BMGT808X: Machine Learning for Business Research
(Term A, Fall 2022)

Logistics

Instructor: Kunpeng Zhang

Time: Friday 11:00 – 14:00

Location: VMH 4333 (conference room)

Learning Goals

By the end of the semester, students should be able to:

1. Understand and evaluate the use of basic machine learning and data science in research papers.
2. Implement data science techniques in Python.
3. Formulate, conduct, and position data science research for publication in business school journals.

Course Description

The design of this class is to work towards a deeper understanding of data science and especially research on data science, based on a combination of readings on fundamental material, such as textbook chapters or classic papers, plus more recent research that uses those fundamentals. We will dig deep into the material in our class discussions, so we may end up spending entire class periods discussing one paper -- meaning discussing the fundamentals underlying what is presented in the paper, plus the particular contributions of the paper, and the general notions of contributions to research in data science, technical information systems, design science, etc., as well as how one actually goes about conducting research (really, not in some ideal world).

We will especially focus on data science and/or method papers that have been published in top business school and computer science journals, and discuss distinguishing characteristics of such papers. We might invite Smith faculty who have used machine learning in their research (across DO&IT, marketing, finance, and possibly others) to guest lecture throughout the semester.

The second component of the class (taught by Jessica Clark, in Term B) will be practical: learning essential tools that data science researchers use to conduct their research. The two components of the class will be taught concurrently, as data science skills are essential to conducting data science research. We will encourage that students complete a project for the class using these tools, ideally one which could lead to a research paper.

Course Outline

- **1: Introduction to Machine Learning**
  - Terminologies and definitions
  - Bias vs. Variance
  - Learning paradigms
  - Evaluation metrics

Readings:

- Chapter 1 & 2 of Probabilistic Machine Learning: An Introduction (PDF version)

- **2: Unsupervised learning**
  - Measuring (dis)similarity and evaluating the output of clustering methods
  - Traditional clustering methods: K-means, DBSCAN
  - Dimension reduction: PCA, Matrix Factorization, and t-SNE

Readings:

- Rui Xu and D. Wunsch, "Survey of clustering algorithms," in *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678
- Yu-Xiong Wang and Yu-Jin Zhang, "Nonnegative matrix factorization: A comprehensive review," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 6, pp. 1336-1353
- van der Maaten, Laurens & Hinton, Geoffrey, "Viualizing data using t-SNE," Journal of Machine Learning Research. 9. 2579-2605.
- Jolliffe Ian T., Cadima Jorge, "Principal component analysis: a review and recent developments," Phil. Trans. R. Soc. A.374:20150202
- Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press. pp. 226–231.

- **3: Deep learning**
  - Introduction
  - Deep neural nets (DNN)
  - Convolutional neural nets (CNN)
  - Sequence learning (recurrent neural nets, RNN)

Readings:

- Part II of Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville

- **4: NLP: text mining**
  - Text representation
  - Conventional ML techniques
  - Word embedding

    - Transformer-based language models (BERT)

Readings:

- Mikolov et al. (2013), "Efficient estimation of word representation in vector space," Proceedings of Workshop at ICLR. 2013.
- Vaswani et al. (2017), "Attention is all you need," *In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17).* Curran Associates Inc., Red Hook, 6000–6010.
- Devlin et al. (2019), "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1, pages 4171–4186

- **5: Generative Models (self-study)**
    - Directed graphical models – Bayes nets
    - Undirected graphical models – Markov networks
    - Inference – (collapsed) MCMC, variational inference
    - Deep generative models: VAE, GAN and Flow-based models

Readings:

- Introduction to probabilistic graphical models, By Michael Jordan
- An introduction to variational autoencoders, by Diederik P. Kingma, Max Welling
- A survey on Generative Adversarial Networks: variants, application, and training, by Abdul Jabbar, Xi Li, and Bourahla Omar
- Normalizing flows: an introduction and review of current methods, by Ivan Kobyzev, Simon J.D. Prince, and Marcus A. Brubaker

- **6: Social Network Analysis**
    - Basic network characteristics
    - Community detection
    - Network representation learning – network embedding: GAT and GCN

Readings:

- Thomas N. Kipf and Max Welling, (2017), "Semi-supervised classification with graph convolutional networks," in Conference at ICLR 2017.
- Velickovic et al. (2018), "Graph attention networks," in Conference at ICLR 2018.
- Graph neural networks – textbook by Wu et al.