90-866, Large Scale Data Analysis for Policy (Fall 2014)

Course Information

90-866, Large Scale Data Analysis for Policy, is a 6 unit course, taught in Mini-2. Classes begin Tuesday October 21st and end Thursday December 4th, with final exam on December 12th.

Class Schedule

Tuesdays and Thursdays, 10:30-11:50am, Hamburg Hall 236

Instructor

Daniel Neill (neill@cs.cmu.edu)
Office hours: Tuesdays 3:00-3:45pm and Thursdays, noon-12:45pm (or by appointment)
Hamburg Hall 2105B, x8-3885

Faculty assistant: Natalia Pascal (npascal@andrew.cmu.edu), Hamburg Hall 2112

Teaching Assistants

Sriram Somanchi (somanchi@cmu.edu)
Office hours: Wednesdays 4:00-5:00pm, Hamburg Hall 3030

Abhinav Maurya (ahmaurya@cmu.edu)
Office hours: Mondays 6:00-7:00pm, Hamburg Hall 3030

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.

**Course Materials**

Lecture slides and supplemental readings are available in the Course Content section of Blackboard. Weka data mining software is freely available and can be downloaded from [this site](#).

**Evaluation Method**

Grades will be based on the following:

- Class participation: 5%
- Project plan: 10%
- Project progress report: 10%
- Project presentation: 10%
- Final project report: 25%
- Final exam: 40%

The projects will be done in groups of three students and will require the application of machine learning methods to real-world policy data. We plan to give each team the flexibility to define their own project, enabling them to focus on policy questions which are most relevant to their own specific interests. However, each project should consist of the following components:

- Define a relevant policy question to be answered using a dataset of your choice. We have provided several example datasets, as well as other suggested sources of publicly available data.
- Frame the problem in terms of one of the ML paradigms discussed in this class. Discuss this problem framework in detail, justify your choice of a problem framework, and report on methods that have been used to solve the problem in past work.
- Choose an appropriate solution method for the problem. Describe the solution method in detail, compare to relate methods, and defend your choice of method.
- Find, or develop, an appropriate software implementation of this method. We encourage you to use pre-existing toolkits such as Weka, though it would also be acceptable to write your own functions in Matlab, R, etc., if desired.
- Evaluate your method, discussing both quantitative performance results (e.g., cross-validation error) and qualitative consideration of the usefulness of the resulting models, explanations, etc., for the given domain.
- Consider extensions and variations of the original method, or alternative methods, and examine/compare their effects on performance.
Project teams will be assigned by the course instructor, but assignments will be based on student preferences. Please e-mail Daniel and Sriram by Friday 10/24 if you either a) have a team of three, or b) have a particular topic you'd like to work on. Typically, all team members will receive the same grade, but we may make exceptions for unevenly distributed workloads. Final project reports should describe the contributions of each team member to the project. Please see the Assignments section for additional details.

Occasionally, we will hand out short practice exercises to reinforce understanding of the course material. You do not need to turn these in. We will post answers with explanations on Blackboard, and these should help you study for the final exam (Friday December 12th, 1:00-2:30pm).

Due Dates

Reminder: all assignments are due at the beginning of your class period except where otherwise noted. Assignments turned in more than five (5) minutes after class starts may be counted as "late" and treated according to the late work policy below.

Project plan due Tuesday 11/4
Project progress report due Tuesday 11/18
Project presentations Tuesday 12/2 and Thursday 12/4
Final project report due Thursday 12/4, 11:59pm.

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.

Grading Scale

Final grades will be curved, and I do not pre-specify the grading scale. Below is how the cutoffs came out for Spring 2014, but this year's cutoffs will depend on the overall distribution of scores.

A+ 95-100%  B+ 82-84%  C+ 73-75%
A 91-94%  B 79-81%  C 70-72%
A- 85-90%  B- 76-78%  C- 67-69%

Cheating and Plagiarism Notice
Projects will be done in groups of three students; we encourage discussion among teams about the projects, but any work that is submitted for grading must be the work of your team alone. Most importantly, your answers on the final exam must reflect your work alone. Sanctions for cheating include lowering your grade including failing the course. In egregious instances, the instructors may recommend the termination of your enrollment at CMU.

**Additional Course Policies**

**Late Work Policy:** You are expected to turn in all work on time (at the start of class on the due date). Because we understand that exceptional circumstances may arise, each team will be permitted to turn in one of their three project reports up to 48 hours late with no penalty. Any other late assignments will not be accepted.

**Re-grade Policy:** Project grading is inherently subjective, and thus we do not generally consider re-grade requests. We will make exceptions to this rule in cases where we have made an error in grading; in this case, requests must be submitted *in writing* to the course instructor, and all resulting decisions are final.

**Course Outline**

**Module I: Prediction**

*(T 10/21) Lecture 1: 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

*(H 10/23) Lecture 2: Prediction, Rule-Based Learning*
The prediction problem (classification and regression)  
Decision trees for classification and regression

*(T 10/28) Lecture 3: Instance-Based Learning*
K-nearest neighbors for classification  
Kernel regression  
Cross-validation

*(H 10/30) Lecture 4: Model-based learning*
Bayesian classification  
The naive Bayes assumption

**Module II: Data Modeling**
(T 11/4) Lecture 5: Representation and Search
Goal-directed search: priority search and A*
State-space search: hill-climbing and simulated annealing
Deliverable: project plans due at beginning of class (10:30am).

(H 11/6) Lecture 6: Clustering for Modeling Groups
Hierarchical clustering
K-means clustering
Leader clustering

(T 11/11) Lecture 7: Bayesian Networks for Modeling Probabilities
Building Bayes Nets
Interpreting Bayes Nets

(H 11/13) Lecture 8: More Bayesian Networks
Inference with Bayes Nets
Learning Bayes Net structure

(T 11/18) Guest lecture by Sriram on Causality
Deliverable: project progress reports due at beginning of class (10:30am).

Module III: Detection

(H 11/20) Lecture 9: Anomaly Detection
Distance-based anomaly detection
Model-based anomaly detection
Detecting anomalies using Bayesian networks

(T 11/25) Lecture 10: Pattern Detection
Detecting patterns of anomalies
Spatial cluster detection
Applications to disease surveillance

(H 11/27 is Thanksgiving: no classes, no office hours!)

(T 12/2) Project Presentations

(H 12/4) Project Presentations, continued
Deliverable: project final reports due 11:59pm.

FINAL EXAM- FRIDAY DECEMBER 12TH, 1:00-2:30pm, ROOMS TBD (currently booked for HBH 1001 and HBH 1511).