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. The course aims to provide a practical survey of modern machine learning techniques that can be applied to make informed business decisions: regression and classification methods, resampling methods and model selection, regularization, perceptron and artificial neural networks, tree-based methods, support vector machines and kernel methods, 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 implement machine learning methods in the R statistical programming language for improved decision-making in real business situations.

The course format is a combination of textbook readings and lecture slides, R Lab video sessions, and group discussions. Weekly quizzes and programming assignments using R will be used to reinforce both machine learning concepts and practice. The final project will involve students applying multiple machine learning methods to solve a practical business problem in marketing.

COURSE GOALS

- 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.

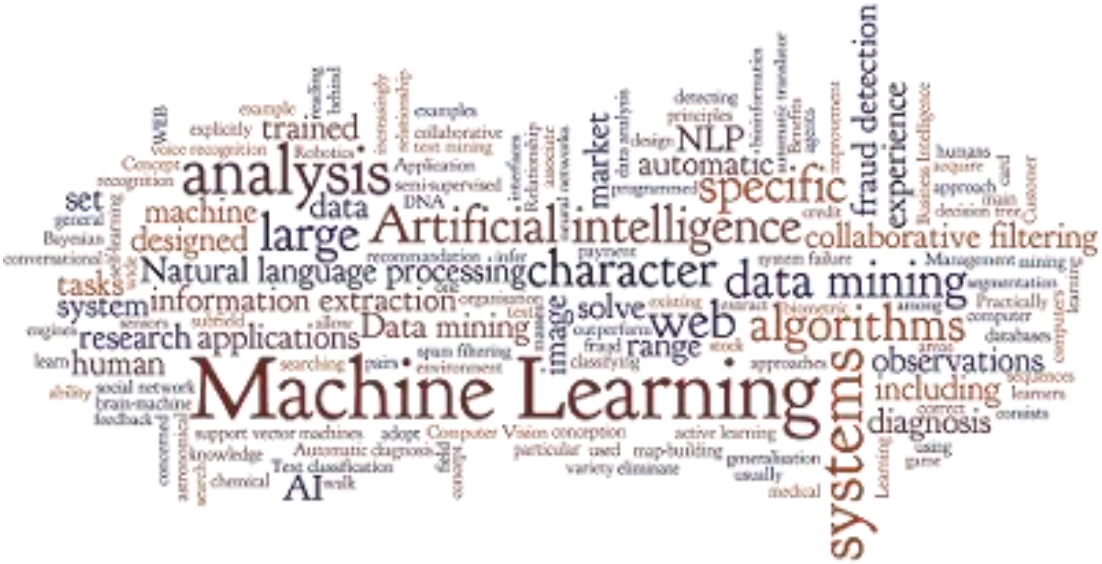
COURSE MATERIALS

Required Textbook: *An Introduction to Statistical Learning, with Applications in R* (2013), by G. James, D. Witten, T. Hastie, and R. Tibshirani.

Note: This textbook is available for *free* download at http://www-bcf.usc.edu/~gareth/ISL/ (Links to an external site.).

Statistical Software: R, which can be downloaded for *free* from http://www.r-project.org (Links to an external site.). Rstudio is the recommended interface for the R statistical programming language software, which can also be downloaded for *free* at http://www.rstudio.org (Links to an external site.).

Please refer to the course syllabus for more details: PREDICT 422 Course Syllabus



COURSE DETAILS
Instructor Name: Anil D. Chaturvedi
Program Name: Master of Science in Predictive Analytics (MSPA)
Course Name: Practical Machine Learning
Course number: PREDICT 422-DL
NU E-mail Address: anil.chaturvedi@northwestern.edu
Response Times: 48 hours
Office Hours: By appointment
Phone: (301) 299-2434

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. The course aims to provide a practical survey of modern machine learning techniques that can be applied to make informed business decisions: regression and classification methods, resampling methods and model selection, regularization, perceptron and artificial neural networks, tree-based methods, support vector machines and kernel methods, 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 implement machine learning methods in the R statistical programming language for improved decision-making in real business situations. The course format is a combination of textbook readings and lecture slides, R Lab video sessions, and group discussions. Weekly quizzes and programming assignments using R will be used to reinforce both machine learning concepts and practice. The final project will involve

students applying multiple machine learning methods to solve a practical business problem in marketing.

COURSE GOALS

- 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.

COURSE OUTLINE

Week	Topic	Reading	Activity
1	Introduction to Machine Learning (The Learning Problem, Assessing Model Accuracy) / Introduction to R Programming	Ch. 1, Ch. 2	Quiz 1, R Lab 1
2	Ordinary Least Squares Linear Regression (Simple, Multiple)	Ch. 3	Quiz 2, R Lab 2
3	Resampling Methods in Machine Learning (Cross-Validation, The Bootstrap)	Ch. 5	Quiz 3, R Lab 3
4	Linear Model Selection and Regularization (Subset Selection, Shrinkage Methods, Dimension Reduction Methods)	Ch. 6	Quiz 4, R Lab 4
5	Non-Linear Models (Polynomial Regression, Regression Splines, Smoothing Splines, Local Regression, Generalized Additive Models)	Ch. 7	Quiz 5, R Lab 5
6	Classification Models (Logistic Regression, Discriminant Analysis, K-Nearest Neighbors)	Ch. 4	Quiz 6, R Lab 6
7	Perceptron Learning Algorithm / Artificial Neural Networks	None	Project Issued

8
Tree-Based Methods (Decision Trees, CART, Bagging,
Random Forests, Boosting)
Ch. 8 Quiz 7, R Lab 7

9
Support Vector Machines and Kernel Methods Ch. 9 Quiz 8, R Lab 8

10
Unsupervised Learning (Principal Components Analysis, KMeans
Clustering, Hierarchical Clustering
Ch.10 Project Due

COURSE MATERIALS

Required Textbook: *An Introduction to Statistical Learning, with Applications in R* (2013), by G.James, D. Witten, T. Hastie, and R. Tibshirani.
Note: This textbook is available for *free* download at http://www-bcf.usc.edu/~gareth/ISL/.

Recommended Textbooks:

- *The Elements of Statistical Learning* (2009), by T. Hastie, R. Tibshirani, and J. Friedman
- *Learning from Data: A Short Course* (2012), by Y. Abu-Mostafa, M. Magdon-Ismail, and H. Lin
- *Machine Learning: A Probabilistic Perspective* (2012), by K. Murphy
- *R and Data Mining: Examples and Case Studies* (2013), Y. Zhao
- *Pattern Recognition and Machine Learning* (2007), C. Bishop

Statistical Software: R, which can be downloaded for *free* from http://www.r-project.org. Rstudio is the recommended interface for the R statistical programming language software, which can also be downloaded for *free* at http://www.rstudio.org.
Course Website: Canvas will be used for posting relevant course materials throughout the term. You need a Northwestern University NetID and password to use Canvas.
Prerequisites: PREDICT 411-DL Generalized Linear Models

COURSE REQUIREMENTS

Evaluation:
- **Weekly Quizzes (20%):**
  - There will be 8 weekly quizzes based on the course material. The quiz is timed for **2 (or 3) hours**, so please allow sufficient time to take each quiz without interruption. I will review the quizzes and release the solutions within 48 hours after the deadline.

- **Weekly Programming Assignments – R Labs (20%):**

There will be 8 weekly programming assignments (R labs). Each lab consists of two parts. The first part is to follow through the textbook's lab on your own time. Once you

feel comfortable with the R code and material discussed in the textbook's lab, then you complete the second part. This second graded part will test your knowledge of the R commands and outputs covered in the textbook's lab. You will have **1 (or 2) hour(s)** to complete this part of the lab, so please allow sufficient time to complete with interruption.

- **Final Project (50%):**
  - There will be 1 individual final project due at the end of the course.

- **Discussion Board Participation (10%):**
  - The final 10% of your grade is evaluated by participation in Canvas discussion forums.

The discussion forum on Canvas is two-fold: one forum is used to discuss a weekly analytics-related video while the other is for question and answer. Questions, tips, and posts helping others count as participation. Postings that are rude or otherwise inappropriate will not receive any credit.

Grading Scale: Your grade will be based on your final weighted average score and the letter grade will
be assigned according to the following table:

Range Grade
[93%, 100%] A
[90%, 93%) A-
[87%, 90%) B+
[83%, 87%) B
[80%, 83%) B-
[77%, 80%) C+
[73%, 77%) C
[70%, 73%) C-
[0%, 70%) F

Discussion Board Etiquette: The purpose of the discussion board in general 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.*
Attendance: This course will not meet at a particular time each week. 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. Please note that any scheduled

synchronous or "live" meetings are considered supplemental and optional. While your attendance is highly encouraged, it is not required and you will not be graded on your attendance or participation.

Late policy: Unless otherwise noted, all work is due on the assigned day by 11:55 PM (Central Time).
This includes programming assignments, quizzes, and participation in the discussions. *Late work is not accepted*.
Academic Integrity at Northwestern University: 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. Some assignments in SCS courses may be required to be submitted through SafeAssign, a plagiarism detection and education tool. You can find an explanation of the tool at http://wiki.safeassign.com/
display/SAFE/How+Does+SafeAssign+Work. In brief, SafeAssign compares the submitted assignment
to millions of documents in large databases. It then generates a report showing the extent to which text
within a paper is similar to pre-existing sources. The user can see how or whether the flagged text is
appropriately cited. SafeAssign also returns a percentage score, indicating the percentage of the submitted paper that is similar or identical to pre-existing sources. High scores are not necessarily bad,
nor do they necessarily indicate plagiarism, since the score does not take into account how or whether
material is cited. If a paper consisted of one long quote that was cited appropriately, it would score
100%. This would not be plagiarism, due to the appropriate citation. However, submitting one long
quote would probably be a poor paper. Low scores are not necessarily good, nor do they necessarily
indicate a lack of plagiarism. If a 50-page paper contained all original material, except for one short

quote that was not cited, it might score around 1%. But, not citing a quotation is still plagiarism, as is

repurposing of one's own work without citation.
SafeAssign includes an option in which the student can submit a paper and see the resultant report
before submitting a final copy to the instructor. This ideally will help students better understand and
avoid plagiarism.
Other Processes and Policies: Please refer to your SCS student handbook at www.scs.northwestern.
edu/grad/information/handbook.cfm for additional course and program processes and policies.