POLSCI 590: Intro to Machine Learning and Text as Data

Scott de Marchi (Professor of Political Science and Director of Decision Science)
demarchi@duke.edu

Time: TBA
Location: online

Overview

In this course, we will cover the introductory material on machine learning and natural language processing (NLP), with an eye toward social science applications. Methodologically, we will work in three main areas. The first is a short introduction to machine learning with a focus on regression-based models. The second is the discovery / creation of latent spaces, where we will focus on PCA, (simple) autoencoders, and embeddings. The third is a general introduction to NLP and text as data, where we will focus on topic models to develop intuition and transformers (preferred for most applications).

Throughout the course, we will rely on datasets of special interest to the social sciences. In particular, we will use legislation in the US Congress and news stories (i.e., event data).

Grading and Assignments

Projects (60%): Most weeks will have an assignment due – these assignments will involve scraping data, writing code, and building statistical models. The expectation is that you will become more competent computer scientists / coders and document your code, learn to write test cases, etc.

Final exam (20%): the exam will cover material from the course and focus on creative applications of this material.

Participation (20%): Different students have different preferences when it comes to participation, but engagement with the material and with your colleagues in class is essential. Accordingly, both in-class, out-of-class, and email participation will be considered in determining this portion of the grade.

Reading: Some of the articles in this course are challenging and contain mathematical notation you may be unfamiliar with. This is by design! If you are going to pursue further studies in this area, reading papers that contain mathematical models is a crucial skill to learn. Take time with the reading and come to class with questions. Be comfortable with the fact that you may not understand every step in these models and that multiple attempts may be required. This course is designed to get your feet wet without drowning you, so *don't panic*.

Most of our readings will be online, but *buy* Geron's book: https://www.amazon.com/gp/product/B07XGF2G87/ref=dbs_a_def_rwt_bibl_vppi_i0 as and the Tunstall, von Werra, and Wolf book: https://www.amazon.com/Natural-Language-Processing-Transformers-Revised-dp-1098136799/dp/1098136799/ref=dp_ob_title_bk. There are short (online) readings from other books and in many cases these were selected to expose you to the central texts in machine learning.

Shadowing and group work: In my experience, the best way to learn this material is if I shadow you. Each week where there is an assignment, I'll also do the assignment. We'll "exchange" our models & code on Friday and see where everyone went wrong. It's OK to work in groups, but a) be clear who you are working with when you turn in assignments and b) don't copy directly from each other. You have to try and implement things yourself. Groups are for bouncing ideas off of each other and getting help on some of the details; it's not joining the borg and copying from each other whole-cloth.

Note: given the pandemic and the fact that this is an online class, it would be helpful if you could send me questions on the reading / assignments early in the week so I can write down responses to share with the class. If you could, send me all questions – big or small – by *noon on Wednesday*.

First assignment (due Monday, 8/24):

1. Install python and anaconda on your machine (https://www.anaconda.com/distribution/) – use the 64-bit version.
2. Launch anaconda / jupyter and write a "hello world" python program. For help, see https://www.amazon.com/Python-Programming-Introduction-Computer-Science/dp/1590282752/ref=sr_1_3?dchild=1&keywords=python+computer+science&qid=1597710154&sr=8-3 or any of the online tutorials on python 3 (*not* python 2).
3. Get a dropbox account (https://www.dropbox.com/). The free one has enough storage for our needs.
4. Get a google Colab account https://colab.research.google.com/.

Readings:

Week 2 (questions due 8/26): **Introduction to machine learning**. Read Chapter 1 2.1, 2.2, 3.1, 3.2 in https://www.statlearning.com/, https://www.cs.ubc.ca/~murphyk/MLbook/pml-intro-22may12.pdf, Watch https://www.youtube.com/watch?v=SjQyLhQIXSM.

Week 3 (9/2): **Regression**. Read 3.3 and do 3.6; read 4.1, 4.2, 4.3, and 7.1 in https://www.statlearning.com/.

Week 4 (9/9): **Regularization**. Read 5.1, 5.2, and do 5.3; read 6.1 and 6.2 in https://www.statlearning.com/, https://web.stanford.edu/~hastie/glmnet/glmnet_beta.html.

Week 5 (9/16): **Unsupervised learning**. Read 10.1, 10.2, and 10.6 in https://www.statlearning.com/, Watch https://www.youtube.com/watch?v=P5mlg91as1c&t=92s.

Week 6 (9/23): **Deep learning and information theory**. Read

Chapter 8 in Geron,
https://arxiv.org/abs/2003.11755,
Pages 71-77 in https://www.deeplearningbook.org/contents/prob.html.

Week 7 (9/30): **Autoencoders and neural nets**. Read
Chapter 17 in Geron up to the section on variational autoencoders,
Chapter 7 in https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf,
If especially interested, chapter 6 in
https://www.deeplearningbook.org/contents/mlp.html.

Week 8 (10/7): **Topic models**. Read
https://www.jmlr.org/papers/v3/blei03a,
https://nyuscholars.nyu.edu/en/publications/text-preprocessing-for-unsupervisedlearning-why-it-matters-when-,
https://www.cambridge.org/core/journals/journalof-public-policy/article/policy-and-the-structure-of-roll-call-voting-in-the-ushouse/E867347CA19B1B6DC34CB7E10F957955
(focus on the online appendix, skim the main article).

Week 9 (10/14): **Classic NLP**. Read
Chapters 3, 8, 20 in https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf.

Week 10 (10/21): **Embeddings**. Read
Chapter 6 in https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf,
https://spacy.io/usage/linguisticfeatures,
https://jalammar.github.io/illustrated-word2vec/,
https://arxiv.org/pdf/1301.3781.pdf.

Week 11 (10/28): **Transformers**. Read
https://jalammar.github.io/illustrated-transformer/,
https://towardsdatascience.com/deconstructing-bert-part-2-visualizing-the-innerworkings-of-attention-60a16d86b5c1,
Chapters 1 and 2 of Tunstall et al.,
https://arxiv.org/abs/1706.03762,
https://arxiv.org/abs/1810.04805,
https://arxiv.org/abs/1907.11692.

Week 12 (11/4): **Zero-shot and Few-shot Learning**. Read
https://discuss.huggingface.co/t/new-pipeline-for-zero-shot-text-classification/681
https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

Week 13 (11/11): no reading, preparation for the exam, advice on next steps.