Stats 170A/B, Project in Data Science

Chen Li\textsuperscript{1} and Vladimir Minin\textsuperscript{2}

\textsuperscript{1}Department of Computer Science
\textsuperscript{2}Department of Statistics
Bren School of Information and Computer Sciences
University of California, Irvine

January 6, 2020
Plan for today

- Introductions
- Class organization and schedule
- Discussion about projects
- Python software
- Data science in the real-world
Introductions

▶ Instructors

– Professor Chen Li
  Research: database systems, data management, … → “Big Data”
  Industry: Start-up experience

– Professor Vladimir Minin
  Research: statistics, stochastic modeling
  Applications: consults/works with Fred Hutchinson Cancer Research Center

▶ Students

– Introduce yourself

– What do you hope to get out of the project class?

– Programming skills you have vs. you want to improve
Philosophy behind this class

▶ Provide an experience of how data science works in the real-world
  
  – Defining a problem
  – Identifying, understanding, exploring relevant data
  – Extracting, cleaning, management of data
  – Exploration and analysis of data
  – Building models from data (e.g., via machine learning)
  – Evaluating models: how well do they predict
  – Communicating your results to others
  
▶ Tie together ideas from different courses you have taken and give you experience in applying these ideas to real-world data
  
  – Databases, software, algorithms, machine learning, statistics
Organizational Items

- **Class Website**
  - Class Canvas page: [https://canvas.eee.uci.edu/courses/22259](https://canvas.eee.uci.edu/courses/22259)
  - This is where to find assignments, links to resources such as software, data sets, project guidelines, etc

- **2-quarter class (Winter and Spring)**
  - Think of it as one 20-week class
  - Will propose and define your project this quarter and work on it in Spring

- **No midterms or final exam**
  - But there will be regular reporting and some presentations
  - Also, individual homework assignments during the first six weeks

- **Textbook and Reading Materials**
  - No official textbook
  - Links to relevant texts (available online via UCI library) on the class wiki page
Textbooks


All of these titles are available for free online via the UCI Library’s subscription to Safari Books Online (http://proquest.safaribooksonline.com/).
Course outline

- **Winter**: Weeks 1 to 6: Lectures and Assignments
  - Review general principles of data science
  - Weeks 1 to 3: databases, data extraction, data cleaning
  - Weeks 4 to 6: text analysis, data exploration, machine learning and statistics
  - Combination of lectures, assignments, and background reading

- **Winter**: Weeks 7 to 10: Project Proposals
  - Project proposals from student teams
  - Feedback from instructors, refine proposal, oral presentation at end of quarter

- **Spring**: Work on Projects
  - Build and use a prototype system/pipeline
  - Develop ideas, implement algorithms, make use of libraries and packages
  - Conduct experiments with real data sets
  - Test and evaluate your system in a systematic manner
  - Communicate your results (presentations and reports)
Grading

- Only one grade, assigned at end of Spring quarter

- Winter quarter (50% of total grade)
  - 50% project proposal
  - 40% homeworks
  - 10% class participation

- Spring quarter (50% of total grade)
  - Distributed across project progress reports, final report, class presentations and participation

- Participation = attending class and participating in class discussion

- No grading of late homeworks
Academic integrity

- Students will be expected to adhere to the UCI and ICS Academic Honesty policies (see https://aisc.uci.edu/policies/academic-integrity/index.php and https://www.ics.uci.edu/ugrad/policies/index.php%23academic_honesty to read their details).

- Any student found to somehow be involved in cheating or aiding others in doing so will be academically prosecuted to the maximum extent possible: that means that you could fail this course in its entirety. (Ask around - it’s happened.) Just say no to cheating!

- This information and associated links are also posted on the class Website
Questions outside class? use Canvas

- Use Canvas discussion board for questions (outside of class time) related to the class
  - Assignments, lectures, projects, data sets, ideas, etc

- Instructors will try to quickly answer questions
  - Students should also feel free to also answer questions
  - If you wish you can use “private mode” to ask questions that only the Professor will see
  - (This way you won’t get lost in our daily faculty e-mail overload)
Class projects

» 2-person teams

– Note that Assignments in weeks 1 to 6 are not team-based; these will be worked on and submitted individually
– For 2-person teams we expect twice as much output and contributions of each individual to be clearly identified in reports

» Each team will propose its own project

– Suggestions for multiple different projects will be provided
– Extensive use of libraries (in addition to writing some of your own code)

» Projects will be graded based on

– Initial proposal
– Weekly updates
– Intermediate and final reports
– In-class presentation

We will discuss all of this in more detail in future lectures
Project expectations

- **Required components**
  - Automatically extract a large-scale data set from Twitter
  - Combine Twitter data with at least one other large-scale data set
  - Make use of data management, cleaning, exploration, visualization tools
  - Develop a prediction/forecasting system using the data sets

- **Software development**
  - You will make use of existing libraries and tools (e.g., PostgreSQL and Python)
  - You are also expected to implement some components of the pipeline yourself

- **Evaluation**
  - You will need to systematically evaluate your prototype
  - E.g., runtime, predictive accuracy, accuracy as a function of data set size, etc.

- **Reporting**
  - You will be required to generate reports, graphs, Jupyter notebooks, etc.
Sources of large data sets that could be used for projects

Text from 4 million Wikipedia articles

Twitter data: large streams of tweets via Twitter API

Dataset Search

Try boston education data or weather site:noaa.gov

Learn more about including your datasets in Dataset Search.

https://toolbox.google.com/datasetsearch
Example of a class project

- Data sources
  - Twitter API: tweets mentioning certain keywords, over time, with metadata
  - Census or government maps of population by US county
  - Weather data over time for US locations
  - Historical data on consumer confidence over time
  - CDC FluView (Weekly U.S. Influenza Surveillance)

- Create query tool that can compute relative popularity of a keyword
  - over time (time-series plot)
  - Over space (tweets are mapped to location)

- Extension 1
  - Predict popularity of a keyword by week, given historical data

- Extension 2
  - Investigate correlation of keywords with weather data and/or Influenza activity (in time and space)

- Extension 3
  - How well can consumer confidence or Influenza activity be predicted from tweet sentiment over time?
Average Happiness for Twitter

https://hedonometer.org/index.html
Another possible class project

- Define a set of entities of interest
  - E.g., movie stars listed in the IMDB data set (see Homework 1)
  - E.g., sports stars, musicians, etc, from Wikipedia
  - Weather data over time for US locations
  - E.g., products and brands (e.g., cars, shoes, phones, apps)

- Crawl Twitter for historical mentions of these entities
  - E.g., for all of 2014-2018

- Build a system that can answer queries and display results
  - E.g., how many tweets per week did entity A get versus entity B in state X
  - E.g., how many positive versus negative tweets did entity A get over time

- Use machine learning/statistics to forecast
  - Popularity (number of tweets) for any entity for week T, given data to T-1
  - Or predict tweet sentiment (proportion positive/negative) for an entity
Tweets mentioning Coke (green) and Pepsi (red)

from
chimpler.wordpress.com
Projects from last year

- Algorithmic Passive Investing
- Can Health Predict Violent Crimes?
- Tracing Fake News & Fact Checks on Reddit
- Rating Differences between Yelp and Google
- Language and Partisanship: Predicting Partisanship with Tweets
- Impaired Water Quality Across the US
- Modeling Change in Song Topics Against Economic Data Using a Variety of Bayesian Approaches
Software for Future Assignments and Projects

▶ Python
  – Python will be the primary language we will use in much of this class
  – Assume that all students have a working knowledge of Python 3

▶ Packages and Libraries
  – We will make extensive use of additional packages and libraries in Python, e.g.,
    ▶ Pandas for data manipulation
    ▶ Scikit-learn: machine learning library
    ▶ Scientific computing/graphs/etc: matplotlib, numpy, scipy, etc

You should download and install the Anaconda package: it contains many packages you need for this class
Screenshot of the Spyder IDE

```python
1 # simple demo of NLTK functions applied to a Web page
2 # start up NLTK (usually not needed, i.e., already installed)
3 # import nltk
4 # nltk.download()
5 # import the packages that we will need
6
7 import nltk, re, pprint
8 from nltk import word_tokenize
9 from urllib import request
10 from bs4 import BeautifulSoup
11
12 # read the text from the Web page for Assignment 1
13 url = "http://www.ics.uci.edu/~smyth/courses/new175/assignment1.xht"
14 html = request.urlopen(url).read().decode('utf8')
15
16 # strip out most of the HTML stuff...
17 raw = BeautifulSoup(html).get_text()
18
19 # segment the text into sentences (useful for tokenizing - it's not necessary to do
20 sent = nltk.sent_tokenize(raw)
21
22 # extract tokens
23 tokens = word_tokenize(raw)
24 tokens[1:20]
25 tokens[21:50]
26
Python 3.4.1 |Anaconda 2.1.0 (64-bit)| (default, Sep 24 2014,
18:32:42) [MSC v.1600 64 bit (AMD64)]
Type "copyright", "credits" or "license" for more information.

IPython 2.2.0 -- An enhanced Interactive Python.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and
https://binstar.org
%quickref → Introduction and overview of IPython's features.
%quickref → Quick reference.
help → Python's own help system.
object? → Details about 'object', use 'object??' for extra
details.
%guirf → A brief reference about the graphical user interface.

In [1]:
```
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
An open-source software library for Machine Intelligence

Eager Execution
We're announcing eager execution, an imperative, define-by-run interface to TensorFlow. Check out the README to get started today.
LEARN MORE

TensorFlow 1.3 has arrived!
We're excited to announce the release of TensorFlow 1.3! Check out the release notes for all the latest.
UPGRADE NOW

The 2017 TensorFlow Dev Summit
Thousands of people from the TensorFlow community participated in the first flagship event. Watch the keynote and talks.
WATCH VIDEOS
It looks like there is a clear decision boundary between the two classes. Now we need to implement logistic regression so we can train a model to predict the outcome. The equations implemented in the following code samples are detailed in "ex2.pdf" in the "exercises" folder.

Figure from http://nbviewer.jupyter.org/github/jdwittenauer/ipython-notebooks/blob/master/notebooks/ml/ML-Exercise2.ipynb
What is Data Science?

- Data science involves the full lifecycle of data: from messy unstructured data to predictions and decisions

- Data science is broader than just databases, statistics, ML, algorithms, but these are all critical components

- Key aspects of data science include
  - Domain knowledge and problem definition
  - Data preparation/organization/management
  - Understanding of uncertainty (statistics)
  - Computing, algorithms, fitting models, machine learning
  - Iterative exploration and experimentation
  - Human judgement and interpretation
Components of Data Science

Hiring managers need to know they are making a tradeoff between these three skills when choosing a candidate. Skimp on hacking and your employee will always be waiting for someone to get them data. Skimp on stats and they'll misinterpret noise as signal. Skimp on expertise (domain knowledge) and they'll overlook important business drivers. If at all possible, remove one of these requirements. You'll hire candidates who are much stronger in the other two.

Notes
1. gr33ndata liked this
2. metricadb posted this
Hiring managers need to know they are making a tradeoff between these three skills when choosing a candidate. Skimp on hacking and your employee will always be waiting for someone to get them data. Skimp on stats and they’ll misinterpret noise as signal. Skimp on expertise (domain knowledge) and they’ll overlook important business drivers. If at all possible, remove one of these requirements. You’ll hire candidates who are much stronger in the other two.
Components of Data Science

Statistics (Mathematical and Probabilistic Foundations)

Computing (Algorithms and Software)

Applications (Analyzing Real Data)

Data Science
Data pipeline

Unstructured Data → Extracted Data → Transformed Data → Data for Modeling → Predictive Model → Predictions/Decisions
How is Data Science used in these Organizations?

Organizations

- Facebook
- Google
- Amazon
- Spotify
- Disney
- Kaiser Permanente
- Blizzard
- Honda

Data Science Applications

- Online advertising
- Automated recommendations
- Demand forecasting
- Fraud detection
- Churn prediction
- Automated customer support
How does Amazon forecast how many items for its warehouses?

From dailymail.co.uk

From www.formaspace.com

From linkedin.com
How does Facebook predict what content to show you?

MONTHLY USERS ON FACEBOOK 2004-2017

The Friendship graph

500M users each connect to an average of 130 other users = ~ 60 Billion Edges

Over 30 billion pieces of content shared every month

Over 3 billion photos uploaded each month

Graphics from Lars Backstrom, ESWC 2011
How do companies decide what ads to show you?

Putin, Flashing Disdain, Defends Action in Crimea

By STEVEN LEETHERS
69 minutes ago

President Vladimir V. Putin’s first public remarks on the political upheaval in Ukraine were aimed at both international and domestic audiences, defending Russia from the fury of global criticism and rallying support at home.

NEWS ANALYSIS

No Easy Way Out of Ukraine Crisis

By PETER BAKER 54 minutes ago

White House officials are weighing their options, knowing that reversing the occupation of Crimea would be difficult, if not impossible, in the short run.

Crimea’s Pro-Russian Leader Says Region Is Secure

By DAVID M. HERSHEYHORN 8:21 PM ET

The prime minister of the autonomous region offered the assurance on Tuesday even as armed standoffs continued.

Related Coverage

- Kerry Takes Offer of Aid to Ukraine 33 minutes ago
- Cyberattacks Rise as Crisis Spills to Internet 6:47 PM ET
- Russia: Confrontation in Crimea
How do public health workers predict infectious disease outbreaks?

Influenza Observations and Forecast

Data for Los Angeles, CA, week ending: Sat Jan 05 20
Using observations through week 53

Peak Timing Date: 25 Jan 2019
StDev: 1.00 weeks, Expected Accuracy: 59%
Peak Intensity: 4612 cases:
StDev: 746, Expected Accuracy: 0%
Onset: 25 Nov 2018, StDev: 0.00 weeks
Duration: 17.00 weeks, StDev: 2.00 weeks

https://cpid.iri.columbia.edu
Questions?