INFO 7390 Advances in Data Sciences and Architecture

The Art and Science of Data:

Mastering Data with Statistics, Visualization and Generative AI

Course Syllabus

Course Information

Professor: Nik Bear Brown
Email: ni.brown@neu.edu
Office: 505A Dana Hall
Office Hours: Zoom Only

Note: I am also a master's student at Northeastern. Do not send e-mail to my student e-mail brown.ni@husky.neu.edu I almost never read that e-mail.

Course Prerequisites

INFO 6105 - There will be an early assessment of what was learned in INFO 6105. Knowledge of basic statistical learning is essential to the course.

Course Description

Garbage-In Garbage Out (GIGO) may be the most widely used maxim in machine learning, but how does one assess the quality of each step in an analysis pipeline? This course teaches students how to understand their data, models and pipelines using visualization.

Part I - Understanding Data

The first part of the course covers understanding the statistical properties of a data set visually, how to fix issues with their data, and how to graphically demonstrate how the data was improved. The choice of the right chart for a particular question is covered. The principles of visual design, including typography, contrast, balance, emphasis, movement, white space, proportion, hierarchy, repetition, rhythm, pattern, unity, and variety are covered.

1A - Data Analysis and Improvement

In the initial segment of the course, students will be introduced to visual understanding of a dataset's statistical properties. This includes identifying potential problems within the data and implementing corrective measures. We'll also delve into effective graphical methods to showcase the enhancement and transformation of data. An essential part of this section is selecting the most suitable chart type to address specific questions or insights about the dataset.

1B- - Principles of Visual Design in Data Presentation:

In this segment, we put emphasis on the art and science of visual design as it pertains to data presentation. Students will explore various fundamental principles, including typography, contrast, and balance. Additionally, we'll discuss advanced design concepts such as emphasis, movement, the strategic use of white space, proportion, hierarchy, and more. We'll also delve into the significance of repetition, rhythm, pattern, unity, and variety, ensuring that the data is not only accurate but also aesthetically compelling and easily comprehensible.

Part II - Generative AI for Data

2A - Understanding Generative AI

Dive deep into the world of generative AI and its impressive capability to produce content. This section introduces its practical uses and constraints. We will differentiate between traditional machine learning models, generative AI, and artificial general intelligence (AGI). Additionally, we'll uncover the primary elements fueling the progress of generative AI.

2B - Building Generative AI Systems

This segment details the crucial procedures involved in crafting generative AI systems. Topics covered include research, design, data gathering, model training, and assessment. Emphasis is placed on the importance of varied datasets and cutting-edge training strategies. We will also explore the different evaluation techniques, highlighting their advantages and drawbacks.

2C - Employing Generative AI for Synthetic Data Creation

In this section, we venture into the practical application of generative AI in fabricating synthetic data. Whether it's text, visuals, videos, or soundscapes, generative AI offers innovative solutions for generating authentic-seeming content. We'll break down the mechanics behind these processes, demonstrating how cutting-edge models can craft content that's nearly indistinguishable from real-world data, and discuss the potential advantages and challenges of using synthetic data across various sectors.

2D - Leveraging Generative AI for Data Verification

In this segment, we explore how generative AI can play a pivotal role in data verification. By harnessing the power of large language models (LLMs) and cross-referencing information across them, we can enhance the accuracy and reliability of our data. We'll discuss the methodologies behind this innovative approach, detailing how multiple LLMs can be employed in tandem to validate the authenticity of a piece of information. Additionally, we'll touch upon the benefits of this process, emphasizing the increased trustworthiness of data and the reduction in misinformation.

Part III - Causal Inference

The third part of the course covers visualizing causal relationships in data. The emphasis is on understanding visual techniques for separating causal relationships for correlation.

3A - Visual Techniques in Causal Data

In this segment of the course, the primary focus will be on the visualization of causal relationships within datasets. We aim to provide a robust understanding of visual methods that can be employed to distinguish genuine causal connections from mere correlations. This involves diving deep into concepts such as confounding, causal graphs, and the intricate relationship between Directed Acyclic Graphs (DAGs) and probability distributions. The segment will also highlight the significance of paths and associations, along with the idea of conditional independence through d-separation.

3B Advanced Causal Analysis Techniques:

Venturing further into the realm of causality, this segment dives into the practical aspects and methodologies used in observational studies. Techniques and concepts such as optimal matching, sensitivity analysis, and Inverse Probability of Treatment Weighting (IPTW) will be detailed. The course will subsequently move into the nuances of marginal structural models, providing insights on IPTW estimation and the meticulous process of causal effect identification and estimation. This part of the course aims to equip students with the advanced tools and knowledge necessary for in-depth causal analyses in complex scenarios.

Learning Objectives

Upon completion of this course, students will be able to:

General Understanding:

- Recognize the importance of data quality in machine learning and analysis pipelines.
- Utilize visualization techniques to gain insights into data, models, and pipelines.

Understanding Data:

- Visually interpret the statistical properties of datasets.
- Identify and rectify issues within datasets.
- Select appropriate chart types to address specific analytical questions.
- Apply principles of visual design, including but not limited to typography, contrast, balance, and hierarchy, to enhance data presentation.

Data Analysis and Improvement:

- Implement corrective measures to address potential problems within datasets.
- Use graphical methods to demonstrate the enhancement and transformation of data.

Principles of Visual Design in Data Presentation:

- Apply foundational and advanced design principles in data visualization.
- Ensure data presentation is both accurate and aesthetically compelling.

Generative AI for Data:

- Differentiate between traditional machine learning models, generative AI, and AGI.
- Understand the applications and limitations of generative AI.

Building Generative AI Systems:

- Recognize the importance of diverse datasets in generative AI model creation.
- Understand advanced training techniques and evaluation methods.
- Appreciate the strengths and weaknesses of different evaluation techniques in generative AI.

Employing Generative AI for Synthetic Data Creation:

- Understand the mechanics behind generating synthetic data using generative AI.
- Analyze the pros and cons of using synthetic data in various applications.

Leveraging Generative AI for Data Verification:

- Utilize large language models for data verification and validation.
- Recognize the significance of cross-referencing information across multiple LLMs for enhanced accuracy.

Causal Inference:

- Visualize and interpret causal relationships within datasets.
- Distinguish between genuine causal connections and correlations.

Visual Techniques in Causal Data:

- Understand concepts like confounding, causal graphs, and the relationship between DAGs and probability distributions.
- Recognize the importance of paths, associations, and conditional independence in visualizing causality.

Advanced Causal Analysis Techniques:

- Implement methodologies used in observational studies for causal analysis.
- Grasp advanced concepts like IPTW estimation and causal effect identification.
- Apply knowledge in real-world scenarios requiring in-depth causal analyses.

These learning objectives serve as a roadmap for students, indicating the skills and knowledge they will acquire by the end of the course.

Weekly Schedule

Week 1

Information Visualization: Foundations, Data Abstraction
Fundamental Graphs and Data Transformation
Graphical Components and Mapping Strategies
Basic data statistics and Exploratory Data Analysis (EDA)

Week 2

Perception for Information Visualization
Effectiveness of Visual Channels
Identifying Statistical Properties: A Visual Approach
Data Issue Diagnosis and Rectification Strategies

Week 3

Demonstrating Data Improvement through Graphics
Chart Selection for Specific Analytical Questions
Introduction to Principles of Visual Design
Typography, Contrast, and Balance in Data Presentation

Week 4

Advanced Visual Design: Emphasis, Movement, White Space
The Role of Proportion, Hierarchy, Repetition in Visualization
Delving Deeper: Rhythm, Pattern, Unity, and Variety in Design

Week 5

Introduction to Generative AI: Distinguishing Traditional ML, Generative AI, and AGI
Real-world Applications and Limitations of Generative AI

Week 6

Crafting Generative AI Systems: Research and Design Phases
The Importance of Diverse Data Gathering

Week 7

Advanced Training Strategies in Generative AI
Evaluation Techniques: Strengths and Limitations

Week 8

Using Generative AI for Synthetic Data Creation: Text, Images, Videos
Mechanics Behind Generative Processes and Real-world Implications
Generating text

Week 9

Generating numeric data
Generating images

Week 10
Generting audio
Generating video

Week 11
Leveraging Large Language Models for Data Verification
Benefits and Challenges of Cross-referencing Information

Week 13

Introduction to Causal Inference in Data Visualization
Visual Techniques for Understanding Causality: Basics

Week 14

Advanced Causal Analysis Techniques: Observational Studies and Beyond
Wrapping Up: IPTW Estimation, Causal Effect Identification, and Future Trends

Week 15

Review

Communication

Communication between instructor and students is through

- Email via the Canvas distribution list
- Announcements posted on Canvas
- Notes posted on the Canvas discussion board
- email to ni.brown@neu.edu
- Course Slack

DO NOT expect responses for messages sent to LinkedIn, Twitter, Teams, facebook, etc. etc. etc.

Course Structure

- Regularly test students on paper/algorithmic exercises
- Evaluate students' implementation competency, using assignments that require coding on given datasets
- Evaluate ability to setup data, code, and execute using python language
- Exams
- Final project is typically asking and answering a "real world" question of interest using machine learning techniques

Course GitHub

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

https://github.com/nikbearbrown

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 ni.brown@neu.edu. 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 Assignments 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

Typically grades will end up roughly 25% A, 25% A-, 25% B+, 20% B , 5% less than B but that depends on students' performance.

Canvas

You will submit your assignments via Canvas ***and*** Github. Click the title of assignment (Canvas -> assignment -> <Title of Assignment>), to go to the submission page. You will know your score on an assignment, project or test via Canvas. Canvas 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 are midnight on the date assigned.

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/

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

Recommended Texts

An Introduction to Statistical Learning with Applications in R (2013)
Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert
Tibshirani Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-7138-7

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

Beginning Python
From Novice to Professional
Authors: Magnus Lie Hetland 2017
ISBN: 978-1-4842-0029-2 (Print) 978-1-4842-0028-5
https://link.Springer.com/book/10.1007/978-1-4842-0028-5

Python Recipes Handbook

A Problem-Solution
Approach Authors: Joey
Bernard 2016
ISBN: 978-1-4842-0242-5 (Print) 978-1-4842-0241-8
https://link.Springer.com/book/10.1007/978-1-4842-0241-8

Lean Python
Learn Just Enough Python to Build Useful Tools
Authors: Paul Gerrard 2016
ISBN: 978-1-4842-2384-0 (Print) 978-1-4842-2385-7
https://link.Springer.com/book/10.1007/978-1-4842-2385-7

Learn to Program with Python
Authors: Irv Kalb 2016
ISBN: 978-1-4842-1868-6 (Print) 978-1-4842-2172-3
https://link.Springer.com/book/10.1007/978-1-4842-2172-3

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

Beginning Python
From Novice to Professional
Authors: Magnus Lie Hetland 2017
ISBN: 978-1-4842-0029-2 (Print) 978-1-4842-0028-5
https://link.Springer.com/book/10.1007/978-1-4842-0028-5

Deep Learning with Python
A Hands-on Introduction
Authors: Nikhil Ketkar 2017
ISBN: 978-1-4842-2765-7 (Print) 978-1-4842-2766-4
https://link.Springer.com/book/10.1007/978-1-4842-2766-4

Pro Python Best Practices
Debugging, Testing and
Maintenance Authors: Kristian
Rother 2017
ISBN: 978-1-4842-2240-9 (Print) 978-1-4842-2241-6 (Online)
https://link.Springer.com/book/10.1007/978-1-4842-2241-6

Mastering Machine Learning with Python in Six Steps
A Practical Implementation Guide to Predictive Data Analytics Using Python
Authors: Manohar Swamynathan 2017
ISBN: 978-1-4842-2865-4 (Print) 978-1-4842-2866-1
https://link.Springer.com/book/10.1007/978-1-4842-2866-1

Introduction to Data Science
A Python Approach to Concepts, Techniques and Applications
Authors: Laura Igual, Santi Seguí 2017

ISBN: 978-3-319-50016-4 (Print) 978-3-319-50017-1

https://link.Springer.com/book/10.1007/978-3-319-50017-1

Python Recipes Handbook
A Problem-Solution
Approach Authors: Joey
Bernard 2016
ISBN: 978-1-4842-0242-5 (Print) 978-1-4842-0241-8
https://link.Springer.com/book/10.1007/978-1-4842-0241-8

Lean Python
Learn Just Enough Python to Build Useful Tools
Authors: Paul Gerrard 2016
ISBN: 978-1-4842-2384-0 (Print) 978-1-4842-2385-7
https://link.Springer.com/book/10.1007/978-1-4842-2385-7

Learn to Program with Python
Authors: Irv Kalb 2016
ISBN: 978-1-4842-1868-6 (Print) 978-1-4842-2172-3
https://link.Springer.com/book/10.1007/978-1-4842-2172-3

Big Data Made Easy
A Working Guide to the Complete Hadoop Toolset
Authors: Michael Frampton 2015
ISBN: 978-1-4842-0095-7 (Print) 978-1-4842-0094-0
https://link.Springer.com/book/10.1007/978-1-4842-0094-0

Software

python Anaconda

- https://www.continuum.io/anaconda-overview

Python Tutorials

Dive into Python http://diveintopython.org

Python 101 – Beginning Python http://www.rexx.com/~dkuhlman/python_101/python_101.html

The Official Python Tutorial http://www.python.org/doc/current/tut/tut.html

The Python Quick Reference http://rgruet.free.fr/PQR2.3.html

Python Fundamentals Training – Classes http://www.youtube.com/watch?v=rKzZEtxIX14

Python 2.7 Tutorial Derek Banas· http://www.youtube.com/watch?v=UQi-L-_chcc

Python Programming Tutorial - thenewboston http://www.youtube.com/watch?v=4Mf0h3HphEA

Google Python Class http://www.youtube.com/watch?v=tKTZoB2Vjuk

Nice free CS/python book https://www.cs.hmc.edu/csforall/index.html

Deep Learning Tutorials

MIT 6.S191: Introduction to Deep Learning http://introtodeeplearning.com/

Stanford Winter Quarter 2016 class: CS231n: Convolutional Neural Networks for Visual Recognition
https://youtu.be/NfnWJUyUJYU

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/.