



Course Syllabus – INST 414 Section 0102 – Fall 2019

Data Science Techniques

Learning Outcomes

This course will explore approaches to extract insights from large-scale datasets. It will cover the complete analytical funnel from data extraction and cleaning to data analysis and insights interpretation and visualization. The data analysis component will focus on techniques in both supervised and unsupervised learning to extract information from datasets. Topics will include clustering, classification, and regression techniques. Through homework assignments, a project, exams and in-class activities, students will practice working with these techniques and tools to extract relevant information from structured and unstructured data.

Extended Course Description

The course explores the application of data science techniques to unstructured, real-world datasets including social media and open data sources. It will focus on techniques and approaches that allow the extraction of information relevant for experts and non-experts in a wide range of areas including smart cities, transportation or public safety.

We will explore approaches to extract insights from large-scale dataset and will cover the complete analytical funnel from data extraction and cleaning to data analysis and insights interpretation and visualization. The data analysis component will focus on techniques in both supervised and unsupervised learning to extract information from datasets. Topics will include clustering, classification, and regression techniques. Through homework assignments, a project, exams and in-class activities, students will practice working with these techniques and tools to extract relevant information from structured and unstructured data.

After successfully completing this course you will be able to:

- Collect and clean large-scale datasets
- Articulate the math behind supervised and unsupervised techniques
- Execute supervised and unsupervised machine learning techniques
- Select and evaluate various types of machine learning techniques
- Explain the results coming out of the models
- Critically evaluate the accuracy of different algorithms and the appropriateness of a given approach

INST 414

Fall 2019

Instructor:

Aditya Bhat

inst414@gmail.com

Teaching Assistant:

Kalpita Raut

krout@umd.edu

Class Meets

Mondays

& Thursdays

5:30pm – 6:45pm

ATL # 2428

Office Hours

TBA

Prerequisites

INST 314 - Statistics for Information Science

Course Communication

I will use ELMS announcements for time-sensitive communication.

I can NOT guarantee prompt answers to questions sent after 7pm or weekends. If you have not received a reply within 2 days, please message me again - I get a lot of emails and sometimes one gets lost in the shuffle.

To notify me you had an excused absence, email me as promptly as possible. Work missed for excused absences may be made up for full credit.



Required Resources

Course website: elms.umd.edu



The Data Science Handbook
Cady, Field
First edition (2017).
Available on Amazon



You need to bring a laptop to each class. If you do not have a laptop you can bring, contact the instructor before the first week of class ends. Over the course of the semester, you may need to install free software including **Open Refine, Weka, Python, Anaconda Distribution, Tableau, and Gephi**. If you encounter problems with these tool, some of them will be available over the iSchool Virtual Computing Lab as an alternative option you may use.

Campus Policies

It is our shared responsibility to know and abide by the University of Maryland's policies that relate to all courses, which include topics like:

- Academic integrity
- Student and instructor conduct
- Accessibility and accommodations
- Attendance and excused absences
- Grades and appeals
- Copyright and intellectual property

Please visit go.umd.edu/ug-policy for the Office of Undergraduate Studies' full list of campus-wide policies and follow up with me if you have questions.



Accessibility and Learning Support

Students with disabilities should inform me of their needs at the beginning of the semester. Please also contact the Accessibility and Disability Support Office (<http://www.counseling.umd.edu/ADS/>). ADS will make arrangements with the student and me to determine and implement appropriate academic accommodations. Inclusion is one of the iSchool's core values, and we have attempted to make all materials and assignments accessible to people with varying abilities. However, if there is something else I can do to make the class more accessible please schedule a time to come talk to me. This will benefit not only yourself but also future students!

Writing and English Help

Even though all UMD students are either native English speakers or have passed the TOEFL, it is normal to struggle to grasp the tone and style expected in advanced professional writing and speaking. The best resource to help with this issue on campus at this time is the Undergraduate Writing Center which offers help with all aspects of writing. Appointments can be made online at www.english.umd.edu/academics/writingcenter.

Student Health and Time Management

For your own health and sanity, please **stick to the hours of daylight** for as much of your work for this class as possible. If you have to turn in an assignment late, I would rather you go to sleep and submit it at lunch time the next day than stay up late and get it in at 3:30 am. Your grade is the same either way and **YOU DESERVE TO SLEEP**.

Furthermore, **DO NOT COME TO CLASS IF YOU ARE ILL**. You will *avoid exposing your classmates and teacher* to infection. As this class meets twice a week, you may miss non-consecutive classes with a self-signed sick notice emailed to me within 48 hours of missing class. For consecutive classes, a doctor's note will be required. The only exception is the midterm and which requires a doctor's note regardless of duration of illness. There is no limit to the total number of absences that may be excused via official doctor's notes, allowing you to make up all missed work without late penalties.

We may need to adjust the schedule (postponing deadlines for snow days, for example) and if most the class is struggling with a particular issue, I may assign an additional assignment on it, but this will be rare. Please take the time to look over the course calendar. You may even want to use a digital calendar so you can set yourself reminders to start each of these assignments and reminders to submit them.



Get Some Help!

Taking personal responsibility for your own learning means acknowledging when your performance does not match your goals and doing something about it. I hope you will come talk to me so that I can help you find the right approach to success in this course, and I encourage you to visit tutoring.umd.edu to learn more about the wide range of campus resources available to you. In particular, everyone can use some help sharpen their communication skills (and improving their grade) by visiting ter.ps/writing and schedule an appointment with the campus Writing Center. Finally, if you just need someone to talk to, visit counseling.umd.edu.



Everything is free because you have already paid for it, and **everyone needs help...** all you have to do is ask for it.

Names/Pronouns and Self Identifications

The University of Maryland recognizes the importance of a diverse student body, and we are committed to fostering equitable classroom environments. I invite you, if you wish, to tell us how you want to be referred to both in terms of your name and your pronouns (he/him, she/her, they/them, etc.). The pronouns someone indicates are not necessarily indicative of their gender identity. Visit trans.umd.edu to learn more.

Additionally, how you identify in terms of your gender, race, class, sexuality, religion, and dis/ability, among all aspects of your identity, is your choice whether to disclose (e.g., should it come up in classroom conversation about our experiences and perspectives) and should be self-identified, not presumed or imposed. I will do my best to address and refer to all students accordingly, and I ask you to do the same for all your fellow Terps.

Grades

Grades are not given, but earned. Your grade is determined by your performance on the learning assessments in the course and is assigned individually (not curved). If earning a particular grade is important to you, please speak with me at the beginning of the semester so that I can offer some helpful suggestions for achieving your goal.

If you feel you are struggling at any point in the course, on an individual assignment or topic or with getting work completed in general, please TALK TO ME as soon as possible so we can come up with a plan to help you do better.

All scores as well as comments and feedback (except on Midterm) will be posted on ELMS. If you would like to review any of your grades (including the exams), or have questions about how something was scored, please email me to schedule a time for us to meet in my office.



Data Science Techniques

INST 414

Fall 2019

Unless you have reported an excused absence to me in advance of the due date, late work will be subject to a markdown of 10% per day. So please plan to have it submitted well before the scheduled deadline. Also, you get the flexibility to extend the deadline for one project and one exercise for up to 5 extra days for each one them. However, you must let me know about the extension 24 hours before the deadline of that exercise or project. I am happy to discuss any of your grades with you, and if I have made a mistake I will immediately correct it. Any formal grade disputes must be submitted in writing and within one week of receiving the grade.

Category	Assignment	Points
Quizzes	Quiz 1 (Sep 5 from 5:30pm - 5:35pm)	1
	Quiz 2 (Sep 12 from 5:30pm - 5:35pm)	1
	Quiz 3 (Sep 19 from 5:30pm - 5:35pm)	1
	Quiz 4 (Sep 26 from 5:30pm - 5:35pm)	1
	Quiz 5 (Oct 3 from 5:30pm - 5:35pm)	1
	Quiz 6 (Oct 10 from 5:30pm - 5:35pm)	1
	Quiz 7 (Oct 17 from 5:30pm - 5:35pm)	1
	Quiz 8 (Oct 24 from 5:30pm - 5:35pm)	1
	Quiz 9 (Oct 31 from 5:30pm - 5:35pm)	1
	Quiz 10 (Nov 7 from 5:30pm - 5:35pm)	1
	Quiz 11 (Nov 14 from 5:30pm - 5:35pm)	1
	Quiz 12 (Nov 21 from 5:30pm - 5:35pm)	1
Critical Reading Assignments	Reading 1 (Sep 13 th at 6pm)	3
	Reading 2 (Sep 20 th at 6pm)	3
	Reading 3 (Oct 4 th at 6pm)	3
Exercises	Exercise 1 (Sep 6 th at 6pm)	4
	Exercise 2 (Sep 27 th at 6pm)	4
	Exercise 3 (Oct 11 th at 6pm)	4
	Exercise 4 (Nov 8 th at 6pm)	4
	Exercise 5 (Nov 22 nd at 6pm)	4
Portfolio Projects	Project 1 (Oct 15 th at 6pm)	8
	Project 2 (Nov 5 th at 6pm)	8
	Project 3 (Nov 19 th at 6pm)	8
	Project 4 (Dec 5 th at 6pm)	8
Final Project	Problem Statement (Sep 24 th at 6pm)	5
	Data Collection, Formatting, and Cleaning (Oct 22 nd at 6pm)	5
	Model selection, Analysis & Interpretation (Nov 26 th at 6pm)	10
	Final Paper (Dec 16 th at 11:59pm)	5



Data Science Techniques

INST 414

Fall 2019

Exams	Midterm (Oct 24 th 7pm – Oct 30 th 11:59pm)	10
Extra Credits		5
Total		105

Final letter grades are assigned based on the percentage of total assessment points earned. To be fair to everyone I must establish clear standards and apply them consistently, so please understand that being close to a cutoff is not the same as making the cut (89.40 \neq 90.00). It would be unethical to make exceptions for some and not others.

Final Grade Cutoffs							
A+	97-100*	B+	87-89.99	C+	77-79.99	D+	67-69.99
A	93-96.99	B	83-86.99	C	73-76.99	D	63-66.99
A-	90-92.99	B-	80-82.99	C-	70-72.99	D-	60-62.99
		F	<60.0%				

* Note: To receive an A+ you must have demonstrated significant contributions to the class in addition to achieving this numeric grade.

Extra Credit

Will be discussed in the following weeks

Course Schedule

Week	Date	Topic	Activities	Readings (Due BEFORE Class)
1	8/26	Syllabus, Introductions	Discussion Threads: Introductions, What is Data/Data Science? Skills Search	None
	8/29	The Data Science Process		“What is Data Science?” Article
2	9/6	Data Collection	Web Scraping with Python	Files and Links in week 2 Module on ELMS
3	9/9	Data Collection and Code Deployment on Cloud	Scraping with Face Pager and Deploying the scraper on cloud for Data	Files and Links in week 3 Module on ELMS



Data Science Techniques

		Collection		
	9/12	Data Cleaning, Cognitive Bias, Critical Thinking	Cleaning dataset, Project Sign Up	DSH Chapter 4 and Article links on week 3 Module on ELMS
4	9/16	Data Formats and Manipulation, Feature Engineering	Implementing on a Real-World Dataset	DSH Chapters 4.6 & 7
	9/19	Regular Expressions		Article links in week 4 Module on ELMS and DSH Chapter 5
5	9/23	Text Analytics, NLP		DSH Chapter 16
	9/26	Text Analytics (Naïve Bayes Classification and Topic Modeling using LDA)		Article links in week 5 Module on ELMS
6	9/30	Data Visualization and Storytelling	Exploratory Data Analysis	Article links in week 6 Module on ELMS
	10/3	Machine Learning and Artificial Intelligence	Discussion on Types of Algorithms	Article Links in week 6 Module on ELMS
7	10/7	Machine Learning Algorithms, Metrics, Overfitting		DSH Chapter 3.3 and Article Links in week 7 Module on ELMS
	10/10	Probability & Distributions Review		DSH Chapters 18 & 19 (Skim)
8	10/14	Introduction to Gradient Boosting		Article Links in week 8 Module on ELMS
	10/17	Introduction to XGBoost		Article Links in week 8 Module on ELMS
9	10/21	Data Preparation, Evaluation & Visualizing XGBoost Models	Implementing the models along the instructor	



Data Science Techniques

	10/24	Midterm Review		
10	10/28	XGBoost Implementation for Regression and Classification		
	10/31	Advanced XGBoost		
11	11/4	Introduction to WEKA		
	11/7	Machine Learning with WEKA		
12	11/11	Machine Learning with WEKA cont'd		
	11/14	Machine Learning with WEKA cont'd		
13	11/18	Network Analysis	Analyzing Social Media Networks	Article Links in week 13 Module on ELMS
	11/21	Network Analysis	Analyzing Social Media Networks	Article Links in week 13 Module on ELMS
14	11/25	Ethics in Data Science		Article Links in week 14 Module on ELMS
	12/5	Overview of Deep Learning		Article Links in week 14 Module on ELMS

Note: This is a tentative schedule, and subject to change as necessary – monitor the course ELMS page for current deadlines. In the unlikely event of a prolonged university closing, or an extended absence from the university, adjustments to the course schedule, deadlines, and assignments will be made based on the duration of the closing and the specific dates missed.