IST 718: Big Data Analytics

Course information

Catalog description

A broad introduction to analytical processing tools and techniques for information professionals. Students will develop a portfolio of resources, demonstrations, recipes, and examples of various analytical techniques.

Detailed course description

This course will prepare you to participate as a Data Scientist on big data and data analytics projects. Upon the successful completion of this course, you will be able to:

- Translate a business challenge into an analytics challenge;
- Analyze big data, create statistical models, and identify insights that can lead to actionable results;
- Use Python and Apache Spark to build big data analytics pipelines
- Learn classic and state of the art machine learning techniques
- Explain how advanced analytics can be leveraged to create competitive advantage;

Prerequisite knowledge required

Familiarity with command-line interfaces, quantitative skills, including statistics as well as programming skills in Python. Please see https://acuna.io/teaching/IST718/

Textbooks:

We will use parts of 4 textbooks:

- *Python Data Science Handbook* (PDSH) by Jake VanderPlas (Free), https://jakevdp.github.io/PythonDataScienceHandbook/
- *An introduction to Statistical Learning with Applications in R* (ISLR) by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (Free) http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf
- *Deep Learning* (DL) by Ian Goodfellow, Yoshua Bengio, and Aaron Courville (Free) http://www.deeplearningbook.org/
- *Apache Spark: The Definitive Guide (Excerpts)* (ASDG) by Chambers and Zaharia (https://pages.databricks.com/definitive-guide-spark.html)

Course Topics

The following is a tentative outline of topics to be covered in the course:

Week	Topics	Suggested reading	Dates/Events
Aug 26/27	**Overview of the course and review** Linear algebra, calculus, statistics; Python, Jupyter notebook	DL: Ch. 2.1 - 2.6 DL: Ch. 3.1 - 3.9.3 DL: Ch. 4.3 PDSH Ch. 1	- HW 1 released
Aug 29*/Sept 3	**Python Programming** Numpy, Pandas, Matplotlib	PDSH: Ch. 2.2 - 2.7 PDSH: Ch. 3.2 - 3.4 PDSH: Ch. 4.1 - 4.9	
Sept 9/10	**Introduction to Hadoop, MapReduce, and Apache Spark**	ASDG: 3-8	- HW 1 due - HW 2 released - Group formation
Sept 16/17	**Introduction to Spark DataFrames and Spark ML**	ASDG: 44 - 126	
Sept 23/24	**A Statistical Perspective on Machine Learning** Introduction to probability; maximum likelihood estimation; mean square error estimation; gradient descent	ISLR: Ch. 1; Ch 2.1 DL: Ch 1-2	- Project proposal due - HW 2 due - HW 3 released
Sept 30/Oct 1	**Assessing Model Accuracy** Confusion matrix, bias–variance tradeoff, model selection: training, validating, and testing	ISLR: Ch 2.2	
Oct 7/8	**Case 1: Sentiment Analysis of Twitter** Supervised learning, logistic regression, regularized logistic regression, elastic net regularization, model interpretation	ISLR: Ch 6	- HW 3 due - HW 4 released
Oct 14/15	**Case 1 (cont.)**		.
Oct 21/22	**Case 2: A recommendation system for courses** Unsupervised learning, nearest neighbors, dimensionality reduction (Principal Component Analysis, PCA), clustering (*k*-means)	ISLR: Ch. 10, sections.10.1, 10.2 and 10.3	- HW 4 due - HW 5 released
Oct 28/29	**Case 2 (cont.)**		- Project update presentation
Nov 4/5	**Case 3: Predicting Credit Scores with Bagging and Boosting** "wisdom of the crowd", bagging, random forests, gradient boosting, feature importance	ISLR: Ch 8	- HW 5 due - HW 6 released
Nov 11/12	**Case 3 (cont.)**		
Nov 18/19	**Case 4: Object Recognition with Deep Learning** Neural networks, multilayer perceptron, other topics and next steps for data science careers	DL: Ch 6	- HW 6 due - Nov 22: Poster session
Nov 25/26	**Thanksgiving**		
Dec 2/3	**Case 4 (cont.)**		- Project code due on Dec 6

*Special session, tentatively scheduled for August 29 between 7:00 PM and 10:00 PM. To be recorded.

Methods of evaluation:

Assessment	Notes	Points
Quizzes (best 5 of 6)	Covers concepts; unannounced; no make-ups	25 (5 x 5)
Homework (6)	Based on class materials; late submission will be discounted	48 (6 x 8)
Group Project (1)	In groups of only 3 or 4; you choose your teammates	24 (total)
- Project proposal	Beginning of semester	4
- Project update	In-class short overview and update of project	5
- Poster presentation	Poster day	10
- Project code	Final submission	5
Participation		-
- In class and/or forums		3
TOTAL		100

Grading scale

Total Points Earned	Registrar Grade
95 - 100	A
90 - 94	A -
85 - 89	B +
80 - 84	B
75 - 79	B -
70 - 74	C +
65 - 69	C
60 - 64	C -
50 - 59	D
0 - 49	F

Quizzes

Quizzes are individual effort, in-class short tests which measure your understanding of the concepts and terminology covered in class, labs and the assigned readings.

- Quizzes could be issued at the beginning of class. Please report to class on time.
- No make-ups will be given to absent or late-arriving students. Consider quizzes part of your attendance.
- Quizzes are unannounced. There are no quiz dates posted on the syllabus, so expect there will be a quiz for class.
- Quizzes will cover all material up to the point where the quiz is issued,

Homework

Homework is typically released on Tuesdays at 12:30 PM after classes and due on the Sundays at midnight before classes. You need to create a Github.com account and share your Github username with the professor. Then, you will use the Github authentication system to use http://notebook.acuna.io/

While you are encouraged to discuss homework with your classmates, homework programming and writing is individual effort. It is designed to check that you are keeping pace with in class concepts and out-of-class lab activities.

- The intention of homework is to ensure students are keeping pace with the out of class activities.
- For late submissions, the grade of your homework will be multiplied by the following factor, where days (could be fractional) is the number of days submitted late (days > 0):

```
def grade_factor(days):
   if days > 3:
      return 0
   else:
      return 1/(1 + days)**(3/4)
```

  e.g., 1 day late 60% of grade, 2 days late 43%, 2 hours late ~95%
- Submissions later than 72 hours from deadline will zero grade.

Group Project

The group project is your chance to demonstrate what you've learned in the course and apply it to a new scenario. It is expected that each group's project will be novel, and that all will be of the highest quality. Your group will be responsible for finding a data set, analyzing it, and producing visualizations and findings from your analysis.

The group project consists of four elements:

1. **Project proposal**. 2-page preliminary description of the project, including problem statement, solution proposed, techniques, datasets and expected results.
2. **Project update:** During the middle of the semester, your project team is required to present updates about your project. You are expected to have done significant advances in your analysis and have concrete ideas about next steps.
3. **Presentation.** Your group is required to present your findings on the iSchool's joint poster day (November 30, 2018) During this time, your professor and TA will stop by and ask you to present your project.
4. **Project code.** Well-commented code to produce the result of your poster.

There is no late submission on the presentation parts and all members of the group must be able to describe the work during poster day.

Participation

It is expected that you participate in class and Blackboard forum discussions. Also, it is expected that you visit the professor during office hours or set up a time to meet him during the semester. Additionally, it is expected that the group visits the professor or talk to the TA to discuss the project.

General information

Teaching philosophy

My teaching philosophy centers around students as critical thinkers that challenge current beliefs using arguments rooted in strong evidence. There are three concepts to this

- We are all inherently curious about how the world works and have an unbounded set of needs
- We all make mistakes and all questions are valid, however we only realize these blind spots when we critically think and discuss our ideas with others
- Data is only a means to a goal but we are responsible for keeping our analysis, policy recommendations, and conclusions as ethical and compassionate as possible.

Who does well in the course?

This is a relatively heavy load course and an ideal student should follow the following items:

- Study consistently throughout the semester in short burst. Research has shown that pulling all-nighters and studying just before the class or lab will make you forget the contents later in your career. Make an effort to study everyday at least 30 minutes for the class.
- Be active in class and ask questions to the professor. Challenge the materials and try to see all angles of the ideas and conclusions being presented in class. Critical thinking is as important as technical ability in a data science job and it is a highly appreciated skill
- Focus on learning how to program and the pieces involved in developing professional software. Although this class does not assume a large amount of programming experience, the sooner you start learning about Python and the tools taught in this class, the better you are going to do throughout the semester. Data scientists who are excellent critical thinkers AND known how to transform ideas into software are the most prized in the job market.

Academic Integrity Policy

Syracuse University's academic integrity policy reflects the high value that we, as a university community, place on honesty in academic work. The pilot policy in effect at the School of Information Studies defines our expectations for academic honesty and holds students accountable for the integrity of all work they submit. Students should understand that it is their responsibility to learn about course-specific expectations, as well as about university-wide academic integrity expectations. The pilot policy governs appropriate citation and use of sources, the integrity of work submitted in exams and assignments, and the veracity of signatures on attendance sheets and other verification of participation in class activities. The pilot policy also prohibits students from submitting the same work in more than one class without receiving written authorization in advance from both instructors. Under the pilot policy, students found in violation are subject to grade sanctions determined by the course instructor and non-grade sanctions determined by the School or College where the course is offered. SU students are required to read an online summary of the university's academic integrity expectations and provide an electronic signature agreeing to abide by them twice a year during pre-term check-in on MySlice. For more information and the pilot policy, see http://academicintegrity.syr.edu

Disability-Related Accommodations

Syracuse University values diversity and inclusion; we are committed to a climate of mutual respect and full participation. If you believe that you need accommodations for a disability, please contact the Office of Disability Services (ODS), disabilityservices.syr.edu, located at 804 University Avenue, room 309, or call 315.443.4498 for an appointment to discuss your needs and the process for requesting accommodations. ODS is responsible for coordinating disability-related accommodations and will issue "Accommodation Authorization Letters" to students as appropriate. Since accommodations may require early planning and generally are not provided retroactively, please contact ODS as soon as possible. Our goal at the iSchool is to create learning environments that are useable, equitable, inclusive and

welcoming. If there are aspects of the instruction or design of this course that result in barriers to your inclusion or accurate assessment or achievement, please meet with me to discuss additional strategies beyond official accommodations that may be helpful to your success.

Religious Observances Notification and Policy

SU's religious observances policy, found at supolicies.syr.edu/emp_ben/religious_observance.htm, recognizes the diversity of faiths represented in the campus community and protects the rights of students, faculty, and staff to observe religious holy days according to their tradition. Under the policy, students should have an opportunity to make up any examination, study, or work requirements that may be missed due to a religious observance provided they notify their instructors no later than the end of the second week of classes through an online notification form in MySlice listed under **Student Services/Enrollment/My Religious Observances/Add a Notification**.