90-866 Large Scale Data Analysis for Public Policy
Carnegie Mellon University, Heinz College, Spring 2017

Syllabus

Instructor    Dr. Artur Dubrawski
The Robotics Institute, School of Computer Science, Heinz College, CMU
Office: Newell-Simon Hall 3121, Phone: 412-268-6233, E-mail: awd@cs.cmu.edu
Office hours: Mondays, 4:30 – 6:00pm (appointments recommended)

Lectures    Tuesdays and Thursdays, 10:30 - 11:50am, Hamburg Hall 1006

Recitations   Fridays 12:00 - 1:20pm, Hamburg Hall 1202

TAs    Jieshi Chen, Benedikt Boecking, Yingjie Zhang, Zhe Zhang.
Office, Phone, E-mail, Office hours: please refer to data on the course blackboard site.

Prerequisites    None.

Course Description
The past decade has seen the increasing availability of very large scale data sets, arising from the rapid growth of transformative technologies such as the Internet and cellular telephones, along with the development of new and powerful computational methods to analyze such datasets. Such methods, developed in the closely related fields of machine learning, data mining, and artificial intelligence, provide a powerful set of tools for intelligent problem-solving and data-driven policy analysis. These methods have the potential to dramatically improve the public welfare by guiding policy decisions and interventions, and their incorporation into intelligent information systems will improve public services in domains ranging from medicine and public health to law enforcement and security.

This course will provide a basic introduction to large scale data analysis methods, focusing on three main problem paradigms (prediction, modeling, and detection). Students will learn how to translate policy questions into these paradigms, choose and apply the appropriate artificial intelligence and machine learning tools, and correctly interpret, evaluate, and apply the results for policy analysis and decision making. We will emphasize tools that can “scale up” to real-world policy problems involving reasoning in complex and uncertain environments, discovering new and useful patterns, and drawing inferences from large amounts of structured, high-dimensional, and multivariate data. No previous knowledge of artificial intelligence or machine learning is required.

The instructor is a scientist and a practitioner. He has been involved in research towards machine intelligence and its applications for over two and half decades. He has been a technical lead and an executive in the new technology industry. Currently Prof. Dubrawski is faculty at the CMU Robotics Institute where he directs the Auton Lab and leads multiple large data analysis projects in support of industry, government, and non-governmental organizations.

Note: This course has been originally designed and taught for several years by Professor Daniel Neill who is currently on a leave from Heinz College. This is the first time this course will be taught by Prof. Dubrawski. We will rely primarily on the original course materials and scope, but certain topics may take longer to cover than others. We will make attempts to identify and follow the interest of students in allocating lecture time to particular topics.
Course Objectives

Upon completion of this course, the student will be able to:

- Identify large scale data analysis methods, focusing on three main problem paradigms: prediction, modeling, and detection.
- Translate policy questions into paradigms.
- Choose and apply the appropriate artificial intelligence and machine learning tools.
- Interpret, evaluate, and apply the results for policy analysis and decision making.

These objectives will be assessed both through the final exam and the students' course projects.

Course Registration

Pass-fail registration is not allowed in this course. Audit requests will be denied except for extraordinary circumstances.

Grading

Final grades will be determined by scoring performance of each student at:

- Active and constructive participation in classroom activities (10%),
- Written, in-class final examination (30%),
- Analytic projects (60%).

Class projects, conducted in teams of 3 will have specific schedule of deliverables including project proposals, milestone reports and progress presentations, as well as final reports and final presentation videos. Project progress presentations will be conducted during weekly recitations starting from the third week of the course. Their purpose is to showcase progress made by each team, enable constructive in-class discussion, and receive guidance regarding the next steps. All teams will need to be ready to present each week, however only an unannounced a priori subset of teams will be called to actually present during each particular recitation session. Details of project deliverables, due dates, and grading will be provided in a separate document posted under Assignments on the course Blackboard site.

Class participation is considered to be a valuable criterion in grading performance. Empirically, the students who take an active part in lectures and recitations, and those who frequent office hours tend to grasp the taught concepts more effectively than those who resort to a passive approach. The final grade for this course will reflect the sum of the above specified component grades.

Course Materials

Lecture slides and supplemental readings will be available in the Course Content section of Blackboard. R and Weka software, recommended for use in this course, are freely available and can be downloaded from the Internet.

Hard copies of the lecture notes will not be distributed. The notes will be available for download from the course blackboard typically 12 hours before each lecture. The students are encouraged to bring their printed copies of the notes to class.

Occasionally, we will hand out short practice exercises to reinforce understanding of the course material. You will not need to turn these in but working on them should help you study for the final exam. We will occasionally review some of these exercises in the recitations sessions.
**Required Name Tags**

The students are expected to bring their letter-sheet-sized name tags to each of the meetings starting from the second lecture. Class attendance will be recorded at the beginning of each lecture based on the visual account of the displayed tags. Attendance in the first lecture will be taken by each participant signing the attendance sheet.

**Academic Integrity and Classroom Habits**

The students are expected to strictly follow Carnegie Mellon University rules of academic integrity in this course. This means in particular that examinations and quizzes are to be the work of the individual student using only permitted material and without any cooperation of other students or third parties. It also means that usage of work by others is only permitted in the form of quotations and any such quotation must be distinctively marked to enable identification of the student’s own work and own ideas. All external sources used must be properly cited, including author name(s), publication title, year of publication, and a complete reference needed for retrieval. Regarding the group work (projects and blogs), the work should be the work of only the group members. In all their work students should not in any way rely on solutions to problems distributed in prior years or on the work of prior students or other current students. Violations will be penalized to the full extent mandated by the CMU policies. There will be no exceptions.

Usage of electronic equipment such as portable computers during classes is very strongly discouraged, except for meetings specifically designated for hands-on software demonstrations and exercises.

No student may record or tape any classroom activity without the express written consent of the instructor. If a student believes that he/she is disabled and needs to record or tape classroom activities, he/she should contact the CMU Office of Disability Resources to request an appropriate accommodation.

**Lecture Topics**

Note: Many of the topics listed below will take longer than one class meeting, and many of the class meetings will begin with a thorough in-class discussion of the due readings and cases. Preparation for these discussions and active, contributory participation in them will be the key factor used to determine the in-class activity grades.

Module I: Prediction

- Introduction to Machine Learning and Artificial Intelligence for Large Scale Data Analysis
  - Course overview
  - Relevance of ML for policy
  - Common ML paradigms
  - Software tools for ML

- Prediction, Rule-Based Learning
  - The prediction problem (classification and regression)
  - Decision trees for classification and regression

- Instance-Based Learning
  - K-nearest neighbors for classification
  - Kernel regression
  - Cross-validation

- Model-based learning
Bayesian classification
The naive Bayes’ assumption

Module II: Modeling

- Representation and Search
  Goal-directed search: priority search and A*
  State-space search: hill-climbing and simulated annealing

- Clustering for Modeling Groups
  Hierarchical clustering
  K-means clustering
  Leader clustering

- Bayesian Networks for Modeling Probabilities
  Building Bayes Nets
  Interpreting Bayes Nets
  Inference with Bayes Nets
  Learning Bayes Net structure

Module III: Detection

- Anomaly Detection
  Distance-based anomaly detection
  Model-based anomaly detection
  Detecting anomalies using Bayesian networks

- Pattern Detection
  Detecting patterns of anomalies
  Applications

One or two class meetings will be devoted to lectures by invited speakers selected among seasoned practitioners of applied analytics. Exact dates, scopes and topics of these lectures will be provided later.

Final exam

Friday May 12th 8:30-10am, HBH 1002 (note the special time and place).