**CUSP-GX-5004: Applied Data Science**

**Fall 2014**

**Lectures: Mondays 5:40pm – 8:30pm**

**CUSP, 2 MTC, 8th Floor, MAGNET Room TBD**

**Instructors:**

Foundations of Data Science: Predictive Analytics and Decision-Making

Tim Savage, PhD. [timothyhsavagephd@gmail.com](mailto:timothyhsavagephd@gmail.com) 917.439.3175

Applied Data Science Module 1: Video Analytics

Greg Dobler, PhD. [greg.dobler@nyu.edu](mailto:greg.dobler@nyu.edu), 646-997-0526

Applied Data Science Module 2: Sound Analytics

Charlie Mydlarz, PhD. [cmydlarz@nyu.edu](mailto:cmydlarz@nyu.edu), 646-997-0536

**Teaching Assistant:**

Karan Modi [km2971@nyu.edu](mailto:km2971@nyu.edu)

**Office Hours**

TBD

**Course Description**

This course introduces students to the theory, principles and applications of mathematical and computer modeling of data as applied to cities. It will be based on two unified themes: foundations for predictive analytics and decision-making followed by applications in data science. The 1st half of the course will cover predictive modeling using a wide array of examples, including predictive modeling, an advanced treatment of regression, visualization and graphics, and automated analysis for high dimensional data. The second half will introduce students to applications in data science such as analytics of images and video as well as subjective data processing and analysis.

**Course Objectives**

The study of statistics lies at the core of applied data science, but it can quickly devolve into little more than equations. The objective of this course is to move beyond the equations to develop an understanding, ideally broad, of the role of data analysis in actual practice. Software can do the calculations. It cannot, however, form meaningful hypotheses, develop properly specified tests, or interpret results, the latter of which are the objectives of this course.

**Course Requirements**

As noted, the course is software neutral. For either R or Python, students can freely download either package together with a GUI of their choice. Both are free programming environments (at least at the point of provision) for statistics and data manipulation. (R has been around longer than Python and has more extensive statistical libraries, especially for Bayesian analysis.) We will use only a small subset of the capabilities of either package, which will not require extensive training.

**Grading:**

Foundations Module: 50% (Homework: 20%; Lab: 10; Project: 20%)

Applied Data Science Module 1: 25% (Homework: 5%; Project: 20%)

Applied Data Science Module 2: 25% (Homework: 5%; Project: 20%)

**Foundations Project**

This project aims to bring together all of the materials of the foundations module. It also aims to expose you to the task of original research using data analytics. At the end of the third session, you will submit a research proposal of one paragraph outlining a particular topic that you would like to explore. This topic can range from analysis of taxi data to analysis of 311 data. It’s your call. In the proposal, you should address what hypotheses you would like to explore and how you might go about doing so. During the remainder of the foundations module, you will be exposed to a variety of statistical techniques which you should be able to apply to the data you propose to analyze. At the end of the course, you will submit a five-page, double-space paper that describes your research agenda, the data you have gathered, the hypotheses you will to explore, the method you have used, and the results. The results can be both tabular and visual. The paper size is independent of tables and visualizations.

**Applied Data Science Module 1 Project**

The final project for the image processing and computer vision module will consist of submitting an analysis pipeline in python to be executed on a video that the student has not yet seen. Students will be graded based on the accuracy of the output. There will be weekly (ungraded) problem sets designed to familiarize students with the basic principles and techniques that will be used on the final project.

**Applied Data Science Module 2 Project**

The project involves the challenges faced and inferences that can be made when studying urban soundscapes. A soundscape can be defined as the auditory environment which surrounds a listener. The term is analogous with landscape in that it represents an individual’s unique experience of inhabiting an auditory environment, based on their previous experiences and interests. Gathering and making use of data from these environments therefore requires both objective audio data and subjective human response data, so the handling and use of these will be introduced, alongside the use of existing datasets to uncover trends around the noise issues of NYC.

**Applied Data Science Module 3 Project**

The project involves recreating and expanding work in computational social justice. The general thrust of the work will be creating tools that identify instances of human trafficking on the open internet. This will combine all arenas of data science: getting the data, cleaning the data, processing the data, and automatic analysis of the data. During the project we will explore techniques involving image processing, text processing and multivariate linear regression.

**NYUClasses**

You must have access to the class Blackboard site (http://classes.nyu.edu/). All announcements and class-related documents (supplemental and suggested readings, discussion questions, etc.) will be posted there.

**Recommended/Suggested Readings**

**Reading for Foundations Module:**

Angrist and Pischke, *Mostly Harmless Econometrics: An Empiricist’s Companion*, Princeton University Press, 2009.

Hastie, et al., *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd Edition, Springer. (Online: <http://statweb.stanford.edu/~tibs/ElemStatLearn/>)

Kabacoff, *R in Action: Data analysis and graphics with R*, 1st Edition, Manning Publications.

McKinney, *Python for Data Analysis*, O’Reilly Media Inc., 2013.

Sheppard, *Introduction to Python for Econometrics, Statistics and Data Analysis*, August 2014.

(Online: http://www.kevinsheppard.com/images/0/09/Python\_introduction.pdf.)

Zumel and Mount, *Practical Data Science with R*, 1st Edition, Manning Publications Company, March 2014. (Select chapters available for free online: <http://www.manning.com/zumel/>)

**Reading for Applied Data Science Module 1:**

Programming Computer Vision with Python <http://programmingcomputervision.com/downloads/ProgrammingComputerVision_CCdraft.pdf>

In addition, prior to the beginning of the module, students should have installed git (and generated a GitHub account), python, ipython, numpy, scipy, matplotlib, scikit-learn, and opencv (with python bindings).

**Readings for Applied Module Data Science 2:**

Kang, Jian. Urban Sound Environment. CRC Press, 2006.

**Statement of Academic Integrity**

NYU-CUSP values both open inquiry and academic integrity. Full and Part-Time graduate programs and advanced certificate programs are expected to follow standards of excellence set forth by New York University. Such standards include but are not limited to: respect, honesty and responsibility. The program has zero-tolerance for violations to academic integrity. Such violations are deemed unacceptable at NYU and CUSP. Instances of academic misconduct include but are not limited to:

* Plagiarism
* Cheating
* Submitting your own work toward requirements in more than one course without (1) prior documented approval from instructor and (2) proper citation
* Forgery of academic documents with the intent to defraud
* Deliberate destruction, theft, or unauthorized use of laboratory data, research materials, computer resources, or University property
* Disruption of an academic event (lecture, laboratory, seminar, session) and interference with access to classroom, laboratories, or academic offices or programs

Students are expected to familiarize themselves with the University’s policy on academic integrity and CUSP’s policies on plagiarism as they will be expected to adhere to such policies at all times – as a student and an alumni of New York University.

**Course Schedule**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **SESSION #** | **DATE** | **MODULE** | **TOPICS** | **ASSESSMENT** |
| 1 | 9/8 | Foundations Session 1 | Introduction to the course. Probability theory, random variables, and important statistical concepts, such as average, variance, and correlation. | Homework 1 |
| 2 | 9/15 | Foundations Session 2 | Classical hypothesis testing. The bivariate linear regression model in full. Other topics in machine learning. | Homework 2 |
| 3 | 9/22 | Foundations Session 3 | The multivariate regression model and an introduction to probability and generalized linear models. | Homework 3 |
| 4 | 9/29 | Foundations Session 4 | Predictive Modeling Lab |  |
| 5 | 10/06 | Foundations Session 5 | Predictive Modeling Lab |  |
|  | 10/13 | Fall Recess | NO CLASSES |  |
| 6 | 10/20 | Foundations Session 6 | Time Series Analysis. | Homework 4 |
| 8 | 10/27 | Foundations Session 7 | Causation and Identification. | Homework 5 |
| 9 | 11/3 | Foundations Session 8 | Other topics (based on input from class): Maximum likelihood, nonlinear models, method of moments, kernel density estimation, and Bayesian analysis. | Homework 6 |
| 10 | 11/10 | Image Analytics Session 1 | Working with images (2D&3D arrays) in python/numpy. Image segmentation. | Foundations Project Due |
| 11 | 11/17 | Image Analytics Session 2 | Video data including I/O and background subtraction. Introduction to machine learning in python. | Homework 7 |
| 12 | 11/24 | Image Analytics Session 3 | Basic hand written digit recognition techniques. | Image Project Due |
| 13 | 12/1 | Subjective Data Analysis Session 1 | Objective audio data processing and analysis |  |
| 14 | 12/8 | Subjective Data Analysis Session 2 | Subjective data processing and analysis | Homework 8 |
| 15 | 12/15 | Subjective Data Analysis Session 3 | Time series analysis of 311 noise complaint data | Sound Project Due |