Practical Machine Learning (with R) - UC Berkeley Extension

Computer Science x460.5 (Classroom Course - 2 semester units in Computer Science)

Wednesdays, January 24 to March 28, 6:30 PM - 9:30 PM

Course Canvas web site: https://onlinelearning.berkeley.edu/

(**Bring your laptop to every class meeting**)

Instructor:

Allan Miller (a.k.a. Allan, Professor Miller, Dr. Miller)

allan.m.miller@berkeley.edu

Course Description:

This course provides an introduction to machine learning (sometimes also called statistical learning) using R. The course will cover the practical aspects of machine learning. Upon completion of the course you will be able to apply machine learning in R to solve problems in your own work. The course covers (tentative):

Weeks 1 - 2: introduction to modeling and machine learning, linear and logistic regression models

Weeks 2 - 3: introduction to model tuning and performance evaluation

Weeks 4 - 5: classification and regression trees, recursive partitioning; ensemble methods (random forests), boosting and bagging

Weeks 6-7: unsupervised learning, k-means and k-nearest neighbor, neural networks

Week 7-8: more on model tuning and feature engineering

Weeks 8-10: time series analysis, advanced methods

Prerequisites:

- **Working knowledge of base R**, including R workspaces and environments; vectorized programming (including the ifelse() function, recycling; R data structures such as vectors, matrix, data frames, and lists; installing and loading packages; writing and calling R functions; binning data using cut.
- Familiar with the **RStudio** development environment including R Notebooks and knitting basic RMarkdown documents to HTML
- Proficient in **ggplot2** graphics and base R plot
- Proficient using **dplyr** for data wrangling
- Experience using other commonly used tidyverse packages, such as readr, tidyr, and purrr
- **Basic knowledge of statistics** as covered in a first-semester undergraduate statistics course, including measures of central tendency (e.g., mean, median), variability (variance, standard deviation), basic probability distributions (e.g., normal, uniform), correlation, hypothesis testing, confidence intervals, and regression. I assume that everybody has taken at least one

course that covers these topics, but that many (most?) of you are at least a bit rusty on your statistics.

Assignments (to be posted weekly on Canvas)

There will be weekly assignments consisting of DataCamp modules and/or assigned exercises. Assignments are learning exercises and are graded on a Credit/No Credit basis. Assignments will be due at the start of every class meeting, solutions posted (or available on DataCamp) and discussed during class meetings. Note: **no late assignments accepted**. Assignments must be turned in as specified, on time, to receive credit.

Instructional Methodology

- In-class presentation of main topics
- Work through selected assignment solutions and additional examples to illustrate concepts
- In-class, hands-on exercises for students to work on and discuss (bring a laptop with RStudio installed and running to every class meeting)
- Online message and discussion forum (Canvas: onlinelearning.berkeley.edu)

Credit Requirements (for your final course grade):

- Your final grade will be based on an in-class (last day of class, March 28) final exam (open book, open notes) and a take-home applied machine learning challenge problem (assigned March 21, due on March 28). Because of University grading requirements, all work must be completed by the last day of class.
- Attendance is required, sign-in will be taken at the start of every class meeting. If you need to miss a class meeting, be sure to complete your weekly assignment for that week, and let me know ahead of time, if possible.
- Assignments: weekly assignments are learning exercises graded on a credit/no credit basis. Credit will be awarded upon reasonable effort, on formal completion and submission. Solutions will be reviewed during class meetings, posted and/or available on DataCamp.
- You must complete assignments to earn course credit and a letter grade as follows: completing 100% of all assignments will earn full credit; completing completing 90% of assignments will lower your course grade by one level; 80% by two, and 70% by three levels. Completing less than 70% of assignments will result in a no credit/failing class grade.
- Participation in class discussions, both in-class and online, manifesting mastery of course topics, may be used to raise your final grade.

  The final exam will consist of short answer problems, writing some R code (similar to class exercises). The final exam is open book, open notes. **No use of R during the final exam**. The exam will take approximately two hours during the last class meeting.

Grade Options:

- Letter Grade
- Credit/No Credit (earned grade C or above)
- Not For Credit (audit, assignments not required or submitted, challenge problem and final exam)

The Not for Credit grade option is recommended if you wish to participate in the class, but are unable to attend class meetings, complete assignments, and take the final exam. It is strongly recommended for students who have concerns about their academic preparedness for taking the class or outside of class work or personal commitments.

Note: Letter Grade is the default grade option. I will pass around a form in class later in the semester on which you can specify your grade option. If you do not submit an alternative grade option, you will be assigned a letter grade.

Texts:

Readings will be assigned from several texts (and other online sources):

Machine Learning with R (Second Edition), Brett Lanz, Packt Publishers

Applied Predictive Modeling, (2013) Kuhn & Johnson, Springer Verlag

Forecasting Principles and Practice, (Second Edition), Hyndman & Athanasopalous, Otexts
(note: this is a new edition, to be published after class starts - we may end up having to use the first edition)

Recommended:

R for Data Science: Visualize, Model, Transform, Tidy and Import Data by Hadley Wickham, Garrett Grolemund, 2016 O'Reilly Media 1st edition (also available online).
ISBN 9781491910399

Base R:

R in Action, (Second edition), Kobacoff, Manning Publishers

Computing Resources

- Access to a R and RStudio to complete programming assignments: http://www.rstudio.com/
- **Bring your laptop to every class session** with RStudio installed to work on classroom hands-on exercises.
- All students must be officially enrolled through U.C. Berkeley Extension.

Academic Integrity

All students must follow the U.C. Berkeley Extension Student Guidelines on Academic Integrity:

> As a student of UC Berkeley Extension, you are encouraged to reach out to your fellow students in the classroom to avoid isolation, to discuss materials, and to ask each other questions; however, there are limits to this collaboration. We do not allow:

- Cheating
- Plagiarism
- Altering academic documents or transcripts
- Gaining access to materials before they are intended to be available
- Helping another student to gain an unfair academic advantage

Any violations of the above will result in a Fail course grade and possible University disciplinary action.