Mathematics, Data Science, and Industry
Dec 1-2, 2018
Purdue University

Titles and Abstracts of Presentations

Parsa Bakhtary, YouTube

Title: *Problem Formulation in Product Analytics*

Abstract: While statistical and coding skills are important for the industry data scientist, equally important is developing product sense. Often times the questions one wishes to answer are vague or not fully understood, and as a result we must use our intuition about the product and its user base in order to formulate the problem correctly. We present a series of analytical case studies, based on product questions asked by executives in the realms of gaming (Zynga), social media (Facebook), and video (YouTube).

Mireille Boutin, Department of Electrical and Computer Engineering, Department of Mathematics, and Regenstrief Center for Healthcare Engineering, Purdue University

Title: *Clustering Small Data using Random Projection*

Abstract: "Real" data, i.e. data generated as part of a real-life experiment, is often quite organized. So much so that, in many cases, projecting the data onto a random line has a high probability of uncovering a clear division of the data into two well-separated groups. In other words, the data can be clustered with a high probability of success using a hyperplane whose normal vector direction is picked at random. We will discuss how to quantify this phenomenon empirically, and propose ways to exploit it to cluster datasets. The proposed clustering methods are particularly well-suited to find patterns in small data sets. We will discuss some applications to education research problems studying students in a single classroom (20-30 students).

Kaiser Fung, Founder, Principal Analytics Prep; Author, Numbers Rule Your World; Editor, Junk Charts blog; Lecturer, Tufts University

Title: *The Data Science Boom: What is a Myth, What is Real, and Is it a Fit for You?*

Abstract: The data science and data analytics sector is absolutely booming, with tens of thousands of open positions across all industries. I was lucky to have the right

skills at the right time, and in the last dozen of years, have built and managed data teams at both large and small companies. During this talk, I will offer an overview of the field, discuss the drivers of the booming job market, describe specific examples of how mathematical, statistical and scientific reasoning are applied to solve business problems, and preview some practical lessons that most people have to learn on the job.

The popular view of data science is far too narrow: mainstream and social media characterize it as a monolithic job on an engineering team responsible for writing expert computer code that automates processes, with the single-minded objective of maximizing the amount of data consumed in the least amount of time. I will present a much broader overview of the field that encompasses a large variety of jobs. According to the media, data scientists have a cure-all called *big data*, when fed to algorithms, can predict anything from when you will die to who will rob a bank. Rarely does the media ask about the nature of this data *science*, or what proof there is to back up the extraordinary claims of predictive prowess. The answers to these questions may surprise you, and help you decide if this booming sector is where you belong.

Suresh Garimella, Office of the Executive Vice President for Research and Partnership and Department of Mechanical Engineering, Purdue University

Title: *Purdue's Integrative Data Science Initiative: Robust interdisciplinary collaboration to harness data for the greater good*

David Gleich, Department of Computer Science and Department Mathematics, Purdue University

Title: *Principled higher-order and multi-way methods in data science*

Abstract: Much of the data now routinely produced in applications ranging from neuroscience to social networks and text has rich structure that cannot be captured with traditional pairwise-based analysis procedures. We propose a number of new methods that do reveal this information using novel perspectives on higher-order data that generalize matrix-based and Markov chains to more complex stochastic processes. In particular, we will describe methods that partition networks of data based on higher-order motifs instead of edges; we will also describe methods that cluster tensor data (such as higher-order word co-occurence data) and can automatically generate stop-words lists from text; finally, we conclude with some recent advances in predicting user behavior based on low-parameter representations of higher-order Markov chains.

Our overall goal is to support rigorous algorithm engineering in the space of clustering and community detection of complex data that allows users to describe higher-order clustering desiderata and then to produce methods that seek these structures with either worst-case guarantees or a-posteriori approximation guarantees, and we give a few instances of this philosophy.

Papers:
https://arxiv.org/abs/1712.05825 (in WWW2018)
https://arxiv.org/pdf/1612.08447.pdf (in Science, 2016)
https://arxiv.org/abs/1704.05982 (at KDD2017)
https://arxiv.org/abs/1603.00395 (at NIPS2016)

Software:
https://github.com/nveldt/LamCC
https://github.com/wutao27/RHOMP
https://github.com/wutao27/GtensorSC
https://github.com/arbenson/higher-order-organization-julia

Jason Lucas, Metron, Inc.

Title: *A Brief Explanation of Kalman Filters*

Abstract: Developed in the 1950s and 60s, the Kalman filter is a recursive algorithm for estimating the state of unknown variables based on models describing the behavior of those variables and on some source of information, periodically or sporadically updated and frequently corrupted with noise. They are adept at fusing different kinds of data from different kinds of sensors. As such, Kalman filters have found applications in wide-ranging fields, from tracking/navigation to signal processing to economics. In this brief technical talk, we will work through the details of the filter and apply it to a simple example in tracking.

Huda Nassar, Department of Computer Science, Purdue University

Title: *The Julia Programming Language for Data Science*

Abstract: The two-language problem is a conundrum that data scientists and other software developers face when they are forced to choose between writing efficient code and writing code efficiently. In this talk, we will discuss the implications of the two-language problem for data scientists and how Julia, a new open-source language, will help you overcome it.

Jennifer Neville, Department of Computer Science and Department of Statistics, Purdue University

Title: *Deep Learning for Relational Networks*

Abstract: Although deep learning methods have been successfully applied in structured domains comprised of images and natural language, it is difficult to apply the methods directly to graph and network domains due to issues of heterogeneity and long range dependence. In this talk, I will discuss some of our recent work developing deep learning and representation learning for complex network domains, including node classification, knowledge graph embeddings, and motif prediction. The key insights include incorporating network dependencies in both the input features and the model architectures, using randomization to transform heterogeneous sets into sequences, and using network-aware data augmentation to offset sparsity. Our experiments on real work social network data shows that our methods produces significant gains compared to other state-of-the-art methods.

Raghu Pasupathy, Department of Statistics, Purdue University

Title: *Stochastic Gradient Descent*

Abstract: Stochastic Gradient Descent (SGD), also known as *stochastic approximation*, refers to certain simple iterative structures used for solving stochastic optimization and root finding problems. The identifying feature of SGD is that, much like gradient descent for deterministic optimization, each successive iterate in the recursion is determined by adding an appropriately scaled gradient estimate to the prior iterate. Originally introduced in 1951, SGD has undergone numerous rebirths in popularity and has evolved to become the principal enabling machinery within large-scale machine learning and "big data" contexts such as classification and regression. I will start this talk with a discussion of the environment that has led to the rise of SGD. I will then cover the basic results in SGD with an emphasis on modern developments. I will then motivate and outline some of my recent investigations on adaptive SGD supported by numerical experience.

Dejan Slepcev, Department of Mathematics, Carnegie Mellon University

Title: *Optimization problems on random structures and their continuum limits*

Abstract: We will discuss optimization problems arising in machine learning and their limits as the number of data points goes to infinity. Consider point clouds obtained as random samples of an underlying "ground-truth" measure. Graph representing the

point cloud is obtained by assigning weights to edges based on the distance between the points. Many machine learning tasks, such as clustering and semi-supervised learning, can be posed as minimizing functionals on such graphs. We consider functionals involving graph cuts, graph laplacians and their limits as the number of data points goes to infinity. We will discuss the mathematics needed to connect the discrete and continuum worlds and the insights one can obtain from such connection.

Michael W. Trosset, Department of Statistics, Indiana University

Title: *Distances and Dissimilarities in Mathematics and Data Science*

Abstract: The mathematical concept of distance permeates data science, providing a conceptual framework in which to extract information from a set of pairwise dissimilarities. I will describe four examples of how mathematical properties of distance inform the analysis of high-dimensional data:

(1) the interplay between Euclidean distance geometry and the representation of dissimilarity data by classical multidimensional scaling (CMDS);

(2) the relation between CMDS and the use of principal component analysis (PCA) for linear dimension reduction;

(3) low-distortion embedding and the use of random projection to preprocess data for PCA; and

(4) the interplay between Riemannian geometry and nonlinear dimension reduction via "manifold learning".

Yi Wang, Bloomberg

Title: *ML/AI in Financial Data System: Practice and Practitioners*

Abstract: The speaker will talk about ML/AI practices at Bloomberg, and ML/AI practitioners at Bloomberg. Related to those topics, he will present opportunities at Bloomberg for students and industry practitioners. Bloomberg's presence in the ML/AI field and Bloomberg's contribution to the ML/AI community will also be showcased.

Mark Daniel Ward, Department of Statistics and Department of Mathematics, Purdue University

Title: *A Game Theory Problem from a Computational View*

Abstract: I will discuss a game theory problem that I have been working on, investigating this problem using computational tools. As time permits, I will also discuss the seminars I have been giving in recent years, for people at all levels (students, postdocs, faculty, and professionals).

Nick Wegman, Antuit

Title: *Experiences in Data Science Consulting*

Abstract: The potential application and benefit of data science is not limited to specific industries or sectors. As data science moves into fields tangential to mathematics, statistics, and computer science, additional skills become as important as understanding the latest algorithms and techniques. In this talk I will share my experience consulting for the retail and CPG (consumer packaged goods) industries and the skills that have made me successful when communicating with a non-technical audience.

Pete Weigel, McKinsey & Company

Title: *Ensemble Machine Learning and Churn Reduction*

Abstract: We present a case study to illustrate the typical course of a consulting analytics engagement from identifying the business case to structuring the analytical approach, data engineering, modeling, profiling and explaining the results, and finally to building a prescriptive strategy for applying the insights.

Patrick Wolfe, College of Science, Department of Computer Science and Department of Statistics, Purdue University

Title: *Big (Network) Data: Challenges and Opportunities for Data Science*

Abstract: How do we draw sound and defensible conclusions from big data? This question lies at the heart of data science. In this talk I will first describe some of the challenges and opportunities inherent in this rapidly emerging field, and then discuss the current state of the art in one area of particular interest: big network data. Progress in this area includes the development of new large-sample theory that helps us to view and interpret networks as statistical data objects, along with the transformation of this theory into new statistical methods to model and draw inferences from network data in the real world. The insights that result from connecting theory to practice also feed back into pure mathematics and theoretical computer science, prompting new questions at the interface of combinatorics, analysis, probability, and algorithms.

Biosketch and photo: https://signalprocessingsociety.org/professional-development/distinguished-lecturers#dexp-accordion-item–6