CptS 475/575: Data Science
About me

• Name: Assefaw Gebremedhin
• Office: EME B43
• Webpage: www.eecs.wsu.edu/~assefaw
• Joined WSU: Fall 2014
• Research interests: algorithmic data science, network science, high performance computing, bioinformatics
• Lab: Scalable Algorithms for Data Science Laboratory (https://scads.eecs.wsu.edu)
• NSF CAREER project: Fast and Scalable Combinatorial Algorithms for Data Analytics www.eecs.wsu.edu/~assefaw/fascada
• Teaching at WSU:
  • CptS 483: Intro to Data Science  (Fall 2015, 2016, 2017)
  • CptS 475/575: Data Science (Fall 2018)
  • CptS/STAT 424: Data Analytics Capstone (Spring 2019)
• CptS 475/575 Fall 2019:
  • Lectures: MWF, 9:10--10, GTZN 21
  • Office Hours: Wed. 10:30am--12pm (or by appointment)
  • Teaching Assistant: Helen Catanese
  • Email: helen.catanese@wsu.edu
  • Office: Dana 115 (office hours: Wed. 2--4pm)
About Data Science Class of 2019
(What I know so far)

- CptS 475: 25 enrolled
- CptS 575: 45 enrolled, 5 waiting list (50 total)

- CptS 475:
  - Almost all BS in CS

- CptS 575 breakdown:
  - 25 MS
    - Computer Science (more than 20)
    - Electrical Engineering
    - Environmental Engineering
  - 25 PhD
    - Computer Science
    - Electrical and Computer Engineering
    - Material Science and Engineering
    - Engineering Science
    - Civil Engineering
    - Environmental Engineering
    - Economics
    - Crop Science
    - Math
    - Individual Interdisciplinary (IID)
Course websites

• Public course site: https://scads.eecs.wsu.edu/index.php/data-science-f19/
  • Syllabus
  • Overview of schedule (updated after every lecture)
  • Resources

• OSBLE+: https://plus.osble.org
  • Lecture material
  • Assignments
  • Announcements and posts
  • Submissions and feedback

• Currently:
  • 37 added users (no further action)
  • 3 whitelisted (be sure to respond to invitation ASAP)
  • 30 not added (make sure to create an account on OSBLE+ by going to https://plus.osble.org and then Join the course CptS 475/575 Fall 2019)
Course Description

• Data Science is the study of the generalizable extraction of knowledge from data
• Data science requires integrated skill set spanning
  • Computer science
  • Mathematics & Statistics
  • Domain expertise
  + art of problem formulation to engineer effective solutions
• Purpose of this course: introduce basic principles, tools, and general mindset
• Emphasis on breadth rather than depth; and on synthesis of concepts
• Primarily uses the statistical computing language R (will also use/allow Python)
Expectation

• Basic knowledge of algorithms and reasonable programming experience (equivalent to completing CptS 223)
• Familiarity with basic linear algebra
• Basic probability and statistics

• Deficiencies can to a degree be overcome with extra effort
Topics

1. **Introduction: What is Data Science?**
2. **Intro to R (and Python)**
3. **Exploratory Data Analysis and the Data Science Process**
4. **Data Wrangling**
   - Data transformation and manipulation (dplyr); Relational data; Data tidying (tidyr)
5. **Linear Regression**
6. **Classification**
   - Overview, Logistic regression, Linear Discriminate Analysis, k-Nearest Neighbors
   - Decision Trees and Random Forest
7. **Resampling Methods**
   - Cross-validation, The bootstrap
8. **Unsupervised Learning**
   - Principal Component Analysis, K-means clustering, Hierarchical clustering
9. **Data Visualization**
10. **Time Series Data Mining**
    - Distance measures, transformations, algorithms, tools (Matrix Profile, SAX)
11. **Intro to Deep Learning**
12. **Data Science and Ethics**
A few things

• Pre-course survey
  • Your background
  • Level of familiarity with R, Python, MathLab
  • Topics you are excited about
  • Other topics you wish to see covered
  • Complete and submit on OSBLE

• R tutorial

• (Python tutorial)

• Tutorial generally preferred time
Course work and assessment: CptS 475

• Assignments (30%)
  • Total of 5 assignments spread through the semester
  • Completed and submitted individually
  • Each of the assignments carries equal weight

• Semester Project (30%)
  • Team of two or three
  • Option between choosing from a given list OR propose own project
  • Guidelines will be provided

• Exam (30%)
  • Late midterm
  • Designed to cover most material AND complement assignments and semester project

• Class participation (10%)
  • Attendance
  • Active participation
Course work and assessment: CptS 575

• Assignments (25%)
  • Total of 5 spread through the semester
  • Completed and submitted individually
  • Each of the assignments carries equal weight

• Semester Project (30%)
  • Team of two or three
  • Option between choosing from a given list OR propose own project
  • Guidelines will be provided

• Survey Paper (15%)
  • Submitted individually
  • Further explore a specific topic related to the course content
  • Topic to be chosen in consultation with instructor

• Exam (20%)
  • Late midterm
  • Designed to cover most material AND complement assignments and semester project

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

- Describe what Data Science is and the skill sets needed
- Describe the Data Science Process
- Use R to carry out basic statistical modeling and analysis
- Carry out exploratory data analysis (to gain insight)
- Apply basic machine learning algorithms for predictive modeling
- Correctly apply cross-validation to assess model performance
- Apply unsupervised learning methods to discover patterns, trends and anomalies in data
- Use effective data wrangling approaches to manipulate data
- Create effective visualization of data (to communicate or persuade)
- Reason around ethical and private issues in data science and apply ethical practices
- Work effectively in teams on data science projects
- Apply knowledge gained in the course to carry out a project and write technical report
Books

- No required textbook
- Lecture notes (slides) and reading material will be made available on the OSBLE+ page
- References
  - Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013. (Freely available online)
  - Hadley Wickham and Garett Grolemund. *R for Data Science*. 2017. (Freely available online)
Policies

- **Conduct in class**
  - Silence personal electronics
  - Arrive on time and remain throughout the class

- **Correspondence**
  - Happens via OSBLE+

- **Attendance**
  - Required. Make sure absences are cleared with me

- **Missing or late work**
  - Max 48 hrs with 10% penalty per 24 hrs

- **Academic Integrity**
  - Strongly enforced

- Consult syllabus for more details
Welcome again to the course!
I am excited to have you in the class,
and I look forward to your participation and to
a great semester!