95-845 Applied Analytics: the Machine Learning Pipeline (Spring 2018)

Course Information

95-845, Applied Analytics: the Machine Learning Pipeline, will be taught in the Spring semester of 2018. Classes begin 1/18/17 and end 5/3/17. Spring break is observed the week of 3/12. Time: TBD. Room: Hamburg Hall TBD

Instructor

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

Office hours: TBD

Faculty assistant: Carole McCoy, HBH 2102

Course Description

Machine learning is a highly valued set of analytics techniques, a confluence of ideas from computer science, statistics, economics, physics, and others. Machine learning is transforming fields with new capabilities, ways of understanding and visualizing data, and is becoming a key driver in decision making. However, knowing when (and how) to apply appropriate machine learning techniques requires understanding of data, machine learning, and the problem domain. This class seeks to teach students how to address the entire machine learning pipeline, starting from messy data and provisional questions and ending with 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 work with real, messy, structured and unstructured data—including databases, text, and images. We will contrast machine learning methods against what is currently used in health care analytics, and describe the advantages and promise of each.

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. Previous exposure to R, Python or another programming language is highly recommended.

Students must request access to MIMIC and eICU data from mimic.physionet.org with myself as research supervisor. We will use this data set throughout the semester and for your end-of-term research project. This requires conducting completing ethics training and signing a data use agreement. Access request must be completed before the end of the 1st week of class so you will have access to data for the first homework assignment and the assignments throughout the semester.

Evaluation Method

Grades will be based on:

- written assignments, 10%
- programming assignments (x5, lowest score dropped), 40%
- course project, 40%
  - proposal, 10%; code, 5%; paper, 25%;
- class participation, 10%

Course Objectives

- learn and adapt the mathematical formulations of machine learning methods for principled application
- perform end-to-end machine learning analysis, including: data exploration, preparation, cleaning, prediction, validation, visualization, and interpretation
- build working knowledge of a data science pipeline: e.g. R tidyverse (we will use this one for class); e.g. python scikit pandas seaborn
- develop machine learning algorithms tailored to data and business or research question
• understand the strengths and limitations of existing analytic strategies, including: randomized controlled trials, observational studies, Cox proportional hazards, logistic regression
• write a conference-style paper in Latex
• use of github for project code

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, completed homework assignments is to 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 including 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 10% 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.

• 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.

Course Topics

Overview of machine learning
logistic regression
Bayesian networks
support vector machines
neural networks
partition-based methods
ensembling
dimensionality reduction;
data wrangling and visualization
prediction versus attribution
missing data
encoding domain expertise
observation versus intervention
algorithmic evaluation
bias-variance tradeoffs
causality
temporal modeling
relational learning
language modeling

Course materials

There is not a required textbook. Readings will come from many sources and will be provided in Canvas and or in class. Useful references include Bishop's Pattern Recognition and Machine Learning, Murphy's Machine Learning: a Probabilistic Perspective, and James' et al's Introduction to Statistical Learning.

Practicum methods

R, Rstudio, dplyr, purrr, ggplot, debug, Rmarkdown, Tensorflow; git; LaTeX