Modern Text Analysis and Machine Learning for Social Research

[POLS 8500] University of Georgia, Fall 2019
Room 302, Baldwin Hall, 3:30-6:15 PM

✉ ljanastas@uga.edu

🌐 https://anastasopoulos.io

https://github.com/ljanastas/POLS-8500-Machine-Learning-Text-Analysis

Prerequisites

A course in probability and a course in statistical inference. Programming experience will be helpful but not necessary. Machine learning algorithms and data pre–processing will be implemented in `R` but you are free to use `Python` or other languages if you choose.

Course Overview and Objectives

This course will provide an introduction to the theory and applications of machine learning algorithms with a focus on text analysis and processing for social research.

The goals of this course include:

- Introducing natural language processing and text analysis methods.
- Developing a basic understanding of the statistical theory underlying common supervised and unsupervised machine learning algorithms.

Required Texts

Hastie, Tibshirani and Friedman. 2013. *The Elements of Statistical Learning* (2nd ed), 7th Printing. Springer Series in Statistics. Available for free here: https://web.stanford.edu/ hastie/Papers/ESLII.pdf.
Referred to in the schedule as **HTF**.

James, Witten, Hastie and Tibshirani. 2015. *An Introduction to Statistical Learning with Applications in R*. Springer Science. Available for free here: http://www-bcf.usc.edu/ gareth/ISL/.
Referred to in the schedule as **JWHT**.

Monogan III, James E. 2015. *Political Analysis Using R*, Springer.
http://link.springer.com/book/10.1007%2F978-3-319-23446-5.
Referred to in the schedule as **M3**.

In addition to these books assigned readings will be available on the ELC.

Attendance and Participation

The most important content from this class will come from the lectures and group assignments during lecture time. Because of this and the technical nature of this class, attendance and participation in class is important.

Computer, Tablet and Cell Phone Use Policy

Laptop computers and tablets may be used during class sessions for note taking and programming exercises. Cell phones and other electronic devices must remain off and stored out of sight at all times during class.

Academic Honesty and Integrity

As a student at the University of Georgia, you have agreed to abide by the University's academic honesty policy. Lack of knowledge of the academic honesty policy is not a reasonable explanation for a violation. Questions related to course assignments and the academic honesty policy should be directed to the instructor.

Special Accommodations

Students with disabilities who require reasonable accommodations in order to participate in course activities or meet course requirements should contact the instructor and work with the Disability Resource Center (https://drc.uga.edu/) to develop an accommodation plan.

Problem Sets

There a total of five problem sets during the semester covering materials discussed in lectures and in the readings. The format of problem set assignments will vary but will invariably involve a combination of math problems and programming. Unless specifically noted on the problem set, these are **individual** assignments so students will need to show independent work. More information about each assignment will be provided in class the week before it is due.

Final Project

Working together in groups or individually, students will propose a course project which either (1) applies one or more of the machine learning algorithms covered to a substantive problem in your relevant discipline or; (2) proposes a method to improve the performance of a machine learning algorithm for a given problem domain. You will be asked to put together a course project proposal halfway through the academic year and the final course project will be due at the end of the semester.

The final course project will contain two components:

(1) A final presentation of the project to the class and others at the end of the semester and;

(2) A paper which will be handed in for a grade.

For students who are pursuing a PhD and are interested in going into academia, you should use the final project as an opportunity to begin a work that will eventually go out for peer review.

For students in the policy or other applied sciences, you may want to use the final project as an item that you can add to your portfolio when it comes time to apply for a job in industry.

Several students who have taken this class in the past have been able to use the skills taught here to begin successful careers in think tanks and private industry.

Discussion Leaders

Students **must** sign up for the role of “discussion leaders” **twice** for this semester. Discussion leaders will lead the class discussion of an applied paper for that week by reading the paper, preparing a 15-30 minute presentation summarizing the paper and proposing 3–5 discussion points to spark a discussion about the content. Every student MUST participate as a discussion leader **twice**.

You can sign up to be a discussion leader here: https://docs.google.com/spreadsheets/d/1dwy2ek-phfchCUu-4GGIJunUC-fm4jw88WxL8dv1oe0/edit?usp=sharing

Grades

Discussion Leader	20%
Problem Sets	40%
Final Project Proposal	5%
Final Project Presentation	5%
Final Project Paper	30%

Overview of Topics

- From text to data, preparing text for analysis:
    - Introduction to text-as-data, review of R and the *quanteda* package in R.
    - Natural language processing: stemming, tokenization, lemmatization .
    - Documents, corpora and document-feature matrices (DFMs).
    - Weighting for document-feature matrices: tf-idf vs. tf.
    - APIs: collecting and processing social media data.
    - Unstructured data extraction and scraping (XML, JSON and HTML).
- Text data and machine learning:
    - Introduction to statistical learning theory.
    - Supervised and unsupervised learning.
    - The bias-variance tradeoff.
- Supervised learning algorithims:
    - k-nearest neighbors
    - Naive Bayes.
    - Decision trees/random forests.
    - Neural networks.
    - Support vector machines.
- Regularization, model selection and inference.
    - Dimensionality reduction.
    - LASSO and regularized logistic regression.
- Unsupervised learning:
    - Latent semantic analysis.
    - Topic models.
    - K-means clustering.
- Special topics (time permitting).
    - Introduction to image analysis and convolutional neural networks.
    - Causal inference with text data.

Tentative Schedule

Week 1: 08.14 **Overview of the course**

Week 2: 08.21 **From text to data**

- Introduction to text-as-data.
- Review of programming in R: R notebooks and R markdown.
- Text analysis with the *quanteda* package in R.

  Readings

  - Quanteda installation: `https://quanteda.io/`.
  - Quanteda quick start guide: `https://quanteda.io/articles/quickstart.html`.
  - **M3** Chapters 1, 2, 10, 11.1-11,4.

  Discussion leader reading:

  Grimmer, Justin, and Brandon M. Stewart. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21, no. 3 (2013): 267-297.

Week 3: 08.28 **Natural language processing**

- Natural language processing: stemming, tokenization, lemmatization.
- Documents, corpora and document-feature matrices (DFMs).
- Weighting for document-feature matrices: tf-idf vs. tf.

  Readings

  - Introduction to NLP Part I: `https://medium.com/analytics-vidhya/introduction-to-natural-language-processing-part-1-777f972cc7b3`

  Discussion leader reading:

  Gentzkow, Matthew, and Jesse M. Shapiro. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78, no. 1 (2010): 35-71.

Week 4: 09.04 **Acquiring and cleaning unstructured text data**

- APIs: collecting and processing Twitter data in R.
- Unstructured data extraction and scraping (XML, JSON and HTML).

Readings

- ∗ Collecting tweets using R: https://medium.com/@GalarnykMichael/accessing-data-from-twitter-api-using-r-part1-b387a1c7d3e
- ∗ Steinert-Threlkeld, Zachary C. "Spontaneous collective action: Peripheral mobilization during the Arab Spring." *American Political Science Review* 111, no. 2 (2017): 379-403.

Discussion leader readings:

- ∗ Mellon, Jonathan, and Christopher Prosser. "Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users." *Research & Politics* 4, no. 3 (2017): 2053168017720008.

PROBLEM SET 1 DUE

Week 5: 09.11 **Introduction to machine learning for text analysis**

- What is machine learning?
- Supervised & unsupervised learning.

Readings

- ∗ Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S., 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), pp.237-293. https://academic.oup.com/qje/article/133/1/237/4095198.
- ∗ **JWHT** – Introduction, pp 1-15.
- ∗ **SM** – Chapter 1.

Discussion leader reading:

- ∗ Anastasopoulos, L. Jason, and Andrew B. Whitford. "Machine learning for public administration research, with application to organizational reputation." *Journal of Public Administration Research and Theory* 29, no. 3 (2018): 491-510.

Week 6: 09.18 **Introduction to statistical learning theory**

- Training, testing and cross–validation.
- Assessing model accuracy.
- Overfitting.
- Regression vs. classification problems.
- The Bias–Variance tradeoff.

  Readings

  - **JWHT** – Statistical Learning, pp 15-37, 176–184.
  - **SM** – 2.1, 2.2,2.3.

  Discussion leader reading:

  - Mitts, Tamar. "From isolation to radicalization: anti-Muslim hostility and support for ISIS in the West." *American Political Science Review* 113, no. 1 (2019): 173-194.

Text Analysis with Supervised Learning

Week 7: 09.25 **Nearest Neighbors**

- kNN algorithm.

  Readings

  - **JWHT** – pp 39–42.

  PROBLEM SET 2 DUE

Week 8: 10.02 **Naive Bayes**

- Review of probability and Bayes rule.
- Learning with naive Bayes.

  Readings

  - Collins on Naive Bayes
    http://www.cs.columbia.edu/ mcollins/em.pdf

    Discussion leader reading:

    - Cantu, Francisco, and Sebastián M. Saiegh. "Fraudulent democracy? An analysis of Argentina's infamous decade using supervised machine learning." *Political Analysis* 19, no. 4 (2011): 409-433.

Week 9: 10.09 **Decision Trees**

- Decision tree–based methods.
- Model assessment, information gain and “white box” methods.

Readings

- ∗ **JWHT** – Chapter 8.

  Discussion leader reading:

  - · Anastasopoulos, Jason, and Anthony M. Bertelli. “Understanding delegation through machine learning: A method and application to the European Union.” Forthcoming. *American Political Science Review.*

Week 10: 10.16 **Regression I: Prediction**

- Inference vs. prediction.
- Linear regression as a machine learning algorithm.
- Logistic regression as a machine learning algorithm.
- Parameter estimation via gradient descent.

Readings

- ∗ **JWHT** – Chapter 3.
- ∗ Ruder, S., 2016. An overview of gradient descent optimization algorithms. arXiv preprint arXiv:1609.04747.https://arxiv.org/pdf/1609.04747v1.pdf.

Discussion leader reading:

- ∗ Muchlinski, David, David Siroky, Jingrui He, and Matthew Kocher. “Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data.” *Political Analysis* 24, no. 1 (2016): 87-103.

PROBLEM SET 3 DUE

Week 11: 10.23 **Regression II: Model Selection and Inference**

- Feature selection.
- Regularization.
- Shrinkage methods: ridge regression, LASSO, Bayesian LASSO.

Readings

∗ **JWHT** – Chapter 5.1, 203-243.

Discussion leader reading:

∗ Kim, In Song, John Londregan, and Marc Ratkovic. "Estimating spatial preferences from votes and text." *Political Analysis* 26, no. 2 (2018): 210-229.

FINAL PROJECT PROPOSAL DUE

Week 12: 10.30 **Neural Networks**

- Overview of neural networks.
- Fitting neural networks with backpropagation.

Readings

∗ Daume. Neural Networks.

Discussion leader reading:

∗ Anastasopoulos, L. Jason, Dhruvil Badani, Crystal Lee, Shiry Ginosar, and Jake Williams. "Photographic home styles in Congress: a computer vision approach." *arXiv preprint* arXiv:1611.09942 (2016).

PROBLEM SET 4 DUE

Week 13: 11.06 **Support Vector Machines and Classifier Choice**

- Comparing classifiers: performance & interpretability.

Readings

∗ **JWHT**. 337-366.

Discussion leader reading:

∗ D'Orazio, Vito, Steven T. Landis, Glenn Palmer, and Philip Schrodt. "Separating the wheat from the chaff: Applications of automated document classification using support vector machines." *Political analysis* 22, no. 2 (2014): 224-242..

UNSUPERVISED LEARNING

Week 14: 11.13 Unsupervised Learning II: Topic models and latent semantic analysis.

- Latent semantic analysis.
- Topic models and their flavors.

Discussion leader reading:

- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58, no. 4 (2014): 1064-1082.

PROBLEM SET 5 DUE