NORTHWESTERN UNIVERSITY | SCHOOL OF CONTINUING STUDIES

PREDICT 412: Advanced Modeling Techniques Syllabus Fall 2013

Thomas W. Miller, Ph.D.
thomas-miller-0@northwestern.edu

Course Description

Drawing upon previous course work in predictive analytics, modeling, and data mining, this course provides a review of statistical and mathematical programming and advanced modeling techniques. It explores computer-intensive methods for parameter and error estimation, model selection, and model evaluation. The course focuses upon business applications of statistical graphics and data visualization, tree-structured classification and regression, neural networks, smoothing methods, hybrid models, multi-way analysis, and hierarchical models. The course also provides an introduction to business applications of geographical information systems (GIS). This is a case-study- and project-based course with a strong programming component.

Texts

Fox, J. & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). Thousand Oaks, CA: Sage. [ISBN-13 978-1412975148]

Matloff, N. (2011). *The art of R programming: A tour of statistical software design.* San Francisco, CA: No Starch Press. [ISBN-13: 978-1593273842]

Izenman, A. J. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning.* New York, NY: Springer. [ISBN-13: 978-0387781884] [electronic copy available through Northwestern University library]

Miller, T. W. (2014). *Modeling techniques in predictive analytics: Business problems and solutions with R.* Upper Saddle River, N.J.: Pearson.
[print: ISBN-13: 978-0-13-341293-2; electronic: ISBN-13: 978-0-13-341297-0]

[No purchase necessary. Selected sections from the electronic edition are posted to Blackboard.]

Software

This course requires the R software environment. R software is free, open-source software available for PC/Windows and Mac/OSX environments. It may be downloaded from the Comprehensive R Archive Network at <http://cran.r-project.org/>.

Prerequisites

CIS 317, CIS 435, PREDICT 401, PREDICT 410, and PREDICT 411

Learning Goals

The goals of this course are to:

- Evaluate management problems that may be addressed by analytic methods.
- Evaluate alternative analytic methods appropriate for addressing management problems.
- Construct models and programs for estimating models.
- Construct databases and data objects needed for estimating models.
- Apply statistical programming methods.
- Apply mathematical programming methods.
- Evaluate results of analytic methods.
- Present analytic solutions to management.

Evaluation

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

- Session 1 Term Project Assignment: Submission of Bid: 20 pts.
- Session 2 Term Project Checkpoint: 25 pts.
- Session 3 Programming Assignment 1: Statistical Graphics in R: 30 pts.
- Session 4 Term Project Checkpoint: 25 pts.
- Session 4 Programming Assignment 2: Evaluating Regression Models in R: 30 pts.
- Session 5 Programming Assignment 3: Evaluating Classification Models in R: 30 pts.
- Session 6 Term Project Checkpoint: 25 pts.
- Session 6 Programming Assignment 4: Evaluating Hierarchical Models in R: 30 pts.
- Session 7 Programming Assignment 5: Mathematical Programming in R: 30 pts.
- Session 8 Term Project Checkpoint: 25 pts.
- Session 8 Programming Assignment 6: Fitting a Spatial Model in R: 30 pts.
- Session 10 Final Term Project Report: 200 pts.
- Sessions 1–10 Discussion Board Participation: 100 pts.

Total: 600 pts.

Grading Scale

A = 93%–100% (558–600 pts.)
A- = 90%–92% (540–557 pts.)
B+ = 87%–89% (522–539 pts.)
B = 83%–86% (498–521 pts.)
B- = 80%–82% (480–497 pts.)
C+ = 77%–79% (462–479 pts.)
C = 73%–76% (438–461 pts.)
C- = 70%–72% (420–437 pts.)
F = 0%–69% (0–419 pts.)

Discussion Board Etiquette

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. The due date and time for posting to each week's discussion forum is Sunday, 11:55 p.m. (Central Time).

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 Work

Students must provide written notification of late work 24 hours prior to the deadline. One grace day is allowed for those who provide late work notification. Only one grace day without reduction of points is allowed. Standard policy is to apply a 25 percent reduction to the grade for every 12 hours late. (No negative points are applied.)

Work Groups

Work groups are utilized in this course. More information about work groups will be provided by the instructor via the Blackboard course site.

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.

Some assignments in this course 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.

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

Course Schedule

Important Note: Changes may occur to the syllabus at the instructor's discretion.
When changes are made, students will be notified via an announcement in Blackboard.

Session 1

Learning Objectives

After this session, the student will be able to:

- Describe what is meant by a statistical programming environment.
- Perform data and modeling object operations using a graphical user interface to a statistical programming environment.
- Perform data entry into a statistical programming environment, including imports from spreadsheets and relational databases.
- Perform data object manipulation in a statistical programming environment, including data mergers, selects, transformations, data type changes, and missing data actions.
- Construct traditional linear and generalized linear modeling objects from data objects within a statistical programming environment.
- Evaluate the structure of data and modeling objects within a statistical programming environment.
- Perform external software package installation within a statistical programming environment.
- Perform data and modeling object operations using an integrated development environment for a statistical programming environment.

Course Content

Textbook Reading

Matloff, chapters 1–4 (pages 1–100)

Online Reading

Miller, T. W. (2014). *Modeling techniques in predictive analytics: Business problems and solutions with R.* Upper Saddle River, N.J.: Pearson. Chapter 1 (Analytics and Data Science, pages 1–13).
[print: ISBN-13: 978-0-13-341293-2; electronic: ISBN-13: 978-0-13-341297-0] [File posted to Blackboard]

Lecture (Connect Recordings)

R Tutorials

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to each week's discussion forum is Sunday at 11:55 p.m. (Central Time).

Assignments

Term Project Assignment: Submission of Bid is due Sunday, September 29, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

Tuesday, September 24, 2013 at 7:00–9:00 p.m. (Central Time)

Session 2

Learning Objectives

After this session, the student will be able to:

- Construct graphical objects from data objects, producing univariate and bivariate graphs.
- Perform density estimation and smoothing operations on data objects.
- Generate graphical objects for model diagnostics.
- Interpret graphical objects for model diagnostics.
- Generate graphical objects within a structured visualization data framework.
- Interpret graphical objects within a structured visualization data framework.
- Generate graphical objects for three-way and multi-way data display.
- Interpret graphical objects for three-way and multi-way data display.
- Evaluate statistical graphics and data visualization as tools in predictive analytics.

Course Content

Textbook Reading

Matloff, chapters 5–7 (pages 101–188)

Izenman, chapter 1 (pages 1–16)

Lecture (Connect Recordings)

R Tutorials

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

Session 2 Term Project Checkpoint is due Sunday, October 6, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None

Session 3

Learning Objectives

After this session, the student will be able to:

- Generate random data to represent discrete probability distributions.
- Generate random data to represent continuous probability distributions.
- List alternative methods of optimization and statistical estimation.
- Perform resampling-based estimation and hypothesis testing.
- Perform leave-one-out cross-validation for model evaluation.
- Perform internal multi-fold cross-validation for model evaluation.
- Complete the design of a statistical simulation experiment for comparing alternative modeling methods.
- Perform a statistical simulation experiment for comparing alternative modeling methods.
- Interpret results from a statistical simulation experiment.
- Evaluate simulation methods as tools in predictive analytics.
- Discuss differences between statistical simulation experiments and process simulations in business.

Course Content

Textbook Reading

Matloff, chapters 8–12 (pages 189–283)

Izenman, chapter 4 (pages 75–106)

Fox and Weisberg, chapter 3 (pages 107–148)

Lecture (Connect Recordings)

R Tutorials

Online Reading

Hothorn, T. Leisch, F. Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics,* 14(3), 675–699.

Schauerhuber, M. Zeileis, A, Meyer, D., & Hornik, K. (2008). Benchmarking open-source tree learners in R/RWeka. In Preisach, C., Burkhardt, H., Schmidt-Thieme, L. & Decker, R. (eds). *Data analysis, machine learning and applications.* New York: Springer, pp. 389–396.

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

Programming Assignment 1: Statistical Graphics in R is due Sunday, October 13, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None

Session 4

Learning Objectives

After this session, the student will be able to:

- Define regression as a set of tools in predictive analytics.
- Describe examples of regression problems in business.
- Discuss alternative regression modeling techniques, including traditional linear regression, neural networks, and tree-structured regression.
- List alternative criteria for evaluating the performance of regression models.
- Identify multivariate linear regression as distinct from multiple regression.
- Define hybrid and ensemble methods for regression.
- Identify management problems in predictive analytics that may be addressed by regression techniques.
- Complete the design of a statistical simulation experiment for comparing alternative regression modeling methods, including hybrid and/or ensemble methods.
- Perform a statistical simulation experiment for comparing alternative regression modeling methods.
- Interpret results from a statistical simulation experiment, comparing alternative regression modeling methods.
- Construct a recommendation to management regarding the regression methods to be employed in solving business research problems.
- Evaluate regression modeling methods as tools in predictive analytics.

Course Content

Textbook Readings

Izenman, chapter 5 (pages 107–158)

Fox and Weisberg, chapter 4 (pages 149–228)

Online Reading

Miller, T. W. (2014). *Modeling techniques in predictive analytics: Business problems and solutions with R.* Upper Saddle River, N.J.: Pearson. Chapter 2 (Advertising and Promotion, pages 15–28) and Appendix A (Regression, pages 238–239).
[print: ISBN-13: 978-0-13-341293-2; electronic: ISBN-13: 978-0-13-341297-0] [Files posted to Blackboard]

Multimedia

Tree-Structured Regression

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

Session 4 Term Project Checkpoint is due Sunday, October 20, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Programming Assignment 2: Evaluating Regression Models in R is due Sunday, October 20, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None

Session 5

Learning Objectives

After this session, the student will be able to:

- Define classification as a set of tools in predictive analytics.
- Describe examples of classification problems in business.
- Discuss alternative classification techniques, including traditional logistic regression, support vector machines, and tree-structured classification.
- Identify classification as distinct from clustering.
- Define hybrid and ensemble methods for classification.
- List alternative criteria for evaluating the performance of classifiers.
- Interpret the receiver operating characteristic curve for classifiers.
- Identify management problems in predictive analytics that may be addressed by classification techniques.
- Complete the design of a statistical simulation experiment for comparing alternative classification methods, including hybrid and/or ensemble methods.
- Perform a statistical simulation experiment for comparing alternative classification methods.
- Interpret results from a statistical simulation experiment, comparing alternative classification methods.
- Construct a recommendation to management regarding classification methods to be employed in business.
- Evaluate classification methods as tools in predictive analytics.

Course Content

Online Reading

Hastie, T., Tibshirani, R, & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction* (2nd ed). New York: Springer. (Chapter 15: Random Forests, pages 587–604)

Miller, T. W. (2014). *Modeling techniques in predictive analytics: Business problems and solutions with R.* Upper Saddle River, N.J.: Pearson. Chapter 8 (Sentiment Analysis, pages 113–147) and Appendix A (Classification, pages 240–241).
[print: ISBN-13: 978-0-13-341293-2; electronic: ISBN-13: 978-0-13-341297-0] [Files posted to Blackboard]

Textbook Readings

Izenman, chapter 9 (pages 281–314), and chapter 11 (pages 369–406)

Fox and Weisberg, chapter 5 (pages 229–284)

Multimedia

Naive Bayes Classification

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

Programming Assignment 3: Evaluating Classification Models in R is due Sunday, October 27, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

Tuesday, October 22, 2013 at 7:00–9:00 p.m. (Central Time)

Session 6

Learning Objectives

After this session, the student will be able to:

- Define hierarchical models for regression and classification.
- Construct examples of hierarchical models as they might be employed in business applications of predictive analytics.
- Identify management problems in predictive analytics that may be addressed by a hierarchical modeling approach.
- Complete the design of a statistical simulation experiment for comparing hierarchical and non-hierarchical approaches to problems in business research.
- Perform a statistical simulation experiment for comparing hierarchical and non-hierarchical approaches to problems in business research.
- Interpret results from a statistical simulation experiment for comparing hierarchical and non-hierarchical approaches to problems in business research.
- Construct recommendations to management regarding the application of hierarchical statistical methods.
- Evaluate hierarchical modeling methods as tools in predictive analytics.

Course Content

Online Reading

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press. (Chapter 11: Multilevel Structures & Chapter 12: Multilevel Linear Models, pages 237–287)

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

Session 6 Term Project Checkpoint is due Sunday, November 3, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Programming Assignment 4: Evaluating Hierarchical Models in R is due Sunday, November 3, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None

Session 7

Learning Objectives

After this session, the student will be able to:

- Define deterministic versus stochastic approaches to modeling.
- Define mathematical programming as a class of methods for constrained optimization.
- Describe alternative methods for performing constrained optimization.
- Describe applications of mathematical programming in business.
- Describe mathematical programming within a statistical programming environment.
- Identify business problems that may be addressed by mathematical programming.
- Construct a mathematical programming model to address a business problem.
- Apply constrained optimization to solve the mathematical programming problem.
- Interpret results from mathematical programming.
- Construct a recommendation to management regarding a mathematical programming solution to a business research problem.
- Evaluate mathematical programming methods as tools in predictive analytics.

Course Content

Online Reading

Braun, W.J. & Murdoch, D. J. (2007) *A first course in statistical programming with R.* Cambridge: Cambridge University Press. (Chapter 7: Numerical Optimization, pages 132–157)

Miller, T. W. (2014). *Modeling techniques in predictive analytics: Business problems and solutions with R.* Upper Saddle River, N.J.: Pearson. Chapter 6 (Operations Management, pages 67–82).
[print: ISBN-13: 978-0-13-341293-2; electronic: ISBN-13: 978-0-13-341297-0] [Files posted to Blackboard]

Williams, H. P. (1999). *Model building in mathematical programming.* New York: Wiley. (Chapter 1: Introduction & Chapter 5: Applications and Special Types of Mathematical Programming, pages 3–9, 59–92)

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

Programming Assignment 5: Mathematical Programming in R is due Sunday, November 10, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None

Session 8

Learning Objectives

After this session, the student will be able to:

- Define geographic information system.
- Construct examples of spatial data problems in predictive analytics.
- Construct examples of spatio-temporal data problems in predictive analytics.
- Describe geo-referencing and geo-coding, including longitude and latitude referencing.
- Perform relational database queries using geometry and timestamp fields.
- Perform relational database calculations using geometry and timestamp functions.
- Perform data record extraction from a relational database with geometry and timestamp fields.
- Perform spatio-temporal data entry into a statistical programming environment.
- Perform data object manipulation in a statistical programming environment, including spatio-temporal data aggregation, selects, transformations, data type changes, and missing data actions.
- Perform distance calculations in a statistical programming environment using longitude and latitude coordinates.
- Evaluate the structure of spatio-temporal data objects within a statistical programming environment.

Course Content

Online Reading

Bivand, R. S., Pebesma, E. J., & Gomez-Rubio, V. (2008). *Applied spatial data analysis with R.* New York: Springer. (Chapter1: Introducing Spatial Data, & Chapter 2: Classes for Spatial Data, pages 1–15, 21–55)

Lloyd, C. D. (2010). *Spatial data analysis: An introduction for GIS users.* New York: Oxford University Press. (Chapter 2: Key Concepts 1 GIS, pages 6–23 and Chapter 8: Exploring Spatial Patterning in Data Values & References, pages 106–128, 175–179)

Miller, T. W. (2014). *Modeling techniques in predictive analytics: Business problems and solutions with R.* Upper Saddle River, N.J.: Pearson. Chapter 11 (Spatial Data Analysis, pages 209–229).
[print: ISBN-13: 978-0-13-341293-2; electronic: ISBN-13: 978-0-13-341297-0] [Files posted to Blackboard]

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

Session 8 Term Project Checkpoint is due Sunday, November 17, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Programming Assignment 6: Fitting a Spatial Model in R is due Sunday, November 17, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None

Session 9

Learning Objectives

After this session, the student will be able to:

- Identify spatial point processes as a class of statistical modeling problems.
- Identify spatio-temporal point processes as a class of statistical modeling problems.
- Construct maps within a statistical programming environment.
- Identify management problems in predictive analytics that may be addressed by spatial data modeling or spatio-temporal data modeling.
- Construct a spatial or spatio-temporal model to address a business research problem.
- Perform spatial or spatio-temporal model estimation within a statistical programming environment.
- Interpret results from fitting a spatial or spatio-temporal model.
- Construct a recommendation to management regarding the solution of a business research problem involving spatial or spatio-temporal data.
- Evaluate geographic information systems as tools for solving spatial data and spatio-temporal data problems in predictive analytics.

Course Content

Online Reading

Cressie, N. & Wikle, C.K. (2011). *Statistics for spatio-temporal data* (Chapter 1. Space-Time: The Next Frontier, pages 1–16)

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

None.

Sync Session

Tuesday, November 19, 2013 at 7:00–9:00 p.m. (Central Time)

Session 10

Learning Objectives

After this session, the student will be able to:

- No new learning objectives will be introduced.

Course Content

None.

Discussion Board

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. For this session's discussion topic(s), visit the discussion board in Blackboard. The due date and time for posting to the discussion for this session is Sunday at 11:55 p.m. (Central Time).

Assignments

Final Term Project Report is due Sunday, December 8, 2013 at 11:55 p.m. (Central Time). For more information, click Assignments on the left navigation panel in Blackboard, and scroll to this assignment's item.

Sync Session

None