Info 2950, Lecture 1
25 Jan 2018

Intro to Data Science

Instructor: Paul Ginsparg (242 Gates Hall)
Tue/Thu 1:25-2:40
Kimball B11

https://courses.cit.cornell.edu/info2950_2018sp/
Info 2950: Introduction to Data Science

Spring 2018
Tue/Thu 1:25-2:40 PM, Kimball Hall B11
4 credits, S/U Optional

This course teaches basic mathematical methods for information science, with applications to data science. Topics include discrete probability, Bayesian methods, graph theory, power law distributions, Markov models, and hidden Markov models. Uses examples and applications from various areas of information science such as the structure of the web, genomics, social networks, natural language processing, and signal processing. Assignments require python programming.

Final: Tue 22 May 2:00-4:30

Announcements:
- 10 Jan: The Registrar has designated final exam date/time for this course as Tue, 22 May 2:00-4:30PM
- 25 Jan: First class

Professor: Paul Ginsparg (2/12 Gates Hall, ginsparg "at" cornell.edu)
Office hours: Wed 1-2 PM (or by appointment)

TAs:

Course website: http://courses.cit.cornell.edu/info2950_2018sp/ (this page)
Prerequisites: An introductory statistics course (from approved list) and an introductory programming class (e.g., CS 1110), or permission of instructor
Highly Recommended: Math 2310 or a similar linear algebra class (but no longer a pre req)
Textbooks: Rosen, *Discrete Mathematics and Its Applications*, 6th edition, is a basic ref for some of the earlier non-programming material

Syllabus:

<table>
<thead>
<tr>
<th>Date</th>
<th>Lecture</th>
<th>Reading</th>
</tr>
</thead>
</table>
| Week 1       | Thu 1/25 1. Course introduction | Instructions for downloading anaconda python are here
|              |                                  | There will be some python resources listed on the Piazza course site, the python.org site has a tutorial, and there are other online resources, including the book Think Python.
|              |                                  | Most recent class syllabus (just for reference): Spring ’17
| Week 2       | Tue 1/30 2. Probability and counting |                                             |
|              | Thu 2/1 3. Conditional Probability, Bayes’ Theorem |                                             |

Grading:
50% Problem Sets
25% Final exam
20% Midterm exam (in-class)
5% Subjective

Academic integrity policy:
You are expected to abide by the Cornell University Code of Academic Integrity. It is your responsibility to understand and follow these policies. (In particular, the work you submit for course assignments must be your own. You may discuss homework assignments with other students at a high level, for example discussing general methods or strategies to solve a problem, but you must cite the other student in your submission. Any work you submit must be your own understanding of the solution, the details of which you personally and individually worked out, and written in your own words.)

Advising Notes:
This course is a required course for the Information Science (IS) major in the College of Arts and Sciences and the College of Agriculture and Life Sciences. It can also be used to satisfy the math/stats course requirement for the Information Science minor/concentration.

[Note for InfoSci majors: INFO2950 and CS2800 may not both be taken for credit towards the IS major.
If (and only if) for some reason you have already taken CS2110 and CS2800, that combination can be petitioned to be used in place of INFO2950.]

cs 2110/2800
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Lecture</th>
<th>Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Thu</td>
<td>1. Course introduction</td>
<td>Lecture 1 slides, jupyter notebook class demo; icel.ipynb (intro)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Instructions for downloading anacoda python are here</td>
<td>There will be some python resources listed on the Piazza course site, the python.org site has a tutorial, and there are other online resources, including the book Think Python. Rosen ch 2.1-2.2 (online here). Most recent class syllabus (just for reference): Spring '16</td>
</tr>
<tr>
<td></td>
<td>Tue</td>
<td>2. Probability and counting</td>
<td>Lecture 2 slides on probability</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rosen ch 6 'Counting' (link given in piazza)</td>
<td>Re-birthday problem, see nytimes popular version (Strogatz)</td>
</tr>
<tr>
<td></td>
<td>Thu</td>
<td>3. Conditional Probability, Bayes' Thm</td>
<td>Lecture 3 slides on conditional probability and Bayes' thm</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Rosen ch 7.1-7.3, see link in piazza)</td>
<td>Re-birthday problem, see nytimes popular version (Strogatz)</td>
</tr>
<tr>
<td></td>
<td>Tue</td>
<td>4. Bayes' theorem and applications</td>
<td>Lecture 4 slides on applications of Bayes' thm, Bayesian spam filters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rosen ch 7.3</td>
<td>A few more &quot;native Bayes&quot; resources: Bayesian spam filtering, A plan for spam, and nytimes popular Bayes (Strogatz)</td>
</tr>
<tr>
<td></td>
<td>Thu</td>
<td>5. Expected value, Variance, Bernoulli trials</td>
<td>Lecture 5 slides on E[X], Var[X], Bernoulli distribution (Rosen 7.4),</td>
</tr>
<tr>
<td></td>
<td></td>
<td>see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tue</td>
<td>6. Text as data, more on Bernoulli</td>
<td>First went over ps2 and ps2_supp (urlpsi.com, Counter(), refindall(), ..)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thu</td>
<td>7. Statistical significance, p-hacking, and</td>
<td>Discussion of p-value as Bernoulli process, ps1 solutions, how to calculate word word correlations in ps2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>harking</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ash data analysis, more simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tue</td>
<td>8. Geodata, normal distributions, central</td>
<td>Started with basemap demo. Recalled last three slides from lecture 6 on normal distributions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>limit theorem</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thu</td>
<td>9. Visualizing pointwise normal information,</td>
<td>Continue discussion of Gaussian, Mendel, colored iris and airport delays</td>
</tr>
<tr>
<td></td>
<td></td>
<td>simulating normal distributions</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tue</td>
<td>10. Normal &gt; Poisson</td>
<td>Continued discussion of Gaussian, Mendel, colored iris and airport delays</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thu</td>
<td>11. Poisson -&gt; Poll Data</td>
<td>Continued discussion of Gaussian, Mendel, colored iris and airport delays</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tue</td>
<td>13. Time + Graph Theory I</td>
<td>Continued discussion of Gaussian, Mendel, colored iris and airport delays</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thu</td>
<td>14. Graph Theory II</td>
<td>continued graph theory in Lecture 14 slides.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tue</td>
<td>15. Graph Algorithms</td>
<td>discussed graph theory in Lecture 14 slides.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Copyright 2017, see also these simulations</td>
<td></td>
</tr>
</tbody>
</table>
INFO 2950: Introduction to Data Science

https://piazza.com/cornell/spring2018/info2950/home

Description

Teaches basic mathematical methods for information science, with applications to data science. Topics include discrete probability, Bayesian methods, graph theory, power law distributions, Markov models, and hidden Markov models. Uses examples and applications from various areas of information science, such as the structure of the web, genomics, social networks, natural language processing, and signal processing. Assignments require Python programming.

Prerequisite: An introductory statistics course (from the approved list of accepted statistics courses found at http://info.sci.cornell.edu/degrees/ba-college-arts-sciences/degree-requirements/core-requirements) and an introductory programming class (e.g., CS 1110) or permission of instructor.

Highly Recommended: Math 2310 or a similar linear algebra class

General Information

Website
https://courses.cit.cornell.edu/inf2950_2018sp/

Lecture Location
Kamhall B11, Tue/Thu 1:25 - 2:40

Midterm
(TBD)

Final Exam
22 May 2018 2:00PM-4:30PM (Exam Room TBD)

Academic integrity policy
You are expected to abide by the Cornell University Code of Academic Integrity (http://cui.info.cornell.edu/alc.cfm). It is your responsibility to understand and follow these policies. In particular, the work you submit for course assignments must be your own. You may discuss homework assignments with other students at a high level, for example discussing general methods or strategies to solve a problem, but you must write the other student in your submission. Any work you submit must be your own understanding of the solution, the details of which you personally and individually worked out, and written in your own words.)
Rough Syllabus

0. Review of basic python / jupyter notebook
1. Counting and probability (factorial, binomial coefficients, conditional probability, Bayes Theorem)
   Real Data: text classifier, etc. [baby machine learning]
2. Statistics: mean, variance; binomial, Gaussian, Poisson distributions
3. Graph theory (nodes, edges), networks (c.f. Info2040), graph algorithms
4. Power Law data (need exponential and logarithms …)
5. Linear and Logistic regression, Pearson and Spearman correlators
6. Markov and other correlated data

Rosen chapters 2,6,7,10,11
Easley/Kleinberg chpts 3,18
+ many other on-line resources
   [e.g. http://www.cs.cornell.edu/courses/cs1380/2018sp/textbook/, adapted from Berkeley http://data8.org/ (started apr ’16)]
A course for anyone who wants to study data visualization, prediction, machine learning, and programming in Python. We’ll analyze real-world data sets on crime, health, transportation, literature, and more!

CS 1380 + ORIE 1380 + STSCI 1380

Data Science For All

Spring 2018 MWF 10:10-11:00 am

https://tinyurl.com/datascienceforall

No experience required – Open to all – Fulfills MQR-AS
Problem sets will involve both programming and non-programming problems.

Problem sets are not group projects.

You are expected to abide by the Cornell University Code of Academic Integrity. It is your responsibility to understand and follow these policies. (In particular, the work you submit for course assignments must be your own. You may discuss homework assignments with other students at a high level, by for example discussing general methods or strategies to solve a problem, but you must cite the other student in your submission. Any work you submit must be your own understanding of the solution, the details of which you personally and individually worked out, and written in your own words.)

You’ll be penalized if you copy an iPython notebook, OR if yours is copied.
“Problem Set 0”, due Tue 30 Jan 23:59
Will be discussed in section tomorrow:

including instructions for installing anaconda, we'll standardize on python 3
due to minor python 2.7/3.6 compatibility issues (though welcome to use python 2)

course upload site: https://pgcourse.infosci.cornell.edu/cgi-bin/probset.py

N.B.: (former?) known problem with python installations:
cs 1110 unfortunately recommended misconfigured software that violates standard practice by surreptitiously adding environment variables to ~/.bashrc file.

(instructions for removing: https://courses.cit.cornell.edu/info2950_2018sp/resources/bashprob.html )
Definition. A set $S$ is a collection of objects.

The objects of a set are called elements $x$ of the set: $x \in S$, or $x \not\in S$.

Examples:

$X = \{1, 2, 3, 4, 5\}$

$C = \{\text{Ithaca, Boston, Chicago}\}$

Stuff = \{1,\text{snow, Cornell, y}\}$

empty set = \emptyset

Can also be defined by rule or equation:

Example: $E$ is the set of even numbers. $E = \{x \mid x \text{ is even}\}$

Cardinality $|S|$ is the number of elements of $S$

Examples: $|X| = 5$, $|C| = 3$, $|\text{Stuff}| = 4$, $|\emptyset| = 0$
Definition. A set $S$ is a collection of objects.

The objects of a set are called elements $x$ of the set: $x \in S$, or $x \notin S$.

Examples:

$X = \{1, 2, 3, 4, 5\}$

$C = \{\text{Ithaca, Boston, Chicago}\}$

- A subset $T$ of a set $S$ is a set of elements all of which are contained in $S$.

  $T \subset S$ (proper subset) or $T \subseteq S$

  empty set $\emptyset \in S$ for all $S$

Examples:

$C' = \{\text{Ithaca, Chicago}\}$

$C' \subset C$

$X' = \{x \mid x \text{ is a whole number between 2 and 5}\}$ is a subset of $X$
For two sets to be the same, must have the same elements.

\[ A = B \] means that \( \forall x \) we have \( x \in A \) iff \( x \in B \)

(Equivalently \( A = B \) means that \( A \subseteq B \) and \( B \subseteq A \))

The power set \( \mathcal{P}(S) \) of a set \( S \) is the set of all subsets of \( S \).

Example: For the set \( A = \{1, 2, 3\} \),

\[ \mathcal{P}(A) = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 3\}, \{1, 2, 3\}\} \]

For a set \( S \) with \( n \) elements, what is \( |\mathcal{P}(S)| \)?

Cartesian product of two sets \( A \times B = \{(x, y) \mid x \in A \text{ and } y \in B\} \)

Example:

\[ A \times A = \{(1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\} \]
Set Operations

- union of two sets $A \cup B = \{x \mid x \in A \text{ or } x \in B\}$

Examples:

$X \cup \text{Stuff} = \{1, 2, 3, 4, 5, \text{snow}, \text{Cornell}, y\}$

$C \cup \emptyset = \{\text{Ithaca, Boston, Chicago}\}$

$A \cup X = \{1, 2, 3, 4, 5\}$ (In this case, $A \cup X = X$).

- intersection of two sets $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$

Examples:

$X \cap \text{Stuff} = \{1\}$

$C \cap \emptyset = \emptyset$

$X \cap E = \{2, 4\}$

$A \cap X = \{1, 2, 3\}$ (In this case $A \cap X = A$)
- Difference of two sets $A - B = \{x \mid x \in A \text{ and } x \not\in B\}$
  
  Examples:
  
  $X - \text{Stuff} = \{2, 3, 4, 5\}$
  
  $\text{Stuff} - X = \{\text{snow, Cornell, y}\}$
  
  $C - \emptyset = \{\text{Ithaca, Boston, Chicago}\}$
  
  $X - E = \{1, 3, 5\}$
  
- Symmetric difference $A \Delta B = \{x \mid x \in A \text{ or } x \in B, \text{ and } x \not\in A \cap B\}$
  
  Examples:
  
  $X \Delta \text{Stuff} = \{2, 3, 4, 5, \text{snow, Cornell, y}\}$
  
  $C \Delta \emptyset = \{\text{Ithaca, Boston, Chicago}\}$
  
  $X \Delta A = \{4, 5\}$