeric quantity.

2A. **Neural Networks for Regression:** (stochastic) gradient descent, non-linear regression, neural networks and MLP; intro to **deep learning**.
(3 lectures, **B 5.1-5.3**; HTF Ch 11.1-11.5)
Objective: Understand SGD and the power of multilayered non-linear models.

3. Data Pre-Processing (Brief): Transformations, Imputations, Outlier detection
(2 lectures, **KJ: Chapter 3**, JW 10.1, 10.2, B 12.1)
Objective: Understand that good data quality is a pre-requisite for effective models, and study some methods for improving data quality.

4. **Modern Data Visualization** (1 lecture) (notes)
Objective: How to visualize (large) data and model results using modern interactive tools.

5. Classification: Scaling decision trees to big data; Bayes decision theory, Logistic regression; Naïve Bayes and Bayesian networks; LDA; Kernel methods and Support Vector Machines (SVMs) for classification and regression; dealing with class imbalance (5 lectures, **B 1.5, 4.1, 4.3.2, 4.3.4**, **JW 4, 8.1**, 9.1-9.3; HTF Ch 4, 7.10, 9.2, 12, 13.3, KJ: Chapter 11,12,13,14.1,14.2,8.1,16)
Objective: Learn to design, understand and implement predictive models where desired outcome is a class label.

6. Ensemble Methods: Model Averaging, Bagging and Random forests, boosting, Gradient boosting; Bag of Little Bootstraps
(2 lectures; **HTF Ch 8.7, 10.1-10.11**; KJ: Chapter 14.3-14.8)
Objective: Understand the benefits of combining multiple predictive models.

7. More on Deep Learning: DL for Vision; Auto-encoders and GANs, Recurrent Nets and LSTMs.
(2-3 lectures; readings from d2.ai ebook).
Objectives: Understand basics of deep learning and its capabilities/limitations.

8. Specialized/Advanced Topic: (coverage depends on time available and interest of class): *Responsible ML (How to make ML solutions more explainable, fair and robust). Ranking and Recommendation*: Applications to Next Generation Recommender systems.

9. Term Project Presentations and Discussion; Conclusions (3-4 lectures)

10. Wildcards: A couple of classes may be used for invited talks by experts.

NOTICES:

ACADEMIC DISHONESTY AND POLICIES ON CHEATING: Faculty at UT are committed to detecting and responding to all instances of scholastic dishonesty and will pursue cases of scholastic dishonesty in accordance with university policy. Scholastic dishonesty, in all its forms, is a blight on our entire academic community. All parties in our community -- faculty, staff, and students -- are responsible for creating an environment that educates outstanding engineers, and this goal entails excellence in technical skills, self-giving citizenry, and ethical integrity. Industry wants engineers who are competent and fully trustworthy, and both qualities must be developed day by day throughout an entire lifetime. Scholastic dishonesty includes, but is not limited to, cheating, plagiarism, collusion, falsifying academic records, or any act designed to give an unfair academic advantage to the student. The fact that you are in this class as an engineering student is testament to your abilities. Penalties for scholastic dishonesty are severe and can include, but are not limited to, a written reprimand, a zero on the assignment/exam, re-taking the exam in question, an F in the course, or expulsion from the University. Don't jeopardize your career by an act of scholastic dishonesty. Details about academic integrity and what constitutes scholastic dishonesty can be found at the website for the UT Dean of Students Office and the General Information Catalog, Section 11-802.

Disabilities statement: "The University of Texas at Austin provides upon request appropriate academic accommodations for qualified students with disabilities. For more information, contact the Office of the Dean of Students at 471-6259, 471-4641 TTY."

- Students with disabilities may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 471-6259, http://www.utexas.edu/diversity/ddce/ssd/
- A notice regarding academic dishonesty. UT Honor Code and example of what constitutes plagiarism : http://registrar.utexas.edu/catalogs/gi09-10/ch01/index.html
- A notice regarding accommodations for religious holidays. "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, you will be given an opportunity to complete the missed work within a reasonable time after the absence.")

Additional important *Student Rights and Responsibilities*