G5243/GU4243 Fall 2016 Applied Data Science

Department of Statistics, Columbia University

Course Information

- Classes: Wednesdays 2:40-5:25 PM, Pupin 329
- Instructor: Tian Zheng (Office hours: Mondays 12:00 PM - 2:00 PM, plus announced online Q&A or by appointments; Room 1007, SSW). Email: tian.zheng@columbia.edu (@tz33cu)
- TA: Chengliang Tang. ct2747@columbia.edu (@ChengliangTang)
- Course assistants
    - Ruixiong Shi (@shi297096484)
    - Yuhan Sun (@Lanmo77)
    - Jingying Zhou (@MadLily)
- Course websites (all accessible via courseworks or github):
    - Grades and basic course info: http://courseworks.columbia.edu
    - Discussion board: Piazza https://piazza.com/class/ishmhsazsc54y
    - Course materials and repositories: 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 programming experience in R or Python is required.

Description

This course will incorporate knowledge and skills covered in a statistical curriculum with topics and projects in data science. Programming will 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 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, 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:

1. Collaborative R notebook project.
2. Open data visualization project.
3. Predictive analytics of images.
4. Relational (network) data analysis.
5. *Free topic* (multiple data sources will be provided).

Below is a tentative schedule we will follow.

- Week 1 (9/7): 1a+1b
- Week 2 (9/14): 1c
- Week 3 (9/21): 1d+2a
- Week 4 (9/28): 2b+2c
- Week 5 (10/5): 2c
- Week 6 (10/12): 2d+3a
- Week 7 (10/19): 3b+3c
- Week 8 (10/26): 3d
- Week 9 (11/2): 4a+4b
- Week 10 (11/9): 4c
- Week 11 (11/16): 4d+5a
- Week 12 (11/30): 5c
- Week 13 (12/7): 5d

Evaluation

Students' performance will be based on

- Participation (instructors' observation) 10%
- Project products (instructor-reviewed and peer-reviewed, averaged over 5 projects) 90%

Communication

Projects grades are managed in courseworks. We will be using the discussion/announcement tools in courseworks (canvas) for our online class communication and discussion (such as virtual office hours). The system is highly catered to getting you help fast and efficiently from classmates, the TA, and myself. Rather than emailing questions to the teaching staff, I 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 your 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 appriporiately. Plagiarism is one of the most likely forms of cheating in this course. Here are some tips to avoid plagiarism.
- 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.