New York University
Robert F. Wagner Graduate School of Public Service

PADM-GP 4147-001
Large Scale Data Analysis I
Spring 2017 (7W1)

Course Information

Large Scale Data Analysis I is a 1.5 unit course, taught in Spring 2017 (first seven weeks). Classes begin Tuesday January 24th and end Tuesday March 7th.

Class Schedule

Tuesdays, 6:45-8:25pm, Bobst Library # LL149.

Instructor

Professor Daniel Neill (daniel.neill@nyu.edu)
Office hours and location: Tuesdays and Fridays, 4:30-5:15pm, Wagner (Puck Building) #3039.

This is a shared office and I am not present most days. Meetings at other times are by appointment only and will typically be held in my NYU CUSP office in Brooklyn (1 MetroTech Center #1926).

Course Description and Objectives

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.

The LSDA course series will provide a basic introduction to large scale data analysis methods, focusing on four main problem paradigms (prediction, clustering, modeling, and detection). The first course (LSDA I) will focus on prediction (both classification and regression) and clustering (identifying underlying group structure in data), while the second course (LSDA II) will focus on probabilistic modeling using Bayesian networks and on anomaly and pattern detection. LSDA I is a prerequisite for LSDA II, as a number of concepts from classification and clustering will be used in the Bayesian networks and anomaly detection modules, and students are expected to understand these without the need for extensive review.

In both LSDA I and LSDA II, students will learn how to translate policy questions into these paradigms, choose and apply the appropriate machine learning and data mining 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 machine learning or data mining is required, and no knowledge of computer programming is required. We will be using Weka, a freely available and easy-to-use machine learning and data mining toolkit, to analyze data in this course.

Relationship to Other Courses

As noted above, LSDA I is a prerequisite for LSDA II. Both LSDA I and LSDA II assume knowledge of basic probability and statistics, and thus CORE-GP.1011 (Statistical Methods for Public, Nonprofit, and Health Management) is a prerequisite for both courses. Other quantitative methods courses taught at the Wagner School, such as PADM-GP.2902 (Multiple Regression), PADM-GP.2875 (Estimating Impacts), and PADM-GP.2172 (Advanced Empirical Methods) emphasize linear models for regression, with a focus on causal inference (particularly, estimation of causal effects). The LSDA course series focuses on a substantially different set of data analysis methods from machine learning and data mining rather than traditional statistics and econometrics, with an emphasis on predictive rather than causal inference, nonlinear and nonparametric models, and problem paradigms other than regression (including classification, clustering, modeling, and detection). We anticipate that many students in this class will already be well-trained in econometric methods (though this is not a prerequisite) and interested in broadening their set of methodological tools and corresponding applications in the public sector (for example, early warning systems for public health, predictive policing approaches, or data-driven discovery of best practices in health care). We encourage students coming from the economic/econometric perspective to read Einav and Levin's "The data revolution and economic analysis" (http://www.nber.org/papers/w19035.pdf), or Hal Varian's "Big data: new tricks for econometrics" (http://people.ischool.berkeley.edu/~hal/Papers/2013/ml.pdf) for additional motivation.

Course Materials

Lecture slides and supplemental readings are available in NYU Classes. Weka data mining software is freely available and can be downloaded from http://www.cs.waikato.ac.nz/ml/weka.

Evaluation Method

Grades will be based on the following:

Class participation: 5%
Problem set 1 (Classification): 20%
Problem set 2 (Clustering): 20%
Mini-project checkpoint report: 15%

Mini-project final report: 40%

The mini-projects will be done in groups of 2-3 students and will require the application of machine learning methods to real-world policy data. Each project team will be given a different real-world dataset to analyze (teams and datasets will be assigned based on student preferences) and a sample set of open-ended questions to answer. We plan to give each team some flexibility to define their own project, enabling them to focus on policy questions which are most relevant to their own specific interests.

However, each mini-project should consist of the following components:

Define a relevant policy question to be answered using a dataset of your choice.

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 related 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.

Cheating and Plagiarism Notice

Mini-projects will be done in teams of 2-3 students; we encourage discussion among teams, but any work that is submitted for grading must be the work of your team alone. Problem sets must be done individually, i.e., any work that is submitted for grading must be the work of that student alone. Sanctions for cheating include lowering your grade including failing the course. In egregious instances, the instructor may recommend the termination of your enrollment at NYU.

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 student will be permitted to turn in one assignment up to 48 hours late with no penalty. Any other late assignments will not be accepted.

NOTE: Assignments turned in more than **five (5)** minutes after class starts will be counted as "late" and treated according to the Late Work Policy above.

Course Outline

Class 1 (1/24): Introduction to Large Scale Data Analysis for Public Service
Course overview
Relevance of ML for policy and the public sector
Common ML paradigms- prediction, clustering, modeling, detection
Software tools for ML
Weka intro

Class 2 (1/31): Interpretable Classification (and Regression) using Decision Trees
The prediction problem (classification and regression)
Interpretable vs. Black-Box prediction methods
Interpretable Prediction: Rule-based, instance-based, and model-based
Rule-based Learning: Decision trees for classification and regression
Decision trees using Weka (examples on real-world policy data)

Class 3 (2/7): Interpretable Classification (and Regression) using Instance-Based Learning
K-nearest neighbors for classification
Kernel regression
Cross-validation for unbiased evaluation of methods
K-nearest neighbors using Weka (examples on real-world policy data)

Class 4 (2/14): Black-Box Classification Methods
***** Mini-project Checkpoint Report Due *****
Discussion of accuracy vs. interpretability tradeoff
Ensemble methods: from trees to forests (bagging, boosting, random forests)
Support vector machines: from linear to non-linear decision boundaries

Class 5 (2/21): Clustering Part 1
***** Classification Problem Set Due *****
Brief digression: representation and search
Introduction to clustering for modeling group structure: what and why
Hierarchical clustering: bottom-up and top-down
K-means clustering
Clustering in Weka

Class 6 (2/28): Clustering Part 2
***** Mini-project Final Report Due: deadline extended to Friday March 3rd at 11:55pm *****
K-means clustering, continued
EM clustering
Leader clustering for massive streaming data

Class 7 (3/7): Wrap-Up Discussion
***** Clustering Problem Set Due *****

1st half of class: Each student group will share the findings of their mini-project with the class. This will be short (4-5 minutes per group), informal (no slides), and ungraded. You can just summarize: 1) the problem your project addressed, 2) the dataset you used, 3) which techniques you used, and 4) main findings (any results that were particularly informative or surprising, or helped you learn something new about the problem domain?)

2nd half of class: Wrap-up discussion. Revisiting accuracy vs. interpretability, new directions in supervised learning (explanation and visualization), scaling up further, where to go next?