STATS 503: Applied Multivariate Analysis

Winter 2016

Prof. Liza Levina

Logistics

- **Lecture:** TuTh 1-2:30pm, 1372 East Hall.
- **Contact info:** 459 West Hall, 764-3235, elevina@umich.edu. Please ask all course-related questions on piazza or in class, not by email (unless they are about something concerning only you personally).
- **Office Hours:** Tue 2:30-4pm and Fri 4-5pm, 459 West Hall.
- **Course Webpage:** ctools.umich.edu, plus piazza for discussions (access through Ctools). **You are responsible for obtaining all notes and assignments from Ctools.**
- **GSI:** Jesus Arroyo, jarroyor@umich.edu. Office hours: Mon 2-3:30pm and Wed 5-6:30pm, Science Learning Center, 1720 Chem.

Text

None required. Lecture notes will posted on ctools, along with some additional resources for R code. Two additional books recommended for reading selectively, both downloadable free from Springer:

Undergraduate level: G. James, D. Witten, T. Hastie, R. Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R.* Springer.

PhD level: T. Hastie, R. Tibshirani, J. Friedman. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd edition, Springer.

Topics covered

Principal components analysis and other dimension reduction techniques, classification (discriminant analysis, decision trees, nearest neighbor classifiers, logistic regression, support vector machines, ensemble methods), clustering (agglomerative and partitioning methods, model-based methods), categorical data analysis. The objective is to learn what methods are available for modern multivariate data analysis, how to use them, and when they should and should not be applied.

Prerequisites

Linear algebra, introductory probability and mathematical statistics (at the level of Stats 425/426), and Stats 500 or equivalent. If in doubt, please talk to me right away.

Computing

We will be using the statistical programming language R, which can be downloaded for free from www.r-project.org. A variety of tutorials and manuals for R are available freely online. If you prefer a different statistical package, you are welcome to use it for assignments, but sample code and GSI assistance will only be provided for R, and the exams contain R output.

Grading

Homework 20%, midterm 30% (Tue March 8, in class), final 30% (Wed Apr 27, 1:30-3:30pm), group project 18%, participation 2%.

Additional information

Homework

There will be around 6 homework assignments, roughly biweekly. **Late homework is not accepted.** The worst homework score will be dropped. Typically, a homework will consist of a data analysis assignment, to be done in R, and a small derivation or calculation question that can be done on paper, with the data analysis part accounting for most of the homework grade.

You are allowed and encouraged to *discuss* homework with each other, including in open discussion forums on piazza, but ultimately all questions must be solved and written up independently. Identical homeworks will receive no credit and lead to serious consequences. Any form of plagiarism, including from open online sources, in homework or projects will be taken very seriously and can result in failing the course.

The GSI will be grading the homework and any questions about homework grading should be directed to him first.

Project

The project will be done in small groups of 2-4 students. The goal of the project is to choose a topic that interests you, find or collect a relevant dataset that can be used to answer some interesting and challenging questions about the topic, formulate these questions and answer them using at least some of the methods covered in class, prepare a presentation to be given in class, and write a formal report due at the end of term. When possible, statistics students are encouraged to team up with students from other departments who may want to analyze data from their own research projects or disciplines.

The first step for the project is to form a group and write a one-page project proposal, due in class on Feb 25 (last class before the break). Feedback on the proposals will be provided to make sure the group is on the right track.

Exams

Exams are in class and do not involve a computer. The final is not cumulative. Exams consist mostly of questions about a particular dataset, with parts of the data analysis presented in the exam. They do not test the knowledge of R functions, but do require an understanding of typical R output. They also typically include a "paper-and-pencil" derivation question, which is a small part of the exam. Sample exams consisting of questions from prior years will be provided.

Participation

The participation grade will be assigned based on participation in online discussions on piazza, e.g., in answering other students' questions about R or the course material online, and/or asking and answering questions in class. Attendance will not be taken after the first two classes.

Participating in online discussions is a voluntary activity. Thus you will need to register yourself in piazza if you wish to participate. Follow the link from Ctools for details.