CSC392/CSC310: Programming for Data Science

Syllabus – Spring 2018

Time: CSC392 Section 1 TuTh 9:30-10:45, Location: Pastore Hall 259
Webpage: http://homepage.cs.uri.edu/faculty/hamel/courses/2018/spring2018/csc310/
Prerequisites: CSC201 or CSC211

Instructor:
Prof. Lutz Hamel
email: lutzhamel@uri.edu
office: Tyler Hall Rm 251
office hours: TBA

Course Description

Data science exists at the intersection of computer science, statistics, and machine learning. That means writing programs to access and manipulate data so that it becomes available for analysis using statistical and machine learning techniques is at the core of data science. Data scientists use their data and analytical ability to find and interpret rich data sources; manage large amounts of data despite hardware, software, and bandwidth constraints; merge data sources; ensure consistency of datasets; create visualizations to aid in understanding data; build mathematical models using the data; and present and communicate the data insights/findings.

This course provides a survey of data science. Topics include data driven programming in Python; data sets, file formats and meta-data; descriptive statistics, data visualization, and foundations of predictive data modeling and machine learning; accessing web data and databases; distributed data management. You will work on weekly substantial programming problems such as accessing data in database and visualize it or build machine learning models of a given data set. For the midterm and the final exam you will undertake either an individual or team based programming project which you will define and implement.

Goals

The primary aim of CSC 310 is to introduce you to programming in the context of data science and statistical thinking by providing a survey of the major technologies and techniques that are currently being employed.
The objectives of CSC 310 are

- To provide an introduction to data sets, file formats, and meta-data.
- To provide an introduction to database systems such as MySQL.
- To provide a basic overview of data manipulation, statistical data summary techniques, and visualization.
- To provide an introduction to data modeling techniques, in particular computational techniques usually referred as “machine learning”.
- To provide an introduction to high-performance data frame works such as Hadoop and Spark.

Learning Outcomes

At the end of this course, students will be able to:

- Describe what data science is with a detailed view
- Access and visualize data
- Build and evaluate models of data
- Solve problems in data science using standard tools and techniques

Required Texts

Python Data Science Handbook, Jake VanderPlas, O'Reilly, 2017.

Software

We will be using open source software available through the Anaconda3 package.

Grading

Attendance, Labs, Homework	40%
Midterm Project	30%
Final Project	30%

Grade	Minimum %
A	95
A-	90
B+	87
B	83
B-	80
C+	77
C	73
C-	70
D+	67
D	60
F	0

Policies

- Check the website (often)! I will try to keep the website as up-to-date as possible.
- Class **attendance** is mandatory and will be checked.
- Class **promptness, participation,** and **adequate preparation** for each class are expected. If you are absent, it is your responsibility to find out what you missed (e.g. handouts, announcements, assignments, new material, etc.)
- **Late assignments** will not be accepted.
- **Make-up quizzes** and **exams** will **not** be given without a valid excuse, such as illness. If you are unable to attend a scheduled examination due to valid reasons, please inform myself, or the department office in Tyler Hall, prior to the exam

time. Under such circumstances, you are not to discuss the exam with any other class member until after a make-up exam has been completed.

- All work is to be the result of your own individual efforts unless explicitly stated otherwise. **Plagiarism, unauthorized cooperation or any form of cheating** will be brought to the attention of the Dean for disciplinary action. See the appropriate sections (8.27) of the University Manual.
- **Software piracy** will be dealt with exactly like stealing of university or departmental property. Any abuse of computer or software equipment will subject to disciplinary action.
- Any student with a documented disability is welcome to contact me as early in the semester as possible so that we may arrange reasonable accommodations. As part of this process, please be in touch with Disability Services for Students Office at 302 Memorial Union, Phone 401-874-2098.

Tentative Schedule

Week 1 & 2

- What is Data Science?
- Python
- Jupyter Notebooks

Week 3

- A Quick Crash Course to get You started
    - Data Sources
        - The CSV File
    - Basic Descriptive Statistics
    - Visualizing Data
    - Model Building/Machine Learning

Week 4 & 5

- More on working with Numeric Data
    - Numpy
    - Pandas Data Frames
    - Data Cleaning/Transforming
- Data Visualization with PyPlot and other Python Modules

Week 6 & 7

- Models of Data
    - Statistical Models
    - Machine Learning
- Evaluating Models
    - Model Selection
    - Model Comparisons
- Model Deployment

Week 8

- Natural Language/Text Processing
    - Bag of Words
    - n-Grams

Week 9

- Databases
    - MySQL

Week 10

- High-Performance Data Science Programming
    - MapReduce
    - Hadoop

Week 11

- Report Writing and Communication

Tentative Lab Schedule

1. Jupyter Notebooks
2. Python Basics
3. Python: Game of Life
4. Data Set Basics - the CSV File: loading, manipulating and summary statistics
5. Machine Learning: Decision Trees
6. Model Evaluation I
7. Working with Python Data Frames
8. Data Visualization
9. Model Evaluation II: Cross-Validation, Confusion Matrices, and Confidence Intervals
10. Additional ML Models: KNN and ANN
11. NLP: “Fake News Detection”