Data Mining (95-791 Z4) Syllabus

Mini 4, Spring 2021

Link to video lectures:

https://heinzcollege.mediasite.com/Mediasite/Catalog/catalogs/95-791a2-data-mining_copyright-2017-carnegie-mellon-university (Links to an external site.)

Note: This is a distance learning course based on Data Mining (95-791) by Professor Alexandra Chouldechova in Fall 2017.

Instructor

Saharsh Agarwal, PhD student, Heinz College (saharshagarwal@cmu.edu)

Office Hours: **Fridays, 5 00 PM - 6 00 PM (**https://cmu.zoom.us/j/495540176 (Links to an external site.))

TA

- Jose Oros (joroscha@andrew.cmu.edu)

Office Hours: **Tuesdays**, **5 00 PM - 6 00 PM**

Prerequisites

1) 95-796 “Statistics for IT Managers” or instructor’s permission based on the student’s knowledge of fundamentals of probability and statistics.
2) R programming experience equivalent to 94-842: Programming in R for Analytics.

Course Description

Data mining (or Data Science in general) is the science of discovering structure and making predictions in large, complex data sets. This course serves as an introduction to data science/data mining methods. Students will learn many commonly used methods for predictive and descriptive analytics tasks. They will also learn to assess the predictive and practical utility of various methods.

Learning Objective

By the end of the class, students will learn to:

- Use R to run many of the commonly used data mining methods
- Understand the advantages and disadvantages of various methods

- Compare the utility of different methods
- Reliably perform model/feature selection
- Use resampling-based approaches to assess model performance and reliability
- Perform analyses of real world data

Textbook

Required textbook

It is available for free at the link below. If you find the textbook to be useful, please show your appreciation by purchasing a copy for personal use.

- Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani (ISLR)

An Introduction to Statistical Learning: with Applications in R Supplemental video lectures/notes is here

http://fs2.american.edu/alberto/www/analytics/ISLRLectures.html (Links to an external site.)

Recommended textbooks

- Kuhn and Johnson, Applied Predictive Modeling (APM)

○ Available for free from the CMU network through SpringerLink:

https://link.springer.com/book/10.1007%2F978-1-4614-6849-3 (Links to an external site.)

○ SpringerLink will print you a black-and-white Softcover version for $24.99

○ Supplementary materials available
here: http://appliedpredictivemodeling.com/ (Links to an external site.)

Helpful R resources

- RStudio (https://www.rstudio.com/ (Links to an external site.))
- R for Data Science (https://r4ds.had.co.nz/ (Links to an external site.)) - written by author for data science R packages tidyverse (https://www.tidyverse.org/ (Links to an external site.))
- swirl: Learn R, in R (https://swirlstats.com/ (Links to an external site.)) - R learning course in R studio environment
- 94-842, R Programming class (http://www.andrew.cmu.edu/user/achoulde/94842/index.htmlLinks to an external site.) -
- Introduction to R Markdown (https://rmarkdown.rstudio.com/articles_intro.html (Links to an external site.))

Assignments and Deadlines

All assignments will be distributed electronically through Canvas. All reports (including homeworks) must be submitted electronically through Canvas. Unless otherwise specified, all assignment will be due on **Wednesdays 11:59 pm Eastern Daylight Time**. Late homeworks will be accepted until 24 hours past the hard deadline, but it will be subject to an automatic 50% grade reduction.

Grading

Grades will be based upon the results of five homework assignments and one analytical project. The analytical projects will be conducted in small groups of students. Each team will analyze specific real-world data. The project will be graded based on the final report.

The final grade for this course will comprise the following:

1. Homework (5 times 15%) 75%
2. Analytical project (in teams of between 1-3 students) 25%

In addition, you will be working on ungraded self-paced lab exercises every week, which are designed to get you familiar with the methods introduced in class.

The weekly quizzes are designed for you to test your understanding of the materials, they are not graded.