CSYE 7470
Advanced Analytics and Causal Inference
Course Syllabus

Course Information

Professor: Nik Bear Brown
Email: ni.brown@neu.edu
Office: 505A Dana Hall
Office hours:
Through Zoom by Appointment

Course Prerequisites

An intro to machine learning course and experience with python.

Course Description

The first part of the course covers data, experimental design and metrics. The examples provided will involve game data, but these techniques apply to data used in AI in general. Garbage-In Garbage Out (GIGO) may be the most widely used maxim in machine learning, but how does one assess the quality of data? Metrics are measurable traits that act as indicators for engagement, user movement, retention, churn or anything that the developer wishes to measure. How does one know decide what to log and measure so that precision is maximized and specific conclusions can be drawn regarding a hypothesis?

The first part of the course covers experimental design and the statistical analysis of data. The cleaning, sampling, imputation, and normalization of data is covered. Data reduction techniques and the creation of synthetic data will also be covered. Exercises are given that cover the exploration and visualization of data and visualization of the data generating process.

The second part of the covers experimental design including: sample size and confidence/credibility, A/B testing and ANOVA with applications to the web and games, Bandits, robust experimental design and Taguchi methods, response surfaces and optimal design.

The third part of the course builds focuses on understanding models created for the data. Why has a model made a certain decision? In this part we focus on using model interpretability algorithms. . Techniques covered include individual conditional expectation (ICE), leave-one-covariance (LOCO), local feature importance, partial dependency plots, tree-based feature importance, standardized coefficient importance, accumulated local effects (ALE) plots and Shapley values.

The fourth part extends a student's knowledge of probability to be able the create causal models from observational data. Students are taught the theory and practice of causal inference. A special emphasis is placed on the assumptions that underlie all causal inferences, formulating those assumptions, and the conditional nature of all causal and counterfactual claims. Exercises are given to teach student how build a causal model from observational data so that they can ask causal and counterfactual questions.

The fifth part of the course covers Evidence Knowledge Graphs (EKGs) and graph-based machine learning. A Knowledge Graph is a network database using information gathered from a variety of sources to present information on a topic or thing and its relations with other things.

Learning Objectives

Learning objectives for the course are:

- Analytics terminology
- Monitor the performance of games or websites or any system
- Expose and fix critical issues in flows
- Creation and analysis of metrics
- Unsupervised knowledge-discovery
- Data collection pipelines
- Metric design
- Churn analysis
- Visualization
- Expose and fix critical issues in user flows
- Discovering system bottlenecks
- Robust experimental design and Taguchi methods
- Response surfaces and optimal design
- Understand research design, research methods, and effective writing
- Descriptive statistics
- Probability distributions
- Imputing data
- Normalizing and scaling data
- Data reduction
- Sampling, bootstrapping and confidence intervals
- Pseudo-labeling
- Synthetic data
- Error analysis
- Data drift and concept drift
- Charts for comparing values
- Data visualization
- Exploratory data analysis (EDA)
- Compositional charts
- Distribution charts
- Charts for trends
- Charts for relationships
- Principles of visual design
- Bias, fairness, and error analysis
- Model interpretability
- Evidence Knowledge Graphs (EKG)
- Causal inference

Course GitHub

The course GitHub (for all lectures, assignments and projects):

nikbearbrown YouTube channel

Over the course of the semester I'll be making and putting additional data science and machine learning related video's on my YouTube channel.

https://www.youtube.com/user/nikbearbrown

The purpose of these videos is to put additional advanced content as well as supplemental content to provide additional coverage of the material in the course. Suggestions for topics for additional videos are always welcome.

Teaching assistants

The Teaching assistants are:

TBA

Programming questions should first go to the TA's. If they can't answer them then the TA's will forward the questions to the Professor.

Learning Assessment

Achievement of learning outcomes will be assessed and graded through:

● Quizzes
● Exams
● Completion of assignments involving scripting in R or python, and analysis of data
● Completion of a term paper asking and answering a "real world" question of interest using machine learning techniques
● Portfolio piece

Reaching out for help

A student can always reach out for help to the Professor, Nik Bear Brown nikbearbrown@gmail.com. In an online course, it's important that a student reaches out early should he/she run into any issues.

Grading Policies

Students are evaluated based on their performance on assignments, performance on exams, and both the execution and presentation of a final project. If a particular grade is required in this class to satisfy any external criteria—including, but not limited to, employment opportunities, visa maintenance, scholarships, and financial aid—it is the student's responsibility to earn that grade by working consistently throughout the semester. Grades will not be changed based on student need, nor will extra credit opportunities be provided to an individual student without being made available to the entire class.

Grading Rubric

A point system is used. Everything that you are expected to turn in has points. Points can range from 1 point to 1000 points. For example, every class you are expected to make class notes and upload them by 11:59PM the day of the lecture and that is worth 5 points. A quiz can range from 25 to 100 points and an exam might be 250 points. Assignments that are worth 50 points or less get a 50% deduction for each day they are late rounded up. For example, a late 5 point class notes would get 3 points (2.5 rounded to 3). Assignments more than 50 points get a 10% deduction for each day they are late rounded up. Exams cannot be made up unless arraignments are made before the exam.

I expect to use the following grading scale at the end of the semester. You should not expect a curve to be applied; but I reserve the right to use one.

Score	Grade
93 – 100	A
90 – 92	A-
88 – 89	B+
83 – 87	B
80 – 82	B-
78 – 79	C+
73 – 77	C
70 – 72	C-
60 – 69	D
<60	F

Scores in-between grades. For example, 82.5 or 92.3 will be decided based on the exams.

* Note the score is calculated using the grading rubric and IS NOT the average of the assignments that is displayed by BlackBoard.

Blackboard

You will submit your assignments via Blackboard *and* Github. Click the title of assignment (blackboard -> assignment -> <Title of Assignment>), to go to the submission page. You will know your score on an assignment, project or test via BlackBoard. BlackBoard only represents only the raw scores. Not normalized or curved grades. A jupyter notebook file ALONG with either a .DOC or .PDF rendering of that jupyter notebook file must be submitted with each assignment.

Multiple files must be zipped. No .RAR, .bz, .7z or other extensions.

Assignment file names MUST start with students last name then first name OR the groups name and include the class number and assignment number.

Assignment MUST estimate the percentage of code written by the student and that which came from external sources.

Assignment MUST specify a license at the bottom of each notebook turned in.

All code must adhere to a style guide and state which guide was used.

Due dates

Due dates for assignments at midnight on due date of the assignment.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Solutions will be posted the following Monday. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Course Materials

Many of the textbooks are all available for free to NEU students via SpringerLink (http://link.springer.com/) or via the authors website. The textbooks we will be using in this class are:

Reinforcement Learning: An Introduction by Richard S. Sutton and Andrew G. Barto
http://incompleteideas.net/book/bookdraft2017nov5.pdf

Causal Inference in Statistics - A Primer by Judea Pearl
https://www.amazon.com/dp/1119186846/ref=cm_sw_r_tw_dp_U_x_IjayEbNAZYFG5

An Introduction to Causal Inference by Judea Pearl
https://www.amazon.com/dp/1507894295/ref=cm_sw_r_tw_dp_U_x_4fayEbZPY0Z68

Interpretable Machine Learning A Guide for Making Black Box Models Explainable. Christoph Molnar
https://christophm.github.io/interpretable-ml-book/

Serious Games Analytics
Methodologies for Performance Measurement, Assessment, and Improvement
Christian Sebastian Loh, Yanyan Sheng, Dirk Ifenthaler
https://link.springer.com/book/10.1007/978-3-319-05834-4

The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2017)

Authors: Trevor Hastie, Robert Tibshirani and Jerome Friedman
Free online https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf

Deep Learning - Adaptive Computation and Machine Learning series by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
https://github.com/HFTrader/DeepLearningBook

Participation Policy

Participation in discussions is an important aspect on the class. It is important that both students and instructional staff help foster an environment in which students feel safe asking questions, posing their opinions, and sharing their work for critique. If at any time you feel this environment is being threatened—by other students, the TA, or the professor—speak up and make your concerns heard. If you feel uncomfortable broaching this topic with the professor, you should feel free to voice your concerns to the Dean's office.

Collaboration Policies

Students are strongly encouraged to collaborate through discussing strategies for completing assignments, talking about the readings before class, and studying for the exams. However, all work that you turn in to me with your name on it must be in your own words or coded in your own style. Directly copied code or text from any other source MUST be cited. In any case, you must write up your solutions, in your own words. Furthermore, if you did collaborate on any problem, you must clearly list all of the collaborators in your submission. Handing in the same work for more than one course without explicit permission is forbidden.

Feel free to discuss general strategies, but any written work or code should be your own, in your own words/style. If you have collaborated on ideas leading up to the final solution, give each other credit on what you turn in, clearly labeling who contributed what ideas. Individuals should be able to explain the function of every aspect of group-produced work. Not understanding what plagiarism is does not constitute an excuse for committing it. You should familiarize yourself with the University's policies on academic dishonesty at the beginning of the semester. If you have any doubts whatsoever about whether you are breaking the rules – ask!

Any submitted work violating the collaboration policies WILL BE GIVEN A ZERO even if "by mistake." Multiple mistakes *will be sent to OSCCR for disciplinary review.*

To reiterate: **plagiarism and cheating are strictly forbidden. No excuses, no exceptions.** *All incidents of plagiarism and cheating will be sent to OSCCR for disciplinary review.*

Assignment Late Policy

Assignments are due by 11:59pm on the due date marked on the schedule. Late assignments will receive a 5% deduction per day that they are late, including weekend days. It is your responsibility to determine whether or not it is worth spending the extra time on an assignment vs. turning in incomplete work for partial credit without penalty. Any exceptions to this policy (e.g. long-term illness or family emergencies) must be approved by the professor.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Only ONE extension will be granted per semester.

Student Resources

Special Accommodations/ADA: In accordance with the Americans with Disabilities Act (ADA 1990), Northeastern University seeks to provide equal access to its programs, services, and activities. If you will need accommodations in this class, please contact the Disability Resource Center (www.northeastern.edu/drc/) *as soon as possible* to make appropriate arrangements, and please provide the course instructors with any necessary documentation. The University requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, and arrange with the instructor to provide those on your behalf, as needed.

Academic Integrity: All students must adhere to the university's Academic Integrity Policy, which can be found on the website of the Office of Student Conduct and Conflict Resolution (OSCCR), at http://www.northeastern.edu/osccr/academicintegrity/index.html. Please be particularly aware of the policy regarding plagiarism. As you probably know, plagiarism involves *representing anyone else's words or ideas as your own*. It doesn't matter where you got these ideas—from a book, on the web, from a fellow-student, from your mother. It doesn't matter whether you quote the source directly or paraphrase it; if you are not the originator of the words or ideas, *you must state clearly and specifically where they came from*. Please consult an instructor if you have any confusion or concerns when preparing any of the assignments so that together. You can also consult the guide "Avoiding Plagiarism" on the NU Library Website at http://www.lib.neu.edu/online_research/help/avoiding_plagiarism/. If an academic integrity concern arises, one of the instructors will speak with you about it; if the discussion does not resolve the concern, we will refer the matter to OSCCR.

Writing Center: The Northeastern University Writing Center, housed in the Department of English within the College of Social Sciences and Humanities, is open to any member of the Northeastern community and exists to help any level writer, from any academic discipline, become a better writer. You can book face-to-face, online, or same day appointments in two locations: 412 Holmes Hall and 136 Snell Library (behind Argo Tea). For more information or to book an appointment, please visit http://www.northeastern.edu/writingcenter/.