UNC GREENSBORO

CS 405/605 Data Science

Dr. Qianqian Tong

About me

Contact information:

Email: q_tong@uncg.edu
Office: Petty 157
Office Hours: Tuesday/Thursday 15:15 – 16:30 @ Petty 157 or appointment by email

B.S. and M.S in Mathematics

Ph.D. degree in Department of CS from University at Connecticut, UCONN in December 2022

Tenure-track Assistant Professor in Department of CS, UNCG

Research: Machine Learning; Stochastic Optimization; Deep Learning; Differential Privacy; Sparse Learning; Graph Convolutional Network (GCN) and Federated Learning.

UNC GREENSBORO

COURSE INFO

What is the course about?

- **Programming your way into Data Science**
- **Theory –Programming**
- **It is not a Statistics or an AI or a Visualization course**
- **The course contains parts of everything**
- **Learn about lot of tools and how to use them in innovative ways**
- **We will work with real-world data**
- **Hopefully develop some cool projects**

COURSE INFO

Prerequisites

A grade of C+ or better in CSC 330 and (STA 271 or STA 290), or permission of instructor (prior programming and statistics experience is required).

Experience in:

- **Programming skills Python**
  We will go through Introduction to Python
  You would have to work hard in the early weeks to get comfortable with Python
- Linux, Terminal, Command-Line

Textbooks

There is no required text for the course. Class slides will be available for download. Suggested textbooks are: 1) Building Machine Learning Systems with Python (Richert and Coelho), 2) Data Science from Scratch (Joel Grus)

COURSE INFO

Grading Policy

Grade	Max%	to	Min%
A	100%	to	92%
A-	< 92%	to	89%
B+	< 89%	to	86%
B	< 86%	to	83%
B-	< 83%	to	80%
C+	< 80%	to	77%
C	< 77%	to	7%
C-	< 74%	to	70%
D+	< 70%	to	67%
D	< 67%	to	64%
D-	< 64%	to	60%
F	< 60%	to	55%

1. Class Participation: 10%

2. In class quizzes (3): 40%

3. Assignments (3): 15%

4. Project: 35%

No final exam! No curve!

UNC GREENSBORO

COURSE INFO

1. Class Participation: 10%

Attendance is mandatory for all class meetings. If a student is unable to attend an in-person class, they must inform the instructor in advance by providing a valid reason for their absence. This communication should be done through email and must be sent before the class session begins. Failure to notify the instructor prior to the start of class will result in the student losing credit for that absence. It should be noted that attendance records may be taken either at the beginning or the end of the class. Students are advised to ensure their presence throughout the session to avoid any discrepancies in the attendance record.

COURSE INFO

2. In class quizzes (3): 40%

Throughout the course, students will be assessed through **three in-class quizzes**, each designed to evaluate their understanding of the lecture material and class discussions. These quizzes will collectively contribute **40 points** toward the final course grade, distributed as follows:

- **Quiz 1** – 10 points
- **Quiz 2** – 10 points
- **Quiz 3** – 20 points

Each quiz is scored out of 100 points and scaled to its respective weight in the overall grade. All quizzes must be completed **in class**, and students are expected to answer all questions thoroughly. To receive full credit, quizzes must be **submitted within the allotted timeframe**. Late submissions may result in partial or no credit, depending on the circumstances.

COURSE INFO

3. Assignments (3): 15%

Three programming-based assignments will be given covering the utilization of the tools learned in class. Each assignment accounts for 5 points. Absolutely no collaboration on assignments. Students must upload (Notebooks) individual assignments to canvas before deadline. Later submission (per day) will have a 50% deduction, late for more than 2 days will directly have zero grade.

COURSE INFO

4. Project: 35%

The project for this course will involve the complete end-to-end development of an analytical model. It will be organized into the following stages:

- Stage I. Dataset Selection and Project Setup,
- Stage II. Data Analysis, Distributions and Hypothesis Testing,
- Stage III. Machine Learning and Deep Learning Model Development,
- Stage IV. Visualization & Final presentation

COURSE INFO

- This will be a team-based effort, where in the first week of the course the students split into teams of 3-4 students.
- After completing each stage, the teams will have to give a short presentation (10 mins) and a report of their progress with the project.
- The projects will be open-source, and the teams will have to use GitHub as their code repository.
- Upon completion of the project the teams will present their software along with the results in form of a presentation (15-20 minutes).

Each Stage of the Final Project has 100 points. They will be equally weighted for the project final score. Each stage consists of: 1). Report; 2). Code Jupyter/IPython Notebooks; 3). Presentation. To get the full points in each stage you need to finish all the deliverables.

Graduate Students Only: In addition to the final presentation, graduate students are required to submit a project report in IEEE format. The report should be at least 3 pages for a single author, 5 pages for two authors or more, inclusive of figures and references. Please ensure that the report is submitted before the final week of the course.

COURSE INFO

Independent project Option:

Undergraduate and master's students are encouraged to pursue independent projects within the Data Science course. If you have a project idea you're passionate about, you are welcome to develop it as part of this course. Projects aimed at publication are particularly encouraged.

Your project should align with the course content, covering topics such as data cleaning, statistical analysis, machine learning models, advanced deep learning networks, and visualization. The project should be both challenging and innovative, offering something new or building upon previous work.

Given that this project constitutes a significant portion of your course grade, a high standard of quality is expected. You will be required to present your progress and final results at each stage throughout the semester.

Note: If you choose this option, please inform the lecturer during the first week of the course.

COURSE INFO

Course Topics and Schedule (Tentative)

1. **Introduction to Data Science: (Week 1-3)**
    - o Data Science Introduction
    - o Class Project discussion and team formed
    - o Programming prepare
        - 1). Re/Introduction to Python
        - 2). IPython, IPython-Notebook
    - o Data Science Reproducibility
        - 1). Setting up your Repository – Data, Code, and Documentation
        - 2). Using Version Control with Git
    - o Project Review - Stage I

2. **Data Munging, Wrangling, Cleaning (Week 4-5)**
    - o Data Structures
    - o Data Manipulation
        - 1). Selection - Indexing
        - 2). Handling Missing Data
        - 3). Aggregation
        - 4). Descriptive Statistics
        - 5). Merging / Join
        - 6). Working with Date-Time
    - o Assignment 1 due
    - o In class quiz 1

COURSE INFO

3. **Data and Statistics (Week 6-8)**
   - o Distributions
   - o Estimates
   - o Statistical Hypothesis Testing
   - o Correlation
   - o Distribution Estimators: MoM, MLE, KDE
   - o Project Review - Stage II
   - o In class quiz 2

4. **Introduction to Applied Data Modeling: (Weeks 9-12)**
   - o Applied Machine Learning
   - o Regression and Feature Selection
   - o Bias versus Variance
   - o Clustering and Dimensionality Reduction
   - o Validation and Model Performance
   - o Mathematical optimization (if time allowed)
   - o Stochastic thinking (if time allowed)
   - o Invited talk (if time allowed)
   - o Assignment 2 due

COURSE INFO

5. **Data Visualization (Week 13-14)**
    - o Graph Generation
        - 1). Types of Graphs
        - 2). Customizing Plots
        - 3). Visualizing Errors
        - 4). Interactive / Dynamic Graphs
    - o Visualization Best Practices
    - o Project Review - Stage III

6. **Project Presentations: (Week 15–16)**
    - o Assignment 3 due
    - o Project Review - Stage IV – Final presentation
    - o Graduate Students report submission
    - o In class quiz 3

COURSE INFO

MIDTERM GRADES:

The midterm grade due for Fall 2024 occurs on September 26, 2024. During this time, I will assign all **undergraduates** a midterm grade for this course, which you can access in UNCGenie. Your midterm grade in this course is a snapshot of how you are currently performing academically based on the assignments we have had to date. It will let you know if you are on the right track or if you need to take action to do something differently to improve your grade. If you have a D or an F at the midterm, we should definitely talk further about strategies and options for continuing in the class.
You can find more information about midterm grades here: https://spartancentral.uncg.edu/student-records/grades/ Once midterm grades are assigned, reach out to me if you have questions. You should also talk with your academic advisor if you are considering withdrawal from this class.

UNC GREENSBORO

COURSE INFO

- Read the syllabus.
- Take regular notes.
- Class is encouraged to participate and discuss/ask questions

On team projects

- Start early

  Emailing me questions about assignments and projects 2 days before submission will not get you a response.

The team creation can be random or self-assigned

- **Task for today**–Get in touch with class participants and setup groups of 3-4 students. Mix of graduate and undergraduate
- Email me the group list (student names, emails, Github ids) before next week.
- Project presentation, all members should present. If someone does not, they will not be graded for the stage.

COURSE INFO

The course is going to be tough, especially for people with limited programming experience

- Work hard, be rewarded with a good data science experience
- Will talk about the benefits later in course intro

Do not cheat in the course –Result will be an 'F' grade.

- Assignment solutions are unique, differs from student to student. **No collaboration on Assignments.**
- I will run the code through plagiarism detection software -*single incident reporting to honor committee*
- In team project
  - Do not think that you can get away without contributing - *I will be monitoring repositories for work done*
  - Any work done should be reported on the repository – *worked locally on your own computer will not count.*

Utilization of resources found on the Internet is allowed for project accomplishment, with caveats

- Any code/library used should be referenced/cited and thoroughly understood
- If you use code without understanding, that counts as plagiarism

Why Data Science?

1. Demand: Data Scientists are in High Demand
Earlier this year, Coding Dojo named data science as the third most in-demand tech job in the U.S. coming in at No. 4 on the list is machine learning engineer, a mid- to senior-level data science job. With great earning potential, thousands of jobs available, and the utter importance of this career, data scientists are highly sought after, making for a great career choice.

2. High Salary: Data Science Careers Have High Earning Potential
There is a lot of money to be made in data science. With base salaries starting at about $100,000 and averaging out around $150K, a career in data science is not only lucrative but fulfilling, rewarding, and challenging.

3. Job Security: Data Science is a Fast-Growing Field
This year, LinkedIn released its annual U.S. emerging jobs report—and data scientists came out on top at 2nd place with 6.5x growth. While it didn't rank as the #1 fastest-growing field, it certainly cropped up pretty quickly.

Why Data Science?

4. Opportunity: Data Science Has a Range of Job Opportunities

According to the U.S. Bureau of Labor Statistics, studying data science gives you access to a diverse range of job opportunities. Some of these employment options include:

- Data Scientist
- Data Analyst
- Data Engineer
- Data Architect
- Business Analyst
- Software Engineer
- Machine Learning Engineer

5. Flexibility: Data Scientists are Needed in Various Sectors

Data scientists are in high demand across a range of industries and sectors. The need will only grow as businesses integrate more data environments into their operations.

Why Data Science?

Requirements

What you'll do in the role:

- The MLE role overlaps with many disciplines, such as Ops, Modeling, and Data Engineering. In this role, you'll be expected to perform many ML engineering activities, including one or more of the following:
- Design, build, and/or deliver ML models and components that solve real-world business problems, while working in collaboration with the Product and Data Science teams.
- Inform your ML infrastructure decisions using your understanding of ML modeling techniques and issues, including choice of model, data, and feature selection, model training, hyperparameter tuning, dimensionality, bias/variance, and validation).
- Solve complex problems by writing and testing application code, developing and validating ML models, and automating tests and deployment.
- Collaborate as part of a cross-functional Agile team to create and enhance software that enables state-of-the-art big data and ML applications.
- Retrain, maintain, and monitor models in production.
- Leverage or build cloud-based architectures, technologies, and/or platforms to deliver optimized ML models at scale.
- Construct optimized data pipelines to feed ML models.
- Leverage continuous integration and continuous deployment best practices, including test automation and monitoring, to ensure successful deployment of ML models and application code.
- Ensure all code is well-managed to reduce vulnerabilities, models are well-governed from a risk perspective, and the ML follows best practices in Responsible and Explainable AI.
- Use programming languages like Python, Scala, or Java.

Why Data Science?

Requirements

- 2+ years' industry experience in Data Science
- M.S or PhD in Data Science/Machine Learning or closely related areas such as Computer Science, Operations Research, Applied Statistics, and Biomedical Informatics.
- Rigorous academic or experiential knowledge of the mathematical essentials for Data Science, including key concepts in probability and statistics, optimization, time series analysis, linear algebra and discrete math. Sampling and estimation, Bayesian analysis, hypothesis testing, uncertainty estimation, stochastic methods, and graphical methods are particularly important to know.
- Deep grounding in machine learning techniques including regression methods (linear, logistic, lasso, support vector, etc.), classification (tree-based models such as XGBoost and Random Forest, Neural Networks, Deep Learning – CNN, RNN, LSTM, etc.), as well as knowledge of clustering and unsupervised learning, time series forecasting and optimization methods.
- Solid foundation with development of data analytics systems, including data exploration/crawling, feature engineering, model building, performance evaluation, and online deployment of models.
- Proficient with server-side programming in Python/Java.
- Hands-on experience in handling large and distributed datasets on Hadoop, Spark, Hive, etc.
- Strong database skills and experience, including experience with SQL programming.
- Experience with AWS or other cloud-based tools and technologies for data pipelining, model development and deployment

Why Data Science?

Requirements

Let's talk about the role:

- You will research, customize when necessary, and develop of statistical and machine learning algorithms to meet complex product requirements. Your tasks will include defining hypotheses, executing necessary tests and experiments, evaluate, tune and optimize algorithms and methods always with an eye towards implementation ease, scalability, and robustness in a live environment.
- You will work closely with other stakeholders from Product Management, Engineering, and other business stakeholders to create impactful, intelligent features and products.
- You will collaborate closely with other team members including other Data Scientists, Machine Learning Engineers, and Data Engineers and “own” the end to end process.
- You will be given wide authority to develop creative model-based solutions but will also be held to high quality and accountability standards.
- You will mentor and train more junior team members and serve as ago-to expert in your area of statistics and machine learning.
- You will thoroughly and diligently document the model design, experiments, tests, validations, and live metrics and outcomes, typically on Confluence. You may be asked to write documents for use in the preparation of intellectual property and technical publications.