# Machine Learning for Linguistics

(Tentative plan, subject to change)

LING-5444 (Spring 2024, room TBA)
Mon+Wed 11-12:15

**Instructor:**
Amir Zeldes
E-Mail: amir.zeldes@georgetown.edu
Website: https://gucorpling.org/amir/
Office: Poulton Hall 243 (office hours TBA)

**Summary:**

In the past few years, the advent of abundant computing power and data has catapulted machine learning to the forefront of a number of fields of research, including Linguistics and especially Natural Language Processing. At the same time, general machine learning toolkits and tutorials make handling 'default cases' relatively easy, but are much less useful in handling non-standard data, less studied languages, low-resource scenarios and the need for interpretability that is essential for drawing robust inferences from data. This course gives a broad overview of the machine learning techniques most used for text processing and linguistic research. The course is taught in Python, covering both general statistical ML algorithms, such as linear models, decision trees and ensembles, and current deep learning models, such as deep neural net classifiers, recurrent networks, transformers, sequence to sequence LLMs and contextualized continuous meaning representations. The course assumes good command of Python (ability to implement a program from pseudo-code) but does not require previous experience with machine learning.

**Course requirements and final grade breakdown:**

| | |
|---|---|
| Attendance | |
| Homework assignments + presentations | 35% |
| Midterm | 20% |
| Final project | 35% |
| Participation | 10% |

**Course plan** (tentative and very much changeable!)

| Week | Main topic | Sub topics | Assignments (tentative) |
|---|---|---|---|
| 1 | Introduction | | Read Banko & Brill 2001 |
| | | Linear model basics, handling data | L2 proficiency |
| 2 | Classification foundations | Binary classification | |
| | | Multinomial classification | Logistic regression |
| 3 | Gradient Descent and Regularization | SGD, mini batches | |
| | | Regularization, Ridge, Lassoo, ElasticNet | Text classification |
| 4 | SVMs | Margins and decision functions, hinge loss | |
| | | Kernelized SVMs | Discourse parsing |
| 5 | Tree based models | CART, Gini impurity | |
| | | Random forest, permutation importance | Coreference |
| 6 | Ensembles | Ensembles (more RF, adaboost, stacking) | |
| | | Gradient boosting (GBM, xgboost) | Mid term |
| 7 | Neural networks | ANN basics, Word embeddings | Halevy et al. (2010) |
| | | MLPs | Dependency parsing |
| 8 | Hyperparameters | Initialization, drop out and normalization | Reading from Goldberg (2017) |
| | | Optimizers, hyperparameter optimization | Manning (2015) |
| 9 | RNNs | GRU, LSTM | Discourse signals |
| | | Contextualized embeddings (ELMo) | |
| 10 | Transformers | BERT | Training BERT from scratch |
| | | XLNet and other models | |
| 11 | Seq2seq models | T5, FLAN | Prompt-based approaches |
| | | RLHF for LLMs | |
| 12 | Dimensionality reduction | PCA, tSNE | |
| | | Other topics (CNN, GAN … ) | |
| 13 | Conclusion | | |

A major goal of the course is to make data transparent, explorable and open to questions, as opposed to an undisputed given, and to see models for what they are: learned and interpretable deterministic mappings from feature representations to outcomes, rather than black boxes. Throughout the course the focus will be on language data, and going beyond 10-line tutorials that leave little room for modification.

By the end of the course, students should be familiar with major strands of machine learning algorithms as they apply to language data, including numerical regression, classification, regularization, and feature representation. We will discuss feature scaling and scaling-invariance/vulnerability, the importance of technical details such as loss functions, initialization strategies, early stopping and hyperparameter optimization, as well as attention to dataset composition, including stratification, cross-validation and covert overfitting. In the discussion of neural networks we will go over contemporary representations of language, including contextualized embeddings (BERT and other transformers, using the Huggingface implementations, sequence to sequence models) and character-based models, but also the integration of features beyond word embeddings for robust and interpretable learning approaches.

**Literature**

The course does not use a textbook directly, but adapts parts of the following text books, which are recommended as additional reading. Specific papers and excerpt readings will also be placed online during the course.


- Géron, A. (2022) Hands-on Machine Learning with Scikit-Learn & TensorFlow. Third Edition. Sebastopol, CA: O'Reilly.
- Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan Claypool.
- Jha A. R., Pillai, G. (2021) *Mastering PyTorch*. Sebastopol, CA: O'Reilly.
- McMahan, B. & Rao, D. (2019) Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning. Sebastopol, CA: O'Reilly.
- Tunstall et al. (2022). Natural Language Processing with Transformers. Sebastopol, CA: O'Reilly.

**Notice regarding Title IX/Sexual Misconduct:**

Georgetown University and its faculty are committed to supporting survivors and those impacted by sexual misconduct, which includes sexual assault, sexual harassment, relationship violence, and stalking. Georgetown requires faculty members, unless otherwise designated as confidential, to report all disclosures of sexual misconduct to the University Title IX Coordinator or a Deputy Title IX Coordinator. If you disclose an incident of sexual misconduct to a professor in or outside of the classroom (with the exception of disclosures in papers), that faculty member must report the incident to the Title IX Coordinator, or Deputy Title IX Coordinator. The coordinator will, in turn, reach out to the student to provide support, resources, and the option to meet. [Please note that the student is not required to meet with the Title IX coordinator.]. More information about reporting options and resources can be found on the Sexual Misconduct Website: https://sexualassault.georgetown.edu/resourcecenter.

If you would prefer to speak to someone confidentially, Georgetown has a number of fully confidential professional resources that can provide support and assistance. These resources include:
Health Education Services for Sexual Assault Response and Prevention: confidential email sarp@georgetown.edu

Counseling and Psychiatric Services (CAPS): 202.687.6985 or after hours, call (833) 960-3006 to reach Fonemed, a telehealth service; individuals may ask for the on-call CAPS clinician

More information about reporting options and resources can be found on the Sexual Misconduct Website.