Predictive Modeling of Spatial, Textual and Network Data (Statistics 5931)

Zack W. Almquist
Fall Semester, 2017

Class Schedule

Lecture:	Th	2:00 – 4:00 pm	Fordhall Statistics Library

URL: `http://moodle.umn.edu`
Note: Requires UMN login and registration in class to access.
URL: `http://www.github.com`
Note: Requires github login.

Professor

Name:	Zack W. Almquist
Office:	372 Ford Hall
Office Hours:	By Appointment Only
Email:	almquist@umn.edu
Telephone:	612-624-4300 (not recommended)

Course Objectives

This course will focus on learning and developing applied skills in text analysis and machine learning classification techniques. Students will be exposed to key software tools in R and python to engage in applied network and text analysis research. Students will learn to work in teams and engage in scholarly research through developing, analyzing and writing up a 6-10 page conference proceeding. All students will be expected to submit (as a team) their final project to an academic conference.

Prerequisites

This course assumes exposure to the R statistical environment and statistics on par with the Statistics sequence 3011 and 3022. More exposure to statistics or programming is helpful, but not required.

Course Requirements

This course meets once a week on Wednesdays from 2:00 to 4:00pm and students are expected to attend regularly, do the readings, exercises and be engaged during each seminar session. At the end of the semester each student team is expected to submit a paper to a conference based on the research project completed over the semester.

Computers

It is not required that students bring their computers/laptops to lecture and lab (if one is owned), but it is *highly* recommended since both lecture and lab will make extensive use of the computer software `R`. Computer labs are available on campus, please consult with the TA if you have trouble finding the various locations of campus computer labs.

Readings

Weekly readings assignments can be found on the course syllabus. All readings are assumed to be completed before each lecture/seminar. You are expected to read over the class notes each week and make sure you are familiar with the material as the course progresses. Questions are encouraged.

Homework

Homework assignments will normally be administered on a bi-weekly basis and will be due on every other Wednesday. Homework assignments are meant to achieve three results: (1) provide practice with the statistical concepts discussed in class, and (2) provide practice with the computational and statistical programing language `R` and (3) provide a chance to demonstrate your mastery of material and highlight areas where more work is needed. You may work in a group, but all write-ups must be done independently. All collaborators should be appropriately cited in your write up and any detailed R code should also be provided.

Final Project

All students will be expected to work in teams of 2-3 individuals to develop, writing and analyze a project focused on employing methods and models for text and network analysis. This final project will take the form of a 8-10 page conference proceedings. Students are expected to submit their project to an academic conference.

Participation

Individuals are expected to attend every course, to have completed every reading, and to participate with questions and discussion on each topic as presented. If you plan on missing any class period you are responsible for all material and for contacting the instructor in a timely manner.

Grading

Participation:	10%
Homework:	20%
Final Project:	70%

Lectures, readings, labs, and review sessions are provided for each student's benefit. It is the responsibility of the student to take advantage of these opportunities to acquire and demonstrate mastery of course material, so as to achieve his or her desired grade.

Letter grade assignment

A	93%+
A-	90-92.99%
B+	87-89.99%
B	83-86.99%
B-	80-82.99%
C+	77-79.99%
C	73-76.99%
C-	70-72.99%
D	60-69.99%
F	<59.99%

Required Texts

Spatial Analysis

- Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V., & Pebesma, E. J. (2008). Applied spatial data analysis with R (Vol. 747248717). New York: Springer.
- Gaetan, C., & Guyon, X. (2010). Spatial statistics and modeling (Vol. 81). New York: Springer.
- Baddeley, A. (2008, February). Analysing spatial point patterns in R. Technical report, CSIRO, 2010. Version 4. `https://research.csiro.au/software/r-workshop-notes`.

Natural Language Processing

- Manning, Raghavan, and Schutze. 2008. Introduction to Information Retrieval. Cambridge University Press.
- Jurafsky, Daniel and James Martin. 2008. Speech and Language Processing. Prentice Hall.

Machine Learning

- Bishop, Christopher. 2006. Pattern Recognition and Machine Learning. Springer.
- Hastie, Tibshirani, and Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction 2nd edition. Springer.
- McLachlan and Peel. 2000 Finite Mixture Models. Wiley.
- McLachlan and Krishnan. 2008. The EM Algorithm and Extensions. 2nd Edition Wiley.

Computer Languages

- Lutz, Mark. 2010. Programming Python. 4th Edition O?Reilly.
- Lutz, Mark. 2009. Learning Python. 4th Edition O?Reilley.
- Wickham, Hadley. 2017. Advanced R. Chapman & Hall. `http://adv-r.had.co.nz/`.
- Grolemund, Garrett and Hadley Wickham. (2017). R for Data Science. `http://r4ds.had.co.nz/`.

Readings

Be prepared to discuss all readings assigned at anytime in lecture/seminar.

Required Software

We will be using the `R` statistical programming language. `R` can be downloaded at `http://www.r-project.org/`.

RStudio IDE (Integrated Development Environment) is a software application which facilitates interaction with the `R` statistical programming language. It is often preferred to the GUI (Graphic User Interface) made available through CRAN. You can download it at `http://www.rstudio.com/`.

Latex is a word processor and a document markup language. It can be downloaded and installed on Windows (`http://miktex.org/`), OSX (`https://tug.org/mactex/`) or Linux (use the package manager of your choice).

A github account will be required of all students. One can register for a github account at `https://github.com/`. You can find information about how github works with Rstudio at `http://z.umn.edu/rstudiogit`, and github maintains a quite good help-system at `https://help.github.com/`.

Course Policies

Missing Class, etc.

It is expected that each member of the class will attend every lecture/discussion. If there is an appropriate reason to miss class it is expected that the individual will email or discuss in person with the instructor at least one week in advance. For any medical issues please see the UMN website for university policies.

Cheating, etc.

All work is assumed to be your own and all individuals are expected to follow the university policy on cheating and misconduct. If you have any questions please consult the UMN website for university policies.

Assignments

Homework Assignments

Homework will be assigned on a biweekly basis starting on the second Wednesday of the Semester and will be due two weeks later at 5:00pm. There will be a total of six homework assignments. Homework assignments will be graded on a 100 point basis. Each assignment must be turned in through github, no late assignments will be accepted. Homework must be turned in using github, knitr/latex and must include all R code. Your lowest score will be dropped in the final calculation of grades.

Final Team Project

The final project is Due by the end of semester and it is expected that all students will submit the team project to an academic conference.

Reading Assignments

Week 1 (09/07): Spatial Data

- *Readings:*
  - Bivand et al (2008) Chapters 1 to 4
  - Baddeley Chapters 5 to 9
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 2 (09/14): Brief Overview of Spatial Modeling

- *Readings:*
  - Bivand et al (2008) Chapters 7 to 10
  - Gaetan and Guyon (2010) Chapters 4 and 5
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 3 (09/21): Text as Data

- *Readings:*
  - Grimmer, Justin and Brandon Stewart. 2013. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Documents. Political Analysis. 21, 3 267-297.
  - Monroe, Burt and Phil Schrodt. 2008. Introduction to the Special Issue: The Statistical Analysis of Political Text. Political Analysis 16, 4, 351-355.
  - `http://thomasleeper.com/Rcourse/Intro2R/Intro2R.pdf`
  - Berinsky, Adam and Gregory Huber and Gabriel Lenz. 2011. Evaluating Online Labor Markets for Experimental Research: Amazon.com?s Mechanical Turk. Political Analysis 20, 3. 351-368.
  - Porter, MF. 2001. Snowball: A Language for Stemming Algorithms `http://snowball.tartarus.org/texts/introduction.html`
  - NLP Stemming TBD
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 4 (09/28): Dictionary Methods: Measuring Weighted Word Usage

- *Readings:*
  - Soroka, Stuart and Lori Young. 2012. ?Affective News: The Automated Coding of Sentiment in Political Texts? Political Communication 29: 205-231.
  - Dodds, Peter and Christopher Danforth. 2009. Measuring the Happiness of LargeScale Written Expression: Songs, Blogs, and Presidents. Journal of Happiness Studies 11, 4. 441-456.
  - Loughran, Tim and Bill McDonald.2011. ?When is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks? Journal of Finance 66, February 35-65.
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 5 (10/05): Methods for Finding Discriminating Words and Applications

- *Readings:*
  - Mosteller, Frederick and David Wallace. 1963. Inference in an Authorship Problem Journal of the American Statistical Association 58, 302. 275-309.
  - Monroe, Burt, Michael Colaresi, and Kevin Quinn. 2008. Fightin Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Political Analysis 16(4)
  - Taddy, Matt. 2013. Multinomial Inverse Regression for Text Analysis Journal of the American Statistical Association 108, 755-770.)
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 6 (10/12): The Vector Space Model and Geometry of Text

- *Readings:*
  - Linear Algebra Handout
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 7 (10/19): PCA, MDS and Text

- *Readings:*
  - Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer.
  - Spirling, Arthur. US Treaty-Making with American Indians: Institutional Change and Relative Power 1784-1911 American Journal of Political Science 56, 1, 84-97.
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 8 (10/26): Categorical and Dirichlet Distributions and Other Distributions on the Simplex

- *Readings:*
  - Chapter 2 Bishop, Christopher. 2006. Pattern Recognition and Machine Learning (Sections 2.1, 2.2 especially).
  - Katz, Jonathan and Gary King. 1999. A Statistical Model for Multiparty Electoral Data American Political Science Review 93, 1, 15-32.
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 9 (11/02): Clustering Methods 1

- *Readings:*
  - Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer.
  - Chp 9. Bishop, Christopher. 2006. Pattern Recognition and Machine Learning (Sections 2.1, 2.2 especially)
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 10 (11/09): Clustering Methods 2

- *Readings:*
  - Grimmer, Justin and Gary King. 2011. General Purpose Computer-Assisted Clustering and Conceptualization. Proceedings of the National Academy of Sciences 108(7), 2643-2650.
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 11 (11/16): Topic Models 1: LDA

- *Readings:*
  - Blei, David, Andrew Ng, and Michael Jordan. 2003. Latent Dirichlet Allocation Journal of Machine Learning.
  - Blei, David. 2012. Probabilistic Topic Models. Communications of the ACM. 55, 4, 77-84.
  - Wallach, Hanna, David Mimno, and Andrew McCallum. Rethinking LDA: Why Priors Matter. Proceedings of the 23rd Annual Conference on Neural Information Processing.
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 12 (11/23): Topic Models 2: sLDA

[*Thanksgiving*]

- *Readings:*
  - Quinn, Kevin et al. 2010 How to Analyze Political Attention with Minimal Assumptions and Costs. American Journal of Political Science, 54, 1 209-228.
  - Grimmer, Justin. 2010. A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases. Political Analysis, 18(1), 1-35.
  - Chp 5. Wallach, Hanna. Structural Topic Models for Language. `http://people.cs.umass.edu/~wallach/theses/wallach_phd_thesis.pdf`
  - Roberts, Margaret E., et al. Structural Topic Models for Open?Ended Survey Responses. American Journal of Political Science 58.4 (2014): 1064-1082.
  - Roberts, Margaret, Brandon Stewart, and Edo Airoldi. Structural Topic Models. Harvard University.
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 13 (11/30): Classification: Naive Bayes, SVM, Ensemble Classifiers

- *Readings:*
  - Hopkins, Dan and Gary King. 2010. ?A Method of Automated Nonparametric Content Analysis for Social Science? American Journal of Political Science, 54, 1.

- Yu, Bei, Stefan Kaufmann, and Daniel Diermeier. 2008. Classifying Party Affiliation from Political Speech?. Journal of Information, Technology, and Politics. 5(1).
- D orazio et al. Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines. Political Analysis 22, 2 224-242.,
- 7.10. Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer.
- Hillard, Dustin, Stephen Purpura and John Wilkerson. 2007. Computer Assisted Classification for Mixed Methods Social Science Research. Journal of Information, Technology, and Politics.

- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 14 (12/07): Model Fit, Complexity and Cross Validation

- *Readings:*
  - Chp 7. Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer.
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.

Week 15 (12/14): Word Scores and Item Response Theory

- *Readings:*
  - Lowe, Will. 2008. Understanding Wordscores. Political Analysis. 16, 356-371.
  - Jackman, Simon, Joshua Clinton and Doug Rivers. 2004. The Statistical Analysis of Roll Call Data. American Political Science Review 98, 2, 355-370.
  - Slapin, Jonathan and Sven-Oliver Prokschk. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. American Journal of Political Science. 52, 3 705-722.
  - Beauchamp, Nick. 2012. Using Text to Scale Legislatures with Uninformative Voting. Northeastern University
- *Homework/Lab:*
  - R Code and Questions Provided Via Github.