Big Data and Security: Background, Theory, Practice

INTA 8803-JB
Spring 2014
Jeffrey Borowitz
jborowitz@gmail.com

Description

Explores the foundations of big data, including it's foundations and its relationships to a variety of applied fields, with a focus on the fields of cybersecurity, intelligence gathering, and digital forensics. Engages with the social implications of increased knowledge, surveillance, and behavioral prediction made possible by big data, and the ethical tradeoffs faced. While the course includes an analytics project, no prior technical experience is required.

Course Outline

- What is Big Data? (1 week)
- Computers: (2 week) Background on how computers work, how they perform common tasks, which tasks are hard for them to perform vs. easy. Understand where changes in technology are possible and how they will help to solve particular problems. Also understand the role of computers in our lives, and how this generates a trail. Discuss forensics as a field in relation to what computers do.
    - How does a computer work? What does this mean for big data?
        - Hardware
        - Software
    - Our digital remnants as evidence: digital forensics
- Statistical basis for big data: (2 weeks)
    - Two approaches to statistics
        - Pick the best parameters
        - Bayesian statistics
    - Introduction to statistics commonly used for big data
        - Regression
        - Trees
        - Hierarchical modeling
        - Neural Nets
    - Natural Language processing
    - Experiments
- Big Data criticism - a critical look at benefits and limits of big data analysis: 3 weeks
    - Models and Errors - when can statistics or predictions based on big data be wrong? And what happens then?
    - Privacy - why does it creep us out when companies market to us?
    - The government and big data: Big Brother
- Big Data applications: (6 weeks) How can data analysis solve key problems in a variety of areas? Discuss what sources of data are important in each area, what key questions

exist which data can address, and what success has already been realized. There is a particular focus on how big data can support security.

  - Fraud detection
  - Intelligence
  - Anti-Money Laundering
  - Anti-Terrorism
  - Cyber security
  - Other Areas:
    - Medicine
    - Marketing
    - Academic Fields
    - Voice recognition
    - Systems of sensors

Readings

- Various readings will be assigned at least one week in advance throughout the semester. Most readings will come from articles. Two books will be used in the course:
  - Mayer-Schonberger and Cukier 2013 *Big Data: A Revolution that Will Transform How We Live, Work, and Think.*
  - Silver, Nate. 2012. *The Signal and the Noise: Why So Many Predictions Fail-But Some Don't.*

Assessment

Participation and In-Class Labs (20% of grade)

Because big data analysis can be computationally intensive, we will spend class time together working to make sure that everyone can run Python and the Natural Language Tool Kit on the Enron data. This project (and this class) does not require any prior programming ability or knowledge. All required software will be available for download, and support will be available.

Activity: Fraud prediction and it's consequences: the case of Enron (70% of grade)

Enron was a major player in the energy industry in the 1990. In the early 2000s, it became clear that enron was involved in widespread accounting fraud. The SEC investigated, and Enron was eventually forced to file for bankruptcy. Enron's $60 billion in assets were liquidated. As a part of this fallout, the entire email history of the top 150 executives at the company was made public. The Enron email data set is one of the canonical data sets in machine learning because it represents an organic data set.

In this course, you will attempt to detect the rise of this fraud through email using machine learning, natural language processing, and other tools. This is a difficult and substantial task, which will make up your entire graded assignment for the course. There will be three related portions of this assignment which will be turned in during the course of the semester.

- A proposed strategy (20% of grade): Due 3/3/2013
    - You will first submit a proposed strategy. This document will describe your approach to solving this “big data” challenge. Your proposal should be 5-10 pages long and should address the following points:
    - What is the source of wrongdoing that you’re looking for in the emails? What law is broken or what behavior is unethical?
    - What patterns will you look for in email data to detect this potential wrongdoing?
    - What machine learning or other methods do you feel might be most appropriate for this task?
    - What extra information will you collect from news reports, court documents, or other sources will you need to use to inform and aid in your work? Be sure to describe why this information is necessary and how it will augment and inform your analysis.
    - What difficulties do you forsee in completing the analysis, either technically or theoretically? Use examples from the class on the strengths and weaknesses of particular approaches to big data analysis.
- Your own “Big Data” analysis (20% of grade): Due 4/7/2013 (you may work in groups of up to 3)
    - Your goal is to:
        - determine which emails in the Enron corpus, if any, are related to accounting fraud.
        - By analyzing the set of emails which you identify to be fraudulent, determine when fraud began at Enron.
        - Choose a subset of your fraudulent emails at random and read them yourself. What fraction of your predicted fraud related emails were in fact innocuous?
    - From your assignment, you should submit:
        - Any code you used to analyze the data
        - A writeup of your strategy and a description of how your code or other supporting documents implement your strategy
        - An assessment of how successful your method was
- A critical analysis of your big data analysis (30% of grade): Due 4/21/2013
    - Each student should write a 10 page paper which imagines a proposal that your fraud detection mechanism (or perhaps a more thoroughly designed mechanism) would be used to monitor email for fraud detection at other comapnies.
        - What ethical issues arise here? Would you
        - Would you want a system like this monitoring your personal email? What about business email? What role do you see for systems like this in our society?
        - What practical issues might arise in applying a model of this sort? In particular, how well might your model generalize to other contexts? What

factors will determine how well your model will generalize?

- Participation (30% of grade)
    - The class will have a number of components which will require interactive participation
        - Performing assigned readings and actively contributing to discussion in class
        - Completing in-class or out of class assigned programming