INST 414 – Data Science Techniques

Catalog Description

Pre-requisite: INST 314 Statistics for Information Science

Restrictions: Must be in BSIS program; Permission of BSIS program.

This course will explore approaches to extract insights from large-scale datasets. The course 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, and exams, students will practice working with these techniques and tools to extract relevant information from structured and unstructured data.

Extended Course Description This course explores the application of data science techniques to unstructured, real-world datasets including social media and open data sources. The course 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.

This course will explore approaches to extract insights from large-scale datasets. The course 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.

Student Learning Outcomes: Upon completing this course, students 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
- Interpret the results coming out of the models
- Critically evaluate the accuracy of different algorithms and the appropriateness of a given approach

Textbooks and Readings

Think Stats 2e
http://greenteapress.com/wp/think-stats-2e/

Data Science for Business

http://shop.oreilly.com/product/0636920028918.do

https://www.amazon.com/Data-Science-Business-Data-Analytic-Thinking/dp/1449361323

Learn Python the Hard Way (Recommended if new to Python)

https://learnpythonthehardway.org/

Other course materials will include class notes and slides provided by the instructor on course webpage.

Required Technology and Background

- **Laptop –** We will do live exercises in class. Please bring your laptop to class. If you do not have one, contact the professor before the first class.

- **Python Software**. Freely available at:

  https://conda.io/docs/user-guide/install/index.html#regular-installation

Mathematical maturity: We will work extensively with probabilities and mathematical functions such as logarithms and differentiation. You should be comfortable manipulating these concepts algebraically. You should also be able to argue why mathematical statements are true.

We will make extensive use of the Python 3 programming language. **It is assumed that you know or will quickly learn how the program in Python. There will be only a brief introduction to this skill-set.** You will need to be able to understand object oriented programming in Python. You can satisfy this requirement by completing a programming course that uses Python and object-oriented techniques.

The computer-based aspects of this course will be oriented toward Unix-like operating systems (Linux, OS X). It may be possible to complete the course using other operating systems, but you will be responsible for troubleshooting any issues you encounter.

Course Activities

a. Homework Assignments

You will have periodic assignments that are designed to provide you with practice and to reinforce the topics and techniques covered previously in the course. These assignments will provide students with feedback on your understanding of the concept covered in the course. The goal of homework assignments is for students to check their comprehension and target areas for improvement.

You may work with your classmates to figure out the underlying concepts but are expected to work *individually* to answer the specific problems that are assigned. Timely submission of the completed assignments is essential. The due date of each assignment will be stated clearly in the assignment description. If an assignment due date is a religious holiday for you, please let the instructor know at least one week in advance, so an alternate due date can be set.

Assignments more than two days late may not be graded.

Assignments are not worth the same number of points: some will be more difficult than others.

b. Group Project

For your group project you will form teams of 3-5 people and prepare a data-related analytic project. The project will be graded on your ability to articulate an appropriate question, prepare the data, identify and perform reasonable methodology and study design, justify the appropriateness of certain machine learning approaches, articulate and conduct evaluations, analyze and interpret the results and create appropriate visualizations. You will be required to analyze your dataset using Python.

c. Exams

In this course, the assignments provide you with opportunity to experiment with and learn about the ideas, concepts, and techniques associated with data science. The exams complement this by providing you with feedback about how well you have learned them and whether you have successfully developed the ability to apply those concepts and techniques in different settings.

There will be a project-based midterm. The midterm will require students to demonstrate their understanding of the material in the previous lectures. The project will be graded on your ability to articulate an appropriate question, prepare the data, identify and perform reasonable methodology and study design, justify the appropriateness of certain machine learning approaches, articulate and conduct evaluations, analyze and interpret the results and create appropriate visualizations. You will be required to analyze your dataset using Python.

We will discuss the full vision for the midterm during class.

The final exam will build on the midterm project, and will give students an opportunity to synthesize concepts from the entire course.

d. Participation

Each class is critical to your learning experience, and I expect you to come to class prepared (having read all assigned readings and ready to engage). I also expect active participation, not passive reception of the material. Your energy in contributing to class discussions and hands-on exercises will make this class an enjoyable experience for all of us.

Grading

Your final grade for the course is computed as the sum of your scores on the individual elements below (100 possible points total), converted to a letter grade:

A+ 97-100*	**B+ 87-89.99**	**C+ 77-79.99**	**D+ 67-69.99**	**F 0-59.99**
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**	

*** Note: To receive an A+ you must have demonstrated significant contributions to the class in addition to achieving this numeric grade. We reserve the right to curve grades upward (but will not curve grades downward).**

COURSE SCHEDULE

Week	Topics
1	Introduction to Data Science and Python
2	Computer Science Basics w/ Python
3	Probability review 1
4	Probability review 2
5	Discrete and Continuous Distributions
6	Maximum Likelihood Estimation
7	Data wrangling

8 Data Visualization

9 Linear Regression

10 Logistic Regression

11 Feature Engineering

12 Clustering

13 Support Vector Machines
14 Neural Networks
(Optional)

15 Ensemble Methods
(Optional)

This schedule is for planning purposes and may change. See course webpage for current information and deadlines.

Policy on Academic Misconduct Cases of academic misconduct will be referred to the Office of Student Conduct irrespective of scope and circumstances, as required by university rules and regulations. It is crucial to understand that the instructors do not have a choice of following other courses of actions in handling these cases. There are severe consequences of academic misconduct, some of which are permanent and reflected on the student's transcript. For details about procedures governing such referrals and possible consequences for the student please visit http://osc.umd.edu/OSC/Default.aspx.

It is very important that you complete your own assignments, and do not share any Excel or SPSS files or other work. The best course of action to take when a student is having problems with an assignment question is to contact the instructor. The instructor will be happy to work with students while they work on the assignments.

University of Maryland Code of Academic Integrity

"The University of Maryland, College Park has a nationally recognized Code of Academic Integrity, administered by the Student Honor Council. This Code sets standards for academic integrity at Maryland for all undergraduate and graduate students. As a student you are responsible for upholding these standards for this course. It is very important for you to be aware of the consequences of cheating, fabrication, facilitation, and plagiarism. For more

information on the Code of Academic Integrity or the Student Honor Council, please visit http://shc.umd.edu/SHC/Default.aspx.

Special Needs

Students with disabilities should inform the instructor of their needs at the beginning of the semester. Please also contact the Disability Support Services (301-314-7682 or http://www.counseling.umd.edu/DSS/). DSS will make arrangements with the student and the instructor to determine and implement appropriate academic accommodations. Students encountering psychological problems that hamper their course work are referred to the Counseling Center (301-314-7651 or http://www.counseling.umd.edu/) for expert help.