**PM591 -- Machine Learning for the Health Sciences**

****

**Units: 4**

**Spring 2022; Day:** Thursday; **Time:** 1-4PM

**Location:** SSB 114 (but online Jan 12 and Jan 17)

**Instructor:** Juan Pablo Lewinger, PhD

**Office Hours:** By appointment

**Contact Info:** [lewinger@usc.edu](mailto:lewinger@usc.edu)

**Teaching Assistant**: Kaili Ding

**Office Hours:** By appointment

**Contact Info**: kailidin@usc.edu

**IT Help:**For Blackboard support please call 213-740-5555 (24hrs a day 365 days a year) or email blackboard@usc.edu.

|  |
| --- |
|  |

**Course Description***.*

### This course introduces students in the Health Sciences (e.g. Biostatistics, Epidemiology, Health Behavior Research) to Machine Learning methods and their Biomedical applications. It is open to students with a solid quantitative background and a working knowledge of traditional statistical methods and data analysis. The course covers the basic theory, algorithms, and applications of key supervised and unsupervised Machine Learning methods. Topics in supervised learning include linear and logistic regression, support vector machines, and others. Topics in unsupervised learning include principal component analysis and hierarchical clustering. Throughout the course, Machine Learning applications in Biological and health related fields such as Genomics, Environmental Epidemiology, Translational Medicine, and Bioinformatics will be highlighted.

**Learning Objectives**

1. To acquire a repertoire of key supervised and unsupervised Machine Learning methods, with a conceptual understanding of how they work, as well as their scope of application, advantages, and limitations.
2. To develop, validate, and test predictive models with binary and continuous outcomes.
3. To understand the impact of high-dimensionality in Machine Learning tasks and the role of regularization and dimension reduction methods.
4. To become a proficient and judicious user of machine learning software for analyzing health science data.
5. To be able to critically assess the quality of applied Biomedical research that uses Machine Learning methods.

**Prerequisites:** PM511A (or permission from Instructor if student has equivalent background)

**Recommended Preparation**: a) A foundation in traditional data analysis methods including linear regression, b) a working knowledge of basic probability and statistics, including common discrete and continuous distributions (e.g. binomial, normal). Familiarity with a high-level programming language such as R or Python is highly recommended.

**Course Notes**

Students will be expected to read the assigned textbook chapters and sections in advance of the Thursday meetings. On Thursday there will a lecture and discussion to clarify and gain perspective on the topic of the day. The second part of each class will be a Lab devoted to analyzing data examples and working on conceptual problems under the guidance of the teaching Assistant.

**Required Textbook**

***An Introduction to Statistical Learning*: *With Applications in R (second edition)*.** Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2021. Springer Publishing Company (freely pdf version available from: https://www.statlearning.com/)

**Additional resources**

Additional readings from ***The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (ESL).** Trevor Hastie, Robert Tibshirani, Jerome Friedman. 2009. Springer Publishing Company (freely available in pdf format from: <https://web.stanford.edu/~hastie/ElemStatLearn/>

**R for Data Science (RDS).** Garrett Grolemund and Hadley Wickham. 2016. O'Reilly (freely available at <http://r4ds.had.co.nz/index.html>)

**Software**

For analysis and implementation of machine learning methods, the R language and environment for statistical computing and graphics will be used (freely available at

<https://cran.r-project.org/>).

**Description of Assessment**

1. biweekly homework assignments. Assignment questions will involve a mix of data analysis, conceptual questions, and simulation.
2. In-class midterm and final examinations will be open book/open note. The midterm will cover the first half of the course, and the final will be cumulative, covering all course topics.
3. A project where students will work individually or in groups choosing from possible topics ranging from applied (e.g. a novel analysis of existing data) to computational (e.g. small simulation study to compare different machine learning methods) projects. Selected groups will be invited to give a short oral presentation on their project.

**Grading Breakdown**

Homework: 40%

Midterm Exam: 20%

Final Project: 15%

Final Exam: 25%

**Assignment Submission Policy**

Assignments and final project are to be submitted electronically through Blackboard.

**IMPORTANT: Additional Policies**

Late assignments and final project will not be accepted.

The homework assignments are designed to give you an opportunity to practice using Machine Learning methods and getting insight into important conceptual issues.We will often reuse problems from previous versions of the course, **BUT YOU MUST NOT TO COPY, REFER TO, OR LOOK AT PREVIOUS SOLUTIONS IN PREPARING YOUR ANSWERS.**

Also, while we encourage collaboration with other students **on the assignment problems, you should write your own code**. **You must indicate on each homework the students with whom you collaborated. IF YOU COPY YOUR ANSWERS FROM PAST OR CURRENT STUDENTS, WE WILL FIND OUT AND WE WILL APPLY THE APPROPRIATE ACADEMIC SANCTIONS. PLEASE DON’T DO IT! SEE THE PLAGIARISM SECTION BELOW.**

**Tentative Course Schedule (subject to change)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Topics/Activities** | **Readings and Homework** | **Deliverables/ Due Dates** |
| Week 1 Jan 13 | Course overview.  Review of basic statistical tools.  Introduction to R | ISLR: Ch 1, Ch 2.1 Assignment 1 out |  |
| Week 2 Jan 20 | Linear Regression for prediction. Least squares. Prediction error. Training and testing sets. | ISLR: Ch 2.2.1, Ch 3.1-3.2**,** Ch 3.3 |  |
| Week 3 Jan 27 | Model Selection. Validation set. Overfitting. Bias-variance trade off. | ISLR: Ch 2.2.2, Ch 5.1 | Assignment 1 due |
| Week 4 Feb 4 | Classification. Linear Discriminant Analysis (LDA)  Misclassification error.  Binary logistic regression. ROC and AUC. | ISLR: Ch 2.1.3, Ch 6.2  Assignment 2 out |  |
| Week 5 Feb 11 | Quadratic discriminant analysis and KNN classification. Cross-validation | ISLR: Ch 6.4 | Assignment 2 due |
| Week 6 Feb 18 | Time to event outcomes. Cox regression. C-index. | ISLR: Ch 6.3, Ch 10.2  Assignment 3 out |  |
| Week 7 Feb 25 | Regularization. Ridge, Lasso, and Elastic-net, linear, logistic and Cox regression. | ISLR: Ch 9.1.1-9.1.4  Assignment 4 out | Assignment 3 due |
| Week 8 March 4 | **Midterm Exam** |  |  |
| Week 9 March 11 | Support vector machines. Kernels. | ISLR: Ch 9.1.5, 9.2-9.6 |  |
| March 18 | Spring Break**. NO CLASS.** |  |  |
| Week 10 March 25 | Tree based methods part 1  Regression and classification trees. | ISLR: Ch 8.1  Assignment 5 out | Assignment 4 due |
| Week 11 April 1 | Tree based methods part 2.  Bagging, random forests, and Boosting | ISLR: Ch 8.2 |  |
| Week 12 April 8 | Methods for survival outcomes | ISLR: Ch 11  Assignment 6 out | Assignment 5 due |
| Week 13April 15 | Dimension reduction methods. Principal components. | ISLR: Ch 12.1-2 |  |
| Week 13 April 22 | K-Means and Hierarchical Clustering | ISLR: Ch 12.4 | Assignment 6 due |
| Week 15 April 29 | Final Project Presentations. Review. |  | Final Project due |
| FINALMay 6 | **Final Exam** |  |  |

**Statement on Academic Conduct and Support Systems**

**Academic Conduct**

Plagiarism – presenting someone else’s ideas as your own, either verbatim or recast in your own words – is a serious academic offense with serious consequences. Please familiarize yourself with the discussion of plagiarism in *SCampus* in Part B, Section 11, “Behavior Violating University Standards” [https://policy.usc.edu/student/scampus/part-b](https://policy.usc.edu/student/scampus/part-b/). Other forms of academic dishonesty are equally unacceptable.  See additional information in *SCampus* and university policies on scientific misconduct, [http://policy.usc.edu/scientific-misconduct](http://policy.usc.edu/scientific-misconduct/).

Discrimination, sexual assault, intimate partner violence, stalking, and harassment are prohibited by the university.  You are encouraged to report all incidents to the *Office of Equity and Diversity*/*Title IX Office* <http://equity.usc.edu> and/or to the *Department of Public Safety* [http://dps.usc.edu](http://dps.usc.edu/). This is important for the health and safety of the whole USC community. Faculty and staff must report any information regarding an incident to the Title IX Coordinator who will provide outreach and information to the affected party. The sexual assault resource center webpage <http://sarc.usc.edu> fully describes reporting options. Relationship and Sexual Violence Services <https://engemannshc.usc.edu/rsvp> provides 24/7 confidential support.

## **Support Systems**

A number of USC’s schools provide support for students who need help with scholarly writing.  Check with your advisor or program staff to find out more.  Students whose primary language is not English should check with the *American Language Institute* [http://ali.usc.edu](http://ali.usc.edu/), which sponsors courses and workshops specifically for international graduate students. *The Office of Disability Services and Programs* [http://dsp.usc.edu](http://dsp.usc.edu/) provides certification for students with disabilities and helps arrange the relevant accommodations. If an officially  declared emergency makes travel to campus infeasible, *USC Emergency Information* <http://emergency.usc.edu>will provide safety and other updates, including ways in which instruction will be continued by means of Blackboard, teleconferencing, and other technology.