1. Description and Goals

Computers have made it possible, even easy, to collect vast amounts of data from a wide variety of sources. It is not always clear, however, how to use those data and how to extract useful information from data. This problem is faced in a tremendous range of scholarly, government, business, medical, and scientific applications. The purpose of this course is to teach some of the best and most general approaches to get the most out of data through clustering, classification, and regression techniques. Students will gain experience analyzing several kinds of data, including document collections, financial data, scientific data, and natural images.

Finally, it is required that you have regular access to a computer and an Internet connection throughout this course. A laptop is preferable. If you have a laptop, it would be useful to bring the laptop to class, especially for working through hands-on examples.

2. Approach

Lectures will be primarily comprise two components: 1) traditional lecture, where I teach the topics accordingly to the syllabus and schedule, 2) a period of time for interactive “lab” exercises, as well as time for working on projects, discussion, and working on homework.

Halfway through the semester, students will complete an in-class midterm that will test high-level understanding of concepts.

In addition, students will work on a project that emphasizes the concepts covered in the class. The project will be in a group of three to four students focusing on a shared problem. This project will have three stages:

- an initial proposal, explaining what they want to do
- a first-step deliverable (due three quarters into the semester) to make sure students are grappling with the problem, e.g. a transformation of the data, performing a preprocessing step, or creating a baseline for comparison
- a final project report
Each assignment will require a small amount of digging into data using the software package R. This software is free and open source software, is available for all major operating systems, and is installed on lab computers. The assignments will take the form of applying already-implemented algorithms to data to understand the processes involved and answering questions to test understanding.

Halfway through the course, students will complete an in-class midterm that will test high-level understanding of concepts.

2.1 Preliminary Topics (Check webpage for current version)

- The challenge of Big Data, introducing R and Rattle (Chapter 1-2)
- Representing Data (Chapter 3-4)
- Exploring Data (Chapter 5, ggplot book)
- Transforming Data (Chapter 7)
- Probability Crash Course
- Supervised Analysis: Regression
- Supervised Analysis: Classification (Chapter 11, 13-14)
- Unsupervised Analysis: Clustering (Chapter 9-10)
- Unsupervised Analysis: Topic Modeling

2.2 Required Background

Mathematical concepts: You should be able to divide, multiply, add, and subtract numbers. You should know what a logarithm and an exponent are (or be willing to relearn that). We’ll use concepts from algebra like representing concepts with variables. This is the extent of the mathematical knowledge we’ll assume. If you can pass the SAT, you should be okay for this course.

However, we will need to use probability. We’ll review this in class, but only for one week. You’re encouraged to ask questions about it that week, as it will be essential for everything else we discuss in the class.

Computer skills: You should be able to read and save files in spreadsheet (such as Excel or OpenOffice) and in a plain text editor.

Part of the goals of this course is interacting with the statistical platform R. This will require you to assign variables, process data, etc. If you have programming skills and are comfortable with a command line, it will likely make your life much easier. However, you should not have to write complicated programs in R.

3. Grading

Components of the final grade are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Homework | 30%
---|---
Midterm | 25%
Final Project | 40%
Participation | 5%
---|---

It is possible to earn extra credit by going above and beyond the expectations of the assignment.

### 3.1 Assignments

There will be a total of three homework assignments due. Together, they are worth 30% of your final grade. Assignments are designed to help you learn the material, so please use them for that! You are allowed to collaborate with others (as many people as you'd like), but you must turn in your own assignment. For example, you could work together in a group, but each person must write up their solutions individually.

Assignments are due on the class day indicated on the syllabus. Late policy: each person has five free late days to be used, no questions asked, during the course (late days can only be used in increments of one day; if your assignment is three minutes or three hours late, that counts as one late day). When turning in a late assignment, clearly mark at the top that you are using a late day. After you use your late days, late assignments will get half credit. Assignments more than two days late may not be graded.

### 3.2 Midterm

There will be an in-class midterm. The midterm will cover material in the previous lectures and will be open notes.

### 3.3 Final Project

More information will be posted on a separate page for the final project.

### 3.4 Class Participation

Each class is critical to your learning experience, and I expect you to come to class prepared (having read all assigned readings, ready to engage). I also expect active participation, not passive reception of the material. Your energy in contributing to class discussions and hands-on exercises will make this class an enjoyable experience for all of us.
We will also be using the online learning platform Piazza. You can get credit for participation by answering and asking useful questions on that platform. Ideally you should be participating both online and in class, however.

4. Academic Integrity

The University of Maryland, College Park has a nationally recognized Code of Academic Integrity, administered by the Student Honor Council. This Code sets standards for academic integrity at Maryland for all undergraduate and graduate students. As a student you are responsible for upholding these standards for this course. It is very important for you to be aware of the consequences of cheating, fabrication, facilitation, and plagiarism. For more information on the Code of Academic Integrity or the Student Honor Council, please visit http://www.shc.umd.edu.

5. Course Policies

The University has a legal obligation to provide appropriate accommodations for students with disabilities. Please inform the professor of any accommodations needed relative to disabilities at the start of the semester.

Also, University of Maryland policy states that students should not be penalized due to observances of their religious beliefs. Please inform the professor of such instances at the start of the semester so that appropriate steps can be taken.