ORGANIZING COMMITTEE
Program Co-Chairs
Cem Iyigun, Middle East Technical University
Ramin Moghaddass, University of Miami
Michele Samorani, Santa Clara University

Management Committee
Tom Au, AT&T Labs
Victoria Chen, University of Texas at Arlington
Kwok-Leung Tsui, City University of Hong Kong
Cynthia Rudin, Duke University
George Runger, Arizona State University

Thank you our Sponsors
Frontline Systems
Salesforce
University of Miami, Department of Industrial Engineering

PROGRAM SCHEDULE
Registration & Breakfast Hilton - 339, Level 3 7:30–8am
Welcome & Keynote - Victoria Chen, Univ. of Texas at Arlington Hilton - 339, Level 3 8–8:45am
Session SA01 - Chair: Shannon Harris Hilton - 337, Level 3 8:45–10:15am
Session SA02 - Chair: Aly Megahed Hilton - 338, Level 3 8:45–10:15am
Break 10:15–10:30am
Keynote - Panos M. Pardalos, Univ. of Florida Hilton - 338, Level 3 10:30–11:15am
Tutorial - Aly Megahed, IBM Research Hilton - 337, Level 3 11:15–12noon
Lunch Hilton - 339, Level 3 12noon–1pm
Tutorial - Michele Samorani, Santa Clara Univ.
Keynote - Galit Shmueli, National Tsing Hua Univ.
Session SB01 - Chair: Angela Zhou, Cornell Tech Hilton - 337, Level 3 1:45–2:30pm
Session SB02 - Chair: Ramin Moghaddas, Univ. of Miami Hilton - 338, Level 3 2:30–4pm
Break 4–4:15pm
Keynote - Endre Boros, Rutgers Univ.
Session SC01 - Chair: Michele Samorani, Santa Clara Univ. Hilton - 337, Level 3 5–6:10pm
Session SC02 - Chair: Cem Iyigun, Middle East Technical Univ. Hilton - 338, Level 3 5–6:10pm
TechnIcal sEsiOns

Saturday, 8:00 - 8:45AM

Keynote
Hilton 339, Level 3
Welcome & Keynote: Statistical Roles in Large-Scale Stochastic Optimization
Contributed Session
Chair: Victoria C. P. Chen, University of Texas at Arlington, Industrial, Manufacturing, & Systems Engr., Campus Box 19017, Arlington, TX, 76019-0017, United States, vchen@uta.edu

1 - Statistical Roles in Large-Scale Stochastic Optimization
Victoria C. P. Chen, University of Texas at Arlington, Industrial, Manufacturing, & Systems Engr., Campus Box 19017, Arlington, TX, 76019-0017, United States, vchen@uta.edu

In complex stochastic optimization problems, analytical solutions are not possible, leading to the use of heuristics or approximate methods. In large-scale problems, even approximate methods can be computationally expensive. In this talk, non-traditional opportunities for data mining methods in decision analytics are presented via two applications to illustrate the numerical solution of such problems. First, an ozone pollution control program is presented that involves a stochastic dynamic program with 524 state variables. To approximate the unknown state transitions and future value functions, a design and analysis of computer experiments approach is implemented using an atmospheric chemistry computer model. Multicollinearity issues are discussed and orthogonalization of the state space is studied. The second example considers adaptive interdisciplinary pain management with 147 state variables. In this case, state transition model development employs actual patient data from the Eugene McDermott Center for Pain Management at the University of Texas Southwest Medical Center at Dallas. The consequence of using observational data from sequential treatment is the presence of time-dependent confounding leading to a form of endogeneity that biases the estimators of statistical model coefficients. The method of inverse-probability-of-treatment weighting is discussed to address the endogeneity. Piecewise-linear network state transition models are incorporated within a two-stage stochastic programming formulation to explore adaptive treatment regimes.

Saturday, 8:45 - 10:15AM

SA01
Hilton 337, Level 3
Contributed Session
Chair: Shannon Harris, The Ohio State University, Columbus, United States, OH, harris.2572@osu.edu

1 - Modeling and Analysis of Chronic Disease Progression using Electronic Health Record (EHR) Data
Vijaya Priya Rama Vijayasarathi, Carnegie Mellon University, Pittsburgh, PA, United States, vijayar@cmu.edu, Reema Padman

Increasing deployment and usage of clinical information systems and the resulting availability of vast amounts of detailed patient data in electronic databases are challenging our ability to utilize them effectively at the point of decision-making. Tracking the progression of disease and identifying high-risk patients with multi-morbid conditions, such as Chronic Kidney Disease (CKD), offer new opportunities for applying innovative data-driven approaches for risk stratification and outcomes prediction. Leveraging the availability of a rich and unique 22-year clinical dataset extracted from the Electronic Health Record (EHR) of a community nephrology practice, we investigate how CKD evolves over time from the onset of a particular stage for an identified patient population. Furthermore, we analyze the different trajectories and their frequencies, the relationship of these trajectories to patient characteristics, and the opportunities for early identification of disease progression profiles of patients in each trajectory group. Applying the novel, semi-parametric, statistical method of group-based trajectory modeling, we detect distinct patient risk groups with differing trajectories that uncover typical progression of CKD within less than two years of an early stage of the disease. Our results suggest that trajectory modeling can shed light on the developmental course of chronic diseases and relevant predictors of group membership to provide interpretable summaries of typical disease progression that can potentially facilitate the development of targeted, proactive interventions for patients in all risk groups.

2 - The Impact of No-Show Predictions on Appointment Scheduling
Michele Samorani, Santa Clara University, 500, El Camino Real, Santa Clara, CA, 95053, United States, msamorani@scu.edu, Shannon Harris

Accurately predicting the attendance behavior of individual clients to clinic appointments makes it possible to overbook appointments more effectively, which can result in monetary savings and higher patient satisfaction. However, there are currently no guidelines on how to determine under which conditions predictive analytics has a significantly positive effect, or on how to select an optimal classification technique and classification threshold when predicting no-shows. In this paper, we provide practical guidelines to implement an appointment scheduling system based upon no-show predictions. We develop a Markov Decision Process model, which, given the confusion matrix corresponding to the no-show predictions, analytically computes the optimal scheduling rules and the average daily cost incurred by the clinic. We find that the improvements resulting from predictive analytics are limited if the no-show rate is low or the predictive performance (measured by the area under the operating characteristic curve) is low. We also find that for high no-show rates, it is near-optimal to choose the classifier that maximizes accuracy; by contrast, for low no-show rates, it is near-optimal to choose the classifier that maximizes the F1 score. Our experiments on three real-world data sets validate our findings.

3 - Does Internet Threaten Prescription Loyalty? A Longitudinal Investigation using PrescriptionData
Utkarsh Shrivastava, University of South Florida, Tampa, FL, United States, utkarsh1@mail.usf.edu, Daniel Zantedeschi, Wolfgang Jank, Philip Stern

Prescription loyalty is an indicator of a physician’s commitment to continue prescribing the drug. Pharmaceutical firms encourage such loyalty because it increases the market shares of their drugs and covers cost of research and development. However, loyalty may be tough to maintain in the digital age as more information about the quality of competing drugs from online sources may reduce the impact of prior experiences and thereby prescription loyalty. Therefore, empirically diagnosing the prescription drug’s loyalty across the physicians would allow firms to optimize expenditure across the range of manufactured drugs and prevent unnecessary costs and overcome the criticism related to over detailing. Using a longitudinal dataset on prescription choices of physicians over a period of ten years we propose a novel approach based on bipartite network of prescriber and prescription nodes to characterize a drug’s loyalty in the market. Two nodes are linked when the physician prescribes the drug and repeat prescription determines the strength of the link. This representation allows to model temporal variations in the loyalty of the prescription drugs. We then investigate the impact of internet diffusion on loyalty of the drug across the physicians. Our findings suggest that internet negatively impacts the loyalty of extremely loyal prescribers while has no impact on the loyalty of prescribers with below average loyalty for the drug. We connect these findings with recent literature on consumer choice in information rich environment.

4 - End-to-End Analytics: Cancer Care Ontario MRI Wait Time Analysis
Alexandra Hayes, University of Waterloo, Waterloo, ON, Canada, ar2hayes@uwaterloo.ca, Christian Gould, Kevin Chu, Srikanth Santhirakumar

End-to-end analytics in this project is aimed at reducing the percentage of late services for Cancer Care Ontario MRI services. Currently, there is no defined method for hospitals to manage the scheduling of patient appointments for MRIs, resulting in exceeded wait time targets. This project alleviates the confusion of patient scheduling through performing end-to-end analytics on historic patient data to provide an innovative heuristic policy specific to each hospital’s complex needs. Key performance indicators (KPIs) are obtained through descriptive analytics to provide insight into the system’s behaviour. Next, additional data sets are generated in the predictive phase using distribution analysis. Finally, a policy recommendation is provided through analysis of a decision matrix that compares the historic KPIs to those of the various heuristic-applied schedules. The aim is to schedule patients effectively in order to meet the ministry’s goal of 90% of patients seen within their priority wait time target. Automation and visualization of the results are utilized to ensure implementation and understanding in a healthcare setting.
12TH INFORMS WORKSHOP ON DATA MINING & DECISION ANALYTICS

SA02
Hilton 338, Level 3
Contributed Session
Chair: Aly Megahed, IBM Research - Almaden, San Jose, CA, United States, aly@gatech.edu

1 - From Correlation to Hierarchy: Practical Topic Modeling via Spectral Inference
Moontae Lee, Cornell University, Ithaca, NY, United States,
moontae@cs.cornell.edu, David Bindel, David Mimno
Topic models were originally applied in text analysis for extracting high-level themes from documents, but they work equally well in any setting where users select items from an inventory. Recent work in spectral topic modeling has provided algorithms that operate on easily-collected summary statistics, rather than exhaustively iterating over the full dataset. The “anchored” algorithms learn topics by decomposing the co-occurrence between pairs of words into a matrix of topics over words and a matrix of relations between topics. While these algorithms provide transparent inference and provable guarantees in addition to scalability, there are several known issues: inference can be infeasible for large vocabularies and cannot learn quality topics on noisy real data with high sensitivity to learning parameters. In this paper, we extend the foundation of anchor-based spectral inference and propose practical algorithms that can efficiently tackle each of these problems within the framework of Joint Stochastic Matrix Factorization. These algorithms prove the provable guarantees and scalability of earlier algorithms, but are more consistent and more stable in identifying quality topics. In addition, this algorithm can also consider and learn meaningful correlations between topics, enabling correlated and hierarchical models. We demonstrate these methods on two text corpora, a corpus of user movie ratings, and a corpus of song playlists.

2 - Estimating an Inverse Mean Subspace
Jiayi Weng, University of Kentucky, 300 Alumni Drive, Apt 179, Lexington, KY, 40503, United States, jiayi.weng@uky.edu,
Yin Xiangrong
Estimating an inverse regression space is especially important in sufficient dimension reduction. It typically requires a tuning parameter, such as a number of slices in the slicing method or bandwidth selection in the kernel estimation approach, which then increases difficulties for multivariate responses. In this paper, we use the Fourier transform idea to avoid such difficulties and easily incorporate multivariate responses. We further develop the Fourier transform method to deal with a sparse issue, categorical predictor variables, and large p, small n data. Asymptotic methods for dimensionality tests are obtained. Simulation studies and a real data analysis show the efficacy of our proposed methods.

3 - Regularization Approach with Nonlinear Dynamical Features for Recognizing Emotional Patterns
Miaolin Fan, 360 Huntington Avenue, Boston, MA, 02115, United States, fan.mia@husky.neu.edu, Chun-An Chou
Emotion recognition is an essential topic in affective science. In the present work, we aim to classify human emotional states by applying machine learning techniques to electroencephalographic (EEG) signals. The emotion recognition problem was formulated as a supervised learning task, which was then accomplished by logistic regression (the original and L1-regularized LR model) to classify the emotional states categorized based on the arousal-valence model [2]. Furthermore, to reduce the complexity of the high-order feature extraction, the feature extraction method used to effectively capture the underlying dynamics behind the complex reaction corresponding to affective phenomena. A benchmark dataset, DEAP [11], was used for our two-fold objectives: (1) to investigate the suitability of ROA measures and regularized learning method for emotion recognition, and (2) to compare the performances as well as topographic patterns of important channels for classifying emotional states with previous studies. The results demonstrated that our proposed method with selected ROA measures has better performance (test accuracy = 75.7% and F1 score = 78.1% on average) compared to previous studies, and L1-regularized model is less over-fitted comparing to the LR.

4 - Querying Online Self-help Forums for Chronic Kidney Disease with Ontology-based Text Mining
Philip Bukchardt, Carnegie Mellon University, Pittsburgh, PA, United States, pgb@andrew.cmu.edu, Rema Padman
Patients with complex health conditions are increasingly participating in online self-help forums. In this paper, we investigate patients’ discussions on chronic kidney disease (CKD) forums to obtain insights about the daily challenges they face, choice of treatments, and interactions with their clinicians and the community. We introduce a text-mining algorithm based on semantic ontologies which allows for fine-grained analyses. We apply this method to a data set extracted from two online forums, Davita and KidneySpace, containing 10,216 posts. Results indicate that most users enter the forums due to change in health or treatment status, with an estimated 20% of users having relatives affected by CKD. Significantly more women than men were active on both forums (p-value < 1e-4). Many discussions revolved around the challenges and advantages of different treatment options. Finally, patients were concerned about nutrition and enquied about the foods and specific recipes that were appropriate. Internet self-help forums are a unique and potentially fruitful medium for learning about the day-to-day issues faced by patients with complex chronic conditions. Insights from the analysis of patient communication may facilitate development of improved interventions such as personalized educational programs and tools to bridge the gaps in care delivery.

Keynote, 10:30 - 11:15AM

Keynote Robust Machine Learning Optimization Models with Data Uncertainties
Panos Pardalos, University of Florida, Gainesville, FL, United States, pardalos@ufl.edu
1 - Robust Machine Learning Optimization Models with Data Uncertainties
Panos Pardalos, University of Florida, Gainesville, FL, United States, pardalos@ufl.edu
This talk presents robust change-constrained support vector machines (SVM) with second-order moment information and obtains equivalent semidefinite programming (SDP) and second-order cone programming (SOCP) reformulations. Three types of estimation errors for mean and covariance matrix are considered and the corresponding formulations and techniques to handle these types of errors are presented. A method to solve robust change-constrained SVM with large scale data is proposed based on a stochastic gradient descent method.

Saturday, 11:15 - 12:00PM

Tutorial
Hilton 337, Level 3
Data Analytics for IBM's IT Service Deals' Solutioning: Methodologies and Practical Implications
Contributed Session
Chair: Aly Megahed, IBM Research - Almaden, 150 Palm Valley Blvd, Apt 2066, San Jose, CA, 95123, United States, aly@gatech.edu

IBM competes in a tender-like process to win highly valued IT service contracts as part of its $20B outsourcing business. Focusing on deals that are $10M and above, the typical negotiation lifecycle is anywhere between 3 to 12 months. Each deal contains IT services like cloud computing, service desk, and databases. The average number of total components/services in such opportunities is 10k. Given this complexity and high value, pricing these deals is a challenging problem. In the first two-thirds of this talk, we present the data analytical approaches that we successfully developed for the business, leading to "revenue increases in the range of millions of dollars each quarter and enabling the pricing of solutions in a tiny fraction of the time that this task used to take us, and in a more accurate and efficient manner", as quoted by our business VP of global solutions. The approach relies on data mining of historical and market data to produce some pricing points that are then fed into a predictive analytics model that outputs the chances of winning a deal at different pricing points. Lastly, we also briefly discuss other data analytics problems in the same framework of IT service deals, such as deal progress monitoring and revenue prediction. In the last third of this presentation, we will go over multiple interesting problems in Cloud Computing, Internet of Things (IoT), and Blockchain technologies that use operations research and analytics to solve them. We will show some of our developments in these areas and illustrate potential research collaborations as well.
**12TH INFORMS WORKSHOP ON DATA MINING & DECISION ANALYTICS**

**SB01**

**Saturday, 2:30 - 4:00PM**

**Tutorial**

**Hilton 337, Level 3**

**Contributed Session**

**SB01**

Chair: Angela Zhou, PhD Student, Cornell University ORIE, 206 Rhodes Hall, Ithaca, NY, 14853, United States, a2434@cornell.edu

1 - Nonparametric Modeling and Prognosis of Condition Monitoring Signals: A Transfer Learning Approach Based on Multivariate Gaussian Convolution Processes

Raed Al Koniar, University of Wisconsin-Madison, 1513 University Avenue, Room 3255, Madison, WI, 53706, United States, alkoniar@wisc.edu, Shiyu Zhou, Chaitanya Sankavaram, Xinyu Du, Yifu Zhang

In this paper, an alternative view on modeling condition monitoring signals is proposed. This view draws its roots from transfer learning and is based on sharing convolved latent functions between training and testing units. Aside from being non-parametric, the flexible and individualistic approach in our model is able to account for heterogeneity in the data and automatically infer the commonalities between the new testing observations and CM signals in the historical dataset. The advantageous features of our method are highlighted with a case study for automotive lead-acid batteries.

### Saturday, 1:00 - 1:45PM

**Tutorial**

**Hilton 337, Level 3**

**Tutorial on Data Science Analysis with Python**

**Contributed Session**

Chair: Michele Samorani, Santa Clara University, 500, El Camino Real, Santa Clara, CA, 95053, United States, msamorani@scu.edu

This brief tutorial will cover the basic use of pandas, the Python package for data manipulation, and sklearn, the Python package for machine learning. Knowledge of Python is not necessary. To follow along, participants are encouraged to install the necessary software on their laptops before the tutorial. The installation instructions are available on github.com/samorani/DM-WORKSHOP-2017.

**Saturday, 1:45 - 2:30PM**

**Keynote**

**Hilton 339, Level 3**

**Keynote: Researcher Dilemmas using Behavioral Big Data in Healthcare**

**Contributed Session**

Chair: Galit Shmueli, National Tsing Hua University, Hsinchu, Taiwan, galit.shmueli@gmail.com

1 - Researcher Dilemmas using Behavioral Big Data in Healthcare

Galit Shmueli, National Tsing Hua University, Hsinchu, Taiwan, galit.shmueli@gmail.com

Behavioral big data (BBD) refers to very large and rich multidimensional data sets on human and social behaviors, actions, and interactions, which have become available to companies, governments, and researchers. A growing number of researchers acquire and analyze BBD for the purpose of extracting knowledge and scientific discoveries. However, the relationships between the researcher, data, human subjects, and research questions differ in the BBD context compared to non-BBD and even traditional behavioral data. Researchers using BBD face not only methodological and technical challenges but also ethical and moral dilemmas. In this talk, I will discuss several dilemmas, challenges, and trade-offs related to acquiring and analyzing BBD in healthcare research.
4 - A Real-Time Prognostic Model Based on Informative Sensor Selection and Multistream Signal Fusion
Xiaolei Fang, Georgia Institute of Technology.
1546 Woodlawn Dr NE, Apt F, Atlanta, GA, 30329, United States, xfang33@gatech.edu, Kartram Paynabari, Naji Gebrail
Advances in sensor technology have facilitated the capability of monitoring the degradation of complex engineering systems through the analysis of multistream degradation signals. However, the varying levels of correlation with physical degradation process for different sensors, high-dimensionality of the degradation signals and cross-correlation among different signal streams pose significant challenges in monitoring and prognostics of such systems. To address the foregoing challenges, we develop a three-step multi-sensor prognostic methodology that utilizes multistream signals to predict residual useful lifetimes of partially degraded systems. We first identify the informative sensors via the penalized functional (log)-location-scale regression. Then, we fuse the degradation signals of the informative sensors using multivariate functional principal component analysis, which is capable of modeling the cross-correlation of signal streams. Finally, the third step focuses on utilizing the fused signal features for prognostics via adaptive penalized (log)-location-scale regression. We validate our multi-sensor prognostic methodology via simulation study as well as a case study of aircraft turbine engines available from NASA repository.

SB02
Hilton 338, Level 3
Contributed Session
Chair: Ramin Moghaddas, University of Miami, Miami, FL, United States, raminn@mit.edu

1 - Multi-scale Gaussian Process for Dynamic Evolution Prediction of Complex Systems
Changqing Cai, Binghamton University, Systems Science and Industrial Engineering, Binghamton University, Binghamton, NY, 13902, United States, cai@binghamton.edu
This paper presents a nonparametric multi-scale Gaussian process model (MGPS) for the dynamic evolution prediction of nonlinear and nonstationary complex systems. Intrinsic time-decomposition (ITD) is first used to iteratively decompose the time series generated from such complex systems into a series of proper rotation components and a baseline trend component. Those components delineate the true time-frequency-energy patterns of the complex systems at different granularity. A Gaussian process (GP) model is then applied on each component to predict the system evolution at each scale. Summation of those individual forecasts represents the overall evolution of the original time series. Case studies using synthetic and real-world data evidenced that the predicted MGPS model significantly outperforms conventional autoregressive models, composite GP model, and support vector regression (SVR) in terms of prediction accuracy, and it is particularly effective for multi-scale forecasting.

2 - A Hybrid Iterative Learning Framework for Anomaly Detection
Chithra Phadke, Bell Labs, Nokia, Murray Hills, NJ, United States, chithra.phadke@nokia-bell-labs.com, Huseyin Uzunalioglu, Jin Cao
Detection of anomalies is an important data analysis technique that can alert users about underlying issues in a system. Most anomaly detection algorithms today rely only on pure statistical approaches, which, although rigorous, cannot take into account many of contextual nuances of the anomaly. This results in generation of many false alerts, wasting valuable time in the subsequent root cause analysis that follows the detected anomaly events. In this paper, we propose a hybrid framework that takes into account human expert feedback, and augments statistical anomaly detection with machine learning to learn the contextual aspects of significant anomalies for which alerts should be generated. Our framework is domain agnostic and independent of the underlying statistical anomaly detection technique or the machine learning algorithm, and therefore easily adaptable to many more real world applications such as detecting network performance degradation or abnormal events observed in video streams. We evaluate our framework with data from a mobile network operator and show that it can better detect anomalies with significance resulting in much smaller number of false positives.

3 - Trading Decisions using EMD-based Stock Index Forecasting Models
Dhanya Jothimani, Indian Institute of Technology Delhi, DMS, IIT Delhi, New Delhi, India, dhanya.jothimani@iitm.ac.in, Ravi Shankar, Surendra S. Yadav
In this paper, an ensemble framework comprising of a non-classical decomposition model and machine learning model is proposed for predicting the stock index. Two ensemble models, namely, Empirical Mode Decomposition-Support Vector Regression (EMD-SVR) and Ensemble Empirical Mode Decomposition (EEMD-SVR) are presented, where decomposed components obtained using EMD and EEMD are predicted using SVR. The models were tested on Nifty data. Error measures and statistical procedure showed that the proposed ensemble framework has superior performance as compared with traditional SVR model (without decomposition). Further, trading decisions based on ensemble models yielded higher return on investments than traditional Buy-and-Hold strategy.

4 - A Hierarchical Framework for Anomaly Detection using Large-Scale Count Data
Ramin Moghaddas, University of Miami, Miami, FL, United States, raminn@mit.edu, Jianhui Wang
Real-time monitoring and control of smart grids is critical to the enhancement of reliability and operational efficiency of power utilities. We develop a real-time anomaly detection framework, which can be built based upon smart meter data collected at the consumers’ premises. The model is designed to detect the occurrence of anomalous events and abnormal conditions at both lateral and customer levels. We propose a generative model for anomaly detection that takes into account the hierarchical structure of the network and the data collected from smart meters. We also address three challenges existing in smart grid analytics: (i) large-scale multivariate count measurements, (ii) missing points, and (iii) variable selection. We present the effectiveness of our approach with numerical experiments.

Saturday, 4:15 - 5:00PM

Keynote:
Hilton 339, Level 3
Keynote: Justifiable and Ethical Learning – A Mathematical View
Contributed Session
Chair: Endre Boros, Rutgers University, Rutgers, Piscataway, NJ, 088, United States, Endre.Boros@rutgers.edu

1 - Justifiable and Ethical Learning – A Mathematical View
Endre Boros, Rutgers University, Piscataway, NJ, United States, Endre.Boros@rutgers.edu
Many learning algorithms are struggling with large data sets, and miss information present in the data simply for computational reasons. A larger and mostly hidden problem is that many algorithms learn (unintentionally and unnoticed) triggering patterns that are not supported by the data. Using such classifiers, we may jump to conclusions that are unjustifiable based on our existing data sets. Both errors imply potential problems: missing important triggers and/or using unsupported ones. This brings up both ethical and legal questions. In this talk we demonstrate these issues with a small example. We propose the notion of a “justifiable” classifier, and on the positive side, we show some results about the existence of learning algorithms that always produce a “justifiable” classifier.
Data Mining & Analytics

Optimization correct 10/12/17 10:56 AM  Page 5

Saturday, 5:00 - 6:10PM

■ SC01

Hilton 337, Level 3

Contributed Session

Chair: Michele Samorani, Santa Clara University, 500 El Camino Real, Santa Clara, CA, 95053, United States, msamorani@scu.edu

1 - Off-Policy Evaluation and Optimization with Continuous Treatments

Nathan Kallus, Cornell University, 111 8th Avenue #302, New York, NY, 10011, United States, kallus@cornell.edu, Angela Zhou

We study the problem of off-policy evaluation and learning in contextual bandits when treatments are continuous, going beyond previous work on discrete treatments. We provide a framework to learn policies that can optimize the expected outcome of continuous actions, even those that have never been observed. Our approach has three novel contributions. First, we develop a new kernel function for continuous treatments. Second, we prove that our approach guarantees non-negative rewards under mild assumptions. Third, we introduce an efficient and practical algorithm that can be used for any problem, including those with non-linear reward structure. We demonstrate the effectiveness of our approach in real-world settings.

2 - Optimal Recovery of Tensor Slices

Vivek Farias, MIT, 100 Main Street, Cambridge, MA, 02142, United States, vivekf@mit.edu, Andrew Li

We consider the problem of large scale matrix recovery given side information in the form of additional matrices of conforming dimension. This is a parsimonious model that captures a number of interesting problems including context and location aware recommendations, personalized ‘tag’ learning, demand learning with side information, etc. Viewing the matrix we seek to recover and the side information we have as slices of a tensor, we consider the problem of Slice Recovery, which is to recover specific slices of a tensor from noisy observations of the tensor. We provide an efficient algorithm to recover slices of structurally simple tensors given noisy observations of the tensor's entries; our definition of simplicity subsumes low-rank tensors for a variety of definitions of tensor rank. Our algorithm is practical for large datasets and provides a significant performance improvement over state of the art incumbent approaches to tensor recovery. We establish theoretical recovery guarantees that under reasonable assumptions are minimax optimal for slice recovery. These guarantees also imply the first minimax optimal guarantees for recovering tensors of low Tucker rank and general noise. Experiments on data from a music streaming service demonstrate the performance and scalability of our algorithm.

3 - The Chinese Voting Process: The Evolution of Online Helpfulness Evaluations in Product Reviews and Question Answers

Moonjac Lee, Cornell University, Ithaca, NY, United States, moonjac@cs.cornell.edu, Seok Hyun Jin, David Mimmo

Helpfulness voting is a popular method for highlighting the usefulness of user-contributed responses such as reviews of products and answers to questions. However such voting is a social process that can gain momentum, biasing toward responses that are already popular. The positional accessibility of individual responses and the presented polarity of existing votes can all affect subsequent votes. Less helpful responses that receive an early boost could block users from accessing more useful future information. Moreover, the momentum may differ between forums, so solutions must be tailored to the micro-cultures of different communities. We propose the Chinese Voting Process (CVP) that can model the evolution of online helpfulness votes as a self-reinforcing system on position and presentation biases. We evaluate this model on Amazon product reviews and more than 80 major StackExchange forums, inferring the intrinsic quality of individual responses and measuring Trendiness and Conformity of different communities.

■ SC02

Hilton 338, Level 3

Contributed Session

Chair: Cem Iyigun, Middle East Technical University, Inonu Blvd, Endustri Muhendisligi, Ankara, 06800, Turkey, iyigun@ie.metu.edu.tr

1 - The Need for Speed: Effects of Uncertainty Reduction in Patenting

Mike Horia Teodorescu, Harvard Business School, 555 South Water Street, Apt 201, Providence, RI, 02903-4370, United States, miketeo@hbs.edu

Patents are essential in commerce to establish property rights for ideas and to give equal protection to firms that develop new technologies. Young firms especially depend on the protection of intellectual property to bring a product from concept to market. However, the market for technology ideas has been recognized as an inefficient market in the management and economics literatures. While information asymmetry and expropriation risks have been studied extensively, the question of the effects of pre-patent grant uncertainty on firm outcomes remains open. This paper introduces a novel analysis based on internal US Patent and Trademark Office databases, exploiting an exogenous shock to startup firms from a previously unstudied executive action involving reduction of patent- pending (time from application to patent decision) for green technology patents. The aim of the paper is to determine whether reduced patent pending improves firm outcomes for startups and to explore its implications. The paper also introduces a novel method for constructing a control group using a classification algorithm rooted in natural language processing, which can be used in conjunction with traditional econometric approaches such as difference-in-differences analysis beyond the topic of this paper.

2 - Learning Interpretable Constructive Algorithms

Michele Samorani, Santa Clara University, 500 El Camino Real, Santa Clara, CA, 95053, United States, msamorani@scu.edu, Manuel Laguna

We propose a method to automatically design effective and interpretable constructive algorithms for any given class of problems. Our method employs machine learning to analyze the optimal solutions of a training set of problem instances. Then, it finds an interpretable rule that can be applied at each step of a constructive algorithm to tackle new instances. We test our method on the knapsack problem and on the capacitated clustering problem. Our results show that in a very short time our method is capable of designing constructive algorithms that are competitive with state-of-the-art methods and present the advantage of being interpretable, which facilitates its implementation and generates insights on the structure of the problem.

3 - Feature Weighting Problems in K-Nearest Neighbor Classifier and Evolutionary Solution Approach

Cem Iyigun, Middle East Technical University, Inonu Blvd, Endustri Muhendisligi, Ankara, 06800, Turkey, iyigun@ie.metu.edu.tr, Nurullah Gulce

The k-Nearest Neighbor (k-NN) algorithm is one of the well-known and most commonly used algorithms for the classification problems. In this study, we have focused on feature weighted k-NN problems. Two different problems are studied. In the first problem, k value and the weights of each feature are optimized to in the form of additional matrices of conforming dimension.