

CSCI 3360 Data Science I

Jaewoo Lee

Fall 2019

E-mail: jwlee@cs.uga.edu Web: www.cs.uga.edu/~jwlee

Office Hours: Mon. 11:30 - 12:30pm (or by appointment)

Office: BOYD 620

Class Hours: TR 2:00 pm - 3:15 pm, M 2:30 pm - 3:20 pm

Class Room: MLC, Room 350

The best way to schedule an appointment outside my office hours is to send me an email with some good dates/times that work for you. I will pick one and reply as quickly as I can.

Course Description

This course is designed as an introductory study of the theory and practice of data science. Data science is about learning from data to extract insight and knowledge. This course introduces computational and statistical tools used in data analysis to answer questions from data. To be specific, we will investigate on tools and methods for

- data collection, data munging, cleaning
- data exploration, hypothesis testing
- data visualization, and communication/interpretation of results
- statistical inference
- making inference on data (regression, classification, and clustering)

Textbook

- An Introduction to Statistical Learning ([pdf](#))
 - Author: Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
 - Publisher: Springer
 - ISBN: 1461471370

Additional materials

- Elements of Statistical Learning, Hastie et al., Springer ([pdf](#))
- Convex Optimization, Boyd and Vandenberghe, Cambridge ([pdf](#))
- Machine Learning: A probabilistic Perspective, Kevin Murphy, MITPress
- Hands-On Machine Learning with Scikit-Learn and TensorFlow, Aurelien Geron, O'Reilly Media
- Deep Learning with Python, Francois Chollet, Manning Publications

Prerequisites/Corequisites

Prerequisites:

CSCI 1302 (Software development), **CSCI 2610** (Discrete math), **CSCI 2720** (Data structures)

- Students are expected to have a working knowledge in at least one programming language (e.g., C/C++, Java, Matlab, Python, or another mainstream language).
- This course will use Python as our main programming language. All class assignments will be in Python (using numpy, scipy, matplotlib, and Keras).
- Undergraduate level **calculus** and **linear algebra**:
We often represent our data as matrices and model parameters as vectors. You should be comfortable with understanding matrix vector operations and notations.

Computing Resource

The main programming language we will be using in this course is PYTHON. All the programming assignments will be done in Python using the Numpy and Scipy package, but prior knowledge of Python is not required. Basic tutorial on how to program in python and how to use numpy and scipy will be taught in class.

- A recommended (and probably the easiest) way of preparing your Python programming environment is to install [Anaconda](#). Installing anaconda on your computer will install both Python and the required libraries.
- The course will be using Python 3. Although installing and using Python 2 won't cause too much trouble, all the provided examples and homework assume that you have installed the Python 3.

Course Objectives

Successful students will be able to:

1. use existing Python tools to read and preprocess raw data of various formats (text, images, binary).

2. choose the most effective visualization to convey the knowledge learned from data.
3. identify and formulate a real-world scientific problem, to design an efficient data science pipeline to solve the problem, and to communicate these solutions effectively.
4. evaluate and compare different learning algorithms.

Grading Policy

Evaluation will 4+ α individual assignments, exams (midterm and final), course project, and random number of quizzes. Each submitted item will be graded on a 100-point scale *without* a curve but 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.

Item	Portion	Description
Homework	40%	4+ α individual assignments
Project	25%	1 page proposal (1%) + formal proposal (4%) + final report (10%) + presentation (10%)
Exams	25%	Midterm 10%, Final 15%
Quizzes	10%	pop quizzes

The grade will be given based on the *total scores*, a weighted sum of collected graded items. It is computed using the following equation:

$$\text{total score} = \frac{\sum_{i=1}^{4+\alpha} \text{HW}_i}{4 + \alpha} \times 0.4 + \text{project} \times 0.25 + \text{midterm} \times 0.1 + \text{final} \times 0.15 + \frac{\sum_{i=1}^? \text{Quiz}_i}{?} \times 0.1$$

Course Policies

Attendance Policy

Attendance is expected in all lectures but is not a part of grade determination. For complete attendance and excused absence policies, please see <https://provost.uga.edu/policies/academic-affairs-policy-manual/4-06-class-attendance/>.

Assignment Submission Policy

All assignments and deliverables are to be submitted via eLC. Allowed file formats are .pdf, .ipynb, and .py. If you need to include images in your answers, you can either embed them into the pdf file or upload the images to cloud storages and include only the links in your solution.

- Email attachments are not considered as an official submission and will not be graded.

Policies on Late Assignments

All assignments are expected to be completed and submitted to the eLC by due date. Normally, assignments are due by 11:59pm on Fridays. Any assignment submitted after 00:01 am on the following day of due date will be considered *late*. Late submission will be penalized by deducting 20% of total marks for the assignment for each day/partial day (including weekend days) beyond the due time. Note that if the assignment is not turned in 5 days after the deadline it will not be accepted.

- All assignment deadlines are *strict*.
- Late assignments will be accepted for no penalty if a valid excuse is communicated to the instructor **at least 2 days before the deadline**; A valid verification record, such as doctor's note, *must* be submitted.

Academic Integrity and Honesty

For all students enrolled in this course, it is assumed that they will abide by UGA's academic honesty policy and procedures. Please refer to UGA's a culture of honesty found at <https://honesty.uga.edu/Academic-Honesty-Policy/Introduction/>. All the linked documents in the url is a part of this syllabus.

For every individual assignment, students are welcome to discuss the problems and share ideas at **high level**. This means that you should not share anything concrete such as write-up, code fragments, or your laptop screen. The submitted item must be a work of yours. For example, you can discuss how to solve a homework problem and share an idea, but you have to write your own answer/code. An egregious violation of these academic honesty codes will result in **F** for the course.

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 Equal Opportunity Office at Suite 119 in Holmes-Hunter Academic building. For more information on UGA's policy on working with students with disabilities, please see <https://eoo.uga.edu/disability-services>.

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 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 UGA's policy and will not be tolerated. Retaliation against any person who complains about discrimination is also prohibited. UGA's policies and regulations covering discrimination, harassment, and retaliation may be accessed at .

Tentative Topics and Schedule

The schedule is tentative and subject to change. Each quiz will test on the material that was taught up to the date of the quiz.

Week 01, 08/12 - 08/15: Introduction

- Overview of data science pipeline
- A case study

Week 02, 08/19 - 08/22: Python Programming

- Basics of numpy and scipy packages
- Scikit learn package
- Drawing plots with matplotlib, Handling data with pandas

Week 03, 08/26 - 08/29: Data Fundamentals and Collection

- What is data?
- Web scraping
- Working with Pandas' DataFrame

Week 04, 09/02 - 09/05: Probability for Computing and Data Analysis

- Probability review
- Random variables and Probability Distributions
- Hypothesis Testing

Week 05, 09/09 - 09/12: Exploratory Data Analysis

- Descriptive Statistics
- Effective Data Visualization

Week 06, 09/16 - 09/19: Statistical Inference

- Parametric VS Non-parametric
- MLE and MAP

Week 07, 09/23 - 09/26: Linear Regression

- Simple Linear Regression
- Multiple Linear Regression
- Overfitting and regularization

Week 08, 09/30 - 10/03: Linear Classification

- What is classification?
- Logistic Regression

Week 09, 10/07 - 10/10: Data Preprocessing

- **Midterm exam** Oct. 7
- Feature encoding
- Feature engineering

Week 10, 10/14 - 10/17: Nonparametric Model

- k -nearest neighbor
- Decision trees

Week 11, 10/21 - 10/24: More algorithms

- Support Vector Machine
- k-means clustering

Week 12, 10/28 - 10/31: Sentiment Analysis

- Bag-of-words model, Naive Bayes
- TFIDF
- NLTK package

Week 13, 11/04 - 11/07: Recommender System

- Collaborative filtering
- Latent factor model

Week 14, 11/11 - 11/14: Fundamentals of Deep Learning

- Multilayer perceptron
- Backpropagation algorithm

Week 15, 11/18 - 11/21: Convolutional Neural Network

- Convolution Operation
- Transfer learning

Week 16, 11/25 - 11/28: Project Presentation I**Week 17, 12/02 - 12/05: Project Presentation II**