94-887 Applied Analytics: the Machine Learning Pipeline (Spring 2021)

Course Information

94-887, Applied Analytics: the Machine Learning Pipeline, will be taught in the Spring semester of 2021. Classes begin 2/1/21. There is no class 4/5/21 (break day). Time: MW 10:10-11:30am. Room: Virtual, TBD

Instructor

Jeremy C. Weiss M.D. Ph.D., Assistant Professor of Health Informatics; jeremyweiss@cmu.edu, OH: TBD.

TA(s): TBD

Please bring your questions to meetings during office hours. Please direct questions to the TA and or instructor by email or on the Canvas discussion board.

Course Description

Machine learning algorithms transform fields with new analytic capabilities, ways of visualizing data, and are key drivers in decision making. But when and how are they useful? Knowing when and how to apply appropriate machine learning techniques requires understanding of the machine learning pipeline, from data to machine learning algorithms to problem domain. This class seeks to teach students how to deal with messy data and provisional questions and turn them into actionable interpretations and insights.

The course will cover discovery, planning, analysis, and interpretation. Discovery involves understanding the data at hand, determining what is and is not answerable, and question generation. Planning involves contrasting the application of the desired machine learning method on ideal clean data with the messy data at hand. Dealing with representation, missing data, and designing appropriate machine learning machinery are all involved in planning. Analysis involves applying the machine learning method, checking model performance and assumptions in a principled and responsible manner. Interpretation involves the transformation of algorithm outputs into meaningful and actionable characterizations of the results. Each part of the pipeline is interconnected and students will learn to anticipate and address limitations through understanding of the pipeline as a whole. Throughout the course we will focus on one vertical, health care, recognizing that the methods developed will generalize to others. We will contrast advanced machine learning methods against simpler methods used in health care analytics, and describe the advantages and limitations of each.

This course will be a mixture of lectures and small group workshop activities culminating in a final project. There will be no final exam.

Course prerequisites

Students should have completed or be concurrently taking Data Mining, Machine Learning for Problem Solving, ML 17-601, ML 17-401 or the equivalent. Experience with R, Python or another programming language is required. We will be using R for this course, and introductory background to R is helpful.

Evaluation Method

Grades will be based on:
- assignments (weekly, first half of semester, x 5), **40%**
- course project (proposal, 10%; code, 5%; paper and deliverables 40%), **55%**
- participation and professionalism, **5%**

Course Objectives

Conceptual understanding and application. You will: - learn and adapt the mathematical formulations of machine learning methods for principled application

- understand the strengths and limitations of existing analytic strategies, including: randomized controlled trials, observational studies, Cox proportional hazards, logistic regression

Practical achievements by the end of the semester include the following. You will: - Perform end-to-end machine learning analysis, including: data exploration, preparation, cleaning, prediction, validation, visualization, and interpretation

- Build working knowledge of the R tidyverse shiny data science pipeline
- Learn to build interactive visualizations of machine learning analyses
- Learn to write a conference-style white paper in Latex

Grading Scale

All grades are tallied and at the end of the course they are scaled to meet the Heinz grading policy.

Cheating and Plagiarism Notice

The project and that is submitted for grading is to be the work of the individual or team alone. Similarly, homework assignments should be your work alone, although you are encouraged to discuss the problems with your classmates. Results that are identical or nearly identical across projects may be regarded as cheating. Penalties for cheating include lowering your grade or failing the course. In extreme cases, the instructors may recommend the termination of your enrollment at CMU.

Additional Course Policies

- Homework Policy: The lowest homework grade will be dropped. If the project grade is lower than any homework grade, all homeworks will be counted and the project grade will count for 20% less of the total grade.
- Late Work Policy: You are expected to turn in all work on time (at the start of class on the due date). Assignments turned in within 48 hours of the deadline will be marked down 20% per day. Additional late assignments will not be accepted.
- Attendance Policy: Attendance is required; please inform the instructor ahead of time if you will be unable to attend. Absence may affect your class participation score.
- Wellness Policy: Take care of yourself and take care of others around you. There are resources to help you both in Heinz and around the University. The Counseling and Psychological Services (CaPS) help line is 412-268-2922. If the situation is life threatening, call the police.

Course Topics

Foundations:
Review of R
Data wrangling: table manipulation, joining, summarization
Model evaluation
Shiny: visualization

Pipeline details:
Prediction versus attribution
Missing data and outliers
Dimensionality reduction
Technical debt

Models:
Generalized linear models
Partition-based methods and ensembling
Neural networks

Course materials

There is not a required textbook. Readings will come from multiple sources and will be provided on Canvas and or in class. Recommended texts include Bishop's Pattern Recognition and Machine Learning (PRML), Murphy's Machine Learning: a Probabilistic Perspective (Murphy), and James' et al's Introduction to Statistical Learning (ISL).

Practicum methods

R, Rstudio, dplyr, purrr/furrr, ggplot, debug, Rmarkdown, keras; git; LaTeX