COLLEGE OF INFORMATION STUDIES
MARYLAND'S iSCHOOL

INST 414 – Advanced Data Science

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, 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

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
- Probability basics
  - Recognize when a function is a valid probability distribution
  - Convert real-concepts to probability distributions
- Conditional probabilities
  - Manipulate conditional and marginal probability distributions
  - Apply Bayes rule to invert conditional probabilities
- Sampling from distributions
  - Describe distributions using standard parameterizations of
  - Use a random number generator to sample from the distribution
- Maximum likelihood estimation

    - Given data and an objective function, derive the form of the maximum likelihood parameters
- Linear Regression
- Logistic Regression
    - Given a parameterization of logistic regression classifiers, output a classification on an example
    - Given training examples, use stochastic gradient to update classifier parameters
- Feature engineering, feature combination
    - Given a regression or classification model, identify data representation failures
    - Correct problems of data representation to improve classification or regression
    - Design training / test data splits that allow effective evaluation of data science algorithms
- Clustering
    - Recognize when a problem can benefit from classification
    - Given a dataset, run k-means clustering or Gaussian mixture models
- SVMs
    - Evaluate a margin-based classifier given a dataset whether it's effective
    - Distinguish when a margin-based classifier needs slack variables
- 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/

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

Every other week you will have an assignment that is designed to assess your mastery of the topics and techniques covered the previous two weeks and provide feedback to improve your understanding of the material. Your lowest homework score will be dropped and your score on the remaining homework assignments will be used to calculate your final course grade.

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.

Late policy: each person has five free late days to be used, no questions asked, during the course (late days can only be used in increments of one day; if your assignment is three minutes or three hours late, that counts as one late day). When turning in a late assignment, clearly mark at the top that you are using a late day. After you use your late days, late assignments will get half credit at grader's discretion. 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. This involves identifying a question, finding or developing a dataset, creating appropriate measures, conducting analyses, and preparing an appropriate information product based on the results. The purpose of the project is to have you go through the steps and processes completing a high-impact data analytics 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 an in-class midterm. The midterm will cover material in the previous lectures and will be open notes (but closed book).

The final exam will be an exam similar to the midterm's structure but comprehensive over the entire course's content. The exam will be held at the time specified by the University unless otherwise stated.

d. Participation

Each class is critical to your learning experience, and I expect you to come to class prepared (having read all assigned readings, viewed the lecture, 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.

We will also be using the online learning platform Piazza. You can get credit for participation by answering and asking useful questions on that platform. Ideally you should be participating both online and in class, however.

e. Quizzes

Quizzes are designed to be easy and to reward doing the reading and attending class. I will drop your lowest quiz scores. Note that this syllabus does not mention the total number of quizzes or the number of scores that will be dropped. You should not use information in Moodle or past iterations of the course as an attempt to guess what will happen with quizzes. Students who have done this in the past have been sorely disappointed, as my goal is to make quizzes unpredictable. Quizzes are also used as a proxy for attendance. Thus, there is no opportunity for making up quizzes (and I don't have the ability to judge what is an excusable absence or not for a large class). Have no fear, however, as I will drop the lowest quiz grades: as long as you do well on a majority of quizzes, you'll be fine.

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).

Graded item	**Percent of final grade**
Homeworks	30
Quizzes / Participation	10
Group Project	20
Midterm	20
Final	20

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
15	Ensemble Methods

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.