Claire Kelling
STAT 540

Investigation of Causal Inference using Machine Learning

2018-02-27
Project Proposal

Introduction: What is Machine/Deep Learning?:

Machine learning is a broad class of algorithms that uses core ideas of Artificial Intelligence to mimic human decisions [8]. *Deep learning* is a subset of machine learning where computers build models based off of a technique that comes naturally to humans: learning by example [12]. Deep learning is often said to be the next advancement in machine learning techniques. While most machine learning techniques require direction on how to proceed with accurate prediction through the use of additional data, deep learning models are able to learn through its own method of computing [4].

LeCun, Bengio, and Hinton develop a review of the state of deep learning in their recent article in Nature, and state why deep learning has the potential to create key contributions [6]. They highlight applications to social and biological sciences, where it has proven to be useful in image recognition, speech recognition, predicting mutations in non-coding DNA, and natural language understanding.

Casual Inference with Machine Learning:

A particular application of machine and deep learning to social science is the use of deep learning to study causal inference. For example, in experimental design, statisticians and scientists often consider blocking observations to improve the estimation of effects. Grimmer highlights two advancements in this work [3]. Higgins and Sekhon use graph theory to create an algorithm for blocking with multiple treatment categories. They block my minimizing the Maximum Within-Block Distance (MWBD) [5]. Moore and Moore study a blocking algorithm for experiments that arrive sequentially [9].

Both of these techniques combine machine learning and experimental design techniques to improve our understanding and estimation of causal effects. However, these techniques are used in the experimental design phase, and this is not possible in numerous social science applications, where we rely on observational data. A very recent paper, by Louizos et al., create a new technique for dealing with observational data by again combining techniques from machine learning and causal inference research, through the use of proxy variables. They introduce a model called the Causal Effect Variational Autoencoder (CEVAE) which is a neural network latent variable model used for estimating causal effects [7]. Through this project, I will study Variational Auto-Encoder algorithms (VAE's) as well as this new VAE for causal inference. Assuming I am able to, I will also implement the CEVAE model (details are not extremely clear in the paper).

VAE's depend on the structure of a neural network and are often trained with stochastic gradient descent. Therefore, this project will involve the understanding of many concepts, including latent variable models, variational Bayesian methods, neural networks, and "minimum description length" coding [2]. I anticipate diving into some of these topics in detail in my presentation/report, but some may be out of the scope of this project for studying in detail.

In Summary:

Through this project, I will investigate the combination of machine learning/deep learning methods on causal inference for observational data. I will learn about how to implement general VAE's as well as the recently developed CAVAE algorithm. I anticipate that I may use multiple VAE's in combination with the CAVAE to compare model performance using simulation studies. For example, I may I compare the typical VAE with the newly developed Ladder VAE in simulation studies, when combined in the CAVAE framework [11]. I will draw conclusions about the use of these models for observational data, through simulation studies. Time allowing, there are many datasets available publicly for applications with machine learning that I could use for analysis of a dataset [1]. Also time allowing, I will review the contributions of Pham and Shen in their recent paper "A Deep Causal Inference Approach to Measuring the Effects of Forming Group Loans in Online Non-profit Microfinance Platform" [10].

References

[1] Skymind team Chris V. Nicholson Adam Gibson. *Open Data for Deep Learning*. URL: `https://deeplearning4j.org/opendata`.

[2] Carl Doersch. "Tutorial on variational autoencoders". In: *arXiv preprint arXiv:1606.05908* (2016).

[3] Justin Grimmer. "We are all social scientists now: how big data, machine learning, and causal inference work together". In: *PS: Political Science & Politics* 48.1 (2015), pp. 80–83.

[4] Brett Grossfeld. *A Simple Way to Understand Machine Learning vs Deep Learning*. July 2017. URL: `https://www.zendesk.com/blog/machine-learning-and-deep-learning/`.

[5] Michael J Higgins and Jasjeet S Sekhon. *Improving Experiments by Optimal Blocking: Minimizing the Maximum Within-block Distance*. Tech. rep. Working Paper, 2013.

[6] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), p. 436.

[7] Christos Louizos et al. "Causal effect inference with deep latent-variable models". In: *Advances in Neural Information Processing Systems*. 2017, pp. 6449–6459.

[8] Bernard Marr. *What Is The Difference Between Deep Learning, Machine Learning and AI?* Dec. 2016. URL: `https://www.forbes.com/sites/bernardmarr/2016/12/08/what-is-the-difference-between-deep-learning-machine-learning-and-ai/#41ae410f26cf`.

[9] Ryan T Moore and Sally A Moore. "Blocking for sequential political experiments". In: *Political Analysis* 21.4 (2013), pp. 507–523.

[10] Thai T Pham and Yuanyuan Shen. "A Deep Causal Inference Approach to Measuring the Effects of Forming Group Loans in Online Non-profit Microfinance Platform". In: *arXiv preprint arXiv:1706.02795* (2017).

[11] Casper Kaae Sønderby et al. "Ladder variational autoencoders". In: *Advances in neural information processing systems*. 2016, pp. 3738–3746.

[12] *What Is Deep Learning? How It Works, Techniques and Applications*. URL: `https://www.mathworks.com/discovery/deep-learning.html`.