ASTR 5900-001: Machine Learning
MW Nielsen Hall Rm 103 3:00-4:15

Instructor:	Prof. Karen M. Leighly
Department:	Physics and Astronomy
Office:	243 Nielsen Hall
Phone:	325-7045
Voice Mail:	325-7045
Email:	leighly@ou.edu
Teaching Assistant / Collaborator	Alex Kerr
Office Hours:	NH 219: Tuesdays 1:30-2:30, Fridays 1:30-2:30 NH 103: Fridays 2:30-4:00

Recommended (not required) Texts:

1. *Data Analysis: A Bayesian Tutorial*, Second Edition, D. S. Silva & J. Skilling
   This book is a very readable introduction to the fundamentals of Bayesian statistics. (Amazon: $41.12)
2. *Pattern Recognition and Machine Learning*, Christopher M. Bishop.
   This is a clear but rather mathematical textbook. The hardcover version is recommended over the paperback (international) version, as it has color figures (and the figures are very good) (Amazon: $43.42)
3. *Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for Analysis of Survey Data,* by Zeljko Ivezic, Andrew J. Connolly, Jacob T. VanderPlas & Alexander Gray.
   This is more of a reference book rather than a textbook. It includes a lot of stuff, as well as example code, so it provides a good starting point for many types of data analysis. Note: OU students and faculty have access to the ebook through the OU library (for online reading, not for download). (Amazon: $51.45)

Motivation:

In the 21st century, astronomy and astrophysics, as well as other scientific fields, are inundated with huge quantities of data. These data sources include not only data from surveys, but also, with the advent of faster, more powerful computers, from simulations. This huge flow of data only expected to increase; for example, the Large Synoptic Survey Telescope is expected to generate terabytes of data per night. Facilities such as LOFAR generate so much data that it is only feasible to save data products, and the raw data is not stored.

Astronomers need sophisticated techniques to analyze these large and complex data sets. So the increase in data flow has been accompanied by a similar increase in interest in statistical and machine learning techniques. For example, national conferences such as those hosted by the American Astronomical Society organize large sessions devoted to software and techniques supporting data analysis.

Statistical, computational, and machine learning techniques are one of the most desirable and potentially transferable skill sets obtainable. Researchers who master these techniques may be able to apply them to fields beyond astronomy, such as finance, pattern recognition, computational, and survey applications.

Course Description and Approach:

This course offers an exploratory, interactive, and applied approach to the field of statistics and machine learning. Applications will be principally agnostic with respect to field, although some examples will be drawn from the fields of astrophysics and physics.

The class is designed to provide students with practical knowledge and experience using these techniques. An overall goal is to help students learn how to ask questions of data, and to then answer their questions using modern statistical and analytical methods. By the end of the course, the student will understand the techniques well enough, and gain enough practical understanding to be able to determine whether a technique will be useful for his or her own research problem, and will have sufficient experience to have a basis from which to begin to implement it. Theoretical development will also be included, but to a lesser degree.

Class Design and Structure:

The first two weeks will focus on the basics that we need for the course: python programming and probability and statistics. Also important is linear algebra, but it is assumed that everyone has had an undergraduate-level class on this topic.

The heart of the course is 6 one or two week modules, each covering a particular topic. Each week will focus on a technique or related set of techniques. The first class period (Wednesdays) will include a formal lecture on the basic theoretical underpinnings. Examples will be included, and homework (see below) will be assigned, due the following Wednesday. The second class period (Mondays) will be open for questions, discussion, and help on the homework or projects, primarily, but may also be devoted to special topics and examples from the literature.

The main portion of the course will culminate in a midterm exam on **November 20**. The exam will focus on conceptual understanding of the techniques that have been discussed (see below).

Projects will also form an important part of this course. Assignments related to the projects will be due throughout the semester (see below), with a presentation and paper due at the end of the semester.

Class website:

This class will be conducted principally through Github Classrooms. The lectures, homework, etc will be posted there (in a public repository), and you will submit your homework to private repositories associated with the class. (A separate document will give instructions how to do that.) Canvas will be used for grades and materials that cannot be posted to a public repository.

Lectures:

The main lecture on each topic will be presented in the form of a Jupyter (iPython) notebook. These will include theoretical development and examples. These will be uploaded to the github website after each lecture, and will be available for downloading after each class. Format for supplemental lectures / examples will be variable.

This class will use python as the programming language. Recognizing that not all students have a background in python, an introduction to the programming language will be given. Note that python is fast becoming a favored scientific programming language, so proficiency is arguably a desirable outcome.

It is intended that the main content of the course will be generic in terms of discipline, so although many of the students in the class are studying astronomy, it should be accessible physics and other students. However, examples may be drawn from astronomy.

Homework:

The homework will be assigned on the day of the relevant lecture, and due (mostly) one week later. They will be predominately in the form of Jupyter (iPython) notebooks. Questions will range from specific (e.g., filling in steps from examples presented in lecture) to open-ended research-type questions (measure some property from a given data set).

Students may work together on the homework, but each student must hand in his or her own and original work.

Homework may be handed in late, with a loss of 10% of the grade per day. The exception will be in the case of documented illness / other excused absence.

Midterm Exam:

In this class, we will study a number of techniques and methods that will be useful for data analysis. Each technique will work well in some circumstances and poorly in others. Each technique has certain built-in assumptions and limitations. In order to effectively use these techniques in your research, it is important to know and understand the strengths, weaknesses, and limitations.

The midterm exam (November 20) will be entirely conceptual. E.g., it may include specific data analysis problems and descriptions, and you will be asked to determine what technique would be useful to use, or whether a conclusion drawn from some particular data may be faulty.

Project:

The project offers the opportunity to apply one of the analysis techniques studied to your own or other data, learn about and apply another technique not covered in the class, or delve deeper in to some technique. More information will be provided about the project later.

The projects will be done individually or in groups of no more than two. If the project is done in as group, the contribution of each individual must be clearly delineated (e.g., each person may apply two different techniques to the same data set). Correspondingly, the paper and presentation will be twice as long.

The timeline for the project will be as follows:

- **Before October 13:** Meet with Alex and Dr. Leighly to discuss your idea (10%)
- **October 13:** A 300 word abstract is due (10%)
- **November 10:** Detailed outline/draft/preliminary results due (20%)
- **December 4, 6, and 13:** Presentation – 12 minutes+3 minutes questions x number of people (30%) and a short paper (5 x number of people pages) (30%) due at the time of the presentation.

Tentative Schedule:

The *tentative* schedule follows:

Date	Topic	Comments
August 21	Introduction	Syllabus, class motivation, computing platform
August 23	Python	Python will be used in this course. It is becoming an extremely common programming language, in astronomy and beyond.
August 28	Probability	Data is probabilistic
August 30	Distributions	Data is distributed according to probability distributions
September 6	1. Statistical Inference	Classical and Bayesian methods for understanding data will be discussed.
September 11		
September 13		
September 18		
September 20	2. Markov Chain Monte Carlo	Markov Chain Monte Carlo modeling provides a concrete example of the power of Bayesian statistics.
September 25		
September 27	3. Cluster Analysis	Unsupervised classification (K-means and Gaussian mixture models)
October 2		
October 4		
October 9		
October 11	4. Regression and Principal Components Analysis	Linear Regression, least squares, curve fitting, decision theory, information theory, dimensionality reduction
October 16		
October 18		
October 23		
October 25	5. Classification	Supervised Classification (e.g., neural net)
October 30		
November 1		
November 6		
November 8	6. Time Series and	Structure Function and Fourier Techniques

November 13	Spatial Analysis	
November 15		
November 20		
November 27	Miscellaneous topics of interest / makeup	
November 29		
December 4	Projects and Presentations	Students will present results of their presentations. Papers will be due at the time of the presentation.
December 6		
December 13		

Evaluation:

Participation	10%
Homework, based on Jupyter notebooks	40%
Project and Presentation	30%
Midterm Exam	20%

Absences: This is a class with a large component to participatory student activities that will be done during class. It is therefore expected that students will attend class unless prevented by, e.g., illness or professional travel. Excessive absences will result in a lower participation grade. In addition, OU policy states that the following comprise excused absences:

- Provost-approved university-sponsored activities such as scholarly competitions, fine arts performances, and academic field trips.
- Legally required activities, such as emergency military service and jury duty.
- Religious holidays.

Ethics: To protect the integrity of your grade in this class and ultimately the integrity of any degree that you may receive from OU, as long as you are enrolled in this class you are expected to conform to an honor code which includes no cheating on tests or any other assignments or exercises in which collaboration or outside help is forbidden. For the record, cheating includes looking on other students' papers or copying answers from other students during an exam, talking to other students during an exam, using any kind of unauthorized auxiliary material (cheat sheets, books, writing on body parts, electronic devices) during an exam, and having someone other than yourself take an exam or answer a clicker question for you. Basically, cheating is representing someone else's work as your own. Please read information about academic misconduct at OU at http://integrity.ou.edu/files/Academic_Misconduct_Code.pdf.

Disabilities: Students requiring academic accommodation should contact the Disability Resource Center for assistance at (405) 325-3852 or TDD: (405) 325-4173. For more information please see the Disability Resource Center website http://www.ou.edu/drc/home.html Any student in this course who has a disability that may prevent him or her from fully demonstrating his or her abilities should contact me personally as soon as possible so we can discuss accommodations necessary to ensure full participation and facilitate your educational opportunities.

Religious holidays: It is the policy of the University to excuse absences of students that result from religious observances and to provide without penalty for the rescheduling of examinations

and additional required classwork that may fall on religious holidays. Students that plan to observe a religious holiday should notify me ahead of time (at least one week; preferably at the beginning of the semester) so that appropriate rescheduling can be made.

Title IX Resources and Reporting Requirement: For any concerns regarding gender-based discrimination, sexual harassment, sexual assault, dating/domestic violence, or stalking, the University offers a variety of resources. To learn more or to report an incident, please contact the Sexual Misconduct Office at 405/325-2215 (8 to 5, M-F) or smo@ou.edu. Incidents can also be reported confidentially to OU Advocates at 405/615-0013 (phones are answered 24 hours a day, 7 days a week). Also, please be advised that a professor/GA/TA is required to report instances of sexual harassment, sexual assault, or discrimination to the Sexual Misconduct Office. Inquiries regarding non-discrimination policies may be directed to: Bobby J. Mason, University Equal Opportunity Officer and Title IX Coordinator at 405/325-3546 or bjm@ou.edu. For more information, visit http://www.ou.edu/eoo.html.