STAT 6500 – Statistical Machine Learning

Term: Spring 2019
Lecture: MWF 11:30AM–12:25PM in CH 240

Instructor: Yoonkyung Lee
Office: 305B Cockins Hall
Office Hours: T 10:30–11:25AM F 2:00–2:55PM or by appointment
Email: yklee@stat.osu.edu

Grader: Shanshan Tu
Office: 454 Mathematics Building
Office Hours: by appointment only
Email: tu.162@osu.edu

Course Website: `https://carmen.osu.edu`

Course Description:
Statistical Machine Learning explores the methodology and algorithms behind modern supervised and unsupervised learning techniques to explore relationships between variables in large, complex datasets. Topics include linear and logistic regression, classification, clustering, resampling methods, model selection and regularization, and non-linear regression. Students will also gain exposure to popular statistical machine learning algorithms implemented in R. A focus will be on understanding the formulation of statistical models and their implementation, and the practical application of learning methods to real-world datasets.

Expected Learning Outcomes:

- Students will understand the statistical learning framework, including core concepts such as loss, learning, and generalization; they will be able to judge when the framework is applicable and be able to formulate problems within this framework.
- Students will recognize the role of statistical models that are appropriate for a variety of statistical learning problems; they will understand the assumptions, formulation, and evaluation of these models.
- Students will understand the rationale and algorithms behind statistical learning methods, and they will know the merits and limitations of these methods.
- Students will be able to quantitatively evaluate and compare different statistical learning methods.
- Students will be able to apply statistical learning methods for data analysis and be able to interpret the results in the context of the application.

Course Prerequisites:
STAT 6450 (Applied Regression Analysis) or permission of instructor. Familiarity with calculus, linear algebra and linear regression analysis will be assumed. Basic proficiency in a programming language, such as `R` is required.

Computing and Software:
One of the goals of the course is to train students in the computing aspects of statistical machine learning and develop the skills to implement machine learning algorithms. Many homework assignments will have a computing and programming component (knowledge of software packages such as `Stata`, `SAS` or `JMP` will *not* be sufficient). There will be example codes provided, primarily written in the language `R`.

Textbooks:
The following textbooks are required for the course and are available to purchase at the university bookstore.

- James, Witten, Hastie, Tibshirani: *An Introduction to Statistical Learning with Applications in R.* (Freely downloadable PDF available at `http://www-bcf.usc.edu/~gareth/ISL`)
- Murphy: *Machine Learning: A Probabilistic Perspective* (An electronic version is available for online reading through the OSU library website)

Coursework:

- **Homework:** 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. A tentative schedule for assignments is shown below. The assignments and assignment solutions will be posted on the course website. Late submissions will NOT be accepted.
- **Midterm:** A take-home midterm exam will be given. The midterm will be completed by each student individually.
- **Group Project:** Students will also complete projects in groups consisting of 4 to 5 members (depending on the enrollment size). The project will consist of selecting a data set (by week 4), performing an exploratory data analysis (EDA, by week 6), making a 5 page proposal (by week 10), presenting (in week 15), and submitting a 10 page final report (by April 30). The proposal should contain a detailed problem statement that includes questions of interest and a description of what methods will be used and how they will be used to answer questions of interest or solve the problem. More details will be given in class.

Tentative Schedule:
Homework due dates and midterm dates are tentatively as follows. Please refer to in-class announcements for official dates.

Week	Description	Week	Description
3	HW1	9	Take-home midterm due
4	Project dataset selection	10	Project proposal
5	HW2	11	HW4
6	Project EDA	13	HW5
7	HW3	15 (April 26)	Project presentation
8	Take-home midterm assigned	April 30	Project final report

Grading:
Grades will be assigned on the basis of the following components.

- Homework (40%)
- Take-home midterm exam (30%)
- Group project (30%)

Tentative Schedule of Topics:

Week	# Lectures	Topic	Description
1	1–2	Overview	Intro to statistical learning
1	1	Linear regression	
2–3	5	Classification	Logistic regression, Gaussian LDA, algorithms for fitting
4	2	Resampling methods	Cross-validation, bootstrap
5–6	6	Model selection and regularization	Overfitting, variable selection, penalization, ridge regression, sparse linear models, LASSO, coordinate descent
7–8	5	Basis function models	Basis expansions, smoothing splines, additive models, backfitting, sparse additive models
8-9	4–5	Support vector machines and kernels	Max margin classification, separating hyperplanes, the kernel trick, comparison with earlier methods, nonlinear decision boundaries, perceptron algorithm
10	3	Tree-based methods	Classification and regression trees, variable importance measures
11–12	4–5	Bagging and boosting	Bagging, random forests, boosting, forward stage-wise additive modeling
12–13	4–5	Unsupervised learning	k-means, Gaussian mixture models and EM, PCA
14	3	Neural networks	Feed-forward NN (multilayer perceptrons), convolutional neural networks, back propagation

Academic Misconduct:
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 Accommodations:

The University strives to make all learning experiences as accessible as possible. If you anticipate or experience academic barriers based on your disability (including mental health, chronic or temporary medical conditions), please let me know immediately so that we can privately discuss options. To establish reasonable accommodations, I may request that you register with Student Life Disability Services. After registration, make arrangements with me as soon as possible to discuss your accommodations so that they may be implemented in a timely fashion. SLDS contact information: slds@osu.edu; 614-292-3307; slds.osu.edu; 098 Baker Hall, 113 W. 12th Avenue.