Machine Learning for Linguistics

LING-504 (Spring 2020, Healy 104)
Mon+Wed 11-12:15

Instructor:
Amir Zeldes
E-Mail: amir.zeldes@georgetown.edu
Website: http://corpling.uis.georgetown.edu/amir
Office: Poulton Hall 243 (office hours Wednesdays, 3:30-5:00)

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, SVMs, decision trees and ensembles, and current deep learning models, such as deep neural net classifiers, recurrent networks 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	CNNs	Basic CNN, text	Morphological analysis
		CNNs and images	
10	RNNs	GRU, LSTM	Discourse signals
		Contextualized embeddings	
11	Reinforcement Learning(?)		TBA
12	Dimensionality reduction	PCA	Stylometry
		tSNE	
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, crossvalidation and covert overfitting. In the discussion of neural networks we will go over contemporary representations of language, including contextualized embeddings 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. (2019) Hands-on Machine Learning with Scikit-Learn & TensorFlow. Sebastopol, CA: O'Reilly.

Goldberg, Y. (2017). *Neural Network Methods for Natural Language Processing*. Morgan Claypool.

McMahan, B. & Rao, D. (2019) Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning

Notice regarding sexual misconduct:
Please know that as a faculty member I am committed to supporting survivors of sexual misconduct, including relationship violence, sexual harassment and sexual assault. University policy also requires me to report any disclosures about sexual misconduct to the Title IX Coordinator, whose role is to coordinate the University's response to sexual misconduct.

Georgetown has a number of fully confidential professional resources who can provide support and assistance to survivors of sexual assault and other forms of sexual misconduct. These resources include:

Jen Schweer, MA, LPC
Associate Director of Health Education Services for Sexual Assault Response and Prevention
(202) 687-0323
jls242@georgetown.edu

Erica Shirley, Trauma Specialist
Counseling and Psychiatric Services (CAPS)
(202) 687-6985
els54@georgetown.edu

More information about campus resources and reporting sexual misconduct can be found at http://sexualassault.georgetown.edu.