Data Science

Course: CSC 405/605 – Data Science
Time: Monday / Wednesday / Friday 12:00 pm – 12:50 pm
Date range: Jan 08, 2024 - Apr 24, 2024
Instructor: Dr. Qianqian Tong
Location: Petty Building 224

Office Hours: Monday/Wednesday 11:00am – 12:00 pm at Petty 152
Email: q_tong@uncg.edu

Course Description

In a world with ever increasing data generated both by humans and machines alike, the field of computer science has seen a transition from computation-intensive solutions to data-intensive ones. Often in such a scenario, solutions to real-world problems can be derived/learned by analyzing disparate, complex, and messy datasets using Data Science methods and approaches.

This course is highly interactive, and will explore the theories, techniques, and the tools necessary to gain insights from such datasets. Using a problem-based learning philosophy, students are expected to make use of such technologies to design data solutions that can process and analyze real-world datasets for a variety of scientific, social, and environmental challenges.

The core topics addressed by the course will be:

- Programming with Data
- Data Mining, Munging, Wrangling
- Statistics, Analytics, Representation, Visualization
- Introduction to Applied Machine-Learning

Prerequisites

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

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 Overview

This course introduces the tools, techniques, and concepts of Data Science. By the end of the course, students will have gained a comprehensive understanding of the data science process, including data collection, cleaning, exploration, feature engineering, modeling, validation, and interpretation.

Course Topics and Schedule (Tentative)

1. **Introduction to Data Science: (Week 1-3)**
    - o Data Science Introduction
    - o Class Project discussion
    - 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 Final Project Discussions - Goals and Requirements

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 submission
    - o Project Review - Stage I
    - o In class quiz

3. **Data and Statistics (Week 6-9)**
    - o Distributions
    - o Estimates
    - o Statistical Hypothesis Testing
    - o Correlation
    - o Distribution Estimators: MoM, MLE, KDE
    - o Project Review - Stage II

4. **Introduction to Applied Data Modeling: (Weeks 10-12)**
    - o Applied Machine Learning

- o Mathematical optimization (if time allowed)
- o Stochastic thinking (if time allowed)
- o Regression and Feature Selection
- o Bias versus Variance
- o Clustering and Dimensionality Reduction
- o Validation and Model Performance
- o Assignment 2 submission
- o Project Review - Stage III
- o In class quiz
- o Invited talk from Industry

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 IV

6**. Project Presentations: (Week 15–16)**
   - o Assignment 3 submission
   - o Project Review - Stage V
   - o Graduate Students report submission

Grading Policy

Grade	Max%	to	Min%
A	100%	to	94%
A-	< 94%	to	90%
B+	< 90%	to	87%
B	< 87%	to	84%
B-	< 84%	to	80%
C+	< 80%	to	77%
C	< 77%	to	74%
C-	< 74%	to	70%
D+	< 70%	to	67%
D	< 67%	to	64%
D-	< 64%	to	60%
F	< 60%	to	55%

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.

2. In class quizzes (2): 10%

Throughout the course, students will be evaluated via two in-class quizzes, each worth 5 points, cumulating to a total of 10 points towards their overall grade. These quizzes are designed to assess the knowledge and understanding students have gained from the lectures and classes. It's imperative that students complete these quizzes in class, ensuring they thoroughly answer all questions. To be considered for full credit, students must submit their responses within the given timeframe. Failure to submit before the deadline may result in a loss of points.

3. Assignments (3): 30%

Three programming-based assignments will be given covering the utilization of the tools learned in class. Each assignment accounts for 10 points. Absolutely no collaboration on assignments. Students must upload (Notebooks) individual assignments to GitHub before deadline. Later submission (per day) will have a 20% deduction, late for 5 days will directly have zero grade.

4. Final Project: 50%

The final project of the class will focus on the end-to-end development of an analytical model. The project will be split into the following stages:

- Stage I. Data/Project Understanding,
- Stage II. Data Modeling,
- Stage III. Distributions and Hypothesis Testing,
- Stage IV. Basic Machine Learning,
- Stage V. Visualization and Dashboard.

This will be a team-based effort, where in the first week of the course the students split into teams of 3-5 students. After completing each stage, the teams will have to give a short presentation (5 mins) and a report (1 page) 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 of the deliverables.

Graduate Students Only: Stage IV has 80 points for your project and 20 points for project report (IEEE format). Minimum 5 pages for single author, 8 for 2 authors, and 12 for 3 authors (figures and references included). The due date should before the final week.

Academic Honesty Policy

The instructor will deal strictly with any violations of academic honesty and integrity in this course. ***Absolutely no discussion, collaboration, copying, and sharing on assignments. This includes coping from the internet. Any student who violates this policy will receive "F" directly in the course. The instructor will report the case to the university.***

Special Needs and/or Disabilities

Students with disabilities should have documentation from the Office of Accessibility Resources & Services. This documentation should be provided to the instructor for review. In the case of major provisions such as separate testing environment or test-readers, the student must make arrangements with Office of Accessibility Resources & Services so that suitable accommodations can be provided.

Midterm Grades

The midterm grade for Spring 2024 dues on Feb 16, 2023. During this time, I will assign 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.

Health and Wellness

Health and well-being have a big impact on your learning and academic success. Throughout your time at UNCG, you may experience a range of concerns that impact your personal and academic success. These might include illnesses, strained relationships, anxiety, high levels of stress, alcohol or drug concerns, crime victimization, feeling down, loss of motivation, or death of a loved one. It is OK TO ASK FOR HELP!

· Student Health Services (SHS) (336-334-5340): For preventative and acute healthcare, SHS offers a primary medical clinic, full pharmacy, and over-the-counter medications.
· Counseling & Psychological Services (336-334-5874): free confidential mental health services
· Spartan Well-Being
· Campus Violence Response Center (336-334-9839)
· Spartan Recovery offers recovery support services (SRP@uncg.edu)