CS676-73434 Algorithms for Data Science

Yiqiao Yin

Spring 2025

E-mail: yyin2@pace.edu
Web: **https://www.y-yin.io/**
Office Hours: TBD
Course Repo: TBD

Course Description

This course delves into essential algorithms for data analytics with a computational emphasis. Students will master Python and R to build algorithms and analyze data. Key topics include data reduction (data mapping, data dictionaries, scalable algorithms, big data), data visualization, regression modeling, and cluster analysis. The course also covers predictive analytics techniques such as k-nearest neighbors, naïve Bayes, time series forecasting, and analyzing streaming data. By the end of the course, students will be proficient in leveraging these algorithms to extract meaningful insights from large datasets.

Required Materials

- Course notes will be released on course repo

Prerequisites/Corequisites

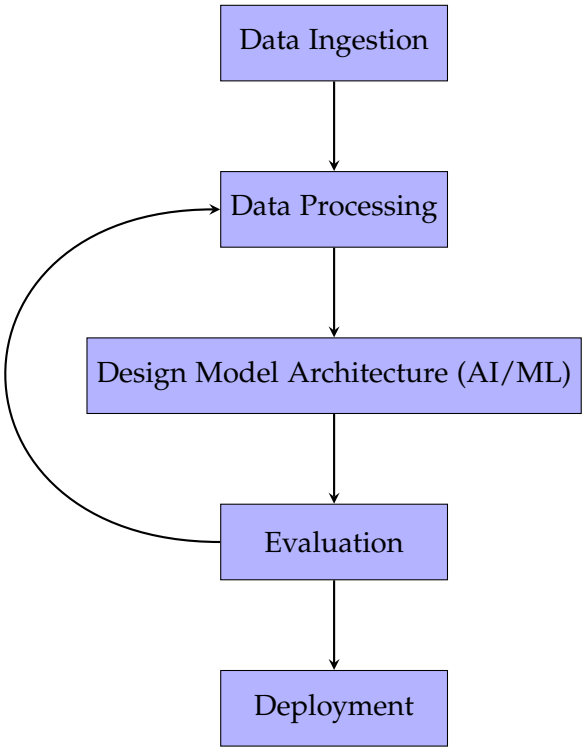
Prerequisites: Open to Data Science Majors.

Course Objectives

Successful students:

1. Develop proficiency in Python and R for data analytics.
2. Implement algorithms for data reduction, including data mapping and data dictionaries.
3. Utilize scalable algorithms to handle big data.
4. Gain insights from data through visualization, regression modeling, and cluster analysis.

Figure 1: **Data Science Roadmap**. This roadmap illustrates the key stages in a data science project, from Data Ingestion through to Deployment, highlighting the iterative process involved in Data Processing and Evaluation.



5. Apply predictive analytics techniques such as k-nearest neighbors, naïve Bayes, and time series forecasting.

6. Analyze and interpret streaming data in real-time.

Course Structure

This course will be conducted in person, allowing for direct interaction and hands-on assistance.

The class is registered for 3 hours per session, but there is flexibility built into the schedule. Each session will be divided into two main parts:

1. The lecture portion will last 1 hour, where key concepts and theoretical foundations will be covered.

2. The coding session will follow, lasting approximately 1.5 hours, depending on the content and complexity of the day's material.

3. A coding component is required for this course. We recommend using Google Colab, which allows students to write and execute Python code in a web-based environment, easily accessible through Google Drive.

Assessments

To successfully qualify as a data scientist, students must demonstrate proficiency in the following areas:

1. Data Engineering: Ability to handle, preprocess, and store large datasets efficiently.
2. Data Visualization: Skills in creating insightful visualizations to communicate data findings.
3. Basic Machine Learning: Understanding and application of fundamental machine learning algorithms or using machine learning tools.
4. Basic API Calls: Competency in making and utilizing API calls to interact with different data services.

Lecture

The lectures are composed of slides and coding sessions. This means that both slides and python notebooks will be used during the lecture. Depending on the content of the materials, slides and coding sessions may be presented in any order.

The slides and coding materials can be found on the course repo.

Final Exam and Class Project

Bla bla **Bla bla**.

Grading Policy

A standard grading policy is adopted. I reserve the right to curve the scale dependent on overall class scores at the end of the semester. Any curve will only ever make it easier to obtain a certain letter grade. The grade will count the assessments using the following proportions:

- **30%** of your grade will be determined by homework assignments. There will be 10 homework assignments, and the 2 lowest scores will be dropped.
- **30%** of your grade will be determined by an open-book, open-source midterm exam. This midterm will be conducted in-person with a computer. Late submissions will have 5 points deducted from the total score of 100 points.
- **40%** of your grade will be determined by the final project. This is a group project where each team will consist of no more than 5 people. The team will pick a project and conduct a data science project. The final deliverable is a presentation, which will be recorded live and sent to external judges.

Course Policies

During Class

This is a technical class with focus on today's modern day technology. I encourage the use of AI tools including ChatGPT and Copilot. You can build your own chatbot if you desire. The class session would be open book and open laptop. I encourage you to use any AI tools to take notes.

Attendance Policy

We do not take attendence.

Policies on Incomplete Grades and Late Assignments

We drop 2 lowest homework grades. There will be no make-up sessions for midterm and final, because they are conducted during class sessions.

Academic Integrity and Honesty

Students are required to comply with the university policy on academic integrity found in the Code of Student Conduct.

Accommodations for Disabilities

Reasonable accommodations will be made for students with verifiable disabilities. In order to take advantage of available accommodations, students must register with the Disability Services Office.

Discrimination based on race, color, religion, creed, sex, national origin, age, disability, veteran status, or sexual orientation is a violation of state and federal law and/or university policy and will not be tolerated. Harassment of any person (either in the form of quid pro quo or creation of a hostile environment) based on race, color, religion, creed, sex, national origin, age, disability, veteran status, or sexual orientation also is a violation of state and federal law and/or NC State University policy and will not be tolerated. Retaliation against any person who complains about discrimination is also prohibited.

Schedule and weekly learning goals

The schedule is tentative and subject to change. The learning goals below should be viewed as the key concepts you should grasp after each week, and also as a study guide before each exam, and at the end of the semester. Each exam will test on the material that was taught up until 1 week prior to the exam (i.e. vorticity will not be tested until exam 2). The applications in the second half of the semester tend to build on the concepts in the first half of the semester though, so it is still important to at least review those concepts throughout the semester. Depending on the school calendar and schedule, the terms "week" and "session" are used interchangeably.

Session 01, 01/08 - 01/12: Introduction

- Overview of the course
- Importance of data science
- Introduction to Python and R

Session 02, 01/15 - 01/19: Basics in Statistical Learning

- Understanding statistical learning
- Key concepts and definitions
- Examples of statistical learning applications

Session 03, 01/22 - 01/26: Linear Regression

- Simple linear regression
- Multiple linear regression
- Assessing the accuracy of the model

Session 04, 01/29 - 02/02: Classification

- Logistic regression
- Linear discriminant analysis
- Performance measures for classification

Session 05, 02/05 - 02/09: Sampling and Bootstrap

- Importance of sampling
- Bootstrap methods
- Applications of sampling and bootstrap

Session 06, 02/12 - 02/16: Model Selection & Regularization

- Criteria for model selection
- Ridge regression
- Lasso regression

Session 07, 02/19 - 02/23: Going Beyond Linearity

- Polynomial regression
- Step functions
- Basis functions and splines

Session 08, 02/26 - 03/02: Tree-based Method and **Midterm**

- Decision trees
- Random forests
- Boosting

Session 09, 03/05 - 03/09: Spring Break

Session 10, 03/12 - 03/16: Continuation of Tree-based Methods

- Detailed analysis of random forests
- Advanced boosting techniques

Session 11, 03/19 - 03/23: Support Vector Machine

- Introduction to SVM
- SVM for classification
- SVM for regression

Session 12, 03/26 - 03/30: Deep Learning

- Fundamentals of deep learning
- Neural networks and architectures
- Applications in real-world problems

Session 13, 04/02 - 04/06: Unsupervised Metrics

- Introduction to unsupervised metrics
- Evaluation of clustering methods
- Practical applications of unsupervised metrics

Session 14, 04/09 - 04/13: Capstone Project Preparation

- Project guidelines
- Team formation
- Initial project planning

Session 15, 04/16 - 04/20: Capstone Project Work

Session 16, 04/23 - 04/27: **Final Exam**