Title: Interpretable Machine Learning: Theory and Practice

Speaker: Dr Rajiv Khanna, a Visiting Faculty Researcher at Google Research, and an incoming Assistant Professor at the Department of Computer Science at Purdue University

Date: September 8, 2021.

Time: 10:00 AM

Location: Online

Abstract

Continued and remarkable empirical successes of increasingly complicated machine learning models such as neural networks without a sound theoretical understanding of success and failure conditions can leave a practitioner blind-sided and vulnerable, especially in critical applications such as self-driving cars and medical diagnosis. As such, there has been an enhanced interest in recent times in research on building interpretable models as well as interpreting model predictions. In this talk, I will discuss various theoretical and practical aspects of interpretability in machine learning along both these directions through the lenses of feature attribution and example-based learning.

In the first part of the talk, I will present novel theoretical results to bridge the gap in theory and practice for interpretable dimensionality reduction aka feature selection. Specifically, I will show that feature selection satisfies a weaker form of sub modularity. Because of this connection, for any function, one can provide constant factor approximation guarantees that are solely dependent on the condition number of the function. Moreover, I will discuss that the cost of interpretability accrued because of selecting features as opposed to principal components is not as high as was previously thought to be.

In the second part of the talk, I will discuss the development of a probabilistic framework for example-based machine learning to address ``which training data points are responsible for making given test predictions? “. This framework generalizes the classical influence functions. I will also present an application of this framework to understanding the transfer of adversarial trained neural network models.

Speaker Bio:

Dr. Rajiv Khanna is currently a Visiting Faculty Researcher at Google Research, and an incoming Assistant Professor at the Department of Computer Science at Purdue University. Previously, he was a postdoc at the Department of Statistics at UC Berkeley and was also associated with the Foundations of Data Analytics Institute (FODA) at UC Berkeley. Before that, he was a Research Fellow in the program of Foundations of Data Science at the Simons Institute for the Theory of Computing, also at UC Berkeley.

His research is focussed on elucidating mechanisms of success/failure conditions of machine learning through optimization, learning theory and interpretability. His work on beyond worst-case analysis on the Column Subset Selection won the best paper award at NeurIPS 2020. He earned his PhD in Electrical and Computer Engineering at UT Austin.