DEPARTMENT OF STATISTICS
COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK

STAT G5243/GU4243

Applied Data Science

Department of Statistics, Columbia University
Course Information

- Section 1:
  - Classes: Wednesdays 6:10pm-8:55pm, 313 Fayerweather
  - Instructors: Ying Liu. liuying4490@gmail.com (@yingliug)
  - TA: Chengliang Tang. ct2747@columbia.edu (@ChengliangTang)
  - Course websites (all accessible via courseworks or github):
- Section 2:
  - Classes: Thursdays 1:10pm-3:30pm, 403 International Affairs Building
  - Instructors: Tian Zheng. tian.zheng@columbia.edu (@tz33cu)
  - TA: Jing Wu. jw3233@columbia.edu (@wendy9217)
  - Course websites (all accessible via courseworks or github):
- Grades and basic course info on **Courseworks**: http://courseworks.columbia.edu
- Discussion board on **Piazza**: https://piazza.com/class/jkst2bihetv196
- Course materials and repositories on **GitHub**: http://tzstatsads.github.io

Prerequisites

The pre-requisite for this course includes working knowledge in Statistics and Probability, data mining, statistical modeling and machine learning. Prior **advanced** programming experience in R or Python is required.

Description

This course incorporates knowledge and skills covered in a statistical curriculum with topics and projects in data science. Programming will be covered using existing tools in R, while students can use tools from other languages. Computing best practices will be taught using test-driven development, version control, and collaboration. Students finish the class with a portfolio on GitHub, and deeper understanding of several core statistical/machine-learning algorithms.

This course will be a project-based hands-on course in data science. **No formal instruction on statistics, data science, machine learning will be given**. Project cycles run every 2-3 weeks, where we will have mini- group data projects. Groups will be formed **randomly** and project products will be peer-reviewed, in addition to evaluation by the instructional team.

Course organization

This course will have a total of *five* project cycles. Each project cycle follows a sequence of four types of activities.

a. Dataset release, introduction to data science problem, individual exercises, team forming

b. Lecture/tutorial

c. Brainstorming, live hacking, code sharing

d. Team presentation, peer reviews, within-team peer reviews

Students will be working in teams of 5 students that will be randomly formed. For a meaningful experience in data science, students are expected to collaborate and work together on all the stages of a project. Code sharing and brainstorming are great opportunities to learn from each other.

We will have a total of five project cycles for this course (topics are subject to change):

1. [Individual] R notebook for exploratory data analysis
2. Shiny app for interactive data visualization project.
3. Predictive analytics of images.
4. Research evaluation and reproducibility challenge.
5. *Free topic*.

Below is a tentative schedule we will follow.

- Week 1 (Sep 5/6): 1a+1b
- Week 2 (Sep 12/13): 1c
- Week 3 (Sep 19/20): 1d+2a
- Week 4 (Sep 26/27): 2b+2c
- Week 5 (Oct 3/4): 2c
- Week 6 (Oct 10/11): 2d+3a
- Week 7 (Oct 17/18): 3b+3c
- Week 8 (Oct 24/25): 3b+3c
- Week 9 (Oct 31/Nov 1): 3d+4a
- Week 10 (Nov 7/8): 4b+4c
- Week 11 (Nov 14/15): 4b+4c
- Thanksgiving week: no class
- Week 12 (Nov 28/29): 4d+5c

- Week 13 (Dec 5/6): 5d

Evaluation

Students' performance will be evaluated based on

- [85%] Project products (instructor-reviewed and/or peer-reviewed, averaged over 5 projects). Each project description will have explicit grading rubrics.
- [15%] Individual participation (based on individual tasks and instructors' observation).

  A note on participation evaluation.

  In addition to individual tasks such as peer reviews, for each project, we will enforce formal evaluation of participation as follows.

  - Each project needs to show clear collaboration and task assignments in Piazza discussion using the group discussion function.
  - Teams should try to use GitHub to coordinate code sharing and project development throughout the project. GitHub activities will be used as part of participation evaluation.
  - Students should participate actively in class discussion and piazza discussion.
  - We will give participation score for each project cycle, the average of which will contribute to 15% of your final grade. The participation will be graded on the following curve.
    - A (1.8-2): project leader, major contributor who contribute substantially in every stage of the project and class discussions.
    - A- (1.5-1.8): major contributor who contributed substantially to two stages of the project and some discussions.
    - B+ (1.2-1.5): average participation, participate in the discussion at every stage and contribute substentially in at least one stage of the project and some discussions.
    - B (1-1.2) or lower: below average performance.
  - This is to ensure a positive learning process for all of us.

Communication

Projects grades are managed in courseworks. We will be using the discussion/announcement tools in Piazza (accessible from Courseworks) for our online class communication and discussion. The system is highly catered to getting you help fast and efficiently from classmates, the TAs, and instructors. Rather than emailing questions to the teaching staff, we encourage you to post your questions online.

Textbook

There is not a single required text. As part of this course, we will learn from what we can find online and in academic papers. Here are a couple of recommended reference books.

- Mount and Zumel (2014) Practical data science with R.
- Segaran (2007) Programming collective intelligence: building smart web 2.0 applications.

- Tuffe (2001) The visual display of quantitative information.
- Fung (2013) Numbersense: how to use big data to your advantage.

Class policy

- We learn together through projects. Please stay positive and congenial. Share what you know with your peers and also learn from them.
- Working towards deadlines can be stressful. Remember, emails or online posts do NOT have tones. Be mindful about how you phrase your questions, comments, inquries and suggestions. Also be generous when reading them.
- Academic Integrity is the cornerstone of meaningful teaching and learning. It is especially important for our project-based course. Remember what matters more is how much you learn not what grade you will get. In your project, document references and resources that have been incorporated into your project and accredit them appropriately. Plagiarism is one of the most likely forms of cheating in this course.
- Be a good team member and contribute to each project as much as you can. Don't underestimate the efforts of your teammates. Something seems simple may not be that simple.
- Emails related to learning and projects shall be redirected to our discussion board.
- Students are expected to check emails at least once every 12 hours during the week and every 24 hours over the weekend. Students should make sure not to miss any important class-related announcements sent by emails or posted on Courseworks. Emails will be delivered to the students' official UNI. It is the students' responsibility to ensure that these emails are properly forwarded if they choose to use an alternative email address.