Poster Presenters

1. A Novel Approach for Clustering Mixed-Type Data

Alexander H. Foss, MA[1], Marianthi Markatou, PhD[1], Bonnie Ray, PhD[2]
[1]Department of Biostatistics, University at Buffalo, Buffalo, New York, United States;
[2]Arenadotio, New York, New York, United States

With the rapid growth in the size and complexity of data sets, dealing with heterogeneous data sets has become imperative. Cluster analysis (unsupervised learning) with mixed continuous and categorical variables is a method used in a wide array of fields, from marketing to genomics. However, many of the most common approaches, such as discretization or dummy coding, are surprisingly inefficient or outright flawed in their ability to equitably combine information across distinct variable types. Distance metrics for mixed-type data such as Gower's distance are generally afflicted by similar problems. Parametric methods can potentially resolve some of these challenges, but perform poorly when parametric assumptions are violated. After a brief survey of existing methods and their properties, we propose a novel clustering method KAMILA (KAy-means for MIxed LArge data) that solves many of the existing problems using an original radial kernel density estimation scheme. We delineate additional challenges associated with clustering mixed-type data in very large data sets, and discuss a map-reduce implementation of KAMILA.

2. Corralling a Band of Bandit Algorithms

Alekh Agarwal, PhD[1], **Haipeng Luo**, PhD[1], Behnam Neyshabur, PhD candidate[2], and Robert E. Schapire, PhD[1]
[1]Microsoft Research, New York, New York, United States;
[2]Toyota Technological Institute, Chicago, Illinois, United States

We study the problem of combining multiple bandit algorithms (that is, online learning algorithms with partial feedback) with the goal of creating a master algorithm that performs almost as well as the best base algorithm *if it were to be run on its own*. The main challenge is that when run with a master, base algorithms unavoidably receive much less feedback and it is thus critical that the master not starve a base algorithm that might perform uncompetitively initially but would eventually outperform others if given enough feedback. We address this difficulty by devising a version of Online Mirror Descent with a special mirror map together with a sophisticated learning rate scheme. We show that this approach manages to achieve a more delicate balance between exploiting and exploring base algorithms than previous works yielding superior regret bounds.

Our results are applicable to many settings, such as multi-armed bandits, contextual bandits, and convex bandits. As examples, we present two main applications. The first is to create an algorithm that enjoys worst-case robustness while at the same time performing much better when the environment is relatively easy. The second is to create an algorithm that works simultaneously under different assumptions of the environment, such as different priors or different loss structures.

3. CNN-based Object Segmentation in Urban LIDAR With Missing Points

Allan Zelener, MPhil[1], Ioannis Stamos, PhD[1,2]
[1]The Graduate Center, CUNY, New York, New York, United States;
[2]Hunter College, New York, New York, United States

We examine the task of point-level object segmentation in outdoor urban LIDAR scans. A key challenge in this area is the problem of missing points in the scans due to technical limitations of the LIDAR sensors. Our core contributions are demonstrating the benefit of reframing the segmentation task over the scan acquisition grid as opposed to considering only the acquired 3D point cloud and developing a pipeline for training and applying a convolutional neural network (CNN) to accomplish this segmentation on large scale LIDAR scenes. By labeling missing points in the scanning grid we show that we can train our classifier to achieve a more accurate and complete segmentation mask for the vehicle object category which is particularly prone to missing points. Additionally we show that the choice of input features maps to the CNN significantly effect the accuracy of the segmentation and these features should be chosen to fully encapsulate the 3D scene structure. We evaluate our model on a LIDAR dataset collected by Google Street View cars over a large area of New York City.

4. Artificial Neural Networks Applied to Free Energy Surface Reconstruction in Biomolecules

Elia Schneider, PhD[1], Luke Dai[1], Mark E. Tuckerman, PhD[123]
[1] Department of Chemistry, New York University, New York, New York, United States;
[2] Courant Institute of Mathematical Science, New York University, New York, New York, United States;
[3] NYU-ECNU Center for Computational Chemistry at NYU Shanghai, Shanghai, China

Prediction and elucidation of biomolecular structure are fundamental computational challenges, which, if solved, will lead to a better understanding of the physiological mechanisms underlying certain diseases classes. In the last twenty years, various enhanced sampling (ES) methods have been developed for accurately sampling conformational equilibria of complex systems, including polypeptides and other polymers, and generating the associated free energy surfaces (FESs). Often these FESs are of very high dimension, suggesting that machine learning (ML) methods can be trained on data from ES calculations and then used to represent the FESs compactly and allow facile computation of physical observables from the FESs. Employing artificial neural networks (ANNs) within ES, we have developed an algorithm for the efficient generation of FESs. Inputs to the ANN are any set of collective variables, e.g., backbone dihedral angles, that characterize conformational space, while outputs are sampled free energy values and/or their gradients. Once trained, we used our trained ANNs to compute statistical averages of physical observables via Monte Carlo Integration. As test cases, we generated the FESs of the alanine di- and tripeptides in the gas and solution phases and compared the trained ANN results to traditional fitting approaches. The trained ANNs are then used to compute the spin-spin coupling constants measured in nuclear magnetic resonance experiments as a function of the size of the training set. As a more challenging case, we apply ANNs to obtain the FES and J-couplings of met-enkaphalin, a five-residue oligopeptide present in opioid receptors of the central neural system.

5. Measuring the Metallicity of Galaxies with Machine Learning

Viviana Acquaviva[1,] PhD
[1]CUNY NYC College of Technology, Brooklyn, New York, United States

A galaxy's Spectral Energy Distribution (SED) is a chart of the brightness of a galaxy as a function of wavelength. It contains information about the physical properties of a galaxy, for example, the age, mass, distance from Earth, and star formation history of the galaxy. We focus on recovering information about the amount of elements heavier than helium, which astronomers refer to, quite imprecisely, as metals.
We present the method we have developed to estimate metallicity from the 5-band Sloan Digital Sky Survey photometry using several machine learning algorithms. Using metallicity estimates from spectroscopic surveys as ground truth to train and test our technique, we show that if a modest-sized (few hundred objects), unbiased training sample is available, it is possible to recover metallicity to better than 10%. We create different sample data sets, and show that the typical Root Mean Square Error (RMSE) is 0.085, the fraction of outliers (objects for which the difference between true and predicted metallicity is larger than 20%) is 2-3%, and the RMSE decreases to less than 0.07 if those objects are excluded.

6. Item Embeddings for Demand Estimation in Economics

Francisco J. R. Ruiz, PhD[1,2], Maja Rudolph, MSc[1], Stephan Mandt, PhD[1], Susan Athey, PhD[3], and David M. Blei, PhD[1]
[1]Columbia University, New York, New York, United States;
[2]University of Cambridge, Cambridge, United Kingdom;
[3]Stanford University, Stanford, California, United States

We develop item embeddings, an exponential family embedding model that finds a representation of items in a latent space. We use ideas based on word embeddings to build a discrete choice model over items that can capture item-to-item interactions. We propose two metrics based on these interactions that can reveal substitutable and complementary pairs of items. In our model, we additionally consider the user consumption behavior and the effect of price on the choices. We apply the item embedding model to supermarket data and show that it outperforms matrix factorization approaches in terms of predictive performance. Furthermore, we show several qualitative results of the fitted embeddings.

7. Edward: A Library for Probabilistic Modeling, Inference, and Criticism

Dustin Tran, MS[1], Matthew D. Hoffman, PhD[23], Kevin Murphy, PhD[2], Eugene Brevdo, PhD[2], Rif A. Saurous, PhD[2], David M. Blei, PhD[1]
[1]Columbia University, New York, New York, United States;
[2]Google Brain, Mountain View, California, United States;
[3]Adobe Research, Mountain View, California, United States

Probabilistic modeling is a powerful approach for analyzing empirical information. In this talk, I will provide an overview of Edward, a software library for probabilistic modeling. Formally, Edward is a probabilistic programming system built on computational graphs, supporting compositions of both models and inference for flexible experimentation. For example, Edward makes it easy to fit the same model using a variety of composable inferences, ranging from point estimation, to variational inference, to MCMC. Edward is also integrated into TensorFlow, providing significant speedups over existing probabilistic systems. As examples, I will show how Edward can be leveraged for expanding the frontier of variational inference and deep generative models.

8. Initialization and Coordinate Optimization for Multi-way Matching

Da Tang, MSc[1] and Tony Jebara, PhD[1,2]
[1]Columbia University, New York, New York, United States;
[2]Netflix Inc., Los Gatos, California, United States

We consider the problem of consistently matching multiple sets of elements to each other, which is a common task in fields such as computer vision. To solve the underlying NP-hard objective, existing methods often relax or approximate it, but end up with unsatisfying empirical performances due to their inexact objectives. We propose a coordinate update algorithm that directly solves the exact objective. By using the pairwise alignment information to build an undirected graph and initializing the permutation matrices along the edges of its Maximum Spanning Tree, our algorithm successfully avoids bad local optima. Theoretically, with high probability our algorithm could guarantee to solve this problem optimally on data with reasonable noise. Empirically, our algorithm consistently and significantly outperforms existing methods on several benchmark tasks on real datasets.

9. Distributed Coverage Maximization via Sketching

MohammadHossein Bateni, PhD[1], Hossein Esfandiari, MS[2], Vahab Mirrokni, PhD[1]
[1]Google Inc., New York, New York, United States;
[2]University of Maryland, College Park, Maryland, United States

Maximum coverage and minimum set cover problems have been studied extensively in streaming models. However, previous research not only achieve suboptimal approximation factors and space complexities, but also study a restricted set arrival model which makes an explicit or implicit assumption on oracle access to the sets, ignoring the complexity of reading and storing the whole set at once. In this paper, we address the above shortcomings, and present algorithms with improved approximation factor and improved space complexity, and prove that our results are almost tight. Moreover, unlike most of previous work, our results hold on a more general edge arrival model.

More specifically, we present (almost) optimal approximation algorithms for maximum coverage and minimum set cover problems in the streaming model with an (almost) optimal space complexity of $\tilde O(n)$, i.e., the space is independent of the size of the sets or the the ground set size. These results not only improve over the best known algorithms for the set arrival model, but also are the first such algorithms for the more powerful edge arrival model.

In order to achieve the above results, we introduce a new general sketching technique for coverage functions: This sketching scheme can be applied to convert an α-approximation algorithm for a coverage problem to a $(1-\epsilon)\alpha$-approximation algorithm for the same problem in streaming, MapReduce or RAM models.

Finally, we perform an extensive empirical study of our algorithms on a number of publicly available real data sets, and show that using sketches of size 30 to 600 times smaller than the input, one can solve the coverage maximization problem with quality very close to that of the state-of-the-art single-machine algorithm.

https://arxiv.org/abs/1610.08096
https://arxiv.org/abs/1612.02327

10. A Study of Compact Reserve Pricing Languages

MohammadHossein Bateni, PhD[1], Hossein Esfandiari, MS[2], Vahab Mirrokni, PhD[1], ***Saeed Seddighin***, MS[2]
[1]Google Inc., New York, New York, United States;
[2]University of Maryland, College Park, Maryland, United States

Online advertising allows advertisers to implement fine-tuned targeting of users. While such precise targeting leads to more effective advertising, it introduces challenging multidimensional pricing and bidding problems for publishers and advertisers. In this context, advertisers and publishers need to deal with an exponential number of possibilities. As a result, designing efficient and \emph{compact} multidimensional bidding and pricing systems and algorithms are practically important for online advertisement. Compact bidding languages have already been studied in the context of multiplicative bidding. In this paper, we study the compact pricing problem.

More specifically, we first define the \emph{multiplicative reserve price optimization problem} (\MRPOP) and show that unlike the unrestricted reserve price system, it is NP-hard to find the best reserve price solution in this setting. Next, we present an efficient algorithm to compute a solution for {\MRPOP} that achieves a logarithmic approximation of the optimum solution of the unrestricted setting, where we can set a reserve price for each individual impression type (i.e., one element in the Cartesian product of all features). We do so by characterizing the properties of an optimum solution. Furthermore, our empirical study confirms the effectiveness of multiplicative pricing in practice. In fact, the simulations show that our algorithm obtains 90--98\% of the value of the best solution that sets the reserve prices for each auction individually (i.e., the optimum set of reserve prices).

Finally, in order to establish the tightness of our results in the adversarial setting, we demonstrate that there is no {\em compact} pricing system (i.e., a pricing system using $O(n^{1-\epsilon})$ bits to set n reserve prices) that loses, in the worst case, less than a logarithmic factor compared to the optimum set of reserve prices. Notice that this hardness result is not restricted to the multiplicative setting and holds for any compact pricing system.

In summary, not only does the multiplicative reserve price system show great promise in our empirical study, but it is also theoretically optimal up to a constant factor in the adversarial setting.

11. A New Theory of Exploration in Reinforcement Learning with Function Approximation

Nan Jiang[1], Akshay Krishnamurthy[2], Alekh Agarwal[3], John Langford[3], and Robert E. Schapire[3]
[1]University of Michigan, Ann Arbor, Michigan, United States;
[2]University of Massachusetts, Amherst, Massachusetts, United States;
[3]Microsoft Research, New York, New York, United States

In reinforcement learning, autonomous agents solve sequential decision-making problems by (1) actively exploring the environment to collect data, and (2) improving behavior by learning from the collected data. The recent success of reinforcement learning can largely be attributed to the use of advanced function approximation techniques (e.g., deep neural networks) for the learning component. In contrast, advances in exploration techniques have been rather limited: most existing algorithms that perform systematic exploration only apply to simple problems with small state spaces and cannot accommodate sophisticated function approximation schemes that are critical for solving real-world problems with rich observations.

In this work, we take a unified view of reinforcement learning with rich observations and function approximation through a new model called the Contextual Decision Process (CDP), which encompasses existing models such as MDPs and POMDPs. Several special cases of CDPs are, however, provably intractable in terms of the amount of data needed to find the near-optimal behavior. To overcome this challenge, we identify a structural property, called the Bellman rank, which is naturally small in a range of important reinforcement learning scenarios. We propose a new algorithm, whose sample complexity scales polynomially in the Bellman rank and other relevant parameters, and crucially has no dependence on the number of unique observations. The algorithm uses Bellman error minimization with optimistic exploration, and provides new insights into exploration in complex reinforcement learning problems.

12. Extended Ridge Regression Using an Iterative Solution: Regularization with a Stopping Rule

Kathryn Vasilaky, PhD, Earth Institute, Post-doctoral Fellow, International Research Institute
Columbia University, Lamont Campus

When the design matrix is collinear and/or the number of coefficients to be estimated is large then regularization methods are employed to deal with sensitivity to noisy data and overfitting. Tikhonov regularization, also known as Ridge Regression (RR), is one of the best known regularization methods. In this work, I have extended the Tikhonov Regularization from a one parameter to a two parameter regularization. The additional parameter is an iteration stopping time. The General Cross Validation formula for computing an optimal regularization parameter for Ridge Regression prediction is extended to the two parameter case. The two parameter General Cross Validation, in general, yields smaller Mean Square Prediction Error (MSPE) than Ridge Regression. In addition, I demonstrate that the Mean Square Error (MSE) of the Extended Ridge Regression (ERR) is smaller than the MSE of the RR estimator, which in turn is known to be smaller than the Best Linear Unbiased Estimator (BLUE).

13. Orthogonal Random Features

Felix X. Yu, PhD, Ananda Theertha Suresh, PhD, Krzysztof Choromanski, PhD, Daniel Holtmann-Rice, PhD, Sanjiv Kumar, PhD

Google Research, New York, United States

We present an intriguing discovery related to Random Fourier Features: in Gaussian kernel approximation, replacing the random Gaussian matrix by a properly scaled random orthogonal matrix significantly decreases kernel approximation error. We call this technique Orthogonal Random Features (ORF), and provide theoretical and empirical justification for this behavior. Motivated by this discovery, we further propose Structured Orthogonal Random Features (SORF), which uses a class of structured discrete orthogonal matrices to speed up the computation. The method reduces the time cost from O(d^2) to O(dlogd), where d is the data dimensionality, with almost no compromise in kernel approximation quality compared to ORF. Experiments on several datasets verify the effectiveness of ORF and SORF over the existing methods. We also provide discussions on using the same type of discrete orthogonal structure for a broader range of applications.

14. Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms

Christian A. Naesseth, MSc[1,2], Francisco J. R. Ruiz, PhD[1,3], Scott W. Linderman, PhD[1], David M. Blei, PhD[1]
[1]Columbia University, New York, New York, United States;
[2]Linköping University, Linköping, Sweden
[3]University of Cambridge, Cambridge, United Kingdom

Variational inference using the reparameterization trick has enabled large-scale approximate Bayesian inference in complex probabilistic models, leveraging stochastic optimization to sidestep intractable expectations. The reparameterization trick is applicable when we can simulate a random variable by applying a (differentiable) deterministic function on an auxiliary random variable whose distribution is fixed. For many distributions of interest (such as the gamma or Dirichlet), simulation of random variables relies on rejection sampling. The discontinuity introduced by the accept-reject step means that standard reparameterization tricks are not applicable. We propose a new method that lets us leverage reparameterization gradients even when variables are outputs of a rejection sampling algorithm. Our approach enables reparameterization on a larger class of variational distributions. In several studies of real and synthetic data, we show that the variance of the estimator of the gradient is significantly lower than other state-of-the-art methods. This leads to faster convergence of stochastic optimization variational inference.

15. Unsupervised Learning of Word-Sequence Representations from Scratch via Convolutional Tensor Decomposition

Furong Huang, PhD[1] and Anima Anandkumar, PhD[2]
[1]Microsoft Research, New York, New York, United States;
[2]University of California, Irvine, Irvine, California, United States

Unsupervised text embedding extraction is crucial for text understanding in machine learning. Word2Vec and its variations are successful in mapping words with similar syntactic or semantic meaning to vectors close to each other. However, extracting context-aware word-sequence embedding remains a challenging task. Training over large corpus is difficult as labels are difficult to get. More importantly, it is challenging for pre-trained models to obtain word-sequence embeddings that are universally good for all downstream tasks or for any new datasets.

We propose a two-phased ConvDic+DeconvDec framework to solve the problem by combining a word-sequence dictionary learning model with a word-sequence embedding decode model. We propose a convolutional tensor decomposition mechanism to learn good word-sequence phrase dictionary in the learning phase. It is proved to be more accurate and much more efficient than the popular alternating minimization method. In the decode phase, we introduce a deconvolution framework that is immune to the problem of varying sentence lengths. The word-sequence embeddings we extracted using ConvDic+DeconvDec are universally good for a few downstream tasks we test on. The framework requires neither pre-training nor prior/outside information.

16. Demystifying Multi-Task Deep Neural Networks for Quantitative Structure-Activity Relationships

Yuting Xu[1,2], Junshui Ma[2], Andy Liaw[2], Robert Sheridan[3], and Vladimir Svetnik[2]
[1]Department of Biostatistics, Johns Hopkins University, Baltimore, MD, U.S.A
[2]Biometrics Research Department, MRL, Merck & Co., Inc., Rahway, NJ, U.S.A.
[3]Modeling and Informatics Department, MRL, Merck & Co., Inc., Rahway, NJ, U.S.A.

In the past four years, Deep neural networks (DNNs) generated promising results in quantitative structure-activity relationship (QSAR) tasks. Previous work showed that DNNs can routinely make better predictions than traditional methods, such as random forests, on a large and diverse set of QSAR datasets. It was also found that multi-task DNN models for multiple QSAR tasks simultaneously outperform DNNs for individual dataset, in many but not all tasks. Up to now there is no satisfactory explanation as to why the QSAR model of one task, embedded in a multi-task DNN, can borrow information from other unrelated QSAR tasks. Thus using multi-task DNNs in a way that consistently provides a predictive advantage becomes a challenge. In this work, we explore why multi-task DNNs make a difference in prediction performance. Our results show that, during prediction, a multi-task DNN does borrow "signal" from molecules with similar structures in the training sets of the other tasks. However, whether this borrowing leads to better or worse predictive performance depends on whether correlation of molecular activities exists across those tasks. Based on the discovery, we develop a strategy to use multi-task DNN, and demonstrate its effectiveness on several examples.

17. Oracle-Efficient Learning and Auction Design

Miroslav Didík, PhD[1], **Nika Haghtalab**, MMath[2], Haipeng Luo, PhD[1], Robert E. Schapire, PhD[1], Vasilis Syrgkanis, PhD[3], Jennifer Wortman Vaughan, PhD[1].
[1]Microsoft Research, New York, New York, United States;
[2]Carnegie Mellon University, Pittsburgh, Pennsylvania, United States;
[3]Microsoft Research, Cambridge, Massachusetts, United States

We consider the design of online no-regret algorithms that are computationally efficient, given access to an offline optimization oracle. Our first main contribution is introducing an oracle-based online algorithm and conditions under which it achieves vanishing regret. Our learning algorithm is a generalization of the Follow-The-Perturbed-Leader algorithm of Kalai and Vempala that at every step plays the best-performing action subject to some independent random perturbations. Our design uses a shared source of randomness across all actions that can be efficiently implemented by adding to the history of the game a set of adversary's actions from a carefully designed distribution. Our work extends to oracle-efficient algorithms for contextual learning, learning with Maximal-in-Range approximation algorithms, and no-regret bidding in simultaneous auctions, answering an open problem of Daskalakis and Syrgkanis in the latter case.

Our second main contribution is introducing a new adversarial auction-design framework for revenue maximization and applying our oracle-efficient learning results to adaptive auction design. We give oracle-efficient learning results for: (1) VCG auctions with bidder-specific reserves in single-parameter settings, (2) envy-free item pricing in multi-item auctions, and (3) s-level auctions of Morgenstern and Roughgarden for single-item settings. The last result leads to an approximation of the optimal Myerson auction for the stationary distribution of a Markov process, extending prior work that only gave such guarantees for the i.i.d. setting. We also extend our framework to allow the auctioneer to use side information about the bidders in the design of the optimal auction (contextual learning).

18. Generalized Topic Modeling

Avrim Blum, PhD[1], **Nika Haghtalab**, MMath[1]
[1]Carnegie Mellon University, Pittsburgh, Pennsylvania, United States;

Recently there has been significant activity in developing algorithms with provable guarantees for topic modeling. In standard topic models, a topic (such as sports, business, or politics) is viewed as a probability distribution $\vec{a}_i$ over words, and a document is generated by first selecting a mixture $\vec{w}$ over topics, and then generating words i.i.d. from the associated mixture $A\vec{w}$. Given a large collection of such documents, the goal is to recover the topic vectors and then to correctly classify new documents according to their topic mixture.

In this work we consider a broad generalization of this framework in which words are no longer assumed to be drawn i.i.d. and instead a topic is a complex distribution over sequences of paragraphs. Since one could not hope to even represent such a distribution in general (even if paragraphs are given using some natural feature representation), we aim instead to directly learn a document classifier. That is, we aim to learn a predictor that given a new document, accurately predicts its topic mixture, without learning the distributions explicitly. We present several natural conditions under which one can do this efficiently and discuss issues such as noise tolerance and sample complexity in this model. More generally, our model can be viewed as a generalization of the multi-view or co-training setting in machine learning.

19. Bayesian Nonparametric Latent Factor Analysis with Scalable Algorithm

Ghazal Fazelnia[1], Shaoyang Li[2], Aonan Zhang[1], John Paisley[3], PhD
[1] PhD Student, Electrical Engineering Department, Columbia University, New York, New York, United States;
[2] PhD Student, Electrical Engineering Department, Tsinghua University, Beijing Shi, China;
[3] Assistant Professor, Electrical Engineering Department, Columbia University, New York, New York, United States;

Latent factor analysis plays an important role in discovering the underlying theme and pattern hidden in the data. Broad range of its applications include image and audio processing, text analyzing and population genetics. In this project, our goal is to learn a dictionary from which sparse representation for data is derived. We present a stochastic expectation maximization (EM) algorithm for scalable dictionary learning with the beta-Bernoulli process, a Bayesian nonparametric prior that learns the dictionary size in addition to the sparse coding of each signal. Stochastic extension for handling large data sets is closely related to stochastic variational inference with the stochastic update for parameters. We extend this work to dictionary learning whose factors are unbounded subspaces. The proper number of subspaces as well as dimensionality of each subspace are learned using Gamma process nonparametric prior. Experimental results show significant improvements comparing to current methods such as generalized K-means clustering process (K-SVD), independent component analysis (ICA), and mixture of factor analysis (MFA) on image processing and image denoising applications.

20. Stochastic Variational Unit Tests

Alp Kucukelbir, PhD[1], Christian Naesseth, MSc[2], David M. Blei, PhD[1]
[1]Columbia University, New York, New York, United States;
[2]Linköping University, Linköping, Sweden

A central challenge in Bayesian machine learning is to approximate the posterior density, which is difficult to compute. Variational inference is a powerful technique that approximates the posterior through optimization; it has enabled Bayesian machine learning to scale to massive datasets. However, variational inference offers few theoretical guarantees.

In this work, we propose a three step strategy inspired by software testing to validate the correctness of variational inference. We begin with stochastic unit tests: verifying the building blocks of modern variational inference algorithms. We follow with stochastic integration tests: validating asymptotic correctness of the algorithm, by leveraging the Bayesian central limit theorem. We end with stochastic system tests: validating the accuracy of the variational approximation to the intractable posterior density.

These three procedures are practical stochastic tests that validate any given variational inference method.

21. AdaCluster : Adaptive Clustering for Heterogeneous Data

Mehmet E. Basbug, PhD[1,2], Barbara E. Engelhardt, PhD[3,4]
[1] Lion Cave Capital, Chatham, New Jersey, United States
[2] Department of Electrical Engineering, Princeton University, New Jersey, United States
[3] Department of Computer Science, Princeton University, New Jersey, United States
[4] Center for Statistics and Machine Learning, Princeton University, New Jersey, United States

Clustering algorithms start with a fixed divergence, which captures the possibly asymmetric distance between a sample and a centroid. In the mixture model setting, the sample distribution plays the same role. When all attributes have the same topology and dispersion, the data are said to be *homogeneous*. If the prior knowledge of the distribution is inaccurate or the set of plausible distributions is large, an adaptive approach is essential. The motivation is more compelling for *heterogeneous* data, where the dispersion or the topology differs among attributes. We propose an adaptive approach to clustering using classes of parametrized Bregman divergences. We first show that the density of a steep exponential dispersion model (EDM) can be represented with a Bregman divergence. We then propose *AdaCluster*, an expectation-maximization (EM) algorithm to cluster heterogeneous data using classes of steep EDMs. We compare AdaCluster with EM for a Gaussian mixture model on synthetic data and nine UCI data sets. We also propose an adaptive hard clustering algorithm based on Generalized Method of Moments. We compare the hard clustering algorithm with k-means on the UCI data sets. We empirically verified that adaptively learning the underlying topology yields better clustering of heterogeneous data.

22. Continuous Profile Models in ASL Syntactic Facial Expression Synthesis

Hernisa Kacorri, PhD[1], Matt Huenerfauth, PhD[2]
[1]Carnegie Mellon University, Pittsburgh, Pennsylvania, United States;
[2]Rochester Institute of Technology, Rochester, New York, United States

Technology to automatically synthesize linguistically accurate and natural-looking sign language animations can increase information accessibility for people who are Deaf or hard-of-hearing. We investigate the synthesis of syntactic American Sign Language (ASL) facial expressions, which are grammatically required and essential to the meaning of ASL animations. Specifically, we show that an annotated sign language corpus, including both the manual and non-manual signs, can be used to model and generate linguistically meaningful facial expressions, if it is combined with facial feature extraction techniques, statistical machine learning, and an animation platform with detailed facial parameterization. Our synthesis approach uses recordings of human ASL signers as a basis for generating face and head movements for animation. We train our models with facial expression examples that are represented as MPEG-4 facial action time series extracted from an ASL video corpus using computer vision based face-tracking. To avoid idiosyncratic aspects of a single performance, we model a facial expression based on the underlying trace of movements learned from multiple recordings of different sentences where such expressions occur. Latent traces are obtained using Continuous Profile Models (CPM), which are probabilistic generative models. We assessed our modeling approach through comparison with an alternative centroid approach, where a single representative performance was selected by minimizing DTW distance from the other examples. Through both metric evaluation and an experimental user study with Deaf participants, we found that the facial expressions driven by our CPM models produce high-quality facial expressions that are more similar to human performance of novel sentences.

23. Time Series Prediction and Online Learning: Model Selection and Ensemble Learning

Vitaly Kuznetsov, PhD[1], Mehryar Mohri, PhD[1,2]
[1]Google Research, New York, New York, United States;
[2]Courant Institute, New York, New York, United States

We present a series of theoretical and algorithmic results for time series prediction leveraging recent advances in the statistical learning analysis of this problem and on-line learning. We prove the first generalization bounds for a hypothesis derived by online-to-batch conversion of the sequence of hypotheses output by an online algorithm, in the general setting of a non-stationary non-mixing stochastic process. Our guarantees hold for adapted sequences of hypotheses both for convex and non-convex losses. We give generalization bounds for sequences of hypotheses that may not be adapted but that admit a stability property. Our bounds are given in terms of a discrepancy measure, which we show can be accurately estimated from data under a mild assumption.

We also highlight the application of our results to two related problems: model selection in time series prediction, and the design of accurate ensembles of time series predictors. Model selection for time series prediction appears to be a difficult task: in contrast with the i.i.d. scenario, in time series analysis, there is no straightforward method for splitting a sample into a training and validation sets. Using the most recent data for validation may result in models that ignore the most recent information. Validating over the most distant past may lead to selecting sub-optimal parameters. Any other split of the sample may result in the destruction of important statistical patterns and correlations across time. We show that, remarkably, our on-line-to-batch conversions enable us to use the same time series for both training and model selection.

24. Maximin Variational Inference

Adji B. Dieng, MPhil[1], John Paisley, PhD[1], Stephan Mandt, PhD[2], David M. Blei, PhD[1]
[1]Columbia University, New York, New York, United States
[2]Disney Research, Pittsburg, Pennsylvania, United States

Variational inference is widely used for approximate posterior inference. It casts approximate inference as the optimization of a divergence measure. Different divergence measures have been proposed in the literature, including alpha-divergences and chi-divergences. The choice of the divergence impacts the properties of the approximating variational distribution. For example optimizing chi-divergences is equivalent to optimizing upper bounds to the log marginal likelihood and leads to overdispersed approximations while optimizing alpha-divergences is equivalent to optimizing lower bounds to the log marginal likelihood and leads to underdispersed variational distributions. In this work, we propose two objective functions that combine these bounds into a single objective. We propose two stochastic gradient algorithms for optimization and show their connections to the Importance Weighted Autoencoder. We study the properties of these objectives and their gradients and compare them to existing variational inference approaches in terms of predictive performance and quality of the posterior approximation. Experiments on Bayesian logistic regression and variational autoencoders show the usefulness of the proposed objectives.

25. Stochastic Bouncy Particle Sampler

Ari Pakman[1], Dar Gilboa[1], David Carlson[2], and Liam Paninski[1]
[1] Columbia University, New York, New York, United States;
[2] Duke University, Durham, North Carolina, United States.

We introduce a novel stochastic version of the non-reversible, rejection-free Bouncy Particle Sampler (BPS), a Markov process whose sample trajectories are piecewise linear. The algorithm is based on simulating first arrival times in a doubly stochastic Poisson process using the thinning method, and allows efficient sampling of Bayesian posteriors in big datasets. We prove that in the BPS no bias is introduced by noisy evaluations of the log-likelihood gradient. On the other hand, we argue that efficiency considerations favor a small, controllable bias in the construction of the thinning proposals, in exchange for faster mixing. We introduce a simple regression-based proposal intensity for the thinning method that controls this trade-off. We illustrate the algorithm in several examples in which it outperforms both unbiased, but slowly mixing stochastic versions of BPS, as well as biased stochastic gradient-based samplers.

26. Active Learning for Low-Resource Speech Recognition: Impact of Selection Size and Language Modeling Data

Ali Raza Syed, MSc[1], Andrew Rosenberg, PhD[2], Michael Mandel, PhD[1]
[1]The Graduate Center CUNY, New York, New York, United States;
[2]IBM TJ Watson Research Center, Yorktown Heights, New York, United States

Automatic speech recognition (ASR) requires transcribed speech for training. High quality speech transcription is expensive and time-consuming, especially for low resource languages where expert transcription services are limited. Active learning aims to reduce the time and cost of developing ASR systems by selecting for transcription highly informative subsets from large pools of audio data. In the context of speech recognition on low-resource languages, investigated in the OpenKWS and IARPA BABEL programs, we explored the value of active learning with larger selection sets, and with the introduction of more data for language modeling. We find that, in general, the value of active learning is persistent with larger selections. We see larger gains from active learning over random selection at 4-hour selections than at 2 hours, with smaller but consistent gains at 9 and 14 hours selections. We assess the hypothesis that gains from active learning may be overwhelmed by the introduction of additional language model data and find performance improvements from active learning and additional language model data to be orthogonal and complementary. The impact of additional language data is found to be fairly consistent within language, regardless of the amount of training data or selection strategy.

27. Automating Feature Engineering for Supervised Learning Problems

Udayan Khurana, PhD[1], Fatemeh Nargesian, MS[2], Horst Samulowitz, PhD[1], Deepak Turaga, PhD[1], Elias Khalil, MS[3], Tejaswini Pedapati, MS[1]
[1]IBM TJ Watson Research Center, Yorktown Heights, NY, USA
[2]University of Toronto, Toronta, ON, Canada
[3]Georgia Institute of Technology, Atlanta, GA, USA

Feature Engineering in supervised learning is the task of transforming the set of features in a given dataset to improve the performance of a predictive model. It is a crucial but time-intensive and skillful process, involving a data scientist or a domain expert, or both. It is often the key determinant of the time and cost required to build a model, as well as the effectiveness of the model. As part of the feature engineering task, a data scientist performs transformations, (and their compositions) or subset-selection on given input dataset in an iterative manner while observing their impact on the performance for the desired predictive analysis task. The efficacy of this human driven process is heavily dependent on the domain and statistical expertise of the individual. Also, it is constrained by the time to delivery and prone to person's bias and human error. In this talk, we discuss our system for performing feature engineering in an automated manner using a combination of exploratory and learning techniques. The system explores available choices for data transformations in an efficient manner while optimizing for model performance. At the same time it refines its exploration strategy based on past experiences on previously analyzed datasets. We conclude with a discussion on how one can jointly optimize feature engineering and model selection without explicitly exploring the combinatorial explosion of the choices, and mention our larger charter of an automated data science pipeline with the goal of proliferating the adoption of data science in information systems.

28. The Temporal Neural Coding Network: Towards Lifelong Language Learning

Alexander G. Ororbia II, BS[1], David Reitter, PhD, and C. Lee Giles, PhD[1]
[1]Pennsylvania State University, University Park, Pennsylvania, United States;

We present a novel lifelong neural architecture, the Temporal Neural Coding Network (TNCN), and its learning algorithm, for uncovering multiple levels of distributed representations for language data streams. The TNCN model adapts its parameters iteratively as new samples are observed without resorting to the popular, but expensive back-propagation through time procedure needed to calculate gradients for recurrent neural networks, notably requiring no unrolling of internally defined recurrence relations. Furthermore, we integrate concepts from variational encoder-decoder models, such as those in [1], to easily demonstrate how uncertainty can be characterized in the model's next step predictions. We discuss how the proposed TNCN works specifically on language-based tasks formulated in the streaming setting, compared to various neural models, as well as how the TNCN may be easily adapted to other real-time tasks, such as video sequence modeling, yielding a natural framework for online multi-modal and online semi-supervised learning.

References:
Serban, Iulian; Ororbia II, Alexander G.; Pineau, Joelle; Courville, Aaron. “Multi-modal Variational Encoder-Decoders”. 2017. arXiv:1612.00377 [cs.CL]

29. The Topology of Time Series Networks

Ben Cassidy, PhD[1], DuBois Bowman, PhD[1], Goran Marjanovic, PhD[2], Victor Solo, PhD[2,3]
[1]Columbia University, New York, New York, United States;
[2]The University of New South Wales, Sydney, Australia;
[3]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, Massachusetts, United States

Large high-dimensional networks constructed from time series arise naturally in many areas of machine learning such as genomics, neuroimaging, econometrics, and social networks. But the interpretation and comparison of differently sized networks remains problematic, since many measures assume the networks to be compared are defined on the same node set, or those measures scale with network (graph) size. A further problem is to meaningfully compare between dense and sparse time series networks.
In this work we introduce a new network comparison method based on Persistent Homology, an approach from Topological Signal Processing. Persistent homology studies the mesoscopic architecture within networks, to find distributed patterns that appear or disappear over a range of scales.
We apply the new method to compare networks from neuroimaging (brain activity networks) and econometrics datasets; we show which fundamental topological features are equivalent or different between networks of different sizes, and between sparse and dense networks. In the process we explain why many existing time series network identification methods fail in real world circumstances.

30. AdaNet: Adaptive Structural Learning of Artificial Neural Networks

Corinna Cortes, PhD[1], Xavi Gonzalvo, PhD[1], Vitaly Kuznetsov, PhD[1], Mehryar Mohri, PhD[1,2] and **Scott Yang**, BS[2]
[1]Google Research, New York, New York, United States;
[2]Courant Institute, New York, New York, United States

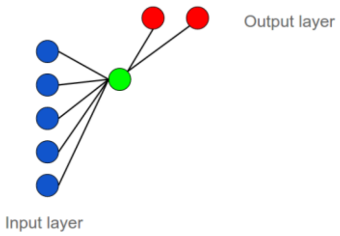
We present a new theoretical framework for building and learning artificial neural networks. Formulating the configuration of a neural network as an effective parametrized model for supervised learning, we design a method for training neural networks that adapts model structure and complexity to the difficulty of the particular problem, with **no pre-defined architecture.**

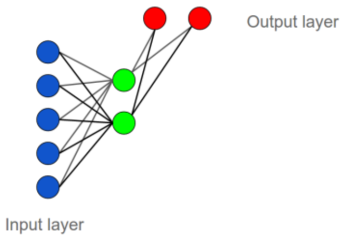
Starting from a simple single layer network, we add additional neurons and layers to the model using rigorous estimates of the generalization ability from statistical learning theory. Specifically, we determine

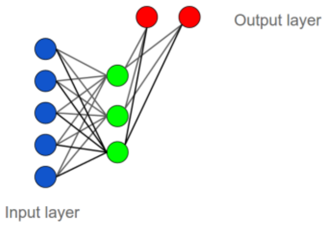
whether to add new nodes to the network by measuring the incremental reduction in empirical error against the additional model complexity.

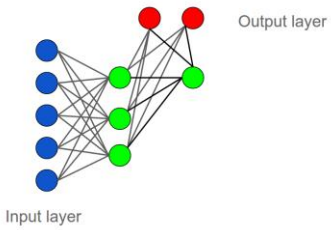
This structural risk minimization-type approach is in contrast to existing work on structure learning, which has been either based on heuristics for growing and pruning the network or focused on designing more efficient methods for large-scale hyperparameter tuning. The former is unable to provide concrete theoretical guarantees, and the latter is wasteful of data and statistically equivalent to standard grid search.

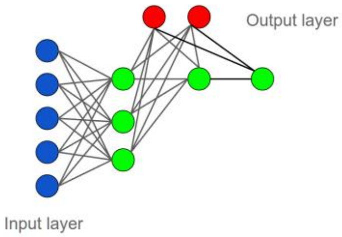
Our structure learning algorithm is based on directly optimizing for generalization performance. This allows us to not only enforce model sparsity but to also **provide strong accompanying generalization bounds**, in contrast to previous work in this area. Remarkably, the resulting method is also convex and hence more stable than many of the current deep learning methodologies employed. Preliminary experiments support and validate our theory and framework.

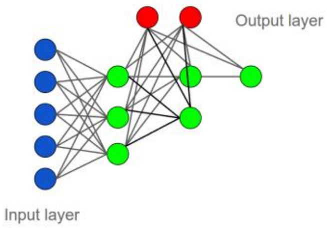












31. Learning with Rejection.

Corinna Cortes, PhD[2], **Giulia DeSalvo**, BS[1], and Mehryar Mohri, PhD[1]
[1]New York University Courant Institute, New York, New York, United States;
[2]Google Research, New York, New York, United States;

We consider a flexible binary classification scenario where the learner is given the option to reject an instance instead of returning its prediction, thereby incurring a cost. We introduce a novel framework for this setting that consists of simultaneously learning two functions: a classifier along with a rejection function. We present a full theoretical analysis of this framework including new data-dependent learning bounds in terms of the Rademacher complexities of the classifier and rejection families as well as consistency and calibration results. These theoretical guarantees guide us in designing new algorithms that can exploit different kernel-based hypothesis sets for the two sets of classifier and rejection functions. We also present a boosting-style algorithm where at each round, it selects a pair of functions, a base classifier and a base rejection function. We give its convergence guarantees along with a linear-time weak-learning algorithm for rejection stumps. We compare and contrast to the special case of confidence-based rejection, devising alternative loss functions and algorithms for this setting as well. We report the results of several experiments showing that our algorithms can yield a notable improvement over the best existing confidence-based rejection algorithm.

32. Scalable Feature Selection via Distributed Diversity Maximization

Sepehr Abbasi Zadeh, BS[1], Mehrdad Ghadiri, BS[1], Vahab Mirrokni, PhD[2], **Morteza Zadimoghaddam**, PhD[2]
[1]Sharif University of Technology, Tehran, Iran;
[2]Google Inc., New York, New York, United States

Feature selection is a fundamental problem in machine learning and data mining. The majority of feature selection algorithms are designed for running on a single machine (centralized setting) and they are less applicable to very large datasets. Although there are some distributed methods to tackle this problem, most of them are distributing the data horizontally which are not suitable for datasets with a large number of features and few number of instances. We introduce a novel vertically distributable feature selection method in order to speed up this process and be able to handle very large datasets in a scalable manner. In general, feature selection methods aim at selecting relevant and non-redundant features (Minimum Redundancy and Maximum Relevance). It is much harder to consider redundancy in a vertically distributed setting than a centralized setting since there is no global access to the whole data. We formalize the feature selection problem as a diversity maximization problem by introducing a mutual-information-based metric distance on the features. We validate our method by performing an extensive empirical study. In particular, we show that our distributed method outperforms state-of-the-art centralized feature selection algorithms on a variety of datasets. From a theoretical point of view, we have proved that the used greedy algorithm in our method achieves an approximation factor of 1/4 for the diversity maximization problem in a distributed setting with high probability. Furthermore, we improve this to 8/25 expected approximation using multiplicity in our distribution.

33. Assessing the Promise of Automated Plant Phenotyping for Comprehensive Genetic Mapping: A Unified Framework of Joint Model Clustering and Estimation for Predictive Modeling and GWAS of Plant Varieties

Addie M. Thompson, PhD[1], Ming Yu, MS[2], Karthikeyan Natesan Ramamurthy, PhD[3], Aurélie C. Lozano, PhD[3], Eunho Yang, PhD[4], Melba M. Crawford, PhD[1], Ayman F. Habib, PhD[1], Edward J. Delp, PhD[1], Clifford F. Weil, PhD[1], Mitchell R. Tuinstra, PhD[1]

[1] Dept. of Agronomy, Purdue University, W. Lafayette, IN, USA
[2] Booth School of Business, The University of Chicago, Chicago, IL, USA
[3] Mathematical Sciences Dept., IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
[4] Dept. of Computer Sciences, KAIST University, Daejeon, South Korea

Traditional genetic mapping of plant traits involve surveying hundreds of plant varieties for the phenotypes of interest (e.g. plant height), using time-consuming hand measurements. The shift to automated, high-throughput remote sensing of phenotypes using airborne and wheel-based systems is beginning to enable much larger scale data collection: more data points per plot and more genotypes.

Can such data provide a more complete description of the genetics contributing to the phenotypic diversity and significantly improve mapping of quantitative traits?

In this work we take the first step in assessing the promise of automated phenotyping. We do so by developing and evaluating 1) predictive models of plant traits using remotely sensed data, 2) GWAS methodologies mapping such derived phenotypes in lieu of hand-measured traits. The presence of multiple varieties with replicates much smaller in number than predictors poses a major challenge: Building separate models for each variety is unrealistic, while a single model does not fit all. We propose a unified framework for 1) and 2) that addresses this challenge by leveraging structured sparsity to automatically discover hierarchical groupings among both varieties and phenotypes, and learn differentiated models for the different groups while sharing information within groups.

We evaluate our approach in a large and diverse sorghum population with features derived from airborne hyperspectral and RGB image data and manually collected trait measurements. Our results suggest that richer genetic mapping can indeed be obtained from automated phenotyping. In addition, our discovered groupings reveal interesting insights from a plant science perspective.

34. Reducing Revenue Optimization to Regression

Andres Munoz Medina, PhD, Sergei Vassilvitskii, PhD.
Google Research

Posted-price and second-price auctions are some of the most popular mechanisms for selling online advertising inventory, and have become popular topics of study in recent years. Given that online advertising generates billions of dollars annually, it comes as no surprise that online companies such as Facebook and Google attempt to maximize revenue obtained through these mechanisms.

In practice, revenue optimization requires some notion of predicting the bidder’s value for a particular impression as a function of some side information or features. The challenge in the prediction problem is that the loss function is not symmetric -- in most applications over-predicting (and attempting to charge too high of a price) is more costly than under-predicting (and charging too little). Prior to our work, the only algorithms that attempt to solve this problem were given in [1,2]. These algorithms however do not provide guarantees on the amount of revenue they obtain.

In this work we show how to avoid the underlying asymmetry of this problem and present a reduction of revenue optimization in posted-price auctions to a simple regression problem.We give the first algorithm with formal approximation guarantees to the optimal revenue attainable with side information. We proceed by first proving a bound on the achievable revenue as a function of the variance of the underlying bid distribution, a result that may be of independent interest. We then show how to combine clustering algorithms with a model trained to minimize regression to achieve an approximately optimal solution.

[1] Mohri, Mehryar, and Andres Munoz Medina. "Learning Theory and Algorithms for revenue optimization in second price auctions with reserve." ICML. 2014.
[2] Cui, Ying, Zhang, Ruofei, Li, Wei, and Mao, Jianchang. Bid landscape forecasting in online ad exchange marketplace. In KDD, pp. 265–273, 2011.

35. Reconstructing Hidden Permutations Using the Average-Precision (AP) Correlation Statistic

Lorenzo De Stefani, PhD[1], Alessandro Epasto, PhD[2] , Eli Upfal, PhD[1], and Fabio Vandin, PhD[3]
[1]Brown University, Providence, Rhode Island, United States;
[2]Google Research, New York, New York, United States;
[3]Univerista degli Studi di Padova, Padova, Italy;

We study the problem of learning probabilistic models for permutations, where the order between highly ranked items in the observed permutations is more reliable (i.e., consistent in different rankings) than the order between lower ranked items, a typical phenomena observed in many applications such as web search results and product ranking. We introduce and study a variant of the Mallows model where the distribution is a function of the widely used Average-Precision (AP) Correlation statistic, instead of the standard Kendall's tau distance.
We present a generative model for constructing samples from this distribution and prove useful properties of that distribution. Using these properties we develop an efficient algorithm that provably computes an asymptotically unbiased estimate of the center permutation, and a faster algorithm that learns with high probability the hidden central permutation for a wide range of the parameters of the model.
We show experimentally the accuracy of our estimators in the reconstruction of the hidden permutation with a limited number of samples. We also show with real data that unsupervised methods based on our model can precisely (and efficiently) identify ground-truth clusters of rankings. Compared to the Kendall's tau based methods, our methods are less affected by noise in low-rank items.
Finally, we show that supervised classification algorithms based on the AP statistic outperform Kendall's tau distance based algorithms by comparing their performance in classifying large genomic expressions from ten publicly available datasets from human cancer research.
Appeared in the proceedings of the Thirtieth (AAAI) Conference on Artificial Intelligence.

36. Interpretable Behavioral Profiling of Patients for Shared Decision Making

Subhro Das, PhD, and Pei-Yun S. Hsueh, PhD
IBM Thomas J. Watson Research Center, Yorktown Heights, New York, United States

In the United States, behavior-related health determinants (McGinnis, 2002) cause nearly 40% of the deaths. In contrast to other health determinants, namely genetics (30%), social/environmental (20%) and access-to-care (10%), behavioral determinants are more likely to be altered by preventive interventions. Recently, owing to the trend of Quantified Self and availability of consumer health devices/sensors, a plethora of patient-centered data are being generated. This data provides a large potential for learning methods to derive goal-oriented patient behavioral insights, which in turn enable better decision support for physicians/practitioners and facilitate positive changes in patient behavior. However, the efforts are often hinged on interpretability of the learning results. To investigate into possibly interpretable learning approaches, we evaluate a method that segments users/patients into different behavioral profiles based on their prior history of behavioral patterns and proxy outcomes. We first apply the locally supervised metric learner (Sun, 2012) of similarity analytics to estimate the outcome-related behavioral distances between the users. Then, based on the adjusted behavioral distances, hierarchical clustering is employed to generate similar patient/user cohorts and learn the key features (user preferences and barriers) that drive the differential behavioral outcomes for goal completion. The derived behavioral profiles, along with prototypical user examples identified for each cohort, yield interpretable explanations behind each patient's behavioral pattern. Finally, predictive models, trained on each behavioral profile, evaluate the effectiveness of various interventions. The predicted interventions and the reasons/explanations behind them enable the practitioners to include their domain-knowledge/experiences resulting in effective shared decision-making.

37. Archeological Feature Recognition from LiDAR Data Using Visual Deep Learnt Representation

C. Albrecht[1], C. Fisher[2], M. Freitag[1], H.F. Hamann[1], S. Pankanti[1], F. Pezzutti[2], F. Rossi[1]

1: IBM TJ Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598. Emails: cmalbrec@us.ibm.com, mfreitag@us.ibm.com, hendrikh@us.ibm.com, sharat@us.ibm.com, francesca.rossi2@ibm.com
2: Clark B224, Department of Anthropology, Colorado State University, Fort Collins, CO 80523. Emails: Ctfisher@colostate.edu,fpezzutt@gmail.com

To understand ancient civilizations, traditionally archeologists perform exhaustive fieldwork and manual mapping. Patterns of past organization can yield important insights that can inform modern policy. Archeologists exploit their domain knowledge of ancient architectural patterning to recognize human constructions and classify them for example as houses, temples, walls, streets, and other elements of past settlements. Traditional archeologist techniques, however, are costly and slow. New remote sensing tools such as light detection and ranging (LiDAR) are revolutionizing the field of archaeology by providing rapid, high resolution scans of ancient cities and landscapes. The problem is that, given the cost and labor intensive nature of traditional methods, archaeologists cannot effectively analyze these big data. In this paper we describe a project where AI techniques are used to scale these methods and speed them up. Using LiDAR data from the ancient Puépecha city of Angamuco, our project exploits domain knowledge to recognize promising geo-spatial areas in combination with a deep learning classifier to understand if they are features of a city (such as houses, temples, or streets). This allows archeologists to identify areas of interest more effectively and to automatically classify and digitize individual features. The performance and accuracy of our method shows great potential and possible applicability also to other scientific areas where 3D or 2D data is given as input.

38. Submodular Optimization over Sliding Windows

Alessandro Epasto, PhD[1], Silvio Lattanzi, PhD[1], Sergei Vassilvitskii, PhD[1], Morteza Zadimoghaddam, PhD[1];
[1]Google Inc., New York, New York, United States

Maximizing submodular functions under cardinality constraints is a fundamental primitive in numerous machine learning applications, including data diversification and summarization, feature selection and coverage problems.
In this work, we study this question in the context of data streams, where elements arrive one at a time, and we want to design low-memory and fast update-time algorithms that maintain a good solution. Specifically, we focus on the sliding window model, where we are asked to maintain a solution that considers only the last W items.

We provide the first algorithm that maintains a provable approximation of the optimum using space sublinear in the size of the window. We give a 1/3 - \epsilon approximation algorithm that uses space polylogarithmic in the spread of the values of the function, and linear in the solution size k for any constant \epsilon > 0, while requiring only polylogarithmic function evaluations per item processed. We also show a different algorithm that, at the cost of using more memory, provides a 1/2 - \epsilon approximation. This algorithm matches the best known approximation guarantees for submodular maximization in insertion-only streams, a less general formulation of the problem.

We demonstrate the efficacy of the algorithms on a number of real world datasets, showing that their practical performance far exceeds the theoretical bounds. The algorithms preserve high quality solutions in streams with millions of items, while storing a negligible fraction of them.

To appear in Proceeding of the International World Wide Web Conference (WWW), 2017.
https://arxiv.org/abs/1610.09984

39. Tensor Decomposition for Single-cell RNA-Sequencing Data

Sandhya Prabhakaran, PhD[2], Ambrose J. Carr, MPhil[1,2], **Kristy Choi**[1,2], and Dana Pe'er, PhD[1,2]
[1]Columbia University, New York, New York, United States;
[2]Memorial Sloan Kettering Cancer Center, New York, New York, United States

Single-cell RNA-sequencing (scRNA-seq) presents exciting opportunities to characterize complex biological systems at an unprecedented resolution. However, extreme levels of sparsity, bias, and noise present in the data distort meaningful signal and pose challenges for downstream analysis. To address this issue, we present a probabilistic framework to better quantify gene expression in scRNA-seq data. scRNA-seq samples from a library of mRNA with replacement, but current methods do not associate the observation's multiplicity with varying confidence Instead, they discard observations by collapsing duplicate mRNA. To utilize the information embedded in this confidence dimension, we pose scRNA-seq data as a third-order tensor of cells, genes, and mRNA counts rather than the traditional count matrix of cells by genes. By using Bayesian Tucker decomposition to infer the compressed core tensor and factor matrices, we learn the data's low-dimensional structure to identify latent cell-gene relationships in the original tensor. We apply our model on a sample of peripheral blood mononuclear cells and demonstrate its ability to: (1) recapitulate the known biology of hematopoiesis and (2) capture block diagonal structures depicting coexpressing genes.

40. Genome-Wide and Transcriptome-Wide Network-Network Analysis of Eqtl Identifying Weak Trans- Signals

Shuo Yang, MSc[1], Dana Pe'er, PhD[2], and Itsik Pe'er, PhD[1]
[1]Columbia University, New York, New York, United States;
[2]Memorial Sloan Kettering Cancer Center, New York, New York, United States

Expression quantitative trait loci (eQTLs) have been extensively studied, as they provide an attractive functional interpretation to sequence variation. Yet, many traditional eQTL studies have been mainly focused on *cis*- analysis of each transcript[1], without drawing on the complete picture of the regulatory repertoire. In this project, we make the landscape of genetic effects on co-expressed genes effectively modeled by separating a network of non-linearly regulated hidden factors from the expression network affected by these factors. Specifically, we build a neural network from genotype (~2.5 million pre-analyzed SNPs) to gene expression with one hidden layer of factors. Each such hidden factor affects a network of co-expressed genes and is determined by a network of regulatory SNPs. This network-network analysis will inform *trans*- signals from *cis*- regulation. The sigmoid function on the hidden layer brings non-linearity to the regulatory effects of SNPs on genes. We further introduce context-specificity to the model, allowing analysis across multiple tissues.

We initialize the model with PCA and LASSO, sparsifying each co-expression network as well as its regulatory SNP network. We solve the model using stochastic gradient descent with backpropagation and L1 regularization, and mitigate computational challenges through GPU computing. We tested our modeling on both simulated and real datasets. We used the GTEx dataset[2] to describe SNP-gene associations with their tissue dynamics, and their enriched association to complex traits.

In conclusion, we model genetics-genomics data in a network-network fashion, trying to identify weak but shared signals of eQTLs, including novel *trans*- associations.

1. Melé et al. Science 348 (6235), 660-665
2. The GTEx Consortium. Science 348 (6235), 648-660

41. Identifying Heterogeneous Treatment Effects In Clinical Trial Data: An Implementation of Causal Forests

Joseph Scarpa, PhD[1], Aaron Baum, PhD[1], Emilie Bruzelius, MPH[1,2], and James H. Faghmous, PhD[1]
[1]Arnhold Institute for Global Health at Mount Sinai School of Medicine, New York, New York, United States;
[2]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York, United States

Randomized controlled trials (RCTs) are considered the gold standard for study design because they estimate the average effect of a treatment on a population of interest. However, effects often vary considerably across populations and identifying the patients for whom a treatment is especially harmful or beneficial is an important priority. Clinical trials pre-specify a limited number of subgroup analyses to minimize multiple hypothesis testing. However, when average treatment effects mask important heterogeneity, limited subgroup analyses may lead to poor clinical or policy conclusions. Machine learning, and especially tree-based methods, are emerging as an important tool for discovering such subpopulations.

Recently, causal forests were proposed as an alternative approach to specification in high dimensional settings. Conceptually, causal forests enable subgroup identification by using a randomly split partition of data for hypothesis generation, while maintaining a hold-out portion for inference. Here, we demonstrate causal forests first application in clinical trials and extend the method--developing a heuristic to select representative subgroups. Using data from Look Ahead (N = 5145), a landmark trial testing whether weight-loss impacts cardiovascular mortality, we estimate that the intervention had significant negative mortality effects for 16% of participants with HbA1C < 6.8% and perceived poor general health, but significant positive mortality effects for the remaining participants (1.84 and 2.33 events per 100 person-years; HR: 0.78; 95% CI: 0.64, 0.93; P=0.020). Our findings suggest that data-driven methods can uncover heterogeneity in experimental data and highlight the potential contributions of machine learning to RCT reanalysis.

42. Proximity Variational Inference

Jaan Altosaar[1], Rajesh Ranganath [1], David M. Blei, PhD[2]
[1]Princeton University, Princeton, New Jersey, United States;
[2]Columbia University, New York, New York, United States

Variational inference is a method for approximating posterior distributions in latent variable models. This technique has enabled the use of Bayesian methodology in settings where it would otherwise be infeasible, yet problems remain. The optimization suffers from sensitivity to initialization and prefers underdispersed distributions. We remedy this by deriving an alternate optimization algorithm for variational inference based on proximal expansions of the variational objective with additional constraints. We derive a scalable variant that runs as fast as variational inference. In our experiments, we design an entropy constraint and show that our approach is less sensitive to initialization. We test the method in a Bernoulli factor model, a sigmoid belief net model, and a deep latent Gaussian model (or variational autoencoder), all trained on MNIST. In the sigmoid belief net and deep latent Gaussian model, our algorithm recovers good posterior predictive distributions where standard variational inference fails.

43. Deep Survival Analysis

Rajesh Ranganath, MA[1], Adler Perotte MD[2], Noémie Elhadad, PhD[2], David Blei PhD[2]
[1]Princeton University, Princeton, New Jersey, United States;
[2]Columbia University, New York, New York, United States

The electronic health record (EHR) provides an unprecedented opportunity to build actionable tools to support physicians at the point of care. In this paper, we introduce deep survival analysis, a hierarchical

generative approach to survival analysis in the context of the EHR. It departs from previous approaches in two main ways: (1) all observations, including covariates, are modeled jointly conditioned on a rich latent structure; and (2) the observations are aligned by their failure time, rather than by an arbitrary time zero as in traditional survival analysis. Further, it handles heterogeneous data types that occur in the EHR. We validate deep survival analysis by stratifying patients according to risk of developing coronary heart disease (CHD) on 313,000 patients corresponding to 5.5 million months of observations. When compared to the clinically validated Framingham CHD risk score, deep survival analysis is superior in stratifying patients according to their risk.

44. Sparse Portfolio Construction by Penalized Regressions

Junyi Zhang, PhD[1], **Ao Lv**, MSc[1]
Paul H. Chook Department of Information Systems and Statistics, Baruch College, CUNY, New York City, New York, United States

Modern portfolio theory has been well developed by Harry Markowitz who introduced the well-known Markowitz mean-variance framework to achieve a family of optimal linear combinations of risky assets. The beauty of Markowitz portfolio construction is lying in revealing the convex boundary, which is now called efficient frontier, on a plane of portfolio mean return versus portfolio mean volatility. Furthermore, since the efficient frontier is convex, the efficient frontier of all risky assets along with a risk-free asset has been shown to be the straight line tangent to the convex boundary. However, for practitioners such as intro-day traders, it is impossible to invest on a portfolio of all risky assets plus several 'risk-free' assets. Therefore a sparse portfolio selection/construction theory is in great demand by financial industry. We use penalized regression methods including least absolute shrinkage and selection operator (LASSO) type of L1 penalty, elastic-net type of L1 plus L2 penalty, smoothly clipped absolute deviation (SCAD) type of penalty to study the convex property of resulting efficient frontier with selected sparse portfolio. In addition, based on the derived efficient frontier of a sparse portfolio, we solve the problem of maximizing Shape's ratio, which can be viewed as risk-adjusted mean return, and compare the results among using various penalization regressions.

45. From Causal Inference to Transfer Learning and Back

Yixin Wang, MA[1], and David M. Blei, PhD[2]
[1]Departments of Statistics, Columbia University, New York, New York, United States;
[2]Departments of Statistics and Computer Science, Columbia University, New York, New York, United States

It has recently been demonstrated that causal inference ideas can facilitate transfer learning: Causal analysis can inspire efficient methods that characterizes knowledge transfer across different environments; The usual covariate shift assumption in transfer learning can also be relaxed by borrowing causal ideas. In this work, however, we study the other direction: how transfer learning could help with causal inference. We take on a novel perspective that views covariate distributions as instruments and transferrable rules as causes. This would enable more accurate estimation of causal effect under the presence of confounders. We illustrate this technique by studying EEG signals and its resulting actions.

46. Clustering Breathing Curves in 4D Radiotherapy by using Multiple Machine Learning Tools: K-Means, Hierarchical Clustering, and Gaussian Mixture Models

Qiongge Li, PhD Candidate[1], Maria Chan, PhD[2] and Chengyu Shi, PhD[2]
[1]The Graduate Center of City University of New York, New York, New York, United States;
[2]The Memorial Sloan Kettering Cancer Center at Basking Ridge, Basking Ridge, New Jersey, United States

Patient respiratory motion introduces a great amount of uncertainty into radiation treatments, which would either need to treat a larger volume of healthy tissues in order not to miss the target, or not catch the whole target and therefore have local recurrence later on. Different patients present different breathing patterns and it is difficult to customize the treatment target. The intent of this research is to use machine learning (ML) techniques to classify patients' breathing patterns into sub-groups, in order to assist physicians to customize the treatment target range. Three ML tools: k-means, hierarchical clustering, and gaussian mixture models were used to analyze 341 breathing curves obtained during 4DCT scanning. After extracting features of these curves (i.e., frequency, amplitude, standard deviation of amplitude, spread of frequency spectrum) and cleaning the data (i.e., correcting the baseline drifting, interpolating the missing data), 74 high-quality datasets were retained for further analysis. The preliminary results were acquired from a small testing sample data set (9 breathing curves) with obvious signal behaviors: one group with fast and shallow breathing (7 curves) and the other group with slow and deep breathing (2 curves). Five features were extracted from time series and frequency spectrum analysis. Three algorithms completed the task successfully (with 100 percent accuracy). This shows that it is possible to cluster the patients' breathing patterns using all three ML models. However, further study is needed to correlate the results with other clinical information of patients before it can be fully benefited at clinical implication.

47. AMOS: An Automated Model Order Selection Algorithm for Spectral Graph Clustering

Pin-Yu Chen, PhD[1], Thibaut Gensollen[2], and Alfred O. Hero III[2], Professor
[1]IBM Thomas J. Watson Research Center, Yorktown Heights, New York
[2]University of Michigan, Ann Arbor, Michigan

One of the longstanding problems in spectral graph clustering (SGC) is the so-called model order selection problem: automated selection of the correct number of clusters. This is equivalent to the problem of finding the number of connected components or communities in an undirected graph. In this paper, we propose AMOS, an automated model order selection algorithm for SGC. Based on a recent analysis of clustering reliability for SGC under the random interconnection model, AMOS works by incrementally increasing the number of clusters, estimating the quality of identified clusters, and providing a series of clustering reliability tests. Consequently, AMOS outputs clusters of minimal model order with statistical clustering reliability guarantees. Comparing to four other automated graph clustering methods on real-world datasets, AMOS shows superior performance in terms of multiple external and internal clustering metrics.

48. Human Teaching by Demonstration: Showing versus Doing Reinforcement Learning Tasks

Mark K Ho, ScM[1], Michael Littman, PhD[1], James MacGlashan, PhD[1], Fiery Cushman, PhD[2], and Joseph L. Austerweil, PhD[3]

[1]Brown University, Providence, Rhode Island, United States of America
[2]Harvard University, Cambridge, Massachusetts, United States of America
[3]University of Wisconsin-Madison, Madison, Wisconsin, United States of America

What is the relationship between doing a task and showing another agent how to do a task? In inverse reinforcement learning, an algorithm attempts to learn a policy or task representation by observing example demonstrations. Typically, these example demonstrations are sampled from the optimal policy for a task. That is, they are samples of an expert "doing" a task. However, when teaching by demonstration, people often do not simply do a task but actively attempt to show others how to do a task. A better understanding of this showing behavior can be used to improve inverse reinforcement learning algorithms.

In this work, we present a computational account of showing as a form of Bayesian pedagogy: optimally selecting sequences of behavior that lead an observer to infer a particular underlying reward function. This formulation can be understood as "planning in a learner's belief space" and extends previous accounts of teaching by example to sequential domains. In several human experiments, we demonstrate that people's behavior when "showing" a task closely matches this model, while their "doing" a task does not. Similarly, we show that these human-generated showing trajectories are more effective for training standard inverse reinforcement learning algorithms. Moreover, this work provides a basis for developing inverse reinforcement learning algorithms that can benefit from reasoning about intentional teaching.

49. PHATE: Potential Heat-diffusion Affinity-based Trajectory Embedding for Visualization of Progression Structure

Kevin R. Moon, PhD*[1], David van Dijk, PhD*[2], Zheng Wang, PhD[1], Tobias Welp, PhD[1], Guy Wolf, PhD[1], Ronald Coifman, PhD[1], Natalia Ivanova, PhD[1], Smita Krishnaswamy, PhD[1]
[1]Yale University, New Haven, Connecticut, United States
[2]Memorial Sloan Kettering, New York, New York, United States

Modern data analysis applications often contain latent notions of time progression of analyzed phenomena. Characterizing such progression paths is crucial in biological data, where cells are actively differentiating or progressing in response to signals. For example, progression of gut bacterial species in patients with autoimmune conditions can reveal the extent of the underlying disease. Another example can be seen in SNP (Single-Nucleotide Polymorphism) data where variations of gene mutations often exhibit gradual progression and branching divergence based on geographic locations. While several unsupervised dimensionality reduction and clustering methods have been used to infer trajectory structures in such data, they are typically inadequate for providing human-interpretable 2D or 3D visualizations of high-dimensional trajectory structures.

We propose an unsupervised low-dimensional (i.e, two- or three-dimensional) visualization of progression structures in high-dimensional data, which we call PHATE (Potential Heat-diffusion Affinity-based Trajectory Embedding). Our method is inspired by word vector embeddings in natural language processing such as Word2vec and Glove. Similarly, we use heat diffusion processes to compute co-occurrence similarities and define distances using their free energy potential. This diffusion-potential geometry preserves high-dimensional trajectory structures while enabling their visualization via a low-dimensional embedding that approximates their geometry. In addition to enabling visualization and manual interpretation, the PHATE method sharply emphasizes trajectories or (time) progressions in the data, while also denoising various technical artifacts and correcting for data collection dropouts. We demonstrate PHATE on a wide variety of big biological datasets including single-cell RNA sequencing, CyTOF data, population genetic data, and gut microbiome data.

* = authors contributed equally

50. Stochastic Bouncy Particle Sampler

Ari Pakman[1], Dar Gilboa[1], David Carlson[2], and Liam Paninski[1]
[1] Columbia University, New York, New York, United States;
[2] Duke University, Durham, North Carolina, United States.

We introduce a novel stochastic version of the non-reversible, rejection-free Bouncy Particle Sampler (BPS), a Markov process whose sample trajectories are piecewise linear. The algorithm is based on simulating first arrival times in a doubly stochastic Poisson process using the thinning method, and allows efficient sampling of Bayesian posteriors in big datasets. We prove that in the BPS no bias is introduced by noisy evaluations of the log-likelihood gradient. On the other hand, we argue that efficiency considerations favor a small, controllable bias in the construction of the thinning proposals, in exchange for faster mixing. We introduce a simple regression-based proposal intensity for the thinning method that controls this trade-off. We illustrate the algorithm in several examples in which it outperforms both unbiased, but slowly mixing stochastic versions of BPS, as well as biased stochastic gradient-based samplers.