CIS 6190: Machine learning for sequences

Winter 2024

Overview:

Machine learning research aims to build computer systems that learn from experience. Learning systems are not directly programmed by a person to solve a problem, but instead they develop their own program based on examples of how they should behave, or from trial-and-error experience trying to solve the problem. These systems require learning algorithms that specify how the system should change its behavior because of experience. Researchers in machine learning develop new algorithms and try to understand which algorithms should be applied in which circumstances. Machine learning is an exciting interdisciplinary field, with historical roots in computer science, statistics, pattern recognition, and even neuroscience and physics. In the past 10 years, many of these approaches have converged and led to rapid theoretical advances and real-world applications. This course emphasizes machine learning for sequential data processing. It covers common challenges and pre-processing techniques for sequential data such as text, biological sequences, and time series data. Students are exposed to machine learning techniques, including classical methods and more recent deep learning models, so that they obtain the background and skills needed to confront real-world applications of sequential data processing. May be offered in conjunction with DATA*6400.

Instructor: Enas Altarawneh

Email: ealtaraw@uoguelph.ca

Lecture Hours: Wed W 2:30 PM - 5:20 PM

Lecture Location: Guelph, ROZH 105

Office Hours: By appointment indvidual or with project teams (online via Zoom)

Class Attendance: Attendance of lectures is expected but not required.

Readings:

There is no required textbook for this course. There are several recommended books.

- Machine Learning Yearning, 2nd edition, by Andrew Ng
- Neural Networks and Deep Learning, by Michael Nielsen.
- Deep Learning, by Ian Goodfellow, Yoshua Bengio and Aaron Courville
- Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd Edition, by Wes McKinney

- Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 2nd Edition

Course requirements and evaluation:

The format of the class will be lecture. Participation is encouraged. The evaluation of the course will be based on projects done by teams (of up to three)

Project Deliverables:

Deliverable	Date due	Weight
Proposal (~2 pages)	1st Feb	10%
Midterm in-class presentation	15th Feb	10%
Midterm progress report (~5 pages)	1st Mar	20%
Final presentation	29th March	20%
Final report (~7 pages)	12th April	40%

Topics covered:

- Data processing techniques. Including data integration, handling dirty and missing data, data transformation, data reduction. The importance of data processing.
- Introduction to supervised learning and classification. Including types of classification such as binary and multi class classification. Applications and how to re-define problems as sequence or time series classification problems.
- Brief introduction on unsupervised learning and clustering. Clustering techniques (such as KNN). Evaluating clustering effectiveness.
- Explaining the inner works of machine learning models. Specifically neural networks and transformers, such as, Convolutional Neural Network (CNN), long short-term memory (LSTM), BERT and Roberta.
- Understanding embedding as a representation of sequence data. How to encode the semantics in embeddings and use in the learning models.
- How to train models. Including data splitting, handling imbalanced data, Loss functions, prediction rules, learning rate and early stopping.
- The evaluation of models including data splitting, establishing baselines, classification reports and the use of appropriate metrics. The effect of training data size on accuracy.

Expected pacing:

Topic	Week	Date
Data processing	1	11th Jan
Supervised learning	2	18th Jan
Unsupervised learning	3	25th Jan
Neural Networks 1	4	1st Feb
Neural networks 2	5	8th Feb

Midterm presentations	6	15th Feb
Reading week	7	22nd Feb
Embeddings and models	8	1st Mar
Concepts in training models	9	15th Mar
Concepts in model evaluation	10	22nd Mar
Final presentations 1	11	29th Mar
Final presentations 2	12	5th April