NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

PREDICT 422: Practical Machine Learning Spring 2017

Jennifer Wightman, PhD
jennifer.wightman@northwestern.edu

Course Description

The rapid advancement of computational methods from machine/statistical learning, data mining, and pattern recognition provides unprecedented opportunities for understanding large, complex datasets. This course takes a practical approach to introduce several machine learning methods with business applications in marketing, finance, and other areas. Topics include regression and classification methods, resampling methods, model selection, regularization, tree-based methods, support vector machines, principal components analysis, and clustering methods.

At the end of this course, students will have a basic understanding of how each of these methods learn from data to find underlying patterns useful for prediction, classification, and exploratory data analysis. Further, each student will learn how to apply machine learning methods using the R statistical programming language for improved decision-making in real business situations.

The course format is a combination of textbook readings, R Lab video sessions, group discussions, and hands-on applications. Weekly programming assignments using R will be used both to reinforce machine learning concepts and to practice.

Learning Objectives

- Demonstrate a practical understanding of the key theoretical concepts of modern computational/analytic methods from machine/statistical learning, data mining, and pattern recognition.
- Identify appropriate machine learning methods to find relationships and structure in data with and without specific output variable(s).
- Apply machine learning methods to build predictive models and discover patterns in data for more informative business decision-making.
- Develop analytic solutions to practical business problems using the R statistical programming language, transforming data into knowledge.

Prerequisites

The prerequisite for PREDICT 422 is PREDICT 411 Generalized Linear Models.

Course Workload

PREDICT 422 is a modeling course with a large programming component. As such, this course carries a heavy workload. Part-time students are strongly urged to take PREDICT 422 as their only course for the current term. Furthermore, if you foresee your work or family life being more demanding than usual this term (e.g. planned travel, starting a new job, moving across the country), then you should consider taking PREDICT 422 at a later time.

Required Texts

[1] James, Witten, Hastie, and Tibshirani (2013). *An Introduction to Statistical Learning*. New York, NY: Springer. [ISBN-13 978-1461471370]

PDF Available Online: <http://www-bcf.usc.edu/~gareth/ISL/>

This textbook will be referred to as **ISLR**.

Optional Texts

[1] Ledolter, Johannes. (2013). *Data Mining and Business Analytics with R*. Hoboken, NJ: John Wiley & Sons, Inc. [ISBN-13 978-1118447147]

This textbook will be referred to as **Ledolter**.

[2] Dalgaard, Peter (2008). *Introductory Statistics with R*. New York, NY: Springer. [ISBN-13 978-0387790534]

PDF Available Online: from the NU Library <http://www.library.northwestern.edu/>

[3] Albert, J. and Rizzo, M. (2012). *R by Example*. New York, NY: Springer. [ISBN-13 978-1461413646]

PDF Available Online: from the NU Library

Software

This course requires the R software environment. R is free, open-source software available for Windows, OSX, and Linux environments. It may be downloaded from <http://www.r-project.org>. RStudio is the recommended interface for R. RStudio can be downloaded for free at <http://www.rstudio.org>.

Evaluation

The student's final grade will be determined as follows:

- Weekly Discussion Board Participation (20%)
- Programming Assignments (40%)
- Charity Project (40%)

Grading Scale

A	= 93.00%–100%
A-	= 90.00%–92.99%
B+	= 87.00%–89.99%
B	= 83.00%–86.99%
B-	= 80.00%–82.99%
C+	= 77.00%–79.99%
C	= 73.00%–76.99%
C-	= 70.00%–72.99%
F	= 0.00%–69.99%

Discussion Board Participation

Each session you are required to participate in the session-specific discussion board forum. Your participation in both posting and responding to other students' comments is graded. The due date and time for posting to each week's discussion forum is Sunday at 11:59 p.m. (Central Time). You are required to make a minimum of three posts each week in order to receive full credit for the week.

The purpose of the discussion boards is to allow students to freely exchange ideas. It is imperative to remain respectful of all viewpoints and positions and, when necessary, agree to respectfully disagree. While active and frequent participation is encouraged, cluttering a discussion board with inappropriate, irrelevant, or insignificant material will not earn additional points and may result in receiving less than full credit. Frequency is not unimportant, but content of the message is paramount. Please remember to cite all sources—when relevant—in order to avoid plagiarism.

Programming Assignments

Programming assignments will be given every one to two weeks and will be due on Sundays at 11:59 p.m. (Central Time). The purpose of the programming assignments is to practice using R for data analysis and to practice applying the learning algorithms and techniques to actual data. The programming assignments will generally take more than a day or two to complete. You are encouraged to start working on the programming assignment in a timely manner to allow sufficient time for you to post questions, receive feedback, and overcome challenges.

Getting stuck is inherent to learning computer programming. There are numerous resources to help you when you are stuck including the course textbooks, R documentation and help pages, and online forums. One of the most important resources is your professor and fellow classmates. You are strongly encouraged to post questions on the Canvas discussion boards when you find yourself stuck in writing code, struggling to understand something, or looking for a better way to do things. In addition, you are permitted to share sections of code when asking or answering a question. The ultimate goal of the programming assignments is for you to learn something, not to test how much you can figure out all on your own.

Charity Project

The third component of this course is a project that addresses a mailing campaign for a charity. For this project you will build several regression and classification models using the charity dataset. The project will broken up into four separate assignments. You will be expected to document your modeling and analysis process and results for each assignment.

Course Outline

Week	Topic	ISLR	Ledolter	Assignment	Sync Session
1	Introduction and Overview / Statistical Learning (The Learning Problem, Assessing Model Accuracy)	Ch. 1, Ch. 2	Ch. 1, Ch. 2	PA 1: R Practice, EDA	Course Intro
2	Unsupervised Learning (Principal Components Analysis, K-Means Clustering, Hierarchical Clustering)	Ch. 10	Ch. 15, Ch. 17		Project Intro
3	Ordinary Least Squares Linear Regression (Simple, Multiple)	Ch. 3	Ch. 3	PA 2: Unsupervised Learning, OLS	
4	Resampling Methods in Machine Learning (Cross-Validation, The Bootstrap)	Ch. 5	Ch. 5	Project: Part 1	
5	Linear Model Selection and Regularization (Subset Selection, Shrinkage Methods, Dimension Reduction Methods)	Ch. 6	Ch. 6, Ch. 18	PA 3: Resampling, Model Select.	
6	Moving Beyond Linearity (Polynomial Regression, Splines, GAMs)	Ch. 7			
7	Classification Models (Logistic Regression, Discriminant Analysis, K-Nearest Neighbors)	Ch. 4	Ch. 7-9, Ch. 12	Project: Part 2	
8	Tree-Based Methods (Decision Trees, CART, Bagging, Random Forests, Boosting)	Ch. 8	Ch. 13, Ch. 14	PA 4: Classification, Trees, Forests	
9	Support Vector Machines	Ch. 9		Project: Part 3	
10	Charity Project		Ch. 16	Project: Part 4	

Attendance

All course goals, session learning objectives, and assessments are supported through classroom elements that can be accessed at any time. To measure class participation (or attendance), your participation in threaded discussion boards is required, graded, and paramount to your success in this class. In addition, there will be two or three scheduled synchronous sessions ("sync sessions" or virtual meetings) held throughout the term via Adobe Connect. The assigned meeting time is Mondays at 7:00 p.m. for Section 57 and Wednesdays at 7:00 p.m. for Section 60 (Central Time). You will be notified of scheduled sync sessions through the Canvas course shell. Attendance of sync sessions is encouraged but not required. All sync sessions will be recorded, and the recordings will be posted in Canvas for later viewing.

Late Work

All assignments and discussion board posts are officially due by 11:59 p.m. (Central Time) on Sunday of the week in which they are assigned. You will be allowed an automatic grace period until the following Monday at noon (Central Time) to have all of your work submitted. If you think you will not be able to submit your work by the end of the grace period, then you must notify the professor in writing by Sunday evening to request an extension. While not strictly required, it is preferred that extensions are not applied to the graded discussion boards (i.e. posts are submitted by the end of the automatic grace period).

Grade penalties for late work (including with approved extensions) will be applied as follows.

- First late assignment: No penalty
- Second late assignment: Grade capped at 90 points
- Third late assignment: Grade capped at 80 points
- Additional late assignments: Grade cap drops 10 points each time

Academic Integrity at Northwestern

Students are required to comply with University regulations regarding academic integrity. If you are in doubt about what constitutes academic dishonesty, speak with your instructor or graduate coordinator before the assignment is due and/or examine the University Web site. Academic dishonesty includes, but is not limited to, cheating on an exam, obtaining an unfair advantage, and plagiarism (e.g., using material from readings without citing or copying another student's paper). Failure to maintain academic integrity will result in a grade sanction, possibly as severe as failing and being required to retake the course, and could lead to a suspension or expulsion from the program. Further penalties may apply. For more information, visit <www.scs.northwestern.edu/student/issues/academic_integrity.cfm>.

Plagiarism is one form of academic dishonesty. Students can familiarize themselves with the definition and examples of plagiarism, by visiting <www.northwestern.edu/uacc/plagiar.html>. A myriad of other sources can be found online.

Other Processes and Policies

Please refer to your SCS student handbook for information about program processes and policies: <http://www.scs.northwestern.edu/program-areas/graduate/student-handbook.php>