CptS 575: Data Science
Syllabus

Course Information

Credit Hours: 3
Semester: Fall 2019
Meeting times and location: MWF, 9:10–10:00, GTZN 21
Course website: https://scads.eecs.wsu.edu/index.php/data-science-f19/

The course website will be used to post relevant course material—including this syllabus—and course related resources. Additionally, the online portal OSBLE+ (https://plus.osble.org) will be used for posting lecture material, assignments, announcements, and messages; and for handling student submissions and instructor feedbacks.

Instructor Information

Instructor: Assefaw Gebremedhin
Office: EME B43
Email: assefaw DOT gebremedhin AT wsu DOT edu
Homepage: www.eecs.wsu.edu/~assefaw

Office hours: Wednesdays 10:30–12pm, or by appointment.

Teaching Assistant: Helen Catanese
Email: helen DOT catanese AT wsu DOT edu
Office: Dana 115
Office Hours: Wednesdays 2–4pm

Course Description

Data Science is the study of the generalizable extraction of knowledge from data. Being a data scientist requires an integrated skill set spanning computer science, mathematics, statistics, and domain expertise along with a good understanding of the art of problem formulation to engineer effective solutions. The purpose of this course is to introduce students to this rapidly growing field and equip them with some of its basic principles and tools as well as its general mindset. The course will primarily use the programming language R and additionally Python as needed.

Topics to be covered include: the data science process, exploratory data analysis, data wrangling, linear regression, classification, clustering, principal component analysis, data visualization, time-series data mining, deep learning, and data and ethics.

The focus in the treatment of these topics is on breadth, rather than depth, and emphasis is placed on integration and synthesis of concepts and their application to solving problems. Necessary theoretical abstractions (mathematical and algorithmic) are introduced as and when needed.

Audience

The course is suitable for graduate students in computer science, engineering, applied mathematics, the sciences, business, and related analytic fields.

The course is offered conjoint with a 400-level course.

Prerequisites

Students are expected to: (i) have taken an introductory course in statistics and probability, (ii) have basic knowledge of algorithms and reasonable programming experience (equivalent to completing a data structures course such as CptS 223), and (iii) have some familiarity with basic linear algebra (e.g. eigenvalue/vector computation).

Course Work

The course cosnists of several elements: lectures (three times a week, 50 min each) that incoprporates labs as needed; a set of assignments; a survey paper; a substantial semester project; and an exam. Below is how the coursework and assessment is broken down.

- **Assignments (25%)**. There will be a total of 5 assignments spread through the semester. Each assignment will have one major topic of emphasis. Assignments are to be completed and submitted individually. Each assignment will carry equal weight. Together all assignments account for 25% of final grade.
- **Semester Project (30%)**. Students, working in teams of two or three, will complete a semester project. A project could take one of several forms: analyzing an interesting dataset using existing methods and software tools; developing new data science methods; careful performance evaluation of known methods; building your own data product; or creating a visualization of a complex dataset. Students will be given an opportunity to choose from a list of projects the instructor provides or propose their own project. Guidelines for what constitutes a project will be provided by the instructor. A project will culminate in a written report and a short (3-min) video presentation in class. General guidelines for how to prepare a report will be provided by the instructor.
- **Survey Paper (15%)**. Each student, individually, will write a survey paper further exploring a specific topic related to the course content. The topic will be chosen in consultation with the instructor. The background material for the paper will be drawn from journal/conference literature reflecting recent research. The format of the paper will be similar to typical journal/conference survey papers. The length of the paper will depend on the nature of the topic chosen, but will typically be around 10 pages.
- **Exam (20%)**. There will be one exam designed to a) cover most of the material in class and b) complement the assignments and semester project.
- **Class participation (10%)**. Attendance and active class participation (in discussions, in-class exercises and thought experiments) is required. It will count to 10% of the final grade.

Expectations for Student Effort

For each hour of lecture equivalent, students should expect to have a minimum of two hours of work outside class.

Grading

Letter grades will be given according to the following ranges:

A (93–100%), A- (90–92.99%), B+ (87–89.99%), B (83–86.99%), B- (80–82.99%), C+ (77–79.99%), C (70–76.99%), C- (67–69.99%), D (60–66.99%), F (less than 60%).

Learning Outcomes and Assessment

Student Learning Outcomes. *By the end of the course, students should be able to:*	**Course Topics/Dates.** *The following topics/dates will address this outcome:*	**Evaluation.** *This outcome will be evaluated primarily by:*
• Describe what Data Science is and the skill sets needed	What is Data Science? (week 1);	Assignments; Exam
• Describe the Data Science Process	EDA and the Data Science Process (week 4)	Assignments; Exam; Project
• Use R (or Python) to carry out statistical modeling and analysis	Intro to R (week 2); Most subsequent topics throughout the semester	Assignments; Project
• Carry out exploratory data analysis	EDA (week 4)	Assignments; Project
• Use effective data wrangling approaches to manipulate data	Data Wrangling (week 5)	Assignments; Project
• Apply machine learning algorithms for predictive modeling	Linear Regression (week 6); Classification (weeks 7 and 8); Deep Learning (week 13, week 15)	Assignments; Project; Exam
• Apply effective methods to assess model performance	Cross-validation (week 9)	Exam; Project
• Apply learning methods to discover patterns, trends and anomalies in data	Unsupervised Learning (week 10); Time Series Data Mining (week 12)	Assignments; Project; Exam
• Create effective visualization of data (to communicate or persuade)	Data Visualization (week 11)	Assignments; Project
• Reason around ethical and privacy issues in data science conduct, and apply ethical practices	Data and Ethics (week 15)	In-class exercise
• Work effectively in teams on data science projects		Project
• Apply knowledge gained in the course to carry out a project and write a technical report		Project

Detailed Topics and Course Outline

1. Introduction: What is Data Science?
   - Big Data and Data Science; Landscape of perspectives; Skill sets needed
2. Intro to R
   - R basics; R graphics; R Markdown
3. Overview of Machine Learning
   - Supervised Learning (canonical examples and real world applications); Unsupervised Learning (canonical examples and real world applications)
4. Exploratory Data Analysis and the Data Science Process
   - Basic tools (plots, graphs and summary statistics) of EDA; Philosophy of EDA; The Data Science Process
5. Data Wrangling
   - Data transformation and manipulation (dplyr); Relational data; Data "tidying" (tidyr)
6. Linear Regression
   - Simple linear regression; Multiple linear regression; Extensions of the linear model
7. Classification
   - Overview of classification; Logistic regression; Linear Discriminate Analysis; Naive Bayes classifier; K-Nearest Neighbors (KNN); Decision Trees and Random Forest
8. Resampling Methods
   - Cross-validation; The Bootstrap
9. Unsupervised Learning
   - Principal Component Analysis (PCA); K-means clustering; Hierarchical clustering
10. Data Visualization
    - Telling story with data; Choosing tools to visualize data; Visualizing patterns over time; Visualizing proportions; Visualizing relationships; Visualizing text information
11. Time Series Data Mining Basics
    - Examples of areas where time series data arise; Distance measures; Algorithms (motif discovery, anomaly detection, segmentation, classification, clustering); Tools: Matrix Profile and SAX
12. Intro to Deep Learning
    - What is deep learning? Overview of deep networks; TensorFlow
13. Data Science and Ethical Issues
    - Discussions on privacy, security, ethics; A look back at Data Science

Books

There is no required "textbook" for this course. Select chapters from the followings books will be used as a starter to guide the discussions, but they will be supplemented with instructor-developed lecture notes and reading assignments from other sources. The lecture notes and reading material will be made available on the OSBLE+ page of the course as the course proceeds.

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R.* Springer, 2013. ISBN 978-1461471370. The book is freely available online at: http://www-bcf.usc.edu/~gareth/ISL/.
- Hadley Wickham and Garett Grolemund. *R for Data Science* 2017. http://r4ds.had.co.nz/
- Avrim Blum, John Hopcroft and Ravindran Kannan. *Foundations of Data Science.* (Note: this is a book currently being written by the three authors. The authors have made a draft of their notes for the book available online https://www.cs.cornell.edu/jeh/book.pdf. The material is intended for a modern theoretical course in computer science.)
- Jure Leskovek, Anand Rajaraman and Jeffrey Ullman. *Mining of Massive Datasets.* v2.1, Cambridge University Press. 2014. The book is freely available online at: http://www.mmds.org/#ver21.
- Jiawei Han, Micheline Kamber and Jian Pei. *Data Mining: Concepts and Techniques.* Third Edition. Morgan Kaufmann Publishers. 2012. ISBN 978-0-12-381479-1.
- Ethem Alpaydin. *Introduction to Machine Learning.* Third Edition. MIT Press, 2014. ISBN 978-0-262-02818-9.
- Nathan Yau. *Visualize This: The FlowingData Guide to Design, Visualization, and Statistics.* Wiley Publications, 2011. ISBN-13: 978-0470944882.
- Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning.* MIT Press, 2016. ISBN 9780262035613. The book is freely available online at: http://www.deeplearningbook.org

Weekly Schedule

See Table 1 for a weekly schedule of topics and assignments.

Policies

Conduct

Students are expected to maintain a professional and respectful classroom environment. In particular, this includes:

- silencing personal electronics
- arriving on time and remaining throughout the class

Correspondence

All class related correspondence with the instructor will be made via OSBLE+. I will check messages sent to my Inbox on a regular basis, and will do my best to respond promptly. Students are encouraged to choose their OSBLE+ settings so that they get emails notifications when messages are sent or posted.

Week	Topics	Assignments
01 (Aug 19)	What is Data Science?	Pre-course survey out
02 (Aug 26)	Intro to R/Python	Survey due, Assignment 1 out
03 (Sep 02)	Overview of Machine Learning	Assignment 1 due
04 (Sep 09)	Exploratory Data Analysis	Assignment 2 out
05 (Sep 16)	Data Wrangling	Assignment 2 due, Assignment 3 out
06 (Sep 23)	Project Setup; Linear Regression (LR)	Assignment 3 due, Project proposal out
07 (Sep 30)	LR II; Classification I	Assignment 4 out
08 (Oct 07)	Classification II	Project proposal due
09 (Oct 14)	Resampling Methods	Assignment 4 due
10 (Oct 21)	Unsupervised Learning	Assignment 5 out
11 (Oct 28)	Data Visualization	Assignment 5 due
12 (Nov 04)	Time Series Data Mining	
13 (Nov 11)	Deep Learning (DL)	Mid-term Exam
14 (Nov 18)	DL II, Ethics, Wrap-up	In-class Exercise
15 (Nov 25)	**Thanksgiving break**	
16 (Dec 02)	Project presentations	Final project report due on Dec 12

Table 1: Tentative week-by-week schedule of topics and assignments.

Attendance

Regular attendance is required. While students may miss class for urgent reasons, absences that are not cleared with the instructor will factor into the Class Participation portion of the semester grade.

Missing or late work

Submissions will be handled via the OSBLE page of the course. Students are expected to submit assignments by the specified due date and time. Assignments turned in up to 48 hours late will be accepted with a 10% grade penalty per 24 hours late. Except by prior arrangement, missing or work late by more than 48 hours will be counted as a zero.

Academic Integrity

Academic integrity is the cornerstone of higher education. As such, all members of the university community share responsibility for maintaining and promoting the principles of integrity in all activities, including academic integrity and honest scholarship. Academic integrity will be strongly enforced in this course. Any student who violates the University's standard of conduct relating to academic integrity will receive an F as a final grade in this course, will not have the option to withdraw from the course and will be reported to the Office of Student Standards and Accountability. Cheating is defined in the Standards for Student Conduct WAC 504-26-010 (3). You can learn more about Academic Integrity on the WSU campus at http://conduct.wsu.edu. Please also read this link carefully: EECS Academic Integrity Policy (http://www.eecs.wsu.edu/~schneidj/Misc/academic-integrity.html). Use these resources to ensure that you do not inadvertently violate WSU's standard of conduct.

Safety on Campus

Washington State University is committed to enhancing the safety of the students, faculty, staff, and visitors. It is highly recommended that you review the Campus Safety Plan (http://safetyplan.wsu.edu/) and visit the Office of Emergency Management web site (http://oem.wsu.edu/) for a comprehensive listing of university policies, procedures, statistics, and information related to campus safety, emergency management, and the health and welfare of the campus community.

WSU Classroom Safety

Classroom and campus safety are of paramount importance at Washington State University, and are the shared responsibility of the entire campus population. WSU urges students to follow the "Alert, Assess, Act" protocol for all types of emergencies and "Run, Hide, Fight" response for an active shooter incident. Remain ALERT (through direct observation or emergency notification), ASSESS your specific situation, and act in most appropriate way to assure your own safety (and the safety of others if you are able).

Please sign up for emergency alerts on your account at MyWSU. For more information on this subject, campus safety and related topics, please view the FBI's Run, Hide, Fight video (https://www.fbi.gov/about-us/cirg/active-shooter-and-mass-casualty-incidents/run-hide-fight-video) and visit the WSU safety portal (https://faculty.wsu.edu/classroom-safety).

Students with Disabilities

Reasonable accommodations are available for students with a documented disability. If you have a disability and need accommodations to fully participate in this class, please either visit or call the Access Center (Washington Building 217; 509-335-3417) to schedule an appointment with an Access Advisor. All accommodations must be approved through the Access Center. For more information, consult the webpage http://accesscenter.wsu.edu or email at Access.Center@wsu.edu.

Important Dates and Deadlines

Students are encouraged to refer to the academic calendar often to be aware of critical deadlines throughout the semester. The academic calendar can be found at http://registrar.wsu.edu/academic-calendar.

Weather Policy

For emergency weather closure policy, consult: http://alert.wsu.edu.

Changes

This syllabus is subject to change. Updates will be posted on the course website.