Proposal for a New Upper-level Undergraduate Course: Introduction to Data Science

Tina Eliassi-Rad[*]
eliassi@cs.rutgers.edu

November 11, 2014

Course description and goals. This course will introduce juniors and seniors (and possibly masters level students) to data science. It will cover topics needed to solve "data" problems, including preparation (collection and integration), characterization and presentation (information visualization), analysis (machine learning and data mining), and products (applications). Similar to other DCS undergraduate courses, it will be for 4 credits with a recitation section.

I taught this class, for the first time, as a special topics course (CS 444) in Fall 2013, and had 27 students. The website from my Fall 2013 course is available at http://eliassi.org/datasci13.html. I am scheduled to teach it again in Spring 2015.

We plan to offer this course every year in the spring semester.[1] Professors Casimir Kulikowski and Vladimir Pavlovic have shown interest in teaching this course.

Due to the popularity of this topic, many top universities (such UC Berkeley and Columbia) have already added a data-science (or similar) course to their undergraduate curriculum. For a non-exhaustive list of such courses, visit http://datascienc.es/resources/.

Prerequisites. 01:198:206 (Introduction to Discrete Structures II) and 01:198:336 (Principles of Information and Data Management).

Syllabus.

- Data visualization (1 week)
- Data wrangling and pre-processing (1 week)
- Map-reduce and the new software stack (1 week)
- Data mining (3 weeks)
    - Finding similar items
    - Mining data streams

[*] Thanks to Vinod Ganapathy for providing the template for this document.

[1] Professor Amélie Marian requested that we move the course from fall to spring so that it does not conflict with the CS 437 course (Database Systems Implementation).

    - Frequent itemsets
    - Link analysis
    - Mining graph data

- Machine learning (6 weeks)

    - Supervised learning: k nearest neighbor, decision trees, naïve Bayes, regression, ensemble methods, perceptron, support vector machines
    - Unsupervised learning: k-means, spectral clustering, hierarchical clustering, dimensionality reduction
    - Evaluation techniques

- Applications (2 weeks)

    - Advertising on the Web
    - Recommendation systems
    - Anomaly detection

Textbook.

- Jure Leskovec, Anand Rajaraman, and Jeff Ullman, *Mining of Massive Datasets*, 2nd Edition, Cambridge University Press, November 2014, ISBN 978-1107077232.

This book is available free-of-charge at `http://www.mmds.org/`.

Grading.

- Mid-term (15%)

- Final exam (25%)

- Three homework assignments (each worth 10% of the grade). Assignments will include both conceptual and programming questions.

- Class project (30%). The project will include solving a data-science problem cover-to-cover: from scraping and cleaning the data, to answering the data-science question with the appropriate learning/mining algorithm, to visualizing and interpreting the results. Students will present their project in-class and submit a final report.[2]

Programming Languages and Tools. For the programming assignments, the students will use Python and its libraries (such as NumPy, SciPy, scikit-learn, pandas, and matplotlib). For the class project, they are allowed to use the language and tool of their choice (such as Java and Weka; or Matlab; or R).

Course number. Based on the Undergraduate Curriculum Committee's suggestion, we are requesting 439 as the number for this course. This numbering matches our existing data courses (namely, 336 and 437).

[2] Students can do the class project individually or in teams of two.