DS-UA 0112: Introduction to Data Science

Credits: 4 credits

Course description and impact

Students will explore the theoretical issues, methods, tools and problems that relate to data-rich issues in the humanities, social sciences, and sciences. Students will learn the core concepts of inference and computing, while working hands-on with real data including humanities data, social science data (ex. geographic data and social networks), and scientific data. We will examine how data analysis technologies can be used to improve decision-making within the liberal arts disciplines, as well as ethical implications.

We will study the fundamental theories, principles and techniques of data science, and we will examine real world examples and cases to place data science techniques in context, to develop data-analytic thinking, and to illustrate that proper application is as much an art as it is a science. In addition, we will work hands-on with the Python programming language and its associated data analysis libraries. Students will also explore theoretical issues related to ethics.

Learning objectives

- After successfully completing the course, students are able to
- Incorporate data science principles to address data-dependent questions in the humanities, social sciences, and sciences.
- Apply basic exploratory analysis to identify abnormalities in data (i.e., missing values, outliers, redundant features, etc.)
- Anticipate and identify ways in which sampled data may be biased
- Prepare data sufficient for answering a range of research questions across liberal arts disciplines.
- Identify instances of data leakage and apply techniques to avoid it
- Perform the appropriate feature transformations for processing categorical data and for making non-linear representations in linear models
- Identify the appropriate set of algorithms (i.e., regression vs. decision tree vs. clustering) for a given problem statement, and give an appropriate analysis of the pros/cons of each for the problem at hand
- Explain relevant data science theories and concepts, such as model regularization and optimization.
- Design and implement an experiment incorporating data science principles.
- Explore ethical implications surrounding privacy, data sharing, and algorithmic decision making for a given data science approach.

Office hours

The course instructor is available two hours per week for one on one meetings with students. ***[Specifics added here for finalized syllabus]***

Topics covered

The course is structured into a sequence of lectures and accompanying assignments.

The assignment consists of labs and homeworks.

- Labs are short exercises done in class and submitted in class.
- Homeworks are longer exercises designed to take a week. The homework problems are based on contemporary, real-world issues.

- Most lectures have accompanying readings from at least one of these two required books:
- PF: *Data Science for Business*, Provost and Fawcett, O'Reilly 2013. Available on-line and in bookstores.
- M: *Python for Data Analysis*, Wes McKinney, O'Reilly , 2012.

The course is structured into logical weeks, which may not correspond to calendar weeks because of holidays in the academic calendar. The first week has its own schedule, but then the weeks typically fall into this pattern:

- The first meeting is a 75-minute lecture covering principles and an overview of techniques that illustrate the principles.
- The second meeting is a 75-minute lecture and lab. The lecture covers the details of the technique that students will learn. The lab allows the students to use the technique to solve a particular problem. Sometimes a more complex problem is assigned as a homework problem. Homeworks are due one week after they are assigned.
- The third meeting is an active problem-solving/lab session. Bring your laptops, work on the homework, and also get help from the instructor and teaching assistants. Ask questions that you find puzzling.

There are no stand-alone essay-type assignments. Instead writing is embedded in the labs and homeworks. During the labs, students write computer programs. The computer programs are complex to write, as often each part must be cohesive with the others. During the homeworks, students also write computer programs, and in addition, they provide short essays on the interpretation of the data and the implications for the data around some decision or problem that the data inform. The writing length would typically be two to three pages for these essays, for a 20 page writing total.

The week by week schedule on a logical basis is just below. After the week number, we name the readings for the week by indicate the authors' initials and chapter or section numbers

1. Week 1: PF 1, 2 (43 pages); W 2 (26 pages), 3 (34 pages)
    a. What is data science;
    b. Introductory Examples; chapters 1 and 2; Homework 01 (understanding data science projects; this and all other homeworks are due in one week)
    c. Tutoring session: assure students have the required software running on their laptops
2. Week 2: PF 3 (39 pages)
    a. Introduction to Predictive Modeling
    b. Decision Trees, Lab 01 (classification with decision trees), Homework 02 (regression with decision trees)
    c. Tutoring session
3. Week 3: PF 4 (31 pages); W 4 (31 pages)
    a. Fitting a model to data
    b. Linear models, Lab 02 (multivariate logistic regression), Homework 03 (multivariate linear regression)
    c. Tutoring session
4. Week 4: PF 5(32 pages); W 5 (42 pages)
    a. Overfitting and its avoidance
    b. Model selection: Lab 03 (validation), Homework 04 (cross validation)
    c. Tutoring session
5. Week 5: PF 6 (48 pages); W 6 (22 pages)
    a. Similarity, neighbors, and clusters
    b. K-means clustering: Lab 04 (fixed K), Homework 05 (finding the best value for K)
    c. Lab session
6. Week 6: PF 7 (24 pages), 11 (12 pages); W 7 (36 pages)
    a. What is a good model?
    b. Expected value: Lab 05 (calculating net present values), no homework
    c. Lab session
7. Week 7: PF 8 (24 pages); W 8 (31 pages)
    a. Visualizing model performance
    b. ROC curves and lift curves: Lab 06 (manual calculations), Homework 06 (programmed calculations)
    c. Lab session
8. Week 8
    a. Review for midterm exam
    b. Mid-term exam
    c. Lab: Q&A on mid-term questions
9. Week 9: PF 9 (17 pages)
    a. Evidence and probabilities
    b. Naive Bayes: Lab 07 (example calculation), Homework 07 (comparing Naive Bayes and logistic regression)

10. Week 10: PF 10 (27 pages)
    a. Representing and mining text
    b. TF-IDF: Lab 08 (jazz musicians, Homework 08 (who wrote the Federalists Papers)
11. Week 11: PF 12 (22 pages)
    a. Other useful models
    b. Co-occurrence: Lab 09 (small example), Homework 09 (big example)
12. Week 12; repeat PF 12
    a. Ensemble methods
    b. Random forests: Lab 10 (small example), Homework 10 (big example)
13. Week 13
    a. Catch up or new topic, at discretion of instructor
14. Week 14; PF 13 (16 pages)), 14 (14 pages)
    a. Fundamental concepts review
    b. Review for final
    c. Review for final
15. Week 15: at discretion of the instructor

Course assessment

All assignments (labs and homeworks) must entirely be the student's own submissions. Any sharing or copying of assignments is considered cheating and will result in an F in the course. A second cheating incident will, by CAS rules, result in a one-semester suspension from the College.

Students accumulate up to 100 points during the course.

- Up to 10 points for completing labs
- Up to 12 points for completing homeworks on time
- Up to 32 points for the midterm
- Up to 46 points for the final.

Grades will be determined using this scale:

Grade in Course	**Points Earned**
A	94 - 100
A-	90 - 93
B+	87 - 89

B	84 - 86
B-	80 - 83
C+	76 - 79
C	72 - 75
C-	70 - 71
D+	66 - 69
D	62 - 65
D-	60 - 61
F	Less than 60

Moses statement

Disability Disclosure Statement: Academic accommodations are available for students with disabilities. The Moses Center website is www.nyu.edu.csd. Please contact the Moses Center for Students with Disabilities (212-998-4980 or mosescsd@nyu.edu) for further information. Students who are requesting academic accommodations are advised to reach out to the Moses Center as early as possible in the semester for assistance.