STAT4620 – 2 CREDIT HOURS
Introduction to Statistical Learning

Term: Fall, 2017
Instructor: Matthew T. Pratola
Email: mpratola@stat.osu.edu
Location: WeFr 9:10-10:05am University Hall 056
Office Hours: Fri 10:30-11:30am CH204D. **I will be available to help answer questions about HOW to perform your chosen analysis, NOT what analysis you should perform.**
Final Exam: Wednesday December 13th, 8:00-9:45am.
Course Website: Carmen (Canvas)
Long course title: Introduction to Statistical Learning

Course description:
The course provides an introduction to the principles of statistical learning and standard learning techniques for regression, classification, clustering, dimensionality reduction, and feature extraction. An outline of topics is:

1. Overview of predictive modeling and model evaluation
2. Penalized regression and nonparametric regression
3. Nearest neighbor methods
4. Classification and regression trees
5. Model selection and validation
6. High-dimensional data and variable selection

Expected Learning Outcomes:
Upon successful completion of the course, students will be able to:

1. Recognize the types of learning problems and understand their statistical formulations.
2. Understand the foundational principles of statistical learning including statistical modeling, computation and evaluation.
3. Comprehend the rationale and algorithms behind statistical learning techniques and know their relative merits and limitations.
4. Evaluate and compare different learning techniques numerically in terms of generalization error.
5. Use statistical learning methods for data analysis and interpret the results in the context of the data problem.

Course Prerequisites:
C- or better in STAT3302 (Statistical Modeling for Discovery II)

Textbooks: The required textbook for the course is *An Introduction to Statistical Learning with Applications in R* by G. James, D. Witten, T. Hastie and R. Tibshirani. The book is available for purchase at the official University bookstore (`ohiostate.bkstore.com`) and is also available for free online in PDF format at `http://www-bcf.usc.edu/~gareth/ISL/`.

Course Requirements:
Students will be required to use the `R` software environment for statistical computing and graphics. `R` can be downloaded for free at `http://www.r-project.org`. Instructions for using the software will be given in class. Many students prefer to use `RStudio`, an IDE deisgned for use with `R`. `RStudio` is available for free at `http://www.rstudio.com`.

Students are responsible for all material covered in class, in the assigned readings and in homework problems. As an introductory course, the quantity of material covered in the lectures is extensive. It is highly recommended that you do not fall behind.

Assignments:
Homework will be assigned (approximately) bi-weekly, will be due on the dates announced in class and will be graded. Assignments will consist of a mix of technical questions to assess students' understanding of the statistical models, and questions asking students to perform analyses of datasets. The grade for the analysis portion of each assignment will be based on both the accurateness and appropriateness of the analysis, as well as the clarity of the description of the analysis and results.

Tentative due dates for the assignments are shown below. The assignments and assignment solutions will be posted on Carmen (Canvas). You must show all your work for all homework problems; do NOT just write the final answer.

Homeworks will be completed in teams of 2 students, however you must work with a different student for each homework (i.e. no pair of students can work together to submit more than one homework). You are encouraged to discuss problems with each other in general terms, but each team must write their own homework solutions. Homework reports must be submitted in hardcopy. Late submissions will NOT be accepted.

Project:
Students will be responsible for completing a team project. Each team will consist of 3 students. Proposals for project ideas will be due mid-way through the semester, and the project will be due near the end of the semester. The project will consist of formulating questions that can be answered with the data, and performing an appropriate analysis to answer the questions.

Exams:
There will be two in-class midterms that cover material from lecture, the assigned readings and homework. A cumulative final examination will be given during the university's examination period. A basic calculator will be necessary for all exams (no cell phone calculators or PDAs). Cellphones must be silenced during class and are not allowed to be on the desk or otherwise accessible during exams. No make-up exams will be given.

Dates:
Homework due dates and midterm dates are tentatively as follows (please refer to in-class announcements for official dates):
HW1 09/08; HW2 09/22;
Midterm I 09/27 (in class);
HW3 10/11; HW4 10/25;
Midterm II 10/27 (in class);
HW5 11/17;
Project 11/29
Last time I checked, the last day to drop the course without a 'W' appearing on your record is September 15th and the last day to drop a course without petition is October 27th. However, please refer to the OSU registrar's

office for official drop guidelines in case these dates change.

Grading:
The final course grade will be based on homework assignments, a project, two midterms and a comprehensive final examination. The weights for each component of the grades are: **15%HW, 20%Midterm 1, 20% Midterm 2, 15% Project, 30% Final**.

(Tentative) Schedule of Topics:

Class	Date	Section	Topic
1	23-Aug	Ch1, Ch2, 3.1,3.2,3.5	Introduction; Linear Regression: SLR,MLR,Geometry and Loss
2	25-Aug	3.2,3.5	Linear Regression continued; Weighted Least Squares
3	30-Aug	3.3 and notes (7.1,7.2)	Beyond Linear Regression; **HW1 posted**
4	1-Sep	4.1,4.2,4.3	Classification: Logistic Regression
5	6-Sep	4.4,4.5	Classification: Linear Discriminant Analysis (LDA) **HW2 posted**
6	8-Sep	5.1	Cross-Validation
7	13-Sep	5.2	The Bootstrap
8	15-Sep	5.3	CV & Bootstrap Examples
9	20-Sep	6.1,6.2.1	Regularization: Ridge Regression
10	22-Sep	6.2.2,6.2.3	Regularization: The LASSO
-	27-Sep	**Midterm 1**	**Midterm 1**
11	29-Sep	6.3	****Regularization: Dimension Reduction/PCA **HW3 posted**
12	4-Oct	6.4	High Dimensional Data Analysis and the Curse of Dimensionality
13	6-Oct	7.3,7.4	Spline Regression
14	11-Oct	7.5	Smoothing Splines **HW4 posted**
15	18-Oct	notes	****Gaussian Processes
16	20-Oct	7.6	Local Regression
17	25-Oct	7.7	Generalized Additive Models (GAM's) (regression & classification)
-	27-Oct	**Midterm 2**	**Midterm 2**
18	1-Nov	8.1	Regression and Classification Trees
19	3-Nov	8.2	Trees: Bagging, Boosting **HW5 posted**
20	8-Nov	8.2	Trees: Random Forests
21	15-Nov	10.3	Clustering: K-means
22	17-Nov	10.3	Clustering: Hierarchical
23	29-Nov	-	Project Presentations
24	1-Dec	9.1,9.2	****Support Vector Machine
25	6-Dec	9.3,9.4	****Support Vector Machine

**** topics to be covered only if time permits.

Academic Misconduct:

ACADEMIC MISCONDUCT OF ANY SORT WILL *NOT* BE TOLERATED.

It is the responsibility of the Committee on Academic Misconduct to investigate or establish procedures for the investigation of all reported cases of student academic misconduct. The term "academic misconduct" includes all forms of student academic misconduct wherever committed; illustrated by, but not limited to, cases of plagiarism and dishonest practices in connection with examinations. Instructors shall report all instances of alleged academic misconduct to the committee (Faculty Rule 3335-5-487). For additional information, see the Code of Student Conduct http://studentlife.osu.edu/csc/.

Special Accomodations:

Students with disabilities that have been certified by the Office for Disability Services will be approporiately accommodated and should inform the instructor as soon as possible of their needs. The Office for Disability Services is located in 150 Pomerene Hall, 1760 Neil Avenue, telephone 292-3307, TDD 292-0901 (or see http://www.ods.ohio-state.edu/).