Statistics 414 – From Predictive AI to Generative AI

Instructor: Guang Cheng Office: Boelter Hall 9404
Office Hour: by appointment Email: guangcheng@ucla.edu
Lecture Time and Location: Thursday 6pm-8:50pm @ MS 6229 [break: 6:50-7:00 & 7:50-8:00]
Reader: Mr. Peiyu Yu (yupeiyu98@g.ucla.edu). Discussion time: 5-5:50 each Thursday @ DODD78

Course description:

This course consists of two components: the first half quarter is contributed to “Predictive AI,” while the second half to “Generative AI.” This also reflects the evolution of AI in the past decades. For predictive AI, I will cover very basic machine learning algorithms from supervised learning to un-supervised learning, together with bagging and boosting. For generative AI, I will focus on the so-called “generative data science” explaining why behind generative data. This is an emerging field in the intersection of large language models, machine/deep learning and data science. Students will be introduced to core concepts of generative data science, covering generation, evaluation, and utilization of generative data. *Digital marketing* will serve as a practical arena for you to implement the diverse data synthesis algorithms learned during the course.

Recommended textbook for Predictive AI:
An Introduction to Statistical Learning
Free Download: https://www.statlearning.com/

No textbook for Generative AI

For your references, I offer 3 levels of graduate courses on Generative AI.

Elementary Level
Stats C161/C261
MAS 414 for Master in Applied Statistics and Data Science

Intermediate Level
C163/C263 [Generative Data Science]. Students are introduced to more comprehensive and in-depth contents in generative data science, and required to do a course project as a team.

Advanced Level
Stats 213 [Synthetic Data Generation]. This is only for PhD students (to be offered in the spring, 2025). The most recent advances in generative AI will be introduced in this course. The course projects done in C163/C263 can be carried over to this course and turned into conference/journal submission, e.g., targeted to NeurIPS in June or ICLR in Oct.

Course Policies

The course is in-person and attendance is required. Copying a group project report or homework is absolutely forbidden and constitutes a violation of the Honor Code; therefore each group must produce their own project report to present and to be handed in and graded.

Lecture notes will be shared in Bruinlearn before each class.

Schedule

Schedule	Detailed Contents & Guest Lecturers
W01 (Jan 9) Introduction and ML Pipeline	• Discuss the bias/variance tradeoff • Model validation • Robust model building. • In-sample, out of sample, and Hold out • Types of Cross-validation. • Defining parametric models • ML Quadrants. Supervised/Unsupervised Discrete/continuous • OLS Regression
W02 (Jan 16) Discrete Classification (Logistic Regression)	• Logistic Regression • Odds, log odds, and probability • Showing the Over/underfitting problem using OOS metrics • Confusion matrices • Statsmodels and scikit learn • Metric-weighted accuracy to target. Tuning parameters for profitability over accuracy
W03 (Jan 23) Linear Models, Regularization, and Hyperparameter Tuning	• Regularization • Lasso • Ridge • Elastic Net • Hyperparameter tuning (using regularization parameters as the examples) • Gradient Descent: Analyze convergence • Case studies and best practices

W04 (Jan 30) Bagging and Boosting	• Motivation for ensemble models • Compare prior Linear regressions and Neural networks to show Bagging/Boosting improvements • Adaboost algorithm for creating strong classier • Random Forest • Model weighted ensembles • Super Stackers
W05 (Feb 6) Unsupervised Learning - Clustering & PCA	• Introduce subspace analysis • Robust PCA: A variant of PCA to deal with noisy high-dimensionality data • Case study to show how it handles noisy and corrupted data • Demonstration of how PCA removes overfitting of wide data • K-means clustering algorithm, • Fuzzy C-means Method • Hierarchical clustering algorithm to generate multi-level partitions of the input data
W06 (Feb 13) Midterm Presentation Intro to generative data science	Midterm presentation (Predictive AI, 10 mins per team including Q&A) Guang Cheng
W07 (Feb 20) Intro to digital marketing Intro to popular data synthesis (GenAI) algorithms (Diffusion, VAE, CTGAN, TabDDPM & LLMs)	i) Chi-hua Wang (chihuawang@g.ucla.edu) ii) Xiaofeng Lin (bernardo1998@g.ucla.edu)
W08 (Feb 27) Evaluation of generative/synthetic data: fidelity, utility and privacy	i) Guang Cheng will give an overview talk on the evaluation of generative data; ii) Lan Tao (lantao@g.ucla.edu) will talk about details of fidelity, utility, and privacy.

W09 (March 6) Deep dive into utility, differential privacy, and data-copying	i) Tomas Kwok (thomask1018999@gmail.com) will talk about high machine learning utility for synthetic data ii) Josh Ward (joshuaward@g.ucla.edu) will talk about data copying and membership inference attacks.
W10 (March 13) Membership inference attack	Final presentation (Generative AI, 15 mins per team including Q&A)

Assessment:

Homework (30%) + Midterm presentation (15%) + Final presentation (25%) + Final Report (30%)

Grading:

Grading of presentation and final report are ranking based.

Homework:

Four homeworks will be assigned each week in the first half quarter. They will be due by next Friday at 5pm. Homework assignments will be announced in Bruinlearn and students will upload their homework there as well. Graders will finish grading within one week after due time, and return with solutions in the Bruinlearn.

Group project:

It is *strictly* required to use the digital marketing dataset (https://www.kaggle.com/datasets/xiaojiu1414/digix-global-ai-challenge), and **midterm presentation** is on data analytics using predictive AI algorithms. **Final presentation** is on synthesizing this digital marketing dataset with suitable evaluations. **For the final presentation, the use of LLM is strongly preferred.**

Each team is recommended to have 3-4 members. Role-wise, some members will work on the programming and literature review parts, while some focus more on the development/evaluation/utilization of predictive AI and synthetic data algorithms.

Final Report

The final report is at least 15 pages, accompanied by source code. If there is sufficient novelty in the course project, the project report will be turned into a ML/AI conference submission (with some extra work after the quarter ends). It is due by 5pm on March 21.

Please feel free to consult the instructor for project design including topic/data/algorithm.

The enrolled students can claim AWS cloud credits for computation support.

Key days:

Team formation is due on Jan 24 (Peiyu will send out a team matching link in the week 2)
Midterm presentation is on Feb 13
Final presentation is on March 13
Final report is due by 5pm on March 21.