samsi
NSF·Duke·NCSU·UNC

Deep Learning Program
Triangle Machine Learning Day
September 20, 2019

SPEAKER TITLES/ABSTRACTS

Nilay Tanik Argon
University of North Carolina

"Data-driven Decision Making in Healthcare Operations"

Researchers have used operations research methods to design and control operations in clinics, hospitals, and pre-hospital settings for decades. Unfortunately, only a small portion of these efforts have successfully translated into practice. The main issue was that most of the prior work on healthcare operations ignored the importance of marriage between data and mathematical models. In recent years, with the availability of abundant data – particularly, electronic health records data -- and growing interest in data analytics, the healthcare operations field has entered a new and exciting era. In this talk, I will provide an overview of my research group's recent efforts in this direction. Specifically, I will explain how we used data-driven approaches to decision making in emergency department operations in collaboration with UNC Health Care, and also discuss future research potentials to incorporate machine learning methods into operational decisions in medical facilities.

Ted Enamorado
Princeton University

"Active Learning for Probabilistic Record Linkage"

Integrating information from multiple sources plays a key role in social science research. However, when a unique identifier that unambiguously links records is not available, merging datasets can be a difficult and error-prone endeavor. Probabilistic record linkage (PRL) aims to solve this problem by providing a framework in which common variables between datasets are used as potential identifiers, with the goal of producing a probabilistic estimate for the unobserved matching status across records. In this paper, I propose an active learning algorithm for PRL, which efficiently incorporates human judgment into the process and significantly improves PRL's performance at the cost of manually labelling a small number of records. Using data from local politicians in Brazil, where a unique identifier is available for validation, I find that the proposed method bolsters the overall accuracy of the merging process. In addition, I examine data from a recent vote validation study conducted for the American National Election Studies (ANES), and I show that the proposed method can recover estimates that are indistinguishable from those obtained from a more extensive, expensive, and time-consuming clerical review.

Sayan Mukherjee
Duke University

"Machine Learning for 3D Imaging"

It has been a longstanding challenge in geometric morphometrics and medical imaging to infer the physical locations (or regions) of 3D shapes that are most associated with a given response variable (e.g. class labels) without needing common predefined landmarks across the shapes, computing correspondence maps between the shapes, or requiring the shapes to be diffeomorphic to each other. In this talk, we introduce SINATRA: the first statistical pipeline for sub-image analysis which identifies physical shape features that explain most of the variation between two classes without the aforementioned requirements. We also illustrate how the problem of 3D sub-image analysis can be mapped onto the well-studied problem of variable selection in nonlinear regression models. Here, the key insight is that tools from integral geometry and differential topology, specifically the Euler characteristic, can be used to transform a 3D mesh representation of an image or shape into a collection of vectors with minimal loss of geometric information.

Crucially, this transform is invertible. The two central statistical, computational, and mathematical innovations of our method are: (1) how to perform robust variable selection in the transformed space of vectors, and (2) how to pullback the most informative features in the transformed space to physical locations or regions on the original shapes. We highlight the utility, power, and properties of our method through detailed simulation studies, which themselves are a novel contribution to 3D image analysis. Finally, we apply SINATRA to a dataset of mandibular molars from four different genera of primates and demonstrate the ability to identify unique morphological properties that summarize phylogeny.

David Page
Duke University

"Machine Learning from De-Identified Coded Electronic Health Records (EHRs)"

This talk begins by showing how accurately 4000 different diagnoses can be predicted in advance for any patient, from one month to twenty years before first occurrence in the patient, using high-throughput machine learning. Shortcomings of this approach motivate ways to turn prediction methods into algorithms for finding causal associations; the resulting algorithms attain high accuracy in tasks of drug repurposing and discovery of adverse drug events, but they do not come with provable guarantees of making correct causal inferences. We then introduce variants of probably-approximately correct (PAC) learning for finding causal associations, that can provide weaker but useful guarantees for such algorithms as these motivated by our experiences with EHR data.

Matthew Phillips
LifeOmic

"Biomedical Image Understanding and EHRs at LifeOmic: Harnessing the Power of the Cloud"

In this talk I'll discuss work in biomedical image and volume segmentation and classification, as well as outcome prediction modeling from insurance claims data that I've pursued at LifeOmic here in the Triangle. In the former case datasets include radiological image volumes, retinal fundus images, and cell images created with fluorescent microscopy. The latter includes MIMIC-III data

represented as FHIR objects. I'll discuss the relative challenges and advantages of doing ML locally vs. on a cloud-based platform.

Xipeng Shen
North Carolina State University

"Adaptive Deep Reuse for Deep Learning"

The speed of Deep Neural Networks (DNN), in both training and inference, is important for its practical usage. This talk presents adaptive deep reuse, a novel optimization to enhance the speed of DNN by efficiently and effectively identifying unnecessary computations in DNN training on the fly. By avoiding these computations, the technique cuts the training time of DNN by 69% and inference time by 50%, with virtually no accuracy loss. The method is fully automatic and ready to be adopted, requiring neither manual code changes nor extra computing resource. It offers a promising way to substantially reduce both the time and energy cost in both the development and deployment of AI products. Since its recent publication, the technique has drawn a lot of interest in media, industry practitioners, and research community.

Wayne Thompson
SAS

"What You Didn't Learn About Machine Learning in School"

Machine learning is all the rage these days. As a data scientist at SAS I thought I would be talking more with customers about machine learning algorithms. You know the stuff we learned in school or by taking online classes. What algorithms to use for what use case? How to tune the model and evaluate for generalization. How to interpret these highly complex nonlinear models? Instead most customers want to know more about putting models into production and managing models once they're deployed. Indeed, model deployment and model management are the two must-have steps to get machine learning right.

A machine learning pipeline is comprised of data wrangling + feature engineering and extraction + model formulae. It may also be layered with rules. Each model includes a lot of data preparation logic. You must aggregate many data sources, include the model formulae, and layer it with rules or policies. Most organizations don't have enough rigor and metadata to re-create the data wrangling phase for scoring. As a result, many of the backward data source dependencies for deriving the new scoring tables get lost. This is the biggest reason why most organizations take too long to put a model to work.

Models begin to degrade as soon as they are deployed. It is important to monitor model drift, retrain champion models and evaluate new challengers. Model fairness and bias must be addressed. Advanced organizations are running model training services at the edge. I believe one of the holy grails of machine learning is being able to orchestrate a continuous learning platform in real-time. One that can adapt as the population is changing. We will discuss these strategies and more to help you get more value out of machine learning.

Susan Xia
Valassis

"Stacking Audience Models -- Using an Ensemble Approach for Predictive Modeling"

In this talk we will discuss a ensemble technique for machine learning known as stacking and how it can be used for predictive modeling. The use case we will focus on is to predict an internet user's likelihood of responding to online advertising. We will describe ensemble learning in general and then stacking more specifically. We will introduce techniques to improve the predictive power of a stacked model. Finally, we will discuss how we deploy our stacked models in production at scale.