Syllabus

Course Name: Modern Statistical Methods for Large and Complex Data Sets
Course Number: ST 538
Course Credits: This course combines approximately 90 hours of instruction, online activities, and assignments for 3 credits.
Term: Winter 2018

Prerequisites: ST 517 Applied Data Analytics I or ST 412/512 Methods of Data Analysis II or ST 552 Statistical Methods II

Lectures: Online
Instructor: Sharmodeep Bhattacharyya
e-mail: bhattash@science.oregonstate.edu
Office: 239 Weniger Hall
Office hours: TBD

Course Description

The goal of this class is to provide the students with the tools and experience to analyze big and messy data and work effectively in a data science team. The course will cover tools to handle big data and answer statistical questions based on the data. The class includes three big data analysis projects that the students will work on in groups. The course will focus on proper use many of the modern data analysis techniques, related to regression, classification and clustering, for data coming from variety of application fields. R will be the lingua franca.

Course Objectives

The goal of this class is to provide the students with the tools and confidence to attack big data and work effectively in a data science team. The focus will be on big and messy data. The course will cover tools to physically handle big data and answer fundamental statistical questions related to it. The class includes three big data analysis projects that the students will work on in groups. By the end of the course the students should be able to properly and judiciously use many of the modern data analysis techniques for data coming from variety of application fields and be aware of the possible statistical pitfalls.

R will be the lingua franca, but students can expect to pick up a little of some other languages along the way.
Course Content

The following Table 1 contains the course content and a tentative schedule. If we deviate from it significantly I’ll make announcements to that effect. I’ll also release more detailed schedules as time goes on.

<table>
<thead>
<tr>
<th>Week</th>
<th>Broad Topics</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction and Getting started with big data</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Handling Data and Regression</td>
<td>First Project given</td>
</tr>
<tr>
<td>3</td>
<td>Regression</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Regression</td>
<td>First Project due</td>
</tr>
<tr>
<td>5</td>
<td>Regression and Classification</td>
<td>Second Project given</td>
</tr>
<tr>
<td>6</td>
<td>Classification</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Classification</td>
<td>Second Project due</td>
</tr>
<tr>
<td>8</td>
<td>Clustering</td>
<td>Third Project given</td>
</tr>
<tr>
<td>9</td>
<td>Clustering</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Networks</td>
<td>Third Project due</td>
</tr>
</tbody>
</table>

Table 1: Course Schedule

Learning Outcomes

- Access and handle big data sets from remote locations for analysis.
- Learn how to formulate fundamental statistical questions given a problem and data.
- Learn to use appropriate statistical methods to answer the statistical question based on the data sensitive to the concerns of data size and computational complexity.
- Learn to apply the methods using statistical programming languages (like R) to large and complex data sets.

Evaluation of Student Performances

Assessment

- 75% projects: 3 projects × 25% each. Your score on each project will be group score adjusted by an individual group citizenship score. Your group citizenship score will be based on self and peer evaluations of your ability to work in a team.
- 25% Quiz: 5 Quiz × 5% each. 6 Quiz will be given, marks for lowest dropped.
Group Guidelines

You will be assigned groups for the first project at the start of the second week. At the completion of that project groups will be rearranged (depending on enrollment), unless there is unanimous agreement from all group members to stay together. Group policies and guidelines will be discussed at the start of the second week.

Project Deliverables

For each project, your group will be responsible for:

• an video/slide presentation in which all members will participate,
• a maximum four page written summary,
• a git repository documenting your work and an R file by running which all your results can be reproduced,

For each project, you will also individually submit:

• a self and peer evaluation form

Learning Resources

Course content material will be available on a week-by-week basis. At the start of each project, there will be a short introduction to the data and your task and you will be assigned a group.

There is no fixed textbook for the class. Readings will be assigned for each topic. Materials will be collected from multiple sources.

Course Website

In Canvas.

University and Department Policies

Disability statement

Accommodations for students with disabilities are determined and approved by Disability Access Services (DAS). If you, as a student, believe you are eligible for accommodations but have not obtained approval please contact DAS immediately at 541-737-4098 or at http://ds.oregonstate.edu. DAS notifies students and faculty members of approved academic
accommodations and coordinates implementation of those accommodations. While not required, students and faculty members are encouraged to discuss details of the implementation of individual accommodations.

Expectations for Student Conduct

Student conduct is governed by the university’s policies, as explained in the Office of Student Conduct and Community Standards (http://oregonstate.edu/studentconduct/offenses-0).

Academic integrity

Academic dishonesty is a serious offense and will be addressed following the guidelines set out in the Academic Regulations of OSU (go to http://catalog.oregonstate.edu, click on Registration Information then Academic Regulations, and read AR 15). The Student Conduct Code defines Academic dishonesty as

... an act of deception in which a Student seeks to claim credit for the work or effort of another person, or uses unauthorized materials or fabricated information in any academic work or research, either through the Students own efforts or the efforts of another.

Examples include, but are not limited to, the following:

- verbatim copying of another students homework assignment
- copying off another students exam
- using prohibited materials (e.g., cell phone, cheat sheet) during an exam
- communicating with another student during an exam
- changing answers on an exam after the exam has been graded
- unattributed use of material copied from an article, textbook, or web site
- continuing to write on an exam after the instructor or TA has asked for the exams to be handed in.

You are responsible for knowing what academic dishonesty is, and for avoiding it. Ignorance of these rules does not absolve you from responsibility.