ICML 北京 2014

International conference on machine learning, 2014

ICML 2014

北京

Abstracts of Papers

TUTORIALS
June 21, 2014
Beijing International Convention Center
Beijing, China

CONFERENCE SESSIONS
June 22 – 24, 2014
Beijing International Convention Center
Beijing, China

WORKSHOPS
June 25 – 26, 2014
Beijing International Convention Center
Beijing, China

ICML is the leading international machine learning conference and is supported by the International Machine Learning Society (IMLS)

The technical program includes 3 keynote speeches and 310 accepted papers, selected from a total of 1238 submissions considered by the program committee.

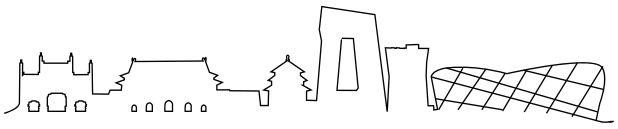


TABLE OF CONTENTS

Organizing Committee 2

Senior Program Committee 2

Message from the Program Chair 3

Sponsors 4

PROGRAM HIGHLIGHTS 6

CONFERENCE MAP 7

SATURDAY, June 21, 2014

Tutorials 11
Abstracts 12

SUNDAY, June 22, 2014

Keynote Session
Eric Horvits, Abstract 16

Conference Sessions
18 Parallel Tracks 17

Poster Sessions
Posters S1 - S106, M1 – M36 23

Sunday Abstracts 29

MONDAY, June 23, 2014

Keynote Session
Michael Kearns, Abstract 66

Conference Sessions
18 Parallel Tracks 67

Monday Abstracts 73

TUESDAY, June 24, 2014

Keynote Session
Michael I. Jordan, Abstract 106

Conference Sessions
18 Parallel Tracks 107

Poster Sessions
Posters M37 – M103, T1 – T101 113

Tuesday Abstracts 121

Workshop 160

Reviewer List 161

Author Index 166

ORGANIZING COMMITTEE

General Chair:	**David McAllester** (Toyota Technological Institute at Chicago)
Program Chazirs:	**Eric P. Xing** (Carnegie Mellon University) **Tony Jebara** (Columbia University)
Tutorial Chair:	**Ruslan Salakhutdinov** (University of Toronto)
Workshop Chair:	**Alex Ihler** (Toyota Technological Institute at Chicago)
Volunteer Chair:	**John Paisley** (Columbia University)
Publication Chairs:	**Fei Sha** (University of Southern California) **Percy Liang** (Stanford University)
Publicity Chair:	**Jingrui He** (Stevens Institute of Technology)
Financial Chairs:	**Artur Dubrawski** (Carnegie Mellon University) **Charles Isbell** (Georgia Institute of Technology)
Workflow Chairs:	**Kui Tang** (Columbia University) **Junming Yin** (Carnegie Mellon University)
Local Chairs:	**Changshui Zhang** (Tsinghua University) **Jun Zhu** (Tsinghua University) **Tie-Yan Liu** (Microsoft Research Asia)

SENIOR PROGRAM COMMITTEE

- Ryan Adams (Harvard University)
- Amr Ahmed (Google)
- Edoardo Airoldi (Harvard University)
- Animashree Anandkumar (University of California, Irvine)
- Peter Auer (University of Leoben)
- Francis Bach (INRIA)
- Drew Bagnell (Carnegie Mellon University)
- Arindam Banerjee (University of Minnesota)
- Mikhail Belkin (Ohio State University)
- Yoshua Bengio (University of Montreal)
- David Blei (Princeton University)
- Ryan Adams (Harvard University)
- Amr Ahmed (Google)
- Edoardo Airoldi (Harvard University)
- Animashree Anandkumar (University of California, Irvine)
- Peter Auer (University of Leoben)
- Francis Bach (INRIA)
- Drew Bagnell (Carnegie Mellon University)
- Arindam Banerjee (University of Minnesota)
- Mikhail Belkin (Ohio State University)
- Yoshua Bengio (University of Montreal)
- David Blei (Princeton University)
- Percy Liang (Stanford University)
- Han Liu (Princeton University)
- Yan Liu (University of Southern California)
- Marina Meila (University of Washington)
- Claire Monteleoni (George Washington University)
- Sayan Mukherjee (Duke University)
- Joelle Pineau (McGill University)
- Massimiliano Pontil (University College London)
- Pascal Poupart (University of Waterloo)
- Doina Precup (McGill University)
- Alan Qi (Purdue University)
- Marc'Aurelio Ranzato (University of Toronto)
- Pradeep Ravikumar (University of Texas, Austin)
- Ashutosh Saxena (Cornell University)
- Tobias Scheffer (University of Potsdam)
- Dale Schuurmans (University of Alberta)
- Fei Sha (University of Southern California)
- Ohad Shamir (Microsoft Research)
- Le Song (Georgia Institute of Technology)
- David Sontag (New York University)
- Suvrit Sra (Max-Planck Institute for Intelligent Systems)
- Nati Srebro (Toyota Technological Institute at Chicago)
- Rich Sutton (University of Alberta)
- Csaba Szepesvari (University of Alberta)
- Nuno Vasconcelos (University of California, San Diego)
- Jean-Philippe (Vert Mines ParisTech)
- S V N Vishwanathan (Purdue University)
- Kilian Weinberger (Washington University St. Louis)
- Sinead Wiliamson (University of Texas at Austin)
- Jieping Ye (University of Arizona)
- Rich Zemel (University of Toronto)
- Tong Zhang (Rutgers University)
- Alice Zheng (Microsoft)
- Zhi-Hua Zhou (Nanjing University)
- Xiaojin (Jerry) Zhu (University of Wisconsin-Madison)
- Jun Zhu (Tsinghua University)

MESSAGE FROM THE PROGRAM CHAIR

Dear ICML attendees,

Welcome to Beijing and the 31st International Conference on Machine Learning (ICML 2014)! We are excited to bring the premiere machine learning conference to China for the very first time. The conference will take place from June 21st to June 26th, 2014 at the Beijing International Convention Center (BICC), overlooking the famous Bird's Nest Olympic Stadium.

At the heart of the conference is the technical program which was implemented in two review cycles this year. A world-class program committee involving 76 area chairs and 597 reviewers evaluated a record-breaking total of 1238 submissions. The committee successfully selected 310 outstanding articles for publication in the proceedings. Along the way, many difficult decisions were made due to space limitations and the competition was fierce. All accepted articles from both Cycle 1 and Cycle 2 are published in the Journal of Machine Learning Research (JMLR) as Volume 32 of their Workshop and Conference Proceedings series. Due to the outstanding quality of all accepted papers, we allocated both an oral presentation as well as a poster presentation for each article. Oral presentations have been organized into sessions within 6 parallel tracks. These are located in nearby auditoriums so that conference participants can easily attend the sessions that are most relevant to their research interests.

In addition to the superb technical program of talks and posters, the conference also offers the following excellent attractions:

- Invited keynote speeches from the field's leading luminaries: Eric Horvitz, Michael Kearns and Michael Jordan. Each will present their insights on exciting machine learning topics.
- A tutorial program with 6 tutorials in currently blossoming areas of machine learning given by leading experts.
- A workshop program with 18 exciting workshops to present late-breaking work in specific areas of machine learning with many opportunities for collaboration and exploration. One of the workshops is co-located with the APSYS conference to promote synergy between machine learning researchers and systems researchers.
- A museum-tour and banquet featuring traditional Chinese cuisine with spectacular cultural performances.
- Best paper awards to honor the top articles by both established researchers and most promising students.

We would like to acknowledge all the people who made exceptional efforts and dedicated their time to bring this conference together; we were honored to work with them.

ICML's world-class technical program could not have been possible without the deep expertise of 76 amazing area chairs which are listed in this booklet. They worked tirelessly to shepherd the review process and find authoritative reviewers to evaluate the submitted articles. The program committee, onsisting of 597 referees which are also listed in this booklet, helped identify the very best papers and justified all decisions with detailed reviews and discussions.

We are grateful to David McAllester as General Chair who provided leadership, direction and mentoring to make the conference a success. A special thanks to Fei Sha and Percy Liang who were excellent as Publications Co-Chairs and skillfully produced the proceedings volume that is now hosted on the JMLR website. Many thanks to Kui Tang and Junming Yin for working tirelessly as Workflow Chairs and doing much of the heavy-lifting behind the scenes. We would also like to acknowledge Laurent Charlin for his help with the Toronto Paper Matching System. We are deeply indebted to Ruslan Salakhutdinov for his work as Tutorials Chair where he curated an amazing program with excellent presenters. Our deep gratitude goes to Alex Ihler, the Workshops Chair, who tirelessly brought together 18 mini-conferences taking place over a two day period. A special thank you to John Paisley for being our Volunteers Chair and recruiting student volunteers to support the conference (and, of course, thanks to the volunteers themselves for doing the leg-work). Our thanks go to Jingrui He who served as our Publicity Chair. We are grateful to Artur Dubrawski and Charles Isbell who were our Financial Chairs. Thank you as well to Priscilla Rasmussen, the IMLS Treasurer, who helped with budgeting. We are extremely grateful to William Cohen, IMLS President, who was our connection to past conferences, helped us with fund-raising, and was a tremendous resource for recruiting great colleagues to serve at ICML 2014.

Changshui Zhang, Jun Zhu, Tie-Yan Liu were all magnificent and gave so much of their time and energy as ICML Local Chairs. They really made the conference a success and took care of a bewildering number of logistics, from securing the conference center, to running the website, to raising extensive sponsor funding and much, much more. We are humbled by their efforts and simply cannot thank them enough. Our sincere thanks also go to Kai Yu and Shiqiang Yang, our Local Sponsor Chairs, they helped secure significant funding that was critical life-support for the conference. We are also indebted to the Local Organization Team: James Kwok, Qing Tao, Junping Zhang, Jian Yu, Qiuxin Hu, Liwei Wang, Jiang Bian, Jie Tang, Hong Chang, Zhiyuan Liu, Ning Chen, and Aonan Zhang. Furthermore, thanks to Tsinghua University and Microsoft Research Asia for providing institutional support.

Finally, our sincere thanks to our distinguished sponsors, ICML 2014 simply would not have been possible without their support: Baidu, TP-LINK, Alibaba, the National Science Foundation, Tencent, Bloomberg, Booking.com, Google, Huawei, Microsoft, NEXGO, TNList, 973 Program, Amazon, Criteo, Facebook, Miaozhen Systems, Yahoo! Labs, University of Rochester and Springer, Springer and Yandex.

Every one of the people above worked hard and gave so much because they believe that machine learning can only be a great field if and only if we continue to have a great machine learning community. Research and progress is bolstered by our collaborations, service and contributions. On behalf of all of us at ICML 2014, enjoy the conference and see you in Beijing!

Tony Jebara and Eric Xing
ICML 2014 Program Co-Chairs

SPONSORS

Diamond Sponsors

Jade Sponsors

Platinum Sponsors

Gold Sponsors

Silver Sponsors

Sponsors

PROGRAM HIGHLIGHTS

Saturday June 21

09:30 – 11:30am Tutorial 1 and 2 (running in parallel)

Lunch (on own)
13:00 – 15:00pm Tutorial 3 and 4 (running in parallel)

Coffee break
15:30 – 17:30pm Tutorial 4 and 5 (running in parallel)

Sunday June 22

8:30 -10:00 Welcome and Keynote by Eric Horvitz

Coffee break
10:30 – 12:30 36×20 minute talks in 6 parallel sessions

Lunch (on own)
14:00 –16:00 35×20 minute talks in 6 parallel sessions

Coffee break
16:20 – 18:20 35×20 minute talks in 6 parallel sessions
19:00 – 23:00 Poster session I.

Monday June 23

8:30 -10:00 Keynote by Michael Kearns

Coffee break
10:30 – 12:30 36×20 minute talks in 6 parallel sessions

Lunch (on own)
14:00 –16:00 36×20 minute talks in 6 parallel sessions

Coffee break
16:20 – 18: 00 31×20 minute talks in 6 parallel sessions
18:10 – Buses begin to load
18:10 – 19:30 Go to the Banquet and Museum
19:30 – 21:00 Banquet
21:00 – 21:30 Go back to BICC

Tuesday June 24

8:30 - 08:50 Awards (Hosted by Baidu)
8:50 - 10:20 Keynote by Michael Jordan (Hosted by Tencent)

Coffee break
10:50 – 12:30 30×20 minute talks in 6 parallel sessions

Lunch (on own) (IMLS Board Members Lunch Meeting)
14:00 – 16:00 35×20 minute talks in 6 parallel sessions

Coffee break
16:20 – 18:20 36×20 minute talks in 6 parallel sessions
18:20 - 19:20 IMLS Business Meeting
19:00 – 23:00 Poster session II

Wednesday June 25

9:00 -10:20 workshops

Coffee break
10:40 – 12: 00 workshops

Lunch (on own)
14:00 –15:20 workshops

Coffee break
15:40 – 17: 00 workshops

Thursday June 26

9:00 -10:20 workshops

Coffee break
10:40 – 12: 00 workshops

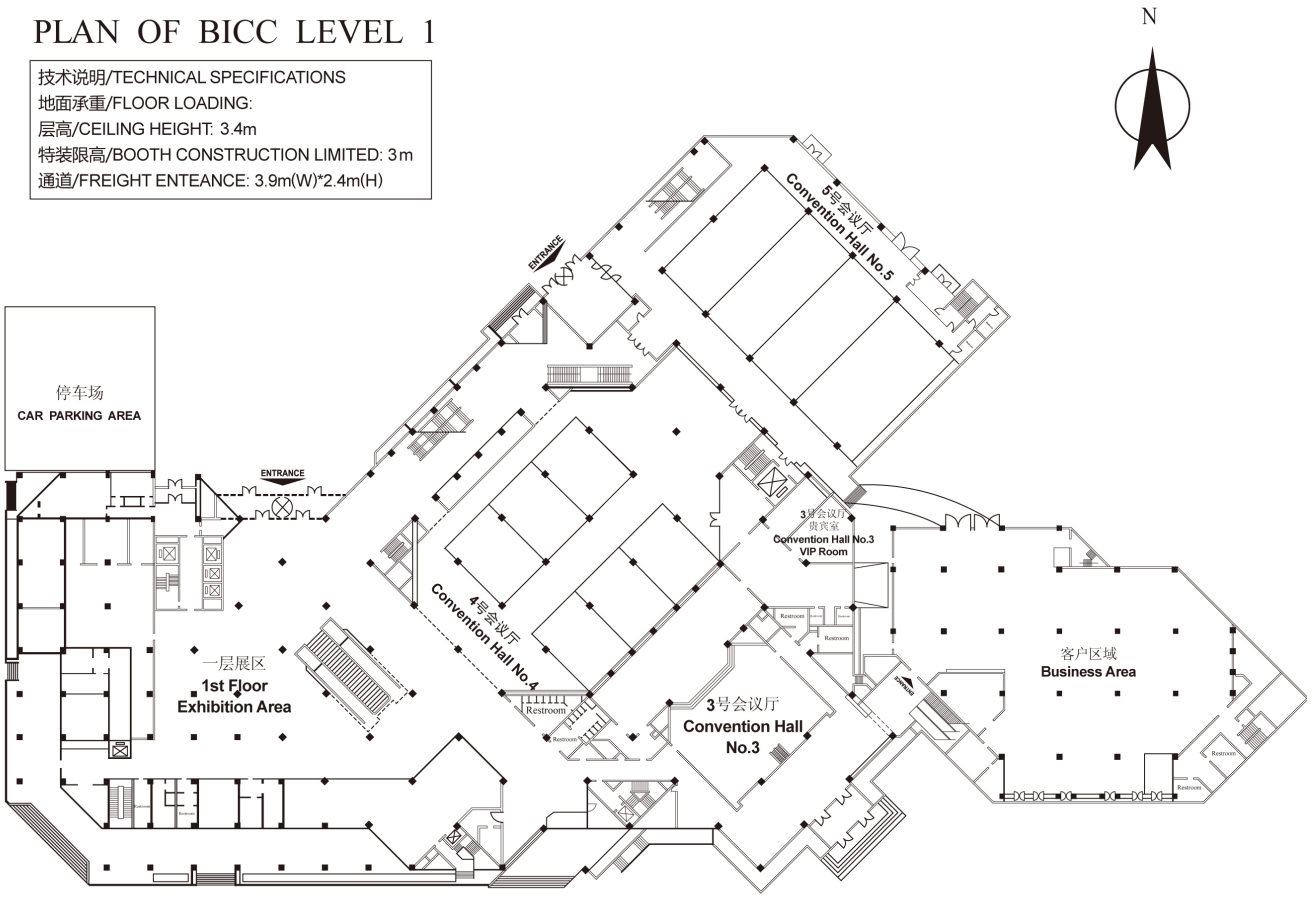
Lunch (on own)
14:00 –15:20 workshops

Coffee break
15:40 – 17: 00 workshops

BICC 一层平面图

PLAN OF BICC LEVEL 1

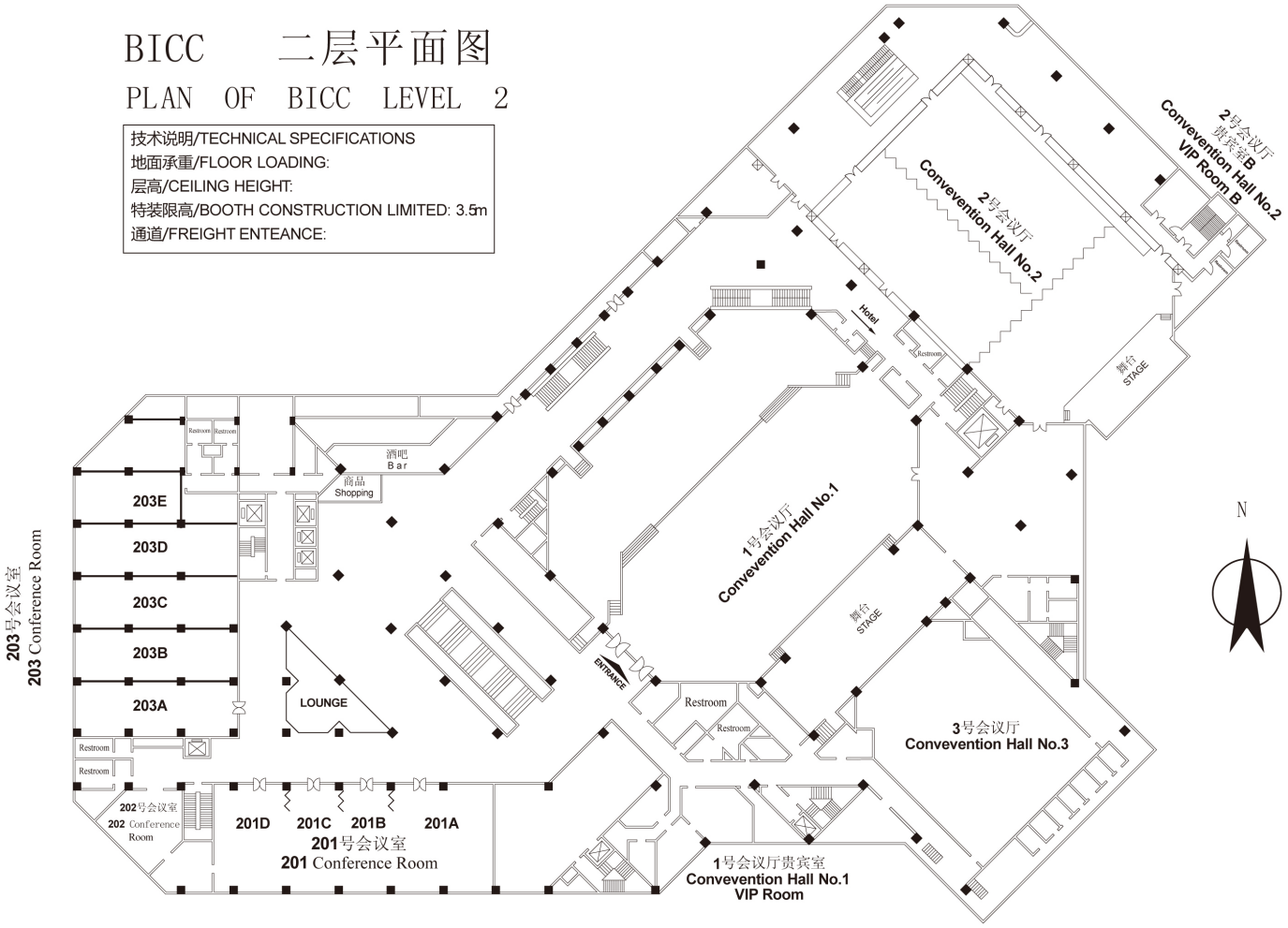
技术说明/TECHNICAL SPECIFICATIONS
地面承重/FLOOR LOADING:
层高/CEILING HEIGHT: 3.4m
特装限高/BOOTH CONSTRUCTION LIMITED: 3m
通道/FREIGHT ENTEANCE: 3.9m(W)*2.4m(H)



BICC 二层平面图

PLAN OF BICC LEVEL 2

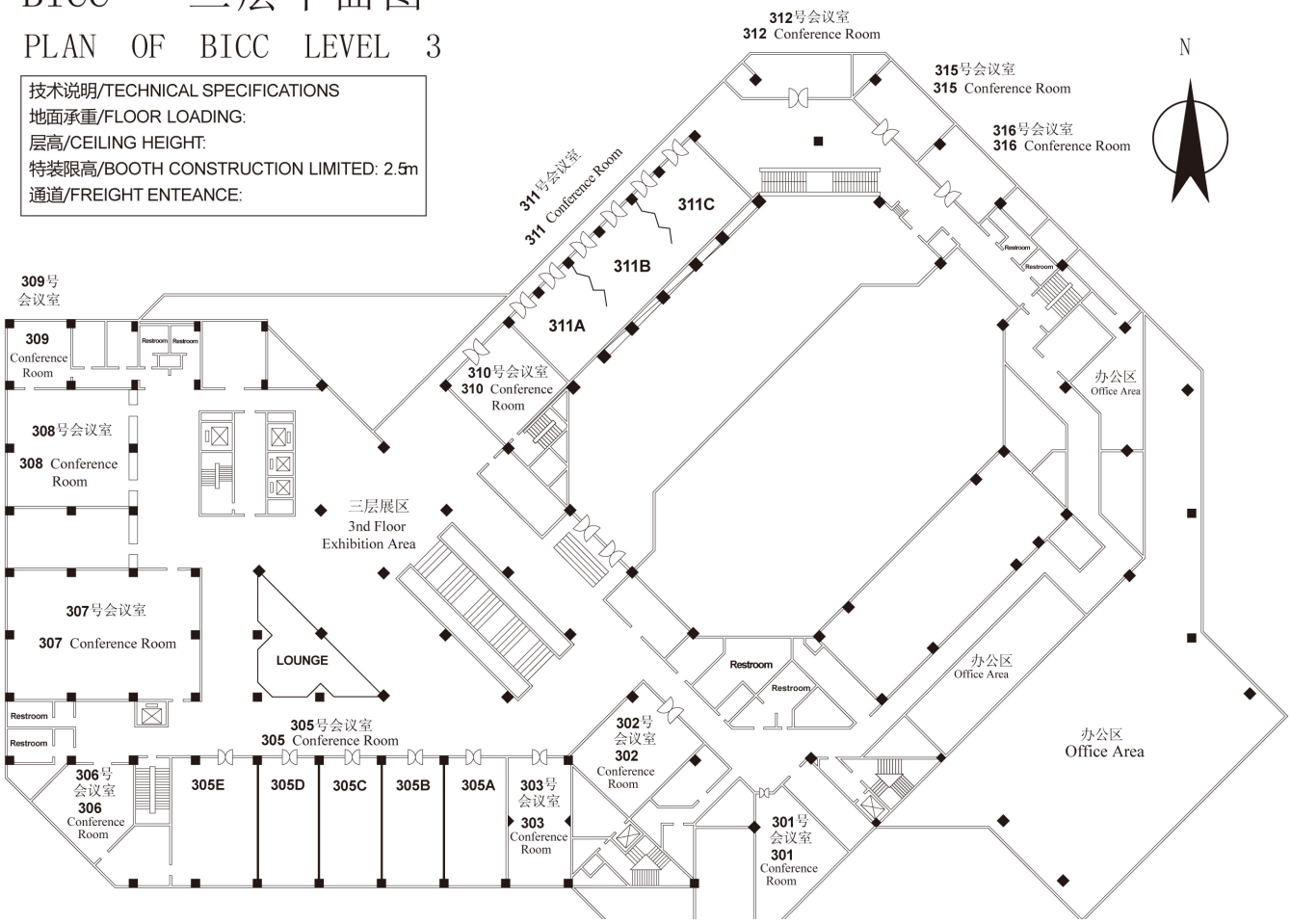
技术说明/TECHNICAL SPECIFICATIONS
地面承重/FLOOR LOADING:
层高/CEILING HEIGHT:
特装限高/BOOTH CONSTRUCTION LIMITED: 3.5m
通道/FREIGHT ENTEANCE:



BICC 三层平面图

PLAN OF BICC LEVEL 3

技术说明/TECHNICAL SPECIFICATIONS
地面承重/FLOOR LOADING:
层高/CEILING HEIGHT:
特装限高/BOOTH CONSTRUCTION LIMITED: 2.5m
通道/FREIGHT ENTEANCE:



SATURDAY

SATURDAY TUTORIALS

Saturday June 21

09:30 – 11:30am Tutorial 1 and 2 (running in parallel).

Tutorial 1: Bayesian Posterior Inference in the Big Data Arena
Room 201

Max Welling and Anoop Balan Korattikara

Tutorial 2: Frank-Wolfe and Greedy Optimization for Learning with Big Data
Room 305
Zaid Harchaoui and Martin Jaggi

Lunch (on own)

13:00 – 15:00pm Tutorial 3 and 4 (running in parallel)

Tutorial 3: Finding Structure with Randomness: Stochastic Algorithms for Numerical Linear Algebra
Room 201
Joel A. Tropp

Tutorial 4: Emerging Systems for Large-Scale Machine Learning
Room 305
Joseph Gonzalez

Coffee break

15:30 – 17:30pm Tutorial 5 and 6 (running in parallel)

Tutorial 5: An introduction to probabilistic programming
Room 201
Vikash Mansinghka and Dan Roy

Tutorial 6:Deep Learning: from Speech Analysis and Recognition to Language and Multi-modal Processing
Room 305
Li Deng

Tutorial 1: Bayesian Posterior Inference in the Big Data Arena

Max Welling, Anoop Balan Korattikara

Abstract: Traditional algorithms for Bayesian posterior inference require processing the entire dataset in each iteration and are quickly getting obsoleted by the data deluge in various application domains. Most successful applications of learning with big data have been with very simple algorithms such as Stochastic Gradient Descent, because they are the only ones that can computationally handle today's large datasets. However, by restricting ourselves to these algorithms, we miss out on all the advantages of Bayesian modeling, such as quantifying uncertainty and avoiding over-fitting. In this tutorial, we will explore recent advances in scalable Bayesian posterior inference. We will talk about a new generation of MCMC algorithms and variational methods that use only a mini-batch of data points per iteration, whether to generate an MCMC sample or update a variational parameter. We will also present applications to various real world problems and datasets.

Tutorial 2: Frank-Wolfe and Greedy Optimization for Learning with Big Data

Zaid Harchaoui, Martin Jaggi

Abstract: We provide a unified overview of several families of algorithms proposed in different settings: Frank-Wolfe aka Conditional Gradient algorithms, greedy optimization methods, and related extensions. Frank-Wolfe methods have been successfully applied to a wide range of large-scale learning and signal processing applications, such as matrix factorization, multi-task learning, image denoising, and structured prediction. On the other hand, greedy optimization algorithms, which underlie several versions of boosting, appear in structured variable selection, metric learning, and training of sum-product networks.

All these algorithms have in common that they rely on the atomic decomposition of the variable of interest, which is expanding it as a linear combination of the elements of a dictionary. In this tutorial, we showcase these algorithms in a unified framework, and present simple proofs of convergence rates and illustrate their underlying assumptions. We show how these families of algorithms relate to each other, illustrate several successful applications, and highlight current challenges.

Tutorial 3: Finding Structure with Randomness: Stochastic Algorithms for Numerical Linear Algebra

Joel A. Tropp

Abstract: Computer scientists have long known that randomness can be used to improve the performance of algorithms. A familiar application is the process of dimension reduction, in which a random map transports data from a high-dimensional space to a lower-dimensional space while approximately preserving some geometric properties. By operating with the compact representation of the data, it is possible to produce approximate solutions to certain large problems very efficiently.

Recently, it has been observed that dimension reduction has powerful applications in numerical linear algebra and numerical analysis. This tutorial will offer a high-level introduction to randomized methods for some of the core problems in this field. In particular, it will cover techniques for constructing standard matrix factorizations, such as the truncated singular value decomposition and the Nystrom approximation. In practice, the algorithms are so effective that they compete with, or even outperform, classical algorithms. These methods are likely to have significant applications in modern large-scale learning systems.

Some of the ideas in this tutorial are documented in the paper.

Tutorial 4: Emerging Systems for Large-Scale Machine Learning

Joseph Gonzalez

Abstract: The need to apply machine learning techniques to vast amounts of data and to train iterative algorithms at scale.increasingly complex models has driven the development of new systems. These systems exploit common patterns in machine learning to leverage advances in hardware and distributed computing, and separate the design of machine learning algorithms from the complexities of systems engineering. By understanding the goals, designs, and limitations of these systems we can develop more scalable machine learning algorithms and influence the direction of systems research.

In the first half of this tutorial we will survey developments in the space of emerging systems to support large-scale machine learning. We will characterize a small set of computational patterns that span a wide range of machine learning algorithms and explore how systems have evolved to support these patterns. From map-reduce to batch processing systems (e.g., Dryad and Spark), we will review the development of traditional data analytics technologies as they adapted to iterative machine learning. Driven by developments in stochastic optimization we will describe the limitations of batch processing systems and how they lead to the emergence of streaming systems (e.g., VW, Hogwild) and subsequently the parameter server.n the second half of the tutorial we will dive into a parallel line of research in graph-processing systems (e.g., Pregel, GraphLab). We will describe how these systems emerged, the space of problems they address, and the design decisions they made which enable them to efficiently execute complex We will then explore more recent developments in the fusion of graph and batch processing systems (e.g., GraphX, GraphLab Create) and how combining systems is essential to scalable machine learning.

Throughout the tutorial we will provide concrete examples demonstrating the process of designing, implementing, and even running machine learning algorithms on the various systems. Along the way we will also elude to new potential research directions and opportunities for the co-design of machine learning algorithms and systems.

Tutorial 5: An introduction to probabilistic programming

Vikash Mansinghka and Dan Roy

Abstract: Probabilistic models and approximate inference algorithms are powerful tools, central to modern artificial intelligence and widely used in fields from robotics to machine learning to statistics. However, simple variations on models and algorithms from the standard machine learning toolkit can be difficult and time-consuming to design, specify, analyze, implement, optimize and debug. Additionally, careful probabilistic treatments of complex problems can seem impractical. The emerging probabilistic programming community aims to address these challenges by developing formal languages and software systems that integrate key ideas from probabilistic modeling and inference with programming languages and Turing-universal computation. This tutorial will provide an introduction to the field, including a survey of languages, current system capabilities/limitations, mathematical foundations, and current research directions.

Probabilistic programming principles will be illustrated via live demonstrations and real-world examples written in Venture, a general-purpose probabilistic programming platform, as well as via languages such as Stan, Church, Figaro, Markov Logic and BLOG.

Tutorial 6: Deep Learning: from Speech Analysis and Recognition to Language and Multi-modal Processing

Li Deng

Abstract: While artificial neural networks have been around for more than half a century and had drawn attentions from speech researchers from time to time, it was not until around 2010-2011 that real impact on speech feature extraction and recognition has been made

by novel machine learning methods. The first part of this tutorial will reflect on the path to this transformative success after providing sufficient background material on speech signal processing and speech pattern recognition for the non-speech machine learning audience. Some historical development in speech recognition technology will be discussed that is relevant to introducing deep neural networks to the speech community around 2009-2010. Roles of well-timed academic-industrial collaboration in deep learning will be highlighted that helped shape the entire speech recognition industry in following years. This tutorial will also review the recent history of how the insights derived jointly from industrial needs for speech technology and from understandings of both the capabilities and limitations of deep neural networks rapidly pushed deep learning into industry-wide deployment. Subsequent research on overcoming such limitations will then be examined.The second part of the tutorial will give an overview of the sweeping achievements of deep learning in speech recognition since 2010, attributing to a number of additional enabling factors. Several key innovations in recent years will be analyzed that have further advanced the state of the art beyond the earlier successes based on the basic architecture and learning methods for deep neural networks. These advances have resulted in across-the-board deployment of deep learning for both research and industrial speech recognition systems, where the deep learning approaches are shown to scale beautifully with big data. Parallels will be drawn and comparisons be made with the no-less-striking impact of deep learning in image recognition and computer vision with the initial success reported in 2012.

The third part of the tutorial will look ahead towards new application challenges of deep learning --- creating systems capable of not only hearing (speech) and seeing (vision), but also thinking and understanding with a "mind"; i.e., reasoning and inference over complex relationships and knowledge sources expressed typically in natural language encompassing a vast number of entities and semantic concepts in the real world.

To this end, researchers are making progress in language and multimodal (jointly text, speech/audio, and image/video) processing, which is in the process of evolving into a new frontier of deep learning. This tutorial will provide a review of recent published studies on the applications of deep learning in this exciting area, emphasizing how discrete symbolic macro-structure in linguistic and cognitive systems can be implemented by deep and recursive neural micro-structure and by continuous distributed representations via semantic symbol embeddings which are operating in a Hilbert space. Supervised and unsupervised learning methods designed in this space will be discussed, with selected applications elaborated.

SUNDAY

June 22, 2014 at 8:30am

Title: People, Decisions, and Cognition: On Deeper Engagements with Machine Learning
Convention Hall No.1

Keynote Speaker, Eric Horvitz, Microsoft

Abstract:
I will share reflections on promising directions for engaging human intellect and effort at multiple touchpoints in machine learning, including mechanisms for interaction, supervision, and decision support. I will frame opportunities with studies on harnessing the complementary intellect of people and machines, guiding human effort for supervision, interactively refining learning and inference, and generating explanations and visualizations. I will conclude with thoughts on aiming machine learning at human cognition,

with eye on applications and services that leverage inferences about cognitive affordances and biases.

Bio:
Eric Horvitz is a distinguished scientist and managing director at the Microsoft Research Lab at Redmond, Washington. His interests span theoretical and practical challenges with machine learning, inference, and decision making. He has been elected a fellow of AAAI, AAAS, the American Academy of Arts and Sciences, and the National Academy of Engineering, and has been inducted into the CHI Academy. He has served as president of the AAAI, chair of the AAAS Section on Information, Computing, and Communications, and on the NSF CISE Advisory Committee. Information on publications and activities can be found at http://research.microsoft.com/~horvitz.

Sunday June 22,

10:30 - Track A - Networks and Graph Based Learning I (Room 305A)

S1 Joint Inference of Multiple Label Types in Large Networks
Deepayan Chakrabarti; Stanislav Funiak; Jonathan Chang; Sofus Macskassy

S2 Learning the Consistent Behavior of Common Users for Target Node Prediction across Social Networks
Shan-Hung Wu; Hao-Heng Chien; Kuan-Hua Lin; Philip Yu

S3 Learning Modular Structures from Network Data and Node Variables
Elham Azizi; Edoardo Airoldi; James Galagan

S4 Weighted Graph Clustering with Non-Uniform Uncertainties
Yudong Chen; Shiau Hong Lim; Huan Xu

S5 Efficient Dimensionality Reduction for High-Dimensional Network Estimation
Safiye Celik; Benjamin Logsdon; Su-In Lee

S6 Discovering Latent Network Structure in Point Process Data
Scott Linderman; Ryan Adams

Sunday June 22,

10:30 - Track B - Reinforcement Learning I (Room 201A)

S7 PAC-inspired Option Discovery in Lifelong Reinforcement Learning
Emma Brunskill; Lihong Li

S8Time-Regularized Interrupting Options (TRIO)
Daniel Mankowitz; Timothy Mann; Shie Mannor

S9 Approximate Policy Iteration Schemes: A Comparison
Bruno Scherrer

S10 Model-Based Relational RL When Object Existence is Partially Observable
Vien Ngo; marc Toussaint

S11 GeNGA: A Generalization of Natural Gradient Ascent with Positive and Negative Convergence Results
Philip Thomas

S12 Scaling Up Robust MDPs using Function Approximation
Aviv Tamar; Shie Mannor; Huan Xu

Sunday June 22,

10:30 - Track C - Bayesian Optimization and Gaussian Processes (Room 201B)

S13 Agnostic Bayesian Learning of Ensembles
Alexandre Lacoste; Mario Marchand; Fran ois Laviolette; Hugo Larochelle

S14 Robust RegBayes: Selectively Incorporating First-Order Logic Domain Knowledge into Bayesian Models
Shike Mei; Jun Zhu; Jerry Zhu

S15 An Efficient Approach for Assessing Hyperparameter Importance
Frank Hutter; Holger Hoos; Kevin Leyton-Brown

S16 Bayesian Optimization with Inequality Constraints
Jacob Gardner; Matt Kusner; Zhixiang; Xu; Kilian Weinberger; John Cunningham

S17 A PAC-Bayesian bound for Lifelong Learning
Anastasia Pentina; Christoph Lampert

S18 Gaussian Processes for Bayesian Estimation in Ordinary Differential Equations
David Barber; Yali Wang

Sunday June 22,

10:30 - Track D - Pca and Subspace Models (Room 307)

S19 Robust Principal Component Analysis with Complex Noise
Qian Zhao; Deyu Meng; Zongben Xu; Wangmeng Zuo; Lei Zhang

S20 Multivariate Maximal Correlation Analysis
Hoang Vu Nguyen; Emmanuel Müller; Jilles Vreeken; Pavel Efros; Klemens Böhm

S21 Discriminative Features via Generalized Eigenvectors
Nikos Karampatziakis; Paul Mineiro

S22 Randomized Nonlinear Component Analysis
David Lopez-Paz; Suvrit Sra; Alex Smola; Zoubin Ghahramani; Bernhard Schoelkopf

S23 Memory and Computation Efficient PCA via Very Sparse Random Projections
Farhad Pourkamali Anaraki; Shannon Hughes

S24 Optimal Mean Robust Principal Component Analysis
Feiping Nie; Jianjun Yuan; Heng Huang

Sunday June 22,

10:30 - Track E - Supervised Learning (Room 201C)

S25 The Coherent Loss Function for Classification
Wenzhuo Yang; Melvyn Sim; Huan Xu

S26 Condensed Filter Tree for Cost-Sensitive Multi-Label Classification
Chun-Liang Li; Hsuan-Tien Lin

S27 Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification
Junfeng Wen; Chun-Nam Yu; Russell Greiner

S28 A Statistical Convergence Perspective of Algorithms for Rank Aggregation from Pairwise Data
Arun Rajkumar; Shivani Agarwa

S29 GEV-Canonical Regression for Accurate Binary Class Probability Estimation when One Class is Rare
Arpit Agarwal; Harikrishna Narasimhan; Shivaram Kalyanakrishnan; Shivani Agarwal

S30 Guess-Averse Loss Functions For Cost-Sensitive Multiclass Boosting
Oscar Beijbom; Mohammad Saberian; David Kriegman; Nuno Vasconcelos

Sunday June 22,

10:30 - Track F - Neural Networks and Deep Learning I (Room 305B)

S31 Structured Recurrent Temporal Restricted Boltzmann Machines Alexandre Roni Mittelman; Benjamin Kuipers; Silvio Savarese; Honglak Lee

S32 A Deep and Tractable Density Estimator
Benigno Uria; Iain Murray; Hugo Larochelle

S33 Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction
Jian Zhou; Olga Troyanskaya

S34 Deep AutoRegressive Networks
Karol Gregor; Ivo Danihelka; Andriy Mnih; Charles Blundell; Daan Wierstra

S35 Stochastic Backpropagation and Approximate Inference in Deep Generative Models
Danilo Jimenez Rezende; Shakir Mohamed; Daan Wierstra

S36 Neural Variational Inference and Learning in Belief Networks
Andriy Mnih; Karol Gregor

Sunday June 22,

14:00 - Track A - Graphical Models I (Room 305A)

S37 Linear and Parallel Learning of Markov Random Fields
Yariv Mizrahi; Misha Denil; Nando De Freitas

S38 Putting MRFs on a Tensor Train
Alexander Novikov; Anton Rodomanov; Anton Osokin; Dmitry Vetrov

S39 Gaussian Approximation of Collective Graphical Models
Liping Liu; Daniel Sheldon; Thomas Dietterich

S40 Scalable Semidefinite Relaxation for Maximum A Posterior Estimation
Qixing Huang; Yuxin Chen; Leonidas Guibas

S41 Globally Convergent Parallel MAP LP Relaxation Solver using the Frank-Wolfe Algorithm
Alexander Schwing; Tamir Hazan; Marc Pollefeys; Raquel Urtasun

S42 Inferning with High Girth Graphical Models
Uri Heinemann; Amir Globerson

Sunday June 22,

14:00 - Track B - Bandits I (Room 201A)

S43 Thompson Sampling for Complex Online Problems
Aditya Gopalan; Shie Mannor; Yishay Mansour

S44 Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms
Richard Combes; Alexandre Proutiere

S45 Reducing Dueling Bandits to Cardinal Bandits
Nir Ailon; Zohar Karnin; Thorsten Joachims

S46 Combinatorial Partial Monitoring Game with Linear Feedback and Its Applications
Tian Lin; Bruno Abrahao; Robert Kleinberg; John Lui; Wei Chen

S47 Online Stochastic Optimization under Correlated Bandit Feedback
Mohammad Gheshlaghi Azar; Alessandro Lazaric; Emma Brunskill

S48 Adaptive Monte Carlo via Bandit Allocation
James Neufeld; Andras Gyorgy; Csaba Szepesvari; Dale Schuurmans

Sunday June 22,

14:00 - Track C - Monte Carlo (Room 307)

S49 Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget
Anoop Korattikara; Yutian Chen; Max Welling

S50Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach
Rémi Bardenet; Arnaud Doucet; Chris Holmes

S51 Distributed Stochastic Gradient MCMC
Sungjin Ahn; Babak Shahbaba; Max Welling

S52 Kernel Adaptive Metropolis-Hastings
Dino Sejdinovic; Heiko Strathmann; Maria Lomeli Garcia; Christophe Andrieu; Arthur Gretton

S53 Stochastic Gradient Hamiltonian Monte Carlo
Tianqi Chen; Emily Fox; Carlos Guestrin

S54 A Compilation Target for Probabilistic Programming Languages
Brooks Paige; Frank Wood

Sunday June 22,

14:00 - Track D - Statistical Methods (Room 201B)

S55 Generalized Exponential Concentration Inequality for Renyi Divergence Estimation
Shashank Singh; Barnabas Poczos

S56 Consistency of Causal Inference under the Additive Noise Model
Samory Kpotufe; Eleni Sgouritsa; Dominik Janzing; Bernhard Schoelkopf

S57 The Falling Factorial Basis and Its Statistical Applications
Yu-Xiang Wang; Alex Smola; Ryan Tibshirani

S58 Concept Drift Detection Through Resampling
Maayan Harel; Shie Mannor; Ran El-Yaniv; Koby Crammer

S59 A Bayesian Wilcoxon signed-rank test based on the Dirichlet process
Alessio Benavoli; Giorgio Corani; Francesca Mangili; Marco Zaffalon; Fabrizio Ruggeri

14:00 - Track E - Structured Prediction (Room 201C)

S60 Marginal Structured SVM with Hidden Variables
Wei Ping; Qiang Liu; Alex Ihler

S61 Scalable Gaussian Process Structured Prediction for Grid Factor Graph Applications
Sebastien Bratieres; Novi Quadrianto; Sebastian Nowozin; Zoubin Ghahramani

S62 High Order Regularization for Semi-Supervised Learning of Structured Output Problems
Yujia Li; Rich Zemel

S63 Spectral Regularization for Max-Margin Sequence Tagging
Ariadna Quattoni; Borja Balle; Xavier Carreras; Amir Globerson

S64 On Robustness and Regularization of Structural Support Vector Machines
Mohamad Ali Torkamani; Daniel Lowd

S65 Structured Prediction of Network Response
Hongyu Su; Aristides Gionis; Juho Rousu

14:00 - Track F - Deep Learning and Vision (Room 305B)

S66 Recurrent Convolutional Neural Networks for Scene Labeling
Pedro Pinheiro; Ronan Collobert

S67 Latent Semantic Representation Learning for Scene Classificatio
Xin Li; Yuhong Guo

S68 DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition
Jeff Donahue; Yangqing Jia; Oriol Vinyals; Judy Hoffman; Ning Zhang; Eric Tzeng; Trevor Darrell

S69 Hierarchical Conditional Random Fields for Outlier Detection: An Application to Detecting Epileptogenic Cortical Malformations
Bilal Ahmed; Thomas Thesen; Karen Blackmon; Yijun Zhao; Orrin Devinsky; Ruben Kuzniecky; Carla Brodley

S70 Stable and Efficient Representation Learning with Nonnegativity Constraints
Tsung-Han Lin; H. T. Kung

S71 Learning by Stretching Deep Networks
Gaurav Pandey; Ambedkar Dukkipati

16:20 - Track A - Matrix Completion and Graphs (Room 305A)

S72 Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery
Cun Mu; Bo Huang; John Wright; Donald Goldfarb

S73 Near-Optimal Joint Object Matching via Convex Relaxation
Yuxin Chen; Leonidas Guibas; Qixing Huang

S74 Coherent Matrix Completion
Yudong Chen; Srinadh Bhojanapalli; Sujay Sanghavi; Rachel Ward

S75 Universal Matrix Completion
Srinadh Bhojanapalli; Prateek Jain

S76 Exponential Family Matrix Completion under Structural Constraints
Suriya Gunasekar; Pradeep Ravikumar; Joydeep Ghosh

S77 A Consistent Histogram Estimator for Exchangeable Graph Models
Stanley Chan; Edoardo Airoldi

16:20 - Track B - Learning Theory I (Room 201A)

S78 Concentration in unbounded metric spaces and algorithmic stability
Aryeh Kontorovich

S79 Heavy-tailed regression with a generalized median-of-means
Daniel Hsu; Sivan Sabato

S80 Learnability of the Superset Label Learning Problem
Liping Liu; Thomas Dietterich

S81 Maximum Margin Multiclass Nearest Neighbors
Aryeh Kontorovich; Roi Weiss

S82 Sample Efficient Reinforcement Learning with Gaussian Processes
Robert Grande; Thomas Walsh; Jonathan How

S83 Scaling Up Approximate Value Iteration with Options: Better Policies with Fewer Iterations
Timothy Mann; Shie Mannor

16:20 - Track C - Clustering and Nonparametrics (Room 307)

S84 Von Mises-Fisher Clustering Models
Siddharth Gopal; Yiming Yang

S85 Online Bayesian Passive-Aggressive Learning
Tianlin Shi; Jun Zhu

S86 Bayesian Nonparametric Multilevel Clustering with Group-Level Contexts
Tien Vu Nguyen; Dinh Phung; Xuanlong Nguyen; Swetha Venkatesh; Hung Bui

S87 Hierarchical Dirichlet Scaling Process
Dongwoo Kim; Alice Oh

S88 Fast Computation of Wasserstein Barycenters
Marco Cuturi; Arnaud Doucet

S89 Max-Margin Infinite Hidden Markov Models
Aonan Zhang; Jun Zhu; Bo Zhang

16:20 - Track D - Active Learning (Room 201B)

S90 Diagnosis determination: decision trees optimizing simultaneously worst and expected testing cost
Ferdinando Cicalese; Eduardo Laber; Aline Medeiros Saettler

S91 Nonmyopic ϵ-Bayes-Optimal Active Learning of Gaussian Processes
Trong Nghia Hoang; Bryan Kian Hsiang Low; Patrick Jaillet; Mohan Kankanhalli

S92 Hard-Margin Active Linear Regression
Zohar Karnin; Elad Hazan

S93 Active Transfer Learning under Model Shift
Xuezhi Wang; Tzu-Kuo Huang; Jeff Schneider

S94 Gaussian Process Optimization with Mutual Information
Emile Contal; Vianney Perchet; Nicolas Vayatis

Sunday June 22,

16:20 - Track E - Optimization I (Room 201C)

S95 An Adaptive Accelerated Proximal Gradient Method and its Homotopy Continuation for Sparse Optimization
Qihang Lin; Lin Xiao

S96 Finito: A faster, permutable incremental gradient method for big data problems
Aaron Defazio; Justin Domke; tiberio Caetano

S97 Asynchronous Distributed ADMM for Consensus Optimization
Ruiliang Zhang; James Kwok

S98 Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods
Jascha Sohl-Dickstein; Ben Poole; Surya Ganguli

S100 Least Squares Revisited: Scalable Approaches for Multi-class Prediction
Alekh Agarwal; Sham Kakade; Nikos Karampatziakis; Le Song; Gregory Valiant

Sunday June 22,

16:20 - Track F - Large-Scale Learning (Room 305B)

S101 Large-scale Multi-label Learning with Missing Labels
Hsiang-Fu Yu; Prateek Jain; Purushottam Kar; Inderjit Dhillon

S102 Dual Query: Practical Private Query Release for High Dimensional Data
Marco Gaboardi; Emilio Jesus Gallego Arias; Justin Hsu; Aaron Roth; Zhiwei Steven Wu

S103 A Highly Scalable Parallel Algorithm for Isotropic Total Variation Models
Jie Wang; Qingyang Li; Sen Yang; Wei Fan; Peter Wonka; Jieping Ye

S104 Buffer k-d Trees: Processing Massive Nearest Neighbor Queries on GPUs
Fabian Gieseke; Justin Heinermann; Cosmin Oancea; Christian Igel

S105 Fast Multi-stage Submodular Maximization
Kai Wei; Rishabh Iyer; Jeff Bilmes

S106 Multi-label Classification via Feature-aware Implicit Label Space Encoding
Zijia Lin; Guiguang Ding; Mingqing Hu; Jianmin Wang

S1	Joint Inference of Multiple Label Types in Large Networks	Papayan Svarabhakti; Stanislav Funiak; Jonathan Chang; Sofus Macskassy
S2	Learning the Consistent Behavior of Common Users for Target Node Prediction across Social Networks	Shan-Hung Wu; Hao-Heng Chien; Kuan-Hua Lin; Philip Yu
S3	Learning Modular Structures from Network Data and Node Variables	Elham Azizi; Edoardo Airoldi; James Galagan
S4	Weighted Graph Clustering with Non-Uniform Uncertainties	Yudong Chen; Shiau Hong Lim; Huan Xu
S5	Efficient Dimensionality Reduction for High-Dimensional Network Estimation	Safiye Celik; Benjamin Logsdon; Su-In Lee
S6	Discovering Latent Network Structure in Point Process Data	Scott Linderman; Ryan Adams
S7	PAC-inspired Option Discovery in Lifelong Reinforcement Learning	Emma Brunskill; Lihong Li
S8	Time-Regularized Interrupting Options (TRIO)	Daniel Mankowitz; Timothy Mann; Shie Mannor
S9	Approximate Policy Iteration Schemes: A Comparison	Bruno Scherrer
S10	Model-Based Relational RL When Object Existence is Partially Observable	Vien Ngo; Marc Toussaint
S11	GeNGA: A Generalization of Natural Gradient Ascent with Positive and Negative Convergence Results	Philip Thomas
S12	Scaling Up Robust MDPs using Function Approximation	Aviv Tamar; Shie Mannor; Huan Xu
S13	Agnostic Bayesian Learning of Ensembles	Alexandre Lacoste; Mario Marchand; Francois Laviolette; Hugo Larochelle
S14	Robust RegBayes: Selectively Incorporating First-Order Logic Domain Knowledge into Bayesian Models	Shike Mei; Jun Zhu; Jerry Zhu
S15	An Efficient Approach for Assessing Hyperparameter Importance	Frank Hutter; Holger Hoos; Kevin Leyton-Brown
S16	Bayesian Optimization with Inequality Constraints	Jacob Gardner; Matt Kusner; Zhixiang; Xu; Kilian Weinberger; John Cunningham
S17	A PAC-Bayesian bound for Lifelong Learning	Anastasia Pentina; Christoph Lampert
S18	Gaussian Processes for Bayesian Estimation in Ordinary Differential Equations	David Barber; Yali Wang
S19	Robust Principal Component Analysis with Complex Noise	Qian Zhao; Deyu Meng; Zongben Xu; Wangmeng Zuo; Lei Zhang
S20	Multivariate Maximal Correlation Analysis	Hoang Vu Nguyen; Emmanuel Müller; Jilles Vreeken; Pavel Efros; Klemens Böhm
S21	Discriminative Features via Generalized	Nikos Karampatziakis; Paul Mineiro

	Eigenvectors	
S22	Randomized Nonlinear Component Analysis	David Lopez-Paz; Suvrit Sra; Alex Smola; Zoubin Ghahramani; Bernhard Schoelkopf
S23	Memory and Computation Efficient PCA via Very Sparse Random Projections	Farhad Pourkamali Anaraki; Shannon Hughes
S24	Optimal Mean Robust Principal Component Analysis	Feiping Nie; Jianjun Yuan; Heng Huang
S25	The Coherent Loss Function for Classification	Wenzhuo Yang; Melvyn Sim; Huan Xu
S26	Condensed Filter Tree for Cost-Sensitive Multi-Label Classification	Chun-Liang Li; Hsuan-Tien Lin
S27	Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification	Junfeng Wen; Chun-Nam Yu; Russell Greiner
S28	A Statistical Convergence Perspective of Algorithms for Rank Aggregation from Pairwise Data	Arun Rajkumar; Shivani Agarwal
S29	GEV-Canonical Regression for Accurate Binary Class Probability Estimation when One Class is Rare	Arpit Agarwal; Harikrishna Narasimhan; Shivaram Kalyanakrishnan; Shivani Agarwal
S30	Guess-Averse Loss Functions For Cost-Sensitive Multiclass Boosting	Oscar Beijbom; Mohammad Saberian; David Kriegman; Nuno Vasconcelos
S31	Structured Recurrent Temporal Restricted Boltzmann Machines	Roni Mittelman; Benjamin Kuipers; Silvio Savarese; Honglak Lee
S32	A Deep and Tractable Density Estimator	Benigno Uria; Iain Murray; Hugo Larochelle
S33	Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction	Jian Zhou; Olga Troyanskaya
S34	Deep AutoRegressive Networks	Karol Gregor; Ivo Danihelka; Andriy Mnih; Charles Blundell; Daan Wierstra
S35	Stochastic Backpropagation and Approximate Inference in Deep Generative Models	Danilo Jimenez Rezende; Shakir Mohamed; Daan Wierstra
S36	Neural Variational Inference and Learning in Belief Networks	Andriy Mnih; Karol Gregor
S37	Linear and Parallel Learning of Markov Random Fields	Yariv Mizrahi; Misha Denil; Nando De Freitas
S38	Putting MRFs on a Tensor Train	Alexander Novikov; Anton Rodomanov; Anton Osokin; Dmitry Vetrov
S39	Gaussian Approximation of Collective Graphical Models	Liping Liu; Daniel Sheldon; Thomas Dietterich
S40	Scalable Semidefinite Relaxation for Maximum A Posterior Estimation	Qixing Huang; Yuxin Chen; Leonidas Guibas
S41	Globally Convergent Parallel MAP LP Relaxation Solver using the Frank-Wolfe Algorithm	Alexander Schwing; Tamir Hazan; Marc Pollefeys; Raquel Urtasun
S42	Inferning with High Girth Graphical Models	Uri Heinemann; Amir Globerson
S43	Thompson Sampling for Complex Online Problems	Aditya Gopalan; Shie Mannor; Yishay Mansour

S44	Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms	Richard Combes; Alexandre Proutiere
S45	Reducing Dueling Bandits to Cardinal Bandits	Nir Ailon; Zohar Karnin; Thorsten Joachims
S46	Combinatorial Partial Monitoring Game with Linear Feedback and Its Applications	Tian Lin; Bruno Abrahao; Robert Kleinberg; John Lui; Wei Chen
S47	Online Stochastic Optimization under Correlated Bandit Feedback	Mohammad Gheshlaghi Azar; Alessandro Lazaric; Emma Brunskill
S48	Adaptive Monte Carlo via Bandit Allocation	James Neufeld; Andras Gyorgy; Csaba Szepesvari; Dale Schuurmans
S49	Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget	Anoop Korattikara; Yutian Chen; Max Welling
S50	Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach	Rémi Bardenet; Arnaud Doucet; Chris Holmes
S51	Distributed Stochastic Gradient MCMC	Sungjin Ahn; Babak Shahbaba; Max Welling
S52	Kernel Adaptive Metropolis-Hastings	Dino Sejdinovic; Heiko Strathmann; Maria Lomeli Garcia; Christophe Andrieu; Arthur Gretton
S53	Stochastic Gradient Hamiltonian Monte Carlo	Tianqi Chen; Emily Fox; Carlos Guestrin
S54	A Compilation Target for Probabilistic Programming Languages	Brooks Paige; Frank Wood
S55	Generalized Exponential Concentration Inequality for Renyi Divergence Estimation	Shashank Singh; Barnabas Poczos
S56	Consistency of Causal Inference under the Additive Noise Model	Samory Kpotufe; Eleni Sgouritsa; Dominik Janzing; Bernhard Schoelkopf
S57	The Falling Factorial Basis and Its Statistical Applications	Yu-Xiang Wang; Alex Smola; Ryan Tibshirani
S58	Concept Drift Detection Through Resampling	Maayan Harel; Shie Mannor; Ran El-Yaniv; Koby Crammer
S59	A Bayesian Wilcoxon signed-rank test based on the Dirichlet process	Alessio Benavoli; Giorgio Corani; Francesca Mangili; Marco Zaffalon; Fabrizio Ruggeri
S60	Marginal Structured SVM with Hidden Variables	Wei Ping; Qiang Liu; Alex Ihler
S61	Scalable Gaussian Process Structured Prediction for Grid Factor Graph Applications	Sebastien Bratieres; Novi Quadrianto; Sebastian Nowozin; Zoubin Ghahramani
S62	High Order Regularization for Semi-Supervised Learning of Structured Output Problems	Yujia Li; Rich Zemel
S63	Spectral Regularization for Max-Margin Sequence Tagging	Ariadna Quattoni; Borja Balle; Xavier Carreras; Amir Globerson
S64	On Robustness and Regularization of Structural Support Vector Machines	Mohamad Ali Torkamani; Daniel Lowd
S65	Structured Prediction of Network Response	Hongyu Su; Aristides Gionis; Juho Rousu
S66	Recurrent Convolutional Neural Networks for Scene Labeling	Pedro Pinheiro; Ronan Collobert
S67	Latent Semantic Representation Learning for Scene Classification	Xin Li; Yuhong Guo
S68	DeCAF: A Deep Convolutional Activation Feature	Jeff Donahue; Yangqing Jia; Oriol Vinyals;

	for Generic Visual Recognition	Judy Hoffman; Ning Zhang; Eric Tzeng; Trevor Darrell
S69	Hierarchical Conditional Random Fields for Outlier Detection: An Application to Detecting Epileptogenic Cortical Malformations	Bilal Ahmed; Thomas Thesen; Karen Blackmon; Yijun Zhao; Orrin Devinsky; Ruben Kuzniecky; Carla Brodley
S70	Stable and Efficient Representation Learning with Nonnegativity Constraints	Tsung-Han Lin; H. T. Kung
S71	Learning by Stretching Deep Networks	Gaurav Pandey; Ambedkar Dukkipati
S72	Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery	Cun Mu; Bo Huang; John Wright; Donald Goldfarb
S73	Near-Optimal Joint Object Matching via Convex Relaxation	Yuxin Chen; Leonidas Guibas; Qixing Huang
S74	Coherent Matrix Completion	Yudong Chen; Srinadh Bhojanapalli; Sujay Sanghavi; Rachel Ward
S75	Universal Matrix Completion	Srinadh Bhojanapalli; Prateek Jain
S76	Exponential Family Matrix Completion under Structural Constraints	Suriya Gunasekar; Pradeep Ravikumar; Joydeep Ghosh
S77	A Consistent Histogram Estimator for Exchangeable Graph Models	Stanley Chan; Edoardo Airoldi
S78	Concentration in unbounded metric spaces and algorithmic stability	Aryeh Kontorovich
S79	Heavy-tailed regression with a generalized median-of-means	Daniel Hsu; Sivan Sabato
S80	Learnability of the Superset Label Learning Problem	Liping Liu; Thomas Dietterich
S81	Maximum Margin Multiclass Nearest Neighbors	Aryeh Kontorovich; Roi Weiss
S82	Sample Efficient Reinforcement Learning with Gaussian Processes	Robert Grande; Thomas Walsh; Jonathan How
S83	Scaling Up Approximate Value Iteration with Options: Better Policies with Fewer Iterations	Timothy Mann; Shie Mannor
S84	Von Mises-Fisher Clustering Models	Siddharth Gopal; Yiming Yang
S85	Online Bayesian Passive-Aggressive Learning	Tianlin Shi; Jun Zhu
S86	Bayesian Nonparametric Multilevel Clustering with Group-Level Contexts	Tien Vu Nguyen; Dinh Phung; Xuanlong Nguyen; Swetha Venkatesh; Hung Bui
S87	Hierarchical Dirichlet Scaling Process	Dongwoo Kim; Alice Oh
S88	Fast Computation of Wasserstein Barycenters	Marco Cuturi; Arnaud Doucet
S89	Max-Margin Infinite Hidden Markov Models	Aonan Zhang; Jun Zhu; Bo Zhang
S90	Diagnosis determination: decision trees optimizing simultaneously worst and expected testing cost	Ferdinando Cicalese; Eduardo Laber; Aline Medeiros Saettler
S91	Nonmyopic ϵ-Bayes-Optimal Active Learning of Gaussian Processes	Trong Nghia Hoang; Bryan Kian Hsiang Low; Patrick Jaillet; Mohan Kankanhalli
S92	Hard-Margin Active Linear Regression	Zohar Karnin; Elad Hazan
S93	Active Transfer Learning under Model Shift	Xuezhi Wang; Tzu-Kuo Huang; Jeff Schneider
S94	Gaussian Process Optimization with Mutual Information	Emile Contal; Vianney Perchet; Nicolas Vayatis

S95	An Adaptive Accelerated Proximal Gradient Method and its Homotopy Continuation for Sparse Optimization	Qihang Lin; Lin Xiao
S96	Finito: A faster, permutable incremental gradient method for big data problems	Aaron Defazio; Justin Domke; Tiberio Caetano
S97	Asynchronous Distributed ADMM for Consensus Optimization	Ruiliang Zhang; James Kwok
S98	Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods	Jascha Sohl-Dickstein; Ben Poole; Surya Ganguli
S99	Least Squares Revisited: Scalable Approaches for Multi-class Prediction	Alekh Agarwal; Sham Kakade; Nikos Karampatziakis; Le Song; Gregory Valiant
S100	A Statistical Perspective on Algorithmic Leveraging	Ping Ma; Michael Mahoney; Bin Yu
S101	Large-scale Multi-label Learning with Missing Labels	Hsiang-Fu Yu; Prateek Jain; Purushottam Kar; Inderjit Dhillon
S102	Dual Query: Practical Private Query Release for High Dimensional Data	Marco Gaboardi; Emilio Jesus Gallego Arias; Justin Hsu; Aaron Roth; Zhiwei Steven Wu
S103	A Highly Scalable Parallel Algorithm for Isotropic Total Variation Models	Jie Wang; Qingyang Li; Sen Yang; Wei Fan; Peter Wonka; Jieping Ye
S104	Buffer k-d Trees: Processing Massive Nearest Neighbor Queries on GPUs	Fabian Gieseke; Justin Heinermann; Cosmin Oancea; Christian Igel
S105	Fast Multi-stage Submodular Maximization	Kai Wei; Rishabh Iyer; Jeff Bilmes
S106	Multi-label Classification via Feature-aware Implicit Label Space Encoding	Zijia Lin; Guiguang Ding; Mingqing Hu; Jianmin Wang
M1	A Discriminative Latent Variable Model for Online Clustering	Rajhans Samdani; Kai-Wei Chang; Dan Roth
M2	Exchangeable Variable Models	Mathias Niepert; Pedro Domingos
M3	Learning Latent Variable Gaussian Graphical Models	Zhaoshi Meng; Brian Eriksson; Al Hero
M4	Latent Variable Copula Inference for Bundle Pricing from Retail Transaction Data	Benjamin Letham; Wei Sun; Anshul Sheopuri
M5	Affinity Weighted Embedding	Jason Weston; Ron Weiss; Hector Yee
M6	Learning the Irreducible Representations of Commutative Lie Groups	Taco Cohen; Max Welling
M7	Covering Number for Efficient Heuristic-based POMDP Planning	Zongzhang Zhang; David Hsu; Wee Sun Lee
M8	Learning Complex Neural Network Policies with Trajectory Optimization	Sergey Levine; Vladlen Koltun
M9	A Physics-Based Model Prior for Object-Oriented MDPs	Jonathan Scholz; Martin Levihn; Charles Isbell
M10	Online Multi-Task Learning for Policy Gradient Methods	Haitham Bou Ammar; Eric Eaton; Paul Ruvolo; Matthew Taylor
M11	Pursuit-Evasion Without Regret, with an Application to Trading	Lili Dworkin; Michael Kearns; Yuriy Nevmyvaka
M12	Adaptivity and Optimism: An Improved Exponentiated Gradient Algorithm	Jacob Steinhardt; Percy Liang

M13	Demystifying Information-Theoretic Clustering	Greg Ver Steeg; Aram Galstyan; Fei Sha; Simon DeDeo
M14	Clustering in the Presence of Background Noise	Shai Ben-David; Nika Haghtalab
M15	Hierarchical Quasi-Clustering Methods for Asymmetric Networks	Gunnar Carlsson; Facundo Mémoli; Alejandro Ribeiro; Santiago Segarra
M16	Local algorithms for interactive clustering	Pranjal Awasthi; Maria Balcan; Konstantin Voevodski
M17	Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance	Simone Romano; James Bailey; Vinh Nguyen; Karin Verspoor
M18	A Single-Pass Algorithm for Efficiently Recovering Sparse Cluster Centers of High-dimensional Data	Jinfeng Yi; Lijun Zhang; Jun Wang; Rong Jin; Anil Jain
M19	Large-Margin Metric Learning for Constrained Partitioning Problems	Rémi Lajugie; Francis Bach; Sylvain Arlot
M20	Robust Distance Metric Learning via Simultaneous L1-Norm Minimization and Maximization	Hua Wang; Feiping Nie; Heng Huang
M21	Efficient Learning of Mahalanobis Metrics for Ranking	Daryl Lim; Gert Lanckriet
M22	Stochastic Neighbor Compression	Matt Kusner; Stephen Tyree; Kilian Weinberger; Kunal Agrawal
M23	Large-margin Weakly Supervised Dimensionality Reduction	Chang Xu; Dacheng Tao; Chao Xu; Yong Rui
M24	Sparse meta-Gaussian information bottleneck	Melanie Rey; Volker Roth; Thomas Fuchs
M25	Fast Stochastic Alternating Direction Method of Multipliers	Wenliang Zhong; James Kwok
M26	Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization	Shai Shalev-Shwartz; Tong Zhang
M27	An Asynchronous Parallel Stochastic Coordinate Descent Algorithm	Ji Liu; Steve Wright; Christopher Re; Victor Bittorf; Srikrishna Sridhar
M28	Towards an optimal stochastic alternating direction method of multipliers	Samaneh Azadi; Suvrit Sra
M29	Stochastic Dual Coordinate Ascent with Alternating Direction Method of Multipliers	Taiji Suzuki
M30	Communication-Efficient Distributed Optimization using an Approximate Newton-type Method	Ohad Shamir; Nati Srebro; Tong Zhang
M31	Multimodal Neural Language Models	Ryan Kiros; Ruslan Salakhutdinov; Rich Zemel
M32	Distributed Representations of Sentences and Documents	Quoc Le; Tomas Mikolov
M33	Learning Character-level Representations for Part-of-Speech Tagging	Cicero Dos Santos; Bianca Zadrozny
M34	Compositional Morphology for Word Representations and Language Modelling	Jan Botha; Phil Blunsom
M35	Towards End-To-End Speech Recognition with Recurrent Neural Networks	Alex Graves; Navdeep Jaitly
M36	A Clockwork RNN	Jan Koutnik; Klaus Greff; Faustino Gomez; Juergen Schmidhuber

Sunday June 22,

10:30 - Track A - Networks and Graph-Based Learning I

S1 Joint Inference of Multiple Label Types in Large Networks

Deepayan Chakrabarti; Stanislav Funiak; Jonathan Chang; Sofus Macskassy

We tackle the problem of inferring node labels in a partially labeled graph where each node in the graph has multiple label types and each label type has a large number of possible labels. Our primary example, and the focus of this paper, is the joint inference of label types such as hometown, current city, and employers, for users connected by a social network. Standard label propagation fails to consider the properties of the label types and the interactions between them. Our proposed method, called EdgeExplain, explicitly models these, while still enabling scalable inference under a distributed message-passing architecture. On a billion-node subset of the Facebook social network, EdgeExplain significantly outperforms label propagation for several label types, with lifts of up to 120% for recall@1 and 60% for recall@3.

S2 Learning the Consistent Behavior of Common Users for Target Node Prediction across Social Networks

Shan-Hung Wu; Hao-Heng Chien; Kuan-Hua Lin; Philip Yu

We study the target node prediction problem: given two social networks, identify those nodes/ users
from one network (called the source network) who are likely to join another (called the target network, with nodes called target nodes). Although this problem can be solved using existing techniques in the field of cross domain classification,we observe that in many real-world situations the cross-domain classifiers perform sub-optimally due to the heterogeneity between source and target networks that prevents the knowledge from being transferred. In this paper, we propose learning the consistent behavior of common users to help the knowledge transfer. We first present the Consistent Incidence Co-Factorization (CICF) for identifying the consistent users, i.e., common users that behave consistently across networks. Then we introduce the Domain-UnBiased (DUB) classifiers that transfer knowledge only through those consistent users. Extensive experiments are conducted and the results show that our proposal copes with heterogeneity and improves prediction accuracy.

S3 Learning Modular Structures from Network Data and Node Variables

Elham Azizi; Edoardo Airoldi; James Galagan

A standard technique for understanding underlying dependency structures among a set of variables posits a shared conditional probability distribution for the variables measured on individuals within a group. This approach is often referred to as module networks, where individuals are represented by nodes in a network, groups are termed modules, and the focus is on estimating the network structuren among modules. However, estimation solely from node-specific variables can lead to spurious dependencies, and unverifiable structural assumptions are often used for regularization. Here,we propose an extended model that leverages direct observations about the network in addition to nodespecific variables.

By integrating complementary data types, we avoid the need for structural assumptions. We illustrate theoretical and practical significance of the model and develop a reversible-jump MCMC learning procedure for learning modules and model parameters. We demonstrate the method accuracy in predicting modular structures from synthetic data and capability to learn regulatory modules in the Mycobacterium tuberculosis gene regulatory network.

S4 Weighted Graph Clustering with Non-Uniform Uncertainties

Yudong Chen; Shiau Hong Lim; Huan Xu

We study the graph clustering problem where each observation (edge or no-edge between a pair of nodes) may have a different level of confidence/uncertainty. We propose a clustering algorithm that is based on optimizing an appropriate weighted objective, where larger weights are given to observations with lower uncertainty. Our approach leads to a convex optimization problem that is efficiently solvable. We analyze our approach under a natural generative model, and establish theoretical guarantees for recovering the underlying clusters. Our main result is a general theorem that applies to any given weight
and distribution for the uncertainty. By optimizing over the weights, we derive a provably {optimal} weighting scheme, which matches the information theoretic lower bound up to logarithmic factors and leads to strong performance bounds in several specific settings. By optimizing over the uncertainty distribution, we show that non-uniform uncertainties can actually help. In particular, if the graph is built by spending a limited amount of resource to take measurement on each node pair, then it is beneficial to allocate the resource in a non-uniform fashion to obtain accurate measurements on a few pairs of nodes, rather than obtaining inaccurate measurements on many pairs. We provide simulation results that validate our theoretical findings.

S5 Efficient Dimensionality Reduction for High-Dimensional Network Estimation

Safiye Celik; Benjamin Logsdon; Su-In Lee

We propose module graphical lasso (MGL), an aggressive dimensionality reduction and network estimation technique for a high-dimensional Gaussian graphical model (GGM). MGL achieves scalability, interpretability and robustness by exploiting the modularity property of many real-world networks.
Variables are organized into tightly coupled modules and a graph structure is estimated to determine the conditional independencies among modules. MGL iteratively learns the module assignment of variables, the latent variables, each corresponding to a module, and the parameters of the GGM of the latent variables. In synthetic data experiments, MGL outperforms the standard graphical lasso and three other methods that incorporate latent variables into GGMs. When applied to gene expression data from ovarian cancer, MGL outperforms standard clustering algorithms in identifying functionally coherent gene sets and predicting survival time of patients. The learned modules and their dependencies provide novel insights into cancer biology as well as identifying possible novel drug targets.

S6 Discovering Latent Network Structure in Point Process Data

Scott Linderman; Ryan Adams

Networks play a central role in modern data analysis, enabling us to reason about systems by

studying the relationships between their parts. Most often in network analysis, the edges are given. However, in many systems it is difficult or impossible to measure the network directly. Examples of latent networks include economic interactions linking financial instruments and patterns of reciprocity in gang violence. In these cases, we are limited to noisy observations of events associated with each node. To enable analysis of these implicit networks, we develop a probabilistic model that combines mutually-exciting point processes with random graph models. We show how the Poisson superposition principle enables an elegant auxiliary variable formulation and a fully-Bayesian, parallel inference algorithm. We evaluate this new model empirically on several datasets.

10:30 - Track B - Reinforcement Learning I

S7 PAC-inspired Option Discovery in Lifelong Reinforcement Learning

Emma Brunskill; Lihong Li

A key goal of AI is to create lifelong learning agents that can leverage prior experience to improve
performance on later tasks. In reinforcement-learning problems, one way to summarize prior experience for future use is through options, which are temporally extended actions (subpolicies) for how to behave. Options can then be used to potentially accelerate learning in new reinforcement learning tasks. In this work, we provide the first formal analysis of the sample complexity, a measure of learning speed, of reinforcement learning with options. This analysis helps shed light on some interesting prior empirical results on when and how options may accelerate learning. We then quantify the benefit of options in reducing sample complexity of a lifelong learning agent. Finally, the new theoretical insights inspire a novel option-discovery algorithm that aims at minimizing overall sample complexity in lifelong reinforcement learning.

S8 Time-Regularized Interrupting Options (TRIO)

Daniel Mankowitz; Timothy Mann; Shie Mannor

High-level skills relieve planning algorithms from low-level details. But when the skills are poorly designed for the domain, the resulting plan may be severely suboptimal. Sutton et al. 1999 made an mimportant step towards resolving this problem by introducing a rule that automatically improves a set of skills called options. This rule terminates an option early whenever switching to another option gives a higher value than continuing with the current option. However, they only analyzed the case where the improvement rule is applied once. We show conditions where this rule converges to the optimal set of options. A new Bellman-like operator that simultaneously improves the set of options is at the core of our analysis. One problem with the update rule is that it tends to favor lower-level skills. Therefore we introduce a regularization term that favors longer duration skills. Experimental results demonstrate that this approach can derive a good set of high-level skills even when the original set of skills cannot solve the problem.

S9 Approximate Policy Iteration Schemes: A Comparison

Bruno Scherrer

We consider the infinite-horizon discounted optimal control problem formalized by Markov Decision Processes. We focus on several approximate variations of the Policy Iteration algorithm: Approximate Policy Iteration, Conservative Policy Iteration (CPI), a natural adaptation of the Policy Search by Dynamic Programming algorithm to the infinite-horizon case (PSDP$_\infty$), and the recently proposed Non-Stationary Policy iteration (NSPI(m)). For all algorithms, we describe performance bounds, and make a comparison by paying a particular attention to the concentrability constants involved, the number of iterations and the memory required. Our analysis highlights the following points: 1) The performance
guarantee of CPI can be arbitrarily better than that of API/API(α), but this comes at the cost of a relative---exponential in $\frac{1}{\epsilon}$---increase of the number of iterations. 2) PSDP$_\infty$ enjoys the best of both worlds: its performance guarantee is similar to that of CPI, but within a number of iterations similar to that of API. 3) Contrary to API that requires a constant memory, the memory needed by CPI and PSDP$_\infty$ is proportional to their number of iterations, which may be problematic when the discount factor γ is close to 1 or the approximation error ϵ is close to 0; we show that the NSPI(m) algorithm allows to make an overall trade-off between memory and performance. Simulations with these schemes confirm our analysis.

S10 Model-Based Relational RL When Object Existence is Partially Observable

Vien Ngo; Marc Toussaint

We consider learning and planning in relational MDPs when object existence is uncertain and new objects may appear or disappear depending on previous actions or properties of other objects. Optimal policies actively need to discover objects to achieve a goal; planning in such domains in general amounts to a POMDP problem, where the belief is about the existence and properties of potential not-yetdiscovered objects. We propose a computationally efficient extension of model-based relational RL methods that approximates these beliefs using discrete uncertainty predicates. In this formulation the belief update is learned using probabilistic rules and planning in the approximated belief space can be achieved using an extension of existing planners. We prove that the learned belief update rules encode an approximation of the exact belief updates of a POMDP formulation and demonstrate experimentally that the proposed approach successfully learns a set of relational rules appropriate to solve such problems.

S11 GeNGA: A Generalization of Natural Gradient Ascent with Positive and Negative Convergence Results

Philip Thomas

Natural gradient ascent (NGA) is a popular optimization method that uses a positive definite metric tensor. In many applications the metric tensor is only guaranteed to be positive semidefinite (e.g., when using the Fisher information matrix as the metric tensor), in which case NGA is not applicable. In our first contribution, we derive generalized natural gradient ascent (GeNGA), a generalization of NGA which allows for positive semidefinite non-smooth metric tensors. In our second contribution we show that, in standard settings, GeNGA and NGA can both be divergent. We then establish sufficient conditions to ensure that both achieve various forms of convergence. In our third contribution we show how several reinforcement learning methods that use NGA without positive definite metric tensors can be adapted to properly use GeNGA.

S12 Scaling Up Robust MDPs using Function Approximation

Aviv Tamar; Shie Mannor; Huan Xu

We consider large-scale Markov decision processes (MDPs) with parameter uncertainty, under the robust MDP paradigm. Previous studies showed that robust MDPs, based on a minimax approach to
handling uncertainty, can be solved using dynamic programming for small to medium sized problems.
However, due to the "curse of dimensionality", MDPs that model real-life problems are typically prohibitively large for such approaches. In this work we employ a reinforcement learning approach to
tackle this planning problem: we develop a robust approximate dynamic programming method based on
a projected fixed point equation to approximately solve large scale robust MDPs. We show that the
proposed method provably succeeds under certain technical conditions, and demonstrate its effectiveness through simulation of an option pricing problem. To the best of our knowledge, this is the
first attempt to scale up the robust MDP paradigm.

10:30 - Track C - Bayesian Optimization and Gaussian Processes

S13 Agnostic Bayesian Learning of Ensembles

Alexandre Lacoste; Mario Marchand; Francois Laviolette; Hugo Larochelle

We propose a method for producing ensembles of predictors based on holdout estimations of their generalization performances. This approach uses a prior directly on the performance of predictors taken from a finite set of candidates and attempts to infer which one is best. Using Bayesian inference, we can thus obtain a posterior that represents our uncertainty about that choice and construct a weighted ensemble of predictors accordingly. This approach has the advantage of not requiring that the predictors be probabilistic themselves, can deal with arbitrary measures of performance and does not assume that the data was actually generated from any of the predictors in the ensemble. Since the problem of finding the best (as opposed to the true) predictor among a class is known as agnostic PAClearning, we refer to our method as agnostic Bayesian learning. We also propose a method to address the case where the performance estimate is obtained from k-fold cross validation. While being efficient and easily adjustable to any loss function, our experiments confirm that the agnostic Bayes approach is state of the art compared to common baselines such as model selection based on k-fold cross-validation or a linear combination of predictor outputs.

S14 Robust RegBayes: Selectively Incorporating First-Order Logic Domain Knowledge into Bayesian Models

Shike Mei; Jun Zhu; Jerry Zhu

Much research in Bayesian modeling has been done to elicit a prior distribution that incorporates domain knowledge. We present a novel and more direct approach by imposing First-Order Logic (FOL) rules on the posterior distribution. Our approach unifies FOL and Bayesian modeling under the regularized Bayesian framework. In addition, our approach automatically estimates the uncertainty of FOL rules when they are produced by humans, so that reliable rules are incorporated while unreliable ones are ignored. We apply our approach to latent topic modeling tasks and demonstrate that by combining FOL knowledge and Bayesian modeling, we both improve the task performance and discover more structured latent representations in unsupervised and supervised learning.

S15 An Efficient Approach for Assessing Hyperparameter Importance

Frank Hutter; Holger Hoos; Kevin Leyton-Brown

The performance of many machine learning methods depends critically on hyperparameter settings. Sophisticated Bayesian optimization methods have recently achieved considerable successes in optimizing these hyperparameters, in several cases surpassing the performance of human experts. However, blind reliance on such methods can leave end users without insight into the relative importance of different hyperparameters and their interactions. This paper describes efficient methods that can be used to gain such insight, leveraging random forest models fit on the data already gathered by Bayesian optimization. We first introduce a novel, linear-time algorithm for computing marginals of random forest predictions and then show how to leverage these predictions within a functional ANOVA framework, to quantify the importance of both single hyperparameters and of interactions between hyperparameters. We conducted experiments with prominent machine learning frameworks and stateof-the-art solvers for combinatorial problems. We show that our methods provide insight into the relationship between hyperparameter settings and performance, and demonstrate that—even in very high-dimensional cases—most performance variation is attributable to just a few hyperparameters.

S16 Bayesian Optimization with Inequality Constraints

Jacob Gardner; Matt Kusner; Zhixiang; Xu; Kilian Weinberger; John Cunningham

Bayesian optimization is a powerful framework for minimizing expensive objective functions while using very few function evaluations. It has been successfully applied to a variety of problems, including hyperparameter tuning and experimental design. However, this framework has not been extended to the inequality-constrained optimization setting, particularly the setting in which evaluating feasibility is just as expensive as evaluating the objective. Here we present constrained Bayesian optimization, which places a prior distribution on both the objective and the constraint functions. We evaluate our method on simulated and real data, demonstrating that constrained Bayesian optimization can quickly find optimal and feasible points, even when small feasible regions cause standard methods to fail.

S17 A PAC-Bayesian bound for Lifelong Learning

Anastasia Pentina; Christoph Lampert

Transfer learning has received a lot of attention in the machine learning community over the last years, and several effective algorithms have been developed. However, relatively little is known about their theoretical properties, especially in the setting of lifelong learning, where the goal is to transfer information to tasks for which no data have been observed so far. In this work we study lifelong learning from a theoretical perspective. Our main result is a PAC-Bayesian generalization bound that offers a unified view on existing paradigms for transfer learning, such as the transfer of parameters or the transfer of low-dimensional representations. We also use the bound to derive two principled lifelong learning algorithms, and we show that these yield results comparable with existing methods.

S18 Gaussian Processes for Bayesian Estimation in Ordinary Differential Equations

David Barber; Yali Wang

Bayesian parameter estimation in coupled ordinary differential equations (ODEs) is challenging due to the high computational cost of numerical integration. In gradient matching a separate data model is introduced with the property that its gradient can be calculated easily. Parameter estimation is achieved by requiring consistency between the gradients computed from the data model and those specified by the ODE. We propose a Gaussian process model that directly links state derivative information with system observations, simplifying previous approaches and providing a natural generative model.

10:30 - Track D - PCA and Subspace Models

S19 Robust Principal Component Analysis with Complex Noise

Qian Zhao; Deyu Meng; Zongben Xu; Wangmeng Zuo; Lei Zhang

The research on robust principal component analysis (RPCA) has been attracting much attention recently. The original RPCA model assumes sparse noise, and use the L_1-norm to characterize the error term. In practice, however, the noise is much more complex and it is not appropriate to simply use a certain L_p-norm for noise modeling. We propose a generative RPCA model under the Bayesian framework by modeling data noise as a mixture of Gaussians (MoG). The MoG is a universal approximator to continuous distributions and thus our model is able to fit a wide range of noises such as Laplacian, Gaussian, sparse noises and any combinations of them. A variational Bayes algorithm is presented to infer the posterior of the proposed model. All involved parameters can be recursively updated in closed form. The advantage of our method is demonstrated by extensive experiments on synthetic data, face modeling and background subtraction.

S20 Multivariate Maximal Correlation Analysis

Hoang Vu Nguyen; Emmanuel Müller; Jilles Vreeken; Pavel Efros; Klemens Böhm

Correlation analysis is one of the key elements of statistics, and has various applications in data analysis. Whereas most existing measures can only detect pairwise correlations between two dimensions, modern analysis aims at detecting correlations in multi-dimensional spaces. We propose MAC, a novel multivariate correlation measure designed for discovering multi-dimensional patterns. It belongs to the powerful class of maximal correlation analysis, for which we propose a generalization to multivariate domains. We highlight the limitations of current methods in this class, and address these with MAC. Our experiments show that MAC outperforms existing solutions, is robust to noise, and discovers interesting and useful patterns.

S21 Discriminative Features via Generalized Eigenvectors

Nikos Karampatziakis; Paul Mineiro

Representing examples in a way that is compatible with the underlying classifier can greatly enhance the performance of a learning system. In this paper we investigate scalable techniques for inducing discriminative features by taking advantage of simple second order structure in the data. We focus on multiclass classification and show that features extracted from the generalized eigenvectors of the class conditional second moments lead to classifiers with excellent empirical performance. Moreover, these features have attractive theoretical properties, such as inducing representations that are invariant to linear transformations of the input. We evaluate classifiers built from these features on three different tasks, obtaining state of the art results.

S22 Randomized Nonlinear Component Analysis

David Lopez-Paz; Suvrit Sra; Alex Smola; Zoubin Ghahramani; Bernhard Schoelkopf

Classical methods such as Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA) are ubiquitous in statistics. However, these techniques are only able to reveal linear relationships in data. Although nonlinear variants of PCA and CCA have been proposed, these are computationally prohibitive in the large scale. In a separate strand of recent research, randomized methods have been proposed to construct features that help reveal nonlinear patterns in data. For basic tasks such as regression or classification, random features exhibit little or no loss in performance, while achieving drastic savings in computational requirements. In this paper we leverage randomness to design scalable new variants of nonlinear PCA and CCA; our ideas extend to key multivariate analysis tools such as spectral clustering or LDA. We demonstrate our algorithms through experiments on real-world data, on which we compare against the state-of-the-art. A simple R implementation of the presented algorithms is provided.

S23 Memory and Computation Efficient PCA via Very Sparse Random Projections

Farhad Pourkamali Anaraki; Shannon Hughes

Algorithms that can efficiently recover principal components in very high-dimensional, streaming, and/or distributed data settings have become an important topic in the literature. In this paper, we propose an approach to principal component estimation that utilizes projections onto very sparse random vectors with Bernoulli-generated nonzero entries. Indeed, our approach is simultaneously efficient in memory/storage space, efficient in computation, and produces accurate PC estimates, while also allowing for rigorous theoretical performance analysis. Moreover, one can tune the sparsity of the random vectors deliberately to achieve a desired point on the tradeoffs between memory, computation, and accuracy. We rigorously characterize these tradeoffs and provide statistical performance guarantees. In addition to these very sparse random vectors, our analysis also applies to more general random projections. We present experimental results demonstrating that this approach allows for simultaneously achieving a substantial reduction of the computational complexity and memory/storage space, with little loss in accuracy, particularly for very high-dimensional data.

S24 Optimal Mean Robust Principal Component Analysis

Feiping Nie; Jianjun Yuan; Heng Huang

10:30 - Track E - Supervised Learning

S25 The Coherent Loss Function for Classification

Wenzhuo Yang; Melvyn Sim; Huan Xu

A prediction rule in binary classification that aims to achieve the lowest probability of misclassification involves minimizing over a non-convex, 0-1 loss function, which is typically a computationally intractable optimization problem. To address the intractability, previous methods consider minimizing the cumulative loss -- the sum of convex surrogates of the 0-1 loss of each sample. In this paper, we revisit this paradigm and develop instead an axiomatic framework by proposing a set of salient properties on functions for binary classification and then propose the coherent loss approach, which is a tractable upper-bound of the empirical classification error over the entire sample set. We show that the proposed approach yields a strictly tighter approximation to the empirical classification error than any convex cumulative loss approach while preserving the convexity of the underlying optimization problem, and this approach for binary classification also has a robustness interpretation which builds a connection to robust SVMs. The experimental results show that our approach outperforms the standard SVM when additional constraints are imposed.

S26 Condensed Filter Tree for Cost-Sensitive Multi-Label Classification

Chun-Liang Li; Hsuan-Tien Lin

Different real-world applications of multi-label classification often demand different evaluation criteria. We formalize this demand with a general setup, cost-sensitive multi-label classification (CSMLC), which takes the evaluation criteria into account during learning. Nevertheless, most existing algorithms can only focus on optimizing a few specific evaluation criteria, and cannot systematically deal with different ones. In this paper, we propose a novel algorithm, called condensed filter tree (CFT), for optimizing any criteria in CSMLC. CFT is derived from reducing CSMLC to the famous filter tree algorithm for costsensitive multi-class classification via constructing the label powerset. We successfully cope with the difficulty of having exponentially many extended-classes within the powerset for representation, training and prediction by carefully designing the tree structure and focusing on the key nodes. Experimental results across many real-world datasets validate that CFT is competitive with special purpose algorithms on special criteria and reaches better performance on general criteria.

S27 Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification

Junfeng Wen; Chun-Nam Yu; Russell Greiner

Many learning situations involve learning the conditional distribution $p(y|x)$ when the training instances are drawn from the training distribution $p_{tr}(x)$, even though it will later be used to predict for instances drawn from a different test distribution $p_{te}(x)$. Most current approaches focus on learning how to reweigh the training examples, to make them resemble the test distribution. However, reweighing does not always help, because (we show that) the test error also depends on the correctness of the underlying model class. This paper analyses this situation by viewing the problem of learning under changing distributions as a game between a learner and an adversary. We characterize when such reweighing is needed, and also provide an algorithm, robust covariate shift adjustment (RCSA), that provides relevant weights. Our empirical studies, on UCI datasets and a real-world cancer prognostic prediction

dataset, show that our analysis applies, and that our RCSA works effectively.

S28 A Statistical Convergence Perspective of Algorithms for Rank Aggregation from Pairwise Data

Arun Rajkumar; Shivani Agarwal

There has been much interest recently in the problem of rank aggregation from pairwise data. A natural question that arises is: under what sorts of statistical assumptions do various rank aggregation algorithms converge to an `optimal' ranking? In this paper, we consider this question in a natural setting where pairwise comparisons are drawn randomly and independently from some underlying probability distribution. We first show that, under a `time-reversibility' or Bradley-Terry-Luce (BTL) condition on the distribution generating the outcomes of the pairwise comparisons, the rank centrality (PageRank) and least squares (HodgeRank) algorithms both converge to an optimal ranking. Next, we show that a matrix version of the Borda count algorithm, and more surprisingly, an algorithm which performs maximal likelihood estimation under a BTL assumption, both converge to an optimal ranking under a `low-noise' condition that is strictly more general than BTL. Finally, we propose a new SVM-based algorithm for rank aggregation from pairwise data, and show that this converges to an optimal ranking under an even more general condition that we term `generalized low-noise'. In all cases, we provide explicit sample complexity bounds for exact recovery of an optimal ranking. Our experiments confirm our theoretical findings and help to shed light on the statistical behavior of various rank aggregation algorithms.

S29 GEV-Canonical Regression for Accurate Binary Class Probability Estimation when One Class is Rare

Arpit Agarwal; Harikrishna Narasimhan; Shivaram Kalyanakrishnan; Shivani Agarwal

We consider the problem of binary class probability estimation (CPE) when one class is rare compared to the other. It is well known that standard algorithms such as logistic regression do not perform well on this task as they tend to under-estimate the probability of the rare class. Common fixes include undersampling and weighting, together with various correction schemes. Recently, Wang & Dey (2010) suggested the use of a parametrized family of asymmetric link functions based on the generalized extreme value (GEV) distribution, which has been used for modeling rare events in statistics. The approach showed promising initial results, but combined with the logarithmic CPE loss implicitly used in their work, it results in a non-convex composite loss that is difficult to optimize. In this paper, we use tools from the theory of proper composite losses (Buja et al, 2005; Reid & Williamson, 2010) to construct a canonical underlying CPE loss corresponding to the GEV link, which yields a convex proper composite loss that we call the GEV-canonical loss; this loss is tailored for the task of CPE when one class is rare, and is easy to minimize using an IRLS-type algorithm similar to that used for logistic regression. Our experiments on both synthetic and real data demonstrate that the resulting algorithm -- which we term GEV-canonical regression -- outperforms common approaches such as under-sampling and weights correction for this problem

S30 Guess-Averse Loss Functions For Cost-Sensitive Multiclass Boosting

Oscar Beijbom; Mohammad Saberian; David Kriegman; Nuno Vasconcelos

Cost-sensitive multiclass classification has recently acquired significance in several applications, through the introduction of multiclass datasets with well-defined misclassification costs.

The design of classification algorithms for this setting is considered. It is argued that the unreliable performance of current algorithms is due to the inability of the underlying loss functions to enforce a certain fundamental underlying property. This property, denoted guess-aversion, is that the loss should encourage correct classifications over the arbitrary guessing that ensues when all classes are equally scored by the classifier. While guess-aversion holds trivially for binary classification, this is not true in the multiclass setting. A new family of cost-sensitive guess-averse loss functions is derived, and used to design new cost-sensitive multiclass boosting algorithms, denoted GEL- and GLL-MCBoost. Extensive experiments demonstrate (1) the general importance of guess-aversion and (2) that the GLL loss function outperforms other loss functions for multiclass boosting.

10:30 - Track F - Neural Networks and Deep Learning I

S31 Structured Recurrent Temporal Restricted Boltzmann Machines

Roni Mittelman; Benjamin Kuipers; Silvio Savarese; Honglak Lee

The Recurrent temporal restricted Boltzmann machine (RTRBM) is a probabilistic model for temporal data, that has been shown to effectively capture both short and long-term dependencies in time-series. The topology of the RTRBM graphical model, however, assumes full connectivity between all the pairs of visible and hidden units, therefore ignoring the dependency structure between the different observations. Learning this structure has the potential to not only improve the prediction performance, but it can also reveal important patterns in the data. For example, given an econometricdataset, we could identify interesting dependencies between different market sectors; given a meteorological dataset, we could identify regional weather patterns. In this work we propose a new class of RTRBM, which explicitly uses a dependency graph to model the structure in the problem and to define the energy function. We refer to the new model as the structured RTRBM (SRTRBM). Our technique is related to methods such as graphical lasso, which are used to learn the topology of Gaussian graphical models. We also develop a spike-and-slab version of the RTRBM, and combine it with our method to learn structure in datasets with real valued observations. Our experimental results using synthetic and real datasets, demonstrate that the SRTRBM can improve the prediction performance of the RTRBM, particularly when the number of visible units is large and the size of the training set is small. It also reveals the structure underlying our benchmark datasets.

S32 A Deep and Tractable Density Estimator

Benigno Uria; Iain Murray; Hugo Larochelle

The Neural Autoregressive Distribution Estimator (NADE) and its real-valued version RNADE are competitive density models of multidimensional data across a variety of domains. These models use a fixed, arbitrary ordering of the data dimensions. One can easily condition on variables at the beginning of the ordering, and marginalize out variables at the end of the ordering, however other inference tasks require approximate inference. In this work we introduce an efficient procedure to simultaneously train a NADE model for each possible ordering of the variables, by sharing parameters across all these models. We can thus use the most convenient model for each inference task at hand, and ensembles of such models with different orderings are immediately available. Moreover, unlike the original NADE, our training procedure scales to deep models. Empirically, ensembles of Deep NADE models obtain state of the art density estimation performance.

S33 Deep Supervised and Convolutional Generative Stochastic Network for Protein Secondary Structure Prediction

Jian Zhou; Olga Troyanskaya

Predicting protein secondary structure is a fundamental problem in protein structure prediction. Here we present a new supervised generative stochastic network (GSN) based method to predict local secondary structure with deep hierarchical representations. GSN is a recently proposed deep learning technique (Bengio&Thibodeau-Laufer, 2013) to globally train deep generative model. We present the supervised extension of GSN, which learns a Markov chain to sample from a conditional distribution, and applied it to protein structure prediction. To scale the model to full-sized, high-dimensional data, like
protein sequences with hundreds of amino-acids, we introduce a convolutional architecture, which allows efficient learning across multiple layers of hierarchical representations. Our architecture uniquely focuses on predicting structured low-level labels informed with both low and high-level representations learned by the model. In our application this corresponds to labeling the secondary structure state of each amino-acid residue. We trained and tested the model on separate sets of non-homologous proteins sharing less than 30% sequence identity. Our model achieves 66.4% Q8 accuracy on the CB513 dataset, better than the previously reported best performance 64.9% (Wang et al., 2011) for this challenging secondary structure prediction problem.

S34 Deep AutoRegressive Networks

Karol Gregor; Ivo Danihelka; Andriy Mnih; Charles Blundell; Daan Wierstra

We introduce a deep, generative autoencoder capable of learning hierarchies of distributed representations from data. Successive deep stochastic hidden layers are equipped with autoregressive connections, which enable the model to be sampled from quickly and exactly via ancestral sampling. We derive an efficient approximate parameter estimation method based on the minimum description length (MDL) principle, which can be seen as maximising a variational lower bound on the log-likelihood, with a feedforward neural network implementing approximate inference. We demonstrate state-of-the-artgenerative performance on a number of classic data sets: several UCI data sets, MNIST and Atari 2600 games.

S35 Stochastic Backpropagation and Approximate Inference in Deep Generative Models

Danilo Jimenez Rezende; Shakir Mohamed; Daan Wierstra

We marry ideas from deep neural networks and approximate Bayesian inference to derive a generalised class of deep, directed generative models, endowed with a new algorithm for scalable inference and learning. Our algorithm introduces a recognition model to represent an approximate posterior distribution and uses this for optimisation of a variational lower bound.We develop stochastic backpropagation -- rules for gradient backpropagation through stochastic variables -- and derive an algorithm that allows for joint optimisation of the parameters of both the generative and recognition models.We demonstrate on several real-world data sets that by using stochastic backpropagation and variational inference, we obtain models that are able to generate realistic samples of data, allow for accurate imputations of missing data, and provide a useful tool for high-dimensional data visualisation.

S36 Neural Variational Inference and Learning in Belief Networks

Andriy Mnih; Karol Gregor

Highly expressive directed latent variable models, such as sigmoid belief networks, are difficult to train on large datasets because exact inference in them is intractable and none of the approximate inference methods that have been applied to them scale well. We propose a fast non-iterative approximate inference method that uses a feedforward network to implement efficient exact sampling from the variational posterior. The model and this inference network are trained jointly by maximizing a variational lower bound on the log-likelihood. Although the naive estimator of the inference network gradient is too high-variance to be useful, we make it practical by applying several straightforward model-independent variance reduction techniques. Applying our approach to training sigmoid belief networks and deep autoregressive networks, we show that it outperforms the wake-sleep algorithm on MNIST and achieves state-of-the-art results on the Reuters RCV1 document dataset.

14:00 - Track A - Graphical Models I

S37 Linear and Parallel Learning of Markov Random Fields

Yariv Mizrahi; Misha Denil; Nando De Freitas

We introduce a new embarrassingly parallel parameter learning algorithm for Markov random fields which is efficient for a large class of practical models. Our algorithm parallelizes naturally over cliques and, for graphs of bounded degree, its complexity is linear in the number of cliques. Unlike its competitors, our algorithm is fully parallel and for log-linear models it is also data efficient, requiring only the local sufficient statistics of the data to estimate parameters.

S38 Putting MRFs on a Tensor Train

Alexander Novikov; Anton Rodomanov; Anton Osokin; Dmitry Vetrov

In the paper we present a new framework for dealing with probabilistic graphical models. Our approach relies on the recently proposed Tensor Train format (TT-format) of a tensor that while being compact allows for efficient application of linear algebra operations. We present a way to convert the energy of a Markov random field to the TT-format and show how one can exploit the properties of the TT-format to attack the tasks of the partition function estimation and the MAP-inference. We provide theoretical guarantees on the accuracy of the proposed algorithm for estimating the partition function and compare our methods against several state-of-the-art algorithms.

S39 Gaussian Approximation of Collective Graphical Models

Liping Liu; Daniel Sheldon; Thomas Dietterich

The Collective Graphical Model (CGM) models a population of independent and identically distributed individuals when only collective statistics (i.e., counts of individuals) are observed. Exact inference in CGMs is intractable, and previous work has explored Markov Chain Monte Carlo (MCMC) and MAP approximations for learning and inference. This paper studies Gaussian approximations to the CGM.

As the population grows large, we show that the CGM distribution converges to a multivariate Gaussian distribution (GCGM) that maintains the conditional independence properties of the original CGM. If the observations are exact marginals of the CGM or marginals that are corrupted by Gaussian noise, inference in the GCGM approximation can be computed efficiently in closed form. If the observations follow a different noise model (e.g., Poisson), then expectation propagation provides efficient and accurate approximate inference. The accuracy and speed of GCGM inference is compared to the MCMC and MAP methods on a simulated bird migration problem. The GCGM matches or exceeds the accuracy of the MAP method while being significantly faster.

S40 Scalable Semidefinite Relaxation for Maximum A Posterior Estimation

Qixing Huang; Yuxin Chen; Leonidas Guibas

Maximum a posteriori (MAP) inference over discrete Markov random fields is a central task spanning a wide spectrum of real-world applications but known to be NP-hard for general graphs. In this paper, we propose a novel semidefinite relaxation formulation (referred to as SDR) to estimate the MAP assignment. Algorithmically, we develop an accelerated variant of the alternating direction method of multipliers (referred to as SDPAD-LR) that can effectively exploit the special structure of SDR. Encouragingly, the proposed procedure allows solving SDR for large-scale problems, e.g. problems comprising hundreds of thousands of variables with multiple states on a grid graph. Compared with prior SDP solvers, SDPAD-LR is capable of attaining comparable accuracy while exhibiting remarkably improved scalability. This contradicts the commonly held belief that semidefinite relaxation can only been applied on small-scale problems. We have evaluated the performance of SDR on various benchmark datasets including OPENGM2 and PIC. Experimental results demonstrate that for a broad class of problems, SDPAD-LR outperforms state-of-the-art algorithms in producing better MAP assignments.

S41 Globally Convergent Parallel MAP LP Relaxation Solver using the Frank-Wolfe Algorithm

Alexander Schwing; Tamir Hazan; Marc Pollefeys; Raquel Urtasun

While MAP inference is typically intractable for many real-world applications, linear programming relaxations have been proven very effective. Dual block-coordinate descent methods are among the most efficient solvers, however, they are prone to get stuck in sub-optimal points. Although subgradient approaches achieve global convergence, they are typically slower in practice. To improve convergence speed, algorithms which compute the steepest ϵ-descent direction by solving a quadratic program have been proposed. In this paper we suggest to decouple the quadratic program based on the Frank-Wolfe approach. This allows us to obtain an efficient and easy to parallelize algorithm while retaining the global convergence properties. Our method proves superior when compared to existing algorithms on a set of spin-glass models and protein design tasks.

S42 Inferning with High Girth Graphical Models

Uri Heinemann; Amir Globerson

Unsupervised learning of graphical models is an important task in many domains.

Although maximum likelihood learning is computationally hard, there do exist consistent learning algorithms (e.g., psuedolikelihood and its variants). However, inference in the learned models is still hard, and thus they are not directly usable. In other words, given a probabilistic query they are not guaranteed to provide an answer that is close to the true one. In the current paper, we provide a learning algorithm that is guaranteed to provide approximately correct probabilistic inference. We focus on a particular class of models, namely high girth graphs in the correlation decay regime. It is well known that approximate inference (e.g, using loopy BP) in such models yields marginals that are close to the true ones. Motivated by this, we propose an algorithm that always returns models of this type, and hence in the models it returns inference is approximately correct. We derive finite sample results guaranteeing that beyond a certain sample size, the resulting models will answer probabilistic queries with a high level of accuracy. Results on synthetic data show that the models we learn indeed outperform those obtained by other algorithms, which do not return high girth graphs.

14:00 - Track B - Bandits I

S43 Thompson Sampling for Complex Online Problems

Aditya Gopalan; Shie Mannor; Yishay Mansour

We consider stochastic multi-armed bandit problems with complex actions over a set of basic arms, where the decision maker plays a complex action rather than a basic arm in each round. The reward of the complex action is some function of the basic arms' rewards, and the feedback observed may not necessarily be the reward per-arm. For instance, when the complex actions are subsets of the arms, we may only observe the maximum reward over the chosen subset. Thus, feedback across complex actions may be coupled due to the nature of the reward function. We prove a frequentist regret bound for Thompson sampling in a very general setting involving parameter, action and observation spaces and a likelihood function over them. The bound holds for discretely-supported priors over the parameter space and without additional structural properties such as closed-form posteriors, conjugate prior structure or independence across arms. The regret bound scales logarithmically with time but, more importantly, with an improved constant that non-trivially captures the coupling across complex actions due to the structure of the rewards. As applications, we derive improved regret bounds for classes of complex bandit problems involving selecting subsets of arms, including the first nontrivial regret bounds for nonlinear MAX reward feedback from subsets. Using particle filters for computing posterior distributions which lack an explicit closed-form, we present numerical results for the performance of Thompson sampling for subset-selection and job scheduling problems.

S44 Unimodal Bandits: Regret Lower Bounds and Optimal Algorithms

Richard Combes; Alexandre Proutiere

We consider stochastic multi-armed bandits where the expected reward is a unimodal function over partially ordered arms. This important class of problems has been recently investigated in (Cope 2009, Yu 2011). The set of arms is either discrete, in which case arms correspond to the vertices of a finite graph whose structure represents similarity in rewards, or continuous, in which case arms belong to a bounded interval. For discrete unimodal bandits, we derive asymptotic lower bounds for the regret achieved under any algorithm, and propose OSUB, an algorithm whose regret matches this lower bound. Our algorithm optimally exploits the unimodal structure of the problem, and surprisingly, its asymptotic regret does not depend on the number of arms. We also provide a regret upper bound for OSUB in nonstationary environments where

the expected rewards smoothly evolve over time.
The analytical results are supported by numerical experiments showing that OSUB performs significantly better than the state-of-the-art algorithms. For continuous sets of arms, we provide a brief discussion. We show that combining an appropriate discretization of the set of arms with the UCB algorithm yields an orderoptimal regret, and in practice, outperforms recently proposed algorithms designed to exploit the unimodal structure.

S45 Reducing Dueling Bandits to Cardinal Bandits

Nir Ailon; Zohar Karnin; Thorsten Joachims

We present algorithms for reducing the Dueling Bandits problem to the conventional (stochastic) Multi-Armed Bandits problem. The Dueling Bandits problem is an online model of learning with ordinal feedback of the form ``A is preferred to B'' (as opposed to cardinal feedback like ``A has value 2.5''), giving it wide applicability in learning from implicit user feedback and revealed and stated preferences. In contrast to existing algorithms for the Dueling Bandits problem, our reductions -- named $\Doubler$, $\MultiSbm$ and $\DoubleSbm$ -- provide a generic schema for translating the extensive body of known results about conventional Multi-Armed Bandit algorithms to the Dueling Bandits setting. For $\Doubler$ and $\MultiSbm$ we prove regret upper bounds in both finite and infinite settings, and conjecture about the performance of $\DoubleSbm$ which empirically outperforms the other two as well as previous algorithms in our experiments. In addition, we provide the first almost optimal regret bound in terms of second order terms, such as the differences between the values of the arms.

S46 Combinatorial Partial Monitoring Game with Linear Feedback and Its Applications

Tian Lin; Bruno Abrahao; Robert Kleinberg; John Lui; Wei Chen

In online learning, a player chooses actions to play and receives reward and feedback from the environment with the goal of maximizing her reward over time. In this paper, we propose the model of combinatorial partial monitoring games with linear feedback, a model which simultaneously addresses limited feedback, infinite outcome space of the environment and exponentially large action space of the player. We present the Global Confidence Bound (GCB) algorithm, which integrates ideas from both combinatorial multi-armed bandits and finite partial monitoring games to handle all the above issues.GCB only requires feedback on a small set of actions and achieves $O(T^{\frac{2}{3}}\log T)$ distribution-independent regret and $O(\log T)$ distribution-dependent regret (the latter assuming unique optimal action), where T is the total time steps played. Moreover, the regret bounds only depend linearly on $\log |X|$ rather than $|X|$, where X is the action space. GCB isolates offline optimization tasks from online learning and avoids explicit enumeration of all actions in the online learning part. We demonstrate that our model and algorithm can be applied to a crowdsourcing application leading to both an efficient learning algorithm and low regret, and argue that they can be applied to a wide range of combinatorial applications constrained with limited feedback.

S47 Online Stochastic Optimization under Correlated Bandit Feedback

Mohammad Gheshlaghi Azar; Alessandro Lazaric; Emma Brunskill dimension.

In this paper we consider the problem of online stochastic optimization of a locally smooth function under bandit feedback. We introduce the high-confidence tree (HCT) algorithm, a novel anytime $\mathcal X$-armed bandit algorithm, and derive regret bounds matching the performance of state-ofthe-art algorithms in terms of the dependency on number of steps and the near-optimality dimension.The main advantage of HCT is that it handles the challenging case of correlated bandit feedback (reward),

HCT also improves on the state-of-the-art in terms of the memory requirement, as well as requiring a weaker smoothness assumption on the mean-reward function in comparison with the existing anytime algorithms. Finally, we discuss how HCT can be applied to the problem of policy search in reinforcement learning and we report preliminary empirical results.

S48 Adaptive Monte Carlo via Bandit Allocation

James Neufeld; Andras Gyorgy; Csaba Szepesvari; Dale Schuurmans

We consider the problem of sequentially choosing between a set of unbiased Monte Carlo estimators to minimize the mean-squared-error (MSE) of a final combined estimate. By reducing this task to a stochastic multi-armed bandit problem, we show that well developed allocation strategies can be used to achieve an MSE that approaches that of the best estimator chosen in retrospect. We then extend these developments to a scenario where alternative estimators have different, possibly stochastic, costs. The outcome is a new set of adaptive Monte Carlo strategies that provide stronger guarantees than previous approaches while offering practical advantages.

14:00 - Track C - Monte Carlo

S49 Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget

Anoop Korattikara; Yutian Chen; Max Welling

Can we make Bayesian posterior MCMC sampling more efficient when faced with very large datasets? We argue that computing the likelihood for N datapoints in the Metropolis-Hastings (MH) test to reach a single binary decision is computationally inefficient. We introduce an approximate MH rule based on a sequential hypothesis test that allows us to accept or reject samples with high confidence using only a fraction of the data required for the exact MH rule. While this method introduces an asymptotic bias, we show that this bias can be controlled and is more than offset by a decrease in variance due to our ability to draw more samples per unit of time.

S50 Towards scaling up Markov chain Monte Carlo: an adaptive subsampling approach

Rémi Bardenet; Arnaud Doucet; Chris Holmes

Markov chain Monte Carlo (MCMC) methods are often deemed far too computationally intensive to be of any practical use for large datasets. This paper describes a methodology that aims to scale up the Metropolis-Hastings (MH) algorithm in this context. We propose an approximate implementation of the accept/reject step of MH that only requires evaluating the likelihood of a random subset of the data, yet is guaranteed to coincide with the accept/reject step based on the full dataset with a probability superior to a user-specified tolerance level. This adaptive subsampling technique is an alternative to the recent approach developed in (Korattikara et al, ICML'14), and it allows us to establish rigorously that the resulting approximate MH algorithm samples from a perturbed version of the target distribution of interest, whose total variation distance to this very target is controlled explicitly. We explore the benefits and limitations of this scheme on several examples.

S51 Distributed Stochastic Gradient MCMC

Sungjin Ahn; Babak Shahbaba; Max Welling

Probabilistic inference on a big data scale is becoming increasingly relevant to both the machine learning and statistics communities. Here we introduce the first fully distributed MCMC algorithm based on stochastic gradients. We argue that stochastic gradient MCMC algorithms are particularly suited for distributed inference because individual chains can draw minibatches from their local pool of data for a flexible amount of time before jumping to or syncing with other chains. This greatly reduces communication overhead and allows adaptive load balancing. Our experiments for LDA on Wikipedia and Pubmed show that relative to the state of the art in distributed MCMC we reduce compute time from 27 hours to half an hour in order to reach the same perplexity level.

S52 Kernel Adaptive Metropolis-Hastings

Dino Sejdinovic; Heiko Strathmann; Maria Lomeli Garcia; Christophe Andrieu; Arthur Gretton

A Kernel Adaptive Metropolis-Hastings algorithm is introduced, for the purpose of sampling from a target distribution with strongly nonlinear support. The algorithm embeds the trajectory of the Markov chain into a reproducing kernel Hilbert space (RKHS), such that the feature space covariance of the samples informs the choice of proposal. The procedure is computationally efficient and straightforward to implement, since the RKHS moves can be integrated out analytically: our proposal distribution in the original space is a normal distribution whose mean and covariance depend on where the current sample lies in the support of the target distribution, and adapts to its local covariance structure. Furthermore, the procedure requires neither gradients nor any other higher order information about the target, making it particularly attractive for contexts such as Pseudo-Marginal MCMC. Kernel Adaptive Metropolis-Hastings outperforms competing fixed and adaptive samplers on multivariate, highly nonlinear target distributions, arising in both real-world and synthetic examples.

S53 Stochastic Gradient Hamiltonian Monte Carlo

Tianqi Chen; Emily Fox; Carlos Guestrin

Hamiltonian Monte Carlo (HMC) sampling methods provide a mechanism for defining distant proposals with high acceptance probabilities in a Metropolis-Hastings framework, enabling more efficient exploration of the state space than standard random-walk proposals. The popularity of such methods has grown significantly in recent years. However, a limitation of HMC methods is the required gradient computation for simulation of the Hamiltonian dynamical system-such computation is infeasible in problems involving a large sample size or streaming data. Instead, we must rely on a noisy gradient estimate computed from a subset of the data. In this paper, we explore the properties of such a stochastic gradient HMC approach. Surprisingly, the natural implementation of the stochastic approximation can be arbitrarily bad. To address this problem we introduce a variant that uses second-order Langevin dynamics with a friction term that counteracts the effects of the noisy gradient, maintaining the desired target distribution as the invariant distribution. Results on simulated data validate our theory. We also provide an application of our methods to a classification task using neural networks and to online Bayesian matrix factorization.

S54 A Compilation Target for Probabilistic Programming Languages

Brooks Paige; Frank Wood

Forward inference techniques such as sequential Monte Carlo and particle Markov chain Monte Carlo for probabilistic programming can be implemented in any programming language by creative use of standardized operating system functionality including processes, forking, mutexes, and shared memory. Exploiting this we have defined, developed, and tested a probabilistic programming language intermediate representation language we call probabilistic C, which itself can be compiled to machine code by standard compilers and linked to operating system libraries yielding an efficient, scalable, portable probabilistic programming compilation target. This opens up a new hardware and systems research path for optimizing probabilistic programming systems.

14:00 - Track D - Statistical Methods

S55 Generalized Exponential Concentration Inequality for Renyi Divergence Estimation

Shashank Singh; Barnabas Poczos

Estimating divergences between probability distributions in a consistent way is of great importance in many machine learning tasks. Although this is a fundamental problem in nonparametric statistics, to the best of our knowledge there has been no finite sample exponential inequality convergence bound derived for any divergence estimators. The main contribution of our work is to provide such a bound for an estimator of Renyi divergence for a smooth Holder class of densities on the d-dimensional unit cube. We also illustrate our theoretical results with a numerical experiment.

S56 Consistency of Causal Inference under the Additive Noise Model

Samory Kpotufe; Eleni Sgouritsa; Dominik Janzing; Bernhard Schoelkopf

We analyze a family of methods for statistical causal inference from sample under the so-called Additive Noise Model. While most work on the subject has concentrated on establishing the soundness of the Additive Noise Model, the statistical consistency of the resulting inference methods has received little attention. We derive general conditions under which the given family of inference methods consistently infers the causal direction in a nonparametric setting.

S57 The Falling Factorial Basis and Its Statistical Applications

Yu-Xiang Wang; Alex Smola; Ryan Tibshirani

We study a novel spline-like basis, which we name the {\it falling factorial basis}, bearing many similarities to the classic truncated power basis. The advantage of the falling factorial basis is that it enables rapid, linear-time computations in basis matrix multiplication and basis matrix inversion. The falling factorial functions are not actually splines, but are close enough to splines that they provably

retain some of the favorable properties of the latter functions. We examine their application in two problems: trend filtering over arbitrary input points, and a higher-order variant of the two-sample Kolmogorov-Smirnov test.

S58 Concept Drift Detection Through Resampling

Maayan Harel; Shie Mannor; Ran El-Yaniv; Koby Crammer

Detecting changes in data-streams is an important part of enhancing learning quality in dynamic environments. We devise a procedure for detecting concept drifts in data-streams that relies on analyzing the empirical loss of learning algorithms. Our method is based on obtaining statistics from the loss distribution by reusing the data multiple times via resampling. We present theoretical guarantees for the proposed procedure based on the stability of the underlying learning algorithms. Experimental results show that the detection method has high recall and precision, and performs well in the presence of noise.

S59 A Bayesian Wilcoxon signed-rank test based on the Dirichlet process

Alessio Benavoli; Giorgio Corani; Francesca Mangili; Marco Zaffalon; Fabrizio Ruggeri

Bayesian methods are ubiquitous in machine learning. Nevertheless, the analysis of empirical results is typically performed by frequentist tests. This implies dealing with null hypothesis significance tests and p-values, even though the shortcomings of such methods are well known. We propose a nonparametric Bayesian version of the Wilcoxon signed-rank test using a Dirichlet process (DP) based prior. We address in two different ways the problem of how to choose the infinite dimensional parameter that characterizes the DP. The proposed test has all the traditional strengths of the Bayesian approach; for instance, unlike the frequentist tests, it allows verifying the null hypothesis, not only rejecting it, and taking decision which minimize the expected loss. Moreover, one of the solutions proposed to model the infinitedimensional parameter of the DP, allows isolating instances in which the traditional frequentist test is guessing at random. We show results dealing with the comparison of two classifiers using real and simulated data.

Sunday June 22,

14:00 - Track E - Structured Prediction

S60 Marginal Structured SVM with Hidden Variables

Wei Ping; Qiang Liu; Alex Ihler

In this work, we propose the marginal structured SVM (MSSVM) for structured prediction with hidden variables. MSSVM properly accounts for the uncertainty of hidden variables, and can significantly outperform the previously proposed latent structured SVM (LSSVM; Yu & Joachims (2009)) and other state-of-art methods, especially when that uncertainty is large. Our method also results in a smoother objective function, making gradient-based optimization of MSSVMs converge significantly faster than for LSSVMs. We also show that our method consistently outperforms hidden conditional random fields (HCRFs; Quattoni et al. (2007)) on both simulated and real-world datasets. Furthermore, we propose a unified framework that includes both our and several other existing methods as special cases, and provides insights into the comparison of different models in practice.

S61 Scalable Gaussian Process Structured Prediction for Grid Factor Graph Applications

Sebastien Bratieres; Novi Quadrianto; Sebastian Nowozin; Zoubin Ghahramani

Structured prediction is an important and well studied problem with many applications across machine learning. GPstruct is a recently proposed structured prediction model that offers appealing properties such as being kernelised, non-parametric, and supporting Bayesian inference (Bratières et al. 2013). The model places a Gaussian process prior over energy functions which describe relationships between input variables and structured output variables. However, the memory demand of GPstruct is quadratic in the number of latent variables and training runtime scales cubically. This prevents GPstruct from being applied to problems involving grid factor graphs, which are prevalent in computer vision and spatial statistics applications. Here we explore a scalable approach to learning GPstruct models based on ensemble learning, with weak learners (predictors) trained on subsets of the latent variables and bootstrap data, which can easily be distributed. We show experiments with 4M latent variables on image segmentation. Our method outperforms widely-used conditional random field models trained with pseudo-likelihood. Moreover, in image segmentation problems it improves over recent state-of-the-art marginal optimisation methods in terms of predictive performance and uncertainty calibration. Finally, it generalises well on all training set sizes.

S62 High Order Regularization for Semi-Supervised Learning of Structured Output Problems

Yujia Li; Rich Zemel

Semi-supervised learning, which uses unlabeled data to help learn a discriminative model, is especially important for structured output problems, as considerably more effort is needed to label its multidimensional outputs versus standard single output problems. We propose a new max-margin framework for semi-supervised structured output learning, that allows the use of powerful discrete optimization algorithms and high order regularizers defined directly on model predictions for the unlabeled examples. We show that our framework is closely related to Posterior Regularization, and the two frameworks optimize special cases of the same objective. The new framework is instantiated on two image segmentation tasks, using both a graph regularizer and a cardinality regularizer. Experiments also demonstrate that this framework can utilize unlabeled data from a different source than the labeled data to significantly improve performance while saving labeling effort.

S63 Spectral Regularization for Max-Margin Sequence Tagging

Ariadna Quattoni; Borja Balle; Xavier Carreras; Amir Globerson

We frame max-margin learning of latent variable structured prediction models as a convex optimization problem, making use of scoring functions computed by input-output observable operator models. This learning problem can be expressed as an optimization involving a low-rank Hankel matrix that represents the input-output operator model. The direct outcome of our work is a new spectral regularization method for max-margin structured prediction. Our experiments confirm that our proposed regularization framework leads to an effective way of controlling the capacity of structured prediction models.

S64 On Robustness and Regularization of Structural Support Vector Machines

Mohamad Ali Torkamani; Daniel Lowd

Previous analysis of binary SVMs has demonstrated a deep connection between robustness to perturbations over uncertainty sets and regularization of the weights. In this paper, we explore the problem of learning robust models for structured prediction problems. We first formulate the problem of learning robust structural SVMs when there are perturbations in the feature space. We consider two different classes of uncertainty sets for the perturbations: ellipsoidal uncertainty sets and polyhedral uncertainty sets. In both cases, we show that the robust optimization problem is equivalent to the non-robust formulation with an additional regularizer. For the ellipsoidal uncertainty set, the additional regularizer is based on the dual norm of the norm that constrains the ellipsoidal uncertainty. For the polyhedral uncertainty set, we show that the robust optimization problem is equivalent to adding a linear regularizer in a transformed weight space related to the linear constraints of the polyhedron. We also show that these constraint sets can be combined and demonstrate a number of interesting special cases. This represents the first theoretical analysis of robust optimization of structural support vector machines. Our experimental results show that our method outperforms the nonrobust structural SVMs on real world data when the test data distributions is drifted from the training data distribution.

S65 Structured Prediction of Network Response

Hongyu Su; Aristides Gionis; Juho Rousu

We introduce the following network response problem: given a complex network and an action, predict the subnetwork that responds to action, that is, which nodes perform the action and which directed edges relay the action to the adjacent nodes. We approach the problem through max-margin structured learning, in which a compatibility score is learned between the actions and their activated subnetworks. Thus, unlike the most popular influence network approaches, our method, called SPIN, is context-sensitive, namely, the presence, the direction and the dynamics of influences depend on the properties of the actions. The inference problems of finding the highest scoring as well as the worst margin violating networks, are proven to be NP-hard. To solve the problems, we present an approximate inference method through a semi-definite programming relaxation (SDP), as well as a more scalable greedy heuristic algorithm. In our experiments, we demonstrate that taking advantage of the context given by the actions and the network structure leads SPIN to a markedly better predictive performance over competing methods.

14:00 - Track F - Deep Learning and Vision

S66 Recurrent Convolutional Neural Networks for Scene Labeling

Pedro Pinheiro; Ronan Collobert

The goal of the scene labeling task is to assign a class label to each pixel in an image. To ensure a good visual coherence and a high class accuracy, it is essential for a model to capture long range pixel) label dependencies in images. In a feed-forward architecture, this can be achieved simply by considering a sufficiently large input context patch, around each pixel to be labeled. We propose an approach that consists of a recurrent convolutional neural network which allows us to consider a large input context while limiting the capacity of the model.

Contrary to most standard approaches, our method does not rely on any segmentation technique nor any task-specific features. The system is trained in an end-to-end manner over raw pixels, and models complex spatial dependencies with low inference cost. As the context size increases with the built-in recurrence, the system identifies and corrects its own errors. Our approach yields state-of-the-art performance on both the Stanford Background Dataset and the SIFT Flow Dataset, while remaining very fast at test time.

S67 Latent Semantic Representation Learning for Scene Classification

Xin Li; Yuhong Guo

The performance of machine learning methods is heavily dependent on the choice of data representation. In real world applications such as scene recognition problems, the widely used low-level input features can fail to explain the high-level semantic label concepts. In this work, we address this problem by proposing a novel patch-based latent variable model to integrate latent contextual representation learning and classification model training in one joint optimization framework. Within this framework, the latent layer of variables bridge the gap between inputs and outputs by providing discriminative explanations for the semantic output labels, while being predictable from the low-level input features. Experiments conducted on standard scene recognition tasks demonstrate the efficacy of the proposed approach, comparing to the state-of-the-art scene recognition methods.

S68 DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition

Jeff Donahue; Yangqing Jia; Oriol Vinyals; Judy Hoffman; Ning Zhang; Eric Tzeng; Trevor Darrell

We evaluate whether features extracted from the activation of a deep convolutional network trained in a fully supervised fashion on a large, fixed set of object recognition tasks can be re-purposed to novel generic tasks. Our generic tasks may differ significantly from the originally trained tasks and there may be insufficient labeled or unlabeled data to conventionally train or adapt a deep architecture to the new tasks. We investigate and visualize the semantic clustering of deep convolutional features with respect to a variety of such tasks, including scene recognition, domain adaptation, and fine-grained recognition challenges. We compare the efficacy of relying on various network levels to define a fixed feature, and report novel results that significantly outperform the state-of-the-art on several important vision challenges. We are releasing DeCAF, an open-source implementation of these deep convolutional activation features, along with all associated network parameters to enable vision researchers to be able to conduct experimentation with deep representations across a range of visual concept learning paradigms.

S69 Hierarchical Conditional Random Fields for Outlier Detection: An Application to Detecting Epileptogenic Cortical Malformations

Bilal Ahmed; Thomas Thesen; Karen Blackmon; Yijun Zhao; Orrin Devinsky; Ruben Kuzniecky; Carla Brodley

We cast the problem of detecting and isolating regions of abnormal cortical tissue in the MRIs of epilepsy patients in an image segmentation framework. Employing a multiscale approach we divide the surface images into segments of different sizes and then classify each segment as being an outlier, by comparing it to the same region across controls. The final classification is obtained by fusing the outlier probabilities obtained at multiple scales using a tree-structured hierarchical conditional random field (HCRF).

The proposed method correctly detects abnormal regions in 90% of patients whose abnormality was detected via routine visual inspection of their clinical MRI. More importantly, it detects abnormalities in 80% of patients whose abnormality escaped visual inspection by expert radiologists.

S70 Stable and Efficient Representation Learning with Nonnegativity Constraints

Tsung-Han Lin; H. T. Kung

Orthogonal matching pursuit (OMP) is an efficient approximation algorithm for computing sparse representations. However, prior research has shown that the representations computed by OMP may be of inferior quality, as they deliver suboptimal classification accuracy on several im- age datasets. We have found that this problem is caused by OMP's relatively weak stability under data variations, which leads to unreliability in supervised classifier training. We show that by imposing a simple nonnegativity constraint, this nonnegative variant of OMP (NOMP) can mitigate OMP's stability issue and is resistant to noise overfitting. In this work, we provide extensive analysis and experimental results to examine and validate the stability advantage of NOMP. In our experiments, we use a multi-layer deep architecture for representation learning, where we use K-means for feature learning and NOMP for representation encoding. The resulting learning framework is not only efficient and scalable to large feature dictionaries, but also is robust against input noise. This framework achieves the state-of-the-art accuracy on the STL-10 dataset.

S71 Learning by Stretching Deep Networks

Gaurav Pandey; Ambedkar Dukkipati

In recent years, deep architectures have gained a lot of prominence for learning complex AI tasks because of their capability to incorporate complex variations in data within the model. However, these models often need to be trained for a long time in order to obtain good results. In this paper, we propose a technique, called `stretching', that allows the same models to perform considerably better with very little training. We show that learning can be done tractably, even when the weight matrix is stretched to infinity, for some specific models. We also study tractable algorithms for implementing stretching in deep convolutional architectures in an iterative manner and derive bounds for its convergence. Our experimental results suggest that the proposed stretched deep convolutional networks are capable of achieving good performance for many object recognition tasks. More importantly, for a fixed network architecture, one can achieve much better accuracy using stretching rather than learning the weights using backpropagation.

Sunday June 22,

16:20 - Track A - Matrix Completion and Graphs

S72 Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery

Cun Mu; Bo Huang; John Wright; Donald Goldfarb

Recovering a low-rank tensor from incomplete information is a recurring problem in signal processing and machine learning. The most popular convex relaxation of this problem minimizes the sum of the nuclear norms (SNN) of the unfolding matrices of the tensor. We show that this approach can be substantially suboptimal: reliably recovering a K-way $n$$\times$$n$$\times$$\cdots$$\times n$ tensor of Tucker rank $(r, r, \ldots, r)$ from Gaussian measurements requires $\Omega(r n^{K-1})$ observations.

In contrast, a certain (intractable) nonconvex formulation needs only $O(r^K + nrK)$ observations. We introduce a simple, new convex relaxation, which partially bridges this gap. Our new formulation succeeds with $O(r^{\lfloor K/2 \rfloor}n^{\lceil K/2 \rceil})$ observations. The lower bound for the SNN model follows from our new result on recovering signals with multiple structures (e.g. sparse, low rank), which indicates the significant suboptimality of the common approach of minimizing the sum of individual sparsity inducing norms (e.g. ℓ_1, nuclear norm). Our new tractable formulation for low-rank tensor recovery shows how the sample complexity can be reduced by designing convex regularizers that exploit several structures jointly.

S73 Near-Optimal Joint Object Matching via Convex Relaxation

Yuxin Chen; Leonidas Guibas; Qixing Huang

Joint object matching aims at aggregating information from a large collection of similar instances (e.g. images, graphs, shapes) to improve the correspondences computed between pairs of objects, typically by exploiting global map compatibility. Despite some practical advances on this problem, from the theoretical point of view, the error-correction ability of existing algorithms are limited by a constant barrier --- none of them can provably recover the correct solution when more than a constant fraction of input correspondences are corrupted. Moreover, prior approaches focus mostly on fully similar objects, while it is practically more demanding and realistic to match instances that are only partially similar to each other. In this paper, we propose an algorithm to jointly match multiple objects that exhibit only partial similarities, where the provided pairwise feature correspondences can be densely corrupted. By encoding a consistent partial map collection into a 0-1 semidefinite matrix, we attempt recovery via a two-step procedure, that is, a spectral technique followed by a parameter-free convex program called MatchLift. Under a natural randomized model, MatchLift exhibits near-optimal error-correction ability, i.e. it guarantees the recovery of the ground-truth maps even when a dominant fraction of the inputs are randomly corrupted. We evaluate the proposed algorithm on various benchmark data sets including synthetic examples and real-world examples, all of which confirm the practical applicability of the proposed algorithm.

S74 Coherent Matrix Completion

Yudong Chen; Srinadh Bhojanapalli; Sujay Sanghavi; Rachel Ward

Matrix completion concerns the recovery of a low-rank matrix from a subset of its revealed entries, and nuclear norm minimization has emerged as an effective surrogate for this combinatorial problem. Here, we show that nuclear norm minimization can recover an arbitrary $n \times n$ matrix of rank r from O(nr log^2(n)) revealed entries, provided that revealed entries are drawn proportionally to the local row and column coherences (closely related to leverage scores) of the underlying matrix. Our results are order-optimal up to logarithmic factors, and extend existing results for nuclear norm minimization which require strong incoherence conditions on the types of matrices that can be recovered, due to assumed uniformly distributed revealed entries. We further provide extensive numerical evidence that a proposed two-phase sampling algorithm can perform nearly as well as local-coherence sampling and without requiring a priori knowledge of the matrix coherence structure. Finally, we apply our results to quantify how weighted nuclear norm minimization can improve on unweighted minimization given an arbitrary set of sampled entries.

S75 Universal Matrix Completion

Srinadh Bhojanapalli; Prateek Jain

The problem of low-rank matrix completion has recently generated a lot of interest leading to several results that offer exact solutions to the problem. However, in order to do so, these methods make assumptions that can be quite restrictive in practice. More specifically, the methods assume that: a) the observed indices are sampled uniformly at random, and b) for every new matrix, the observed indices are sampled \emph{afresh}. In this work, we address these issues by providing a universal recovery guarantee for matrix completion that works for a variety of sampling schemes. In particular, we show that if the set of sampled indices come from the edges of a bipartite graph with large spectral gap (i.e. gap between the first and the second singular value), then the nuclear norm minimization based method exactly recovers all low-rank matrices that satisfy certain incoherence properties.Moreover, we also show that under certain stricter incoherence conditions, $O(nr^2)$ uniformly sampled entries are enough to recover any rank-r $n\times n$ matrix, in contrast to the $O(nr\log n)$ sample complexity required by other matrix completion algorithms as well as existing analyses of the nuclear norm method.

S76 Exponential Family Matrix Completion under Structural Constraints

Suriya Gunasekar; Pradeep Ravikumar; Joydeep Ghosh

We consider the matrix completion problem of recovering a structured matrix from noisy and partial measurements. Recent works have proposed tractable estimators with strong statistical guarantees for the case where the underlying matrix is low--rank, and the measurements consist of a subset, either of the exact individual entries, or of the entries perturbed by additive Gaussian noise, which is thus implicitly suited for thin--tailed continuous data. Arguably, common applications of matrix completion require estimators for (a) heterogeneous data--types, such as skewed--continuous, count, binary, etc., (b) for heterogeneous noise models (beyond Gaussian), which capture varied uncertainty in the measurements, and (c) heterogeneous structural constraints beyond low--rank, such as block--sparsity, or a superposition structure of low--rank plus elementwise sparseness, among others. In this paper, we provide a vastly unified framework for generalized matrix completion by considering a matrix completion setting wherein the matrix entries are sampled from any member of the rich family of \textit{exponential family distributions}; and impose general structural constraints on the underlying matrix, as captured by a general regularizer $\mathcal{R}(.)$. We propose a simple convex regularized M--estimator for the generalized framework, and provide a unified and novel statistical analysis for this general class of estimators. We finally corroborate our theoretical results on simulated datasets.

S77 A Consistent Histogram Estimator for Exchangeable Graph Models

Stanley Chan; Edoardo Airoldi

Exchangeable graph models (ExGM) subsume a number of popular network models. The mathematical object that characterizes an ExGM is termed a graphon. Finding scalable estimators of graphons, provably consistent, remains an open issue. In this paper, we propose a histogram estimator of a graphon that is provably consistent and numerically efficient. The proposed estimator is based on a sorting-and-smoothing (SAS) algorithm, which first sorts the empirical degree of a graph, then smooths the sorted graph using total variation minimization. The consistency of the SAS algorithm is proved by leveraging sparsity concepts from compressed sensing.

16:20 - Track B - Learning Theory I

S78 Concentration in unbounded metric spaces and algorithmic stability

Aryeh Kontorovich

We prove an extension of McDiarmid's inequality for metric spaces with unbounded diameter. To this end, we introduce the notion of the {\em subgaussian diameter}, which is a distribution-dependent refinement of the metric diameter. Our technique provides an alternative approach to that of Kutin and Niyogi's method of weakly difference-bounded functions, and yields nontrivial, dimension-free results in some interesting cases where the former does not. As an application, we give apparently the first generalization bound in the algorithmic stability setting that holds for unbounded loss functions. This yields a novel risk bound for some regularized metric regression algorithms. We give two extensions of the basic concentration result. The first enables one to replace the independence assumption by appropriate strong mixing. The second generalizes the subgaussian technique to other Orlicz norms.

S79 Heavy-tailed regression with a generalized median-of-means

Daniel Hsu; Sivan Sabato

This work proposes a simple and computationally efficient estimator for linear regression, and other smooth and strongly convex loss minimization problems. We prove loss approximation guarantees that hold for general distributions, including those with heavy tails. All prior results only hold for estimators which either assume bounded or subgaussian distributions, require prior knowledge of distributional properties, or are not known to be computationally tractable. In the special case of linear regression with possibly heavy-tailed responses and with bounded and well-conditioned covariates in d-dimensions, we show that a random sample of size $\tilde{O}(d\log(1/\delta))$ suffices to obtain a constant factor approximation to the optimal loss with probability $1-\delta$, a minimax optimal sample complexity up to log factors. The core technique used in the proposed estimator is a new generalization of the median-of-means estimator to arbitrary metric spaces.

S80 Learnability of the Superset Label Learning Problem

Liping Liu; Thomas Dietterich

In the Superset Label Learning (SLL) problem, weak supervision is provided in the form of a {\it superset} of labels that contains the true label. If the classifier predicts a label outside of the superset, it commits a {\it superset error}. Most existing SLL algorithms learn a multiclass classifier by minimizing the superset error. However, only limited theoretical analysis has been dedicated to this approach. In this paper, we analyze Empirical Risk Minimizing learners that use the superset error as the empirical risk measure. SLL data can arise either in the form of independent instances or as multiple-instance bags. For both scenarios, we give the conditions for ERM learnability and sample complexity for the realizable case.

S81 Maximum Margin Multiclass Nearest Neighbors

Aryeh Kontorovich; Roi Weiss

We develop a general framework for margin-based multicategory classification in metric spaces. The basic work-horse is a margin-regularized version of the nearest-neighbor classifier. We prove generalization bounds that match the state of the art in sample size n and significantly improve the dependence on the number of classes k. Our point of departure is a nearly Bayes-optimal finite-sample risk bound independent of k. Although k-free, this bound is unregularized and non-adaptive, which motivates our main result: Rademacher and scale-sensitive margin bounds with a logarithmic dependence on k. As the best previous risk estimates in this setting were of order $\sqrt k$, our bound is exponentially sharper. From the algorithmic standpoint, in doubling metric spaces our classifier may be trained on n examples in $O(n^2\log n)$ time and evaluated on new points in $O(\log n)$ time.

S82 Sample Efficient Reinforcement Learning with Gaussian Processes

Robert Grande; Thomas Walsh; Jonathan How

This paper derives sample complexity results for using Gaussian Processes (GPs) in both model-based and model-free reinforcement learning (RL). We show that GPs are KWIK learnable, proving for the first time that a model-based RL approach using GPs, GP-Rmax, is sample efficient (PAC-MDP). However, we then show that previous approaches to model-free RL using GPs take an exponential number of steps to find an optimal policy, and are therefore not sample efficient. The third and main contribution is the introduction of a model-free RL algorithm using GPs, DGPQ, which is sample efficient and, in contrast to model-based algorithms, capable of acting in real time, as demonstrated on a five-dimensional aircraft simulator.

S83 Scaling Up Approximate Value Iteration with Options: Better Policies with Fewer Iterations

Timothy Mann; Shie Mannor

We show how options, a class of control structures encompassing primitive and temporally extended actions, can play a valuable role in planning in MDPs with continuous state-spaces. Analyzing the convergence rate of Approximate Value Iteration with options reveals that for pessimistic initial value function estimates, options can speed up convergence compared to planning with only primitive actions even when the temporally extended actions are suboptimal and sparsely scattered throughout the state-space. Our experimental results in an optimal replacement task and a complex inventory management task demonstrate the potential for options to speed up convergence in practice. We show that options induce faster convergence to the optimal value function, which implies deriving better policies with fewer iterations.

16:20 - Track C - Clustering and Nonparametrics

S84 Von Mises-Fisher Clustering Models

Siddharth Gopal; Yiming Yang

This paper proposes a suite of models for clustering high-dimensional data on a unit sphere based on Von Mises-Fisher (vMF) distribution and for discovering more intuitive clusters than existing approaches. The proposed models include a) A Bayesian formulation of vMF mixture that enables information sharing among clusters, b) a Hierarchical vMF mixture that provides multi-scale shrinkage and tree structured view of the data and c) a Temporal vMF mixture that captures evolution of clusters in temporal data. For posterior inference, we develop fast variational methods as well as collapsed Gibbs sampling techniques for all three models. Our experiments on six datasets provide strong empirical support in favour of vMF based clustering models over other popular tools such as K-means, Multinomial Mixtures and Latent Dirichlet Allocation.

S85 Online Bayesian Passive-Aggressive Learning

Tianlin Shi; Jun Zhu

Online Passive-Aggressive (PA) learning is an effective framework for performing max-margin online learning. But the deterministic formulation and estimated single large-margin model could limit its capability in discovering descriptive structures underlying complex data. This paper presents online Bayesian Passive-Aggressive (BayesPA) learning, which subsumes the online PA and extends naturally to incorporate latent variables and perform nonparametric Bayesian inference, thus providing great flexibility for explorative analysis. We apply BayesPA to topic modeling and derive efficient online learning algorithms for max-margin topic models. We further develop nonparametric methods to resolve the number of topics. Experimental results on real datasets show that our approaches significantly improve time efficiency while maintaining comparable results with the batch counterparts.

S86 Bayesian Nonparametric Multilevel Clustering with Group-Level Contexts

Tien Vu Nguyen; Dinh Phung; Xuanlong Nguyen; Swetha Venkatesh; Hung Bui

We present a Bayesian nonparametric framework for multilevel clustering which utilizes group-level context information to simultaneously discover low-dimensional structures of the group contents and partitions groups into clusters. Using the Dirichlet process as the building block, our model constructs a product base-measure with a nested structure to accommodate content and context observations at multiple levels. The proposed model possesses properties that link the nested Dirichlet processes (nDP) and the Dirichlet process mixture models (DPM) in an interesting way: integrating out all contents results in the DPM over contexts, whereas integrating out group-specific contexts results in the nDP mixture over content variables. We provide a Polya-urn view of the model and an efficient collapsed Gibbs inference procedure. Extensive experiments on real-world datasets demonstrate the advantage of utilizing context information via our model in both text and image domains.

S87 Hierarchical Dirichlet Scaling Process

Dongwoo Kim; Alice Oh

We present the hierarchical Dirichlet scaling process (HDSP), a Bayesian nonparametric mixed membership model for multi-labeled data. We construct the HDSP based on the gamma representation of the hierarchical Dirichlet process (HDP) which allows scaling the mixture components. With such construction, HDSP allocates a latent location to each label and mixture component in a space, and uses the distance between them to guide membership probabilities. We develop a variational Bayes algorithm for the approximate posterior inference of the HDSP. Through experiments on synthetic datasets as well as datasets of newswire, medical journal articles, and Wikipedia, we show that the HDSP results in better predictive performance than HDP, labeled LDA and partially labeled LDA.

S88 Fast Computation of Wasserstein Barycenters

Marco Cuturi; Arnaud Doucet

We present new algorithms to compute the mean of a set of N empirical probability measures under the optimal transport metric. This mean, known as the Wasserstein barycenter~\citep{agueh2011barycenters,rabin2012}, is the measure that minimizes the sum of its Wasserstein distances to each element in that set. We argue through a simple example that Wasserstein barycenters have appealing properties that differentiate them from other barycenters proposed recently, which all build on kernel smoothing and/or Bregman divergences. Two original algorithms are proposed that require the repeated computation of primal and dual optimal solutions of transport problems. However direct implementation of these algorithms is too costly as optimal transports are notoriously computationally expensive. Extending the work of \citet{cuturi2013sinkhorn}, we smooth both the primal and dual of the optimal transport problem to recover fast approximations of the primal and dual optimal solutions. We apply these algorithms to the visualization of perturbed images and to a clustering problem.

S89 Max-Margin Infinite Hidden Markov Models

Aonan Zhang; Jun Zhu; Bo Zhang

Infinite hidden Markov models (iHMMs) are nonparametric Bayesian extensions of hidden Markov models (HMMs) with an infinite number of states. Though flexible in describing sequential data, the generative formulation of iHMMs could limit their discriminative ability in sequential prediction tasks. Our paper introduces max-margin infinite HMMs (M2iHMMs), new infinite HMMs that explore the max-margin principle for discriminative learning. By using the theory of Gibbs classifiers and data augmentation, we develop efficient beam sampling algorithms without making restricting mean-field assumptions or truncated approximation. For single variate classification, M2iHMMs reduce to a new formulation of DP mixtures of max-margin machines. Empirical results on synthetic and real data sets show that our methods obtain superior performance than other competitors in both single variate classification and sequential prediction tasks.

16:20 - Track D - Active Learning

S90 Diagnosis determination: decision trees optimizing simultaneously worst and expected testing cost

Ferdinando Cicalese; Eduardo Laber; Aline Medeiros Saettler

In several applications of automatic diagnosis and active learning a central problem is the evaluation of a discrete function by adaptively querying the values of its variables until the values read uniquely determine the value of the function. In general reading the value of a variable is done at the expense of some cost (computational or possibly a fee to pay the corresponding experiment). The goal is to design a strategy for evaluating the function incurring little cost (in the worst case or in expectation according to a prior distribution on the possible variables' assignments). We provide an algorithm that builds a strategy (decision tree) with both expected cost and worst cost which are at most an $O(\log n)$ factor away from, respectively, the minimum possible expected cost and the minimum possible worst cost. Our algorithm provides the best possible approximation simultaneously with respect to both criteria. In fact, there is no algorithm that can guarantee $o(\log n)$ approximation, under the assumption that ${\cal P} \neq {\cal NP}$.

S91 Nonmyopic ϵ-Bayes-Optimal Active Learning of Gaussian Processes

Trong Nghia Hoang; Bryan Kian Hsiang Low; Patrick Jaillet; Mohan Kankanhalli

A fundamental issue in active learning of Gaussian processes is that of the exploration-exploitation trade-off. This paper presents a novel nonmyopic ϵ-Bayes-optimal active learning (ϵ-BAL) approach that jointly and naturally optimizes the trade-off. In contrast, existing works have primarily developed myopic/greedy algorithms or performed exploration and exploitation separately. To perform active learning in real time, we then propose an anytime algorithm based on ϵ-BAL with performance guarantee and empirically demonstrate using synthetic and real-world datasets that, with limited budget, it outperforms the state-of-the-art algorithms.

S92 Hard-Margin Active Linear Regression

Zohar Karnin; Elad Hazan

We consider the fundamental problem of linear regression in which the designer can actively choose observations. This model naturally captures various experiment design settings in medical experiments, ad placement problems and more. Whereas previous literature addresses the soft-margin or mean-square-error variants of the problem, we consider a natural machine learning hard-margin criterion. In this setting, we show that active learning admits significantly better sample complexity bounds than the passive learning counterpart, and give efficient algorithms that attain near-optimal bounds.

S93 Active Transfer Learning under Model Shift

Xuezhi Wang; Tzu-Kuo Huang; Jeff Schneider

Transfer learning algorithms are used when one has sufficient training data for one supervised learning

ask (the source task) but only very limited training data for a second task (the target task) that is similar but not identical to the first. These algorithms use varying assumptions about the similarity between the tasks to carry information from the source to the target task. Common assumptions are that only certain specific marginal or conditional distributions have changed while all else remains the same. Alternatively, if one has only the target task, but also has the ability to choose a limited amount of additional training data to collect, then active learning algorithms are used to make choices which will most improve performance on the target task. These algorithms may be combined into active transfer learning, but previous efforts have had to apply the two methods in sequence or use restrictive transfer assumptions. We propose two transfer learning algorithms that allow changes in all marginal and conditional distributions but assume the changes are smooth in order to achieve transfer between the tasks. We then propose an active learning algorithm for the second method that yields a combined active transfer learning algorithm. We demonstrate the algorithms on synthetic functions and a real-world task on estimating the yield of vineyards from images of the grapes.

S94 Gaussian Process Optimization with Mutual Information

Emile Contal; Vianney Perchet; Nicolas Vayatis

In this paper, we analyze a generic algorithm scheme for sequential global optimization using Gaussian processes. The upper bounds we derive on the cumulative regret for this generic algorithm improve by an exponential factor the previously known bounds for algorithms like GP-UCB. We also introduce the novel Gaussian Process Mutual Information algorithm (GP-MI), which significantly improves further these upper bounds for the cumulative regret. We confirm the efficiency of this algorithm on synthetic and real tasks against the natural competitor, GP-UCB, and also the Expected Improvement heuristic.

16:20 - Track E - Optimization I

S95 An Adaptive Accelerated Proximal Gradient Method and its Homotopy Continuation for Sparse Optimization

Qihang Lin; Lin Xiao

We first propose an adaptive accelerated proximal gradient(APG) method for minimizing strongly convex composite functions with unknown convexity parameters. This method incorporates a restarting scheme to automatically estimate the strong convexity parameter and achieves a nearly optimal iteration complexity. Then we consider the ℓ1-regularized least-squares (ℓ1-LS) problem in the high-dimensional setting. Although such an objective function is not strongly convex, it has restricted strong convexity over sparse vectors. We exploit this property by combining the adaptive APG method with a homotopy continuation scheme, which generates a sparse solution path towards optimality. This method obtains a global linear rate of convergence and its overall iteration complexity has a weaker dependency on the restricted condition number than previous work.

S96 Finito: A faster, permutable incremental gradient method for big data problems

Aaron Defazio; Justin Domke; Tiberio Caetano

Recent advances in optimization theory have shown that smooth strongly convex finite sums can be

minimized faster than by treating them as a black box "batch" problem. In this work we introduce a new method in this class with a theoretical convergence rate four times faster than existing methods, for sums with sufficiently many terms. This method is also amendable to a sampling without replacement scheme that in practice gives further speed-ups. We give empirical results showing state of the art performance.

S97 Asynchronous Distributed ADMM for Consensus Optimization

Ruiliang Zhang; James Kwok

Distributed optimization algorithms are highly attractive for solving big data problems. In particular, many machine learning problems can be formulated as the global consensus optimization problem, which can then be solved in a distributed manner by the alternating direction method of multipliers (ADMM) algorithm. However, this suffers from the straggler problem as its updates have to be synchronized. In this paper, we propose an asynchronous ADMM algorithm by using two conditions to control the asynchrony: partial barrier and bounded delay. The proposed algorithm has a simple structure and good convergence guarantees (its convergence rate can be reduced to that of its synchronous counterpart). Experiments on different distributed ADMM applications show that asynchrony reduces the time on network waiting, and achieves faster convergence than its synchronous counterpart in terms of the wall clock time.

S98 Fast large-scale optimization by unifying stochastic gradient and quasi-Newton methods

Jascha Sohl-Dickstein; Ben Poole; Surya Ganguli

We present an algorithm for minimizing a sum of functions that combines the computational efficiency of stochastic gradient descent (SGD) with the second order curvature information leveraged by quasi-Newton methods. We unify these disparate approaches by maintaining an independent Hessian approximation for each contributing function in the sum. We maintain computational tractability and limit memory requirements even for high dimensional optimization problems by storing and manipulating these quadratic approximations in a shared, time evolving, low dimensional subspace. This algorithm contrasts with earlier stochastic second order techniques that treat the Hessian of each contributing function as a noisy approximation to the full Hessian, rather than as a target for direct estimation. Each update step requires only a single contributing function or minibatch evaluation (as in SGD), and each step is scaled using an approximate inverse Hessian and little to no adjustment of hyperparameters is required (as is typical for quasi-Newton methods). We experimentally demonstrate improved convergence on seven diverse optimization problems. The algorithm is released as open source Python and MATLAB packages.

S99 Least Squares Revisited: Scalable Approaches for Multi-class Prediction

Alekh Agarwal; Sham Kakade; Nikos Karampatziakis; Le Song; Gregory Valiant

This work provides simple algorithms for multi-class (and multi-label) prediction in settings where both the number of examples n and the data dimension d are relatively large. These robust and parameter free algorithms are essentially iterative least-squares updates and very versatile both in theory and in practice. On the theoretical front, we present several variants with convergence guarantees.

Owing to their effective use of second-order structure, these algorithms are substantially better than first-order methods in many practical scenarios. On the empirical side, we show how to scale our approach to high dimensional datasets, achieving dramatic computational speedups over popular optimization packages such as Liblinear and Vowpal Wabbit on standard datasets (MNIST and CIFAR-10), while attaining state-of-the-art accuracies.

S100 A Statistical Perspective on Algorithmic Leveraging

Ping Ma; Michael Mahoney; Bin Yu

One popular method for dealing with large-scale data sets is sampling. Using the empirical statistical leverage scores as an importance sampling distribution, the method of algorithmic leveraging samples and rescales rows/columns of data matrices to reduce the data size before performing computations on the subproblem. Existing work has focused on algorithmic issues, but none of it addresses statistical aspects of this method. Here, we provide an effective framework to evaluate the statistical properties of algorithmic leveraging in the context of estimating parameters in a linear regression model. In particular, for several versions of leverage-based sampling, we derive results for the bias and variance, both conditional and unconditional on the observed data. We show that from the statistical perspective of bias and variance, neither leverage-based sampling nor uniform sampling dominates the other. This result is particularly striking, given the well-known result that, from the algorithmic perspective of worst-case analysis, leverage-based sampling provides uniformly superior worst-case algorithmic results, when compared with uniform sampling. Based on these theoretical results, we propose and analyze two new leveraging algorithms: one constructs a smaller least-squares problem with ``shrinked" leverage scores (SLEV), and the other solves a smaller and unweighted (or biased) least-squares problem (LEVUNW). The empirical results indicate that our theory is a good predictor of practical performance of existing and new leverage-based algorithms and that the new algorithms achieve improved performance.

16:20 - Track F - Large-Scale Learning

S101 Large-scale Multi-label Learning with Missing Labels

Hsiang-Fu Yu; Prateek Jain; Purushottam Kar; Inderjit Dhillon

The multi-label classification problem has generated significant interest in recent years. However, existing approaches do not adequately address two key challenges: (a) scaling up to problems with a large number (say millions) of labels, and (b) handling data with missing labels. In this paper, we directly address both these problems by studying the multi-label problem in a generic empirical risk minimization (ERM) framework. Our framework, despite being simple, is surprisingly able to encompass several recent label-compression based methods which can be derived as special cases of our method. To optimize the ERM problem, we develop techniques that exploit the structure of specific loss functions - such as the squared loss function - to obtain efficient algorithms. We further show that our learning framework admits excess risk bounds even in the presence of missing labels. Our bounds are tight and demonstrate better generalization performance for low-rank promoting trace-norm regularization when compared to (rank insensitive) Frobenius norm regularization. Finally, we present extensive empirical results on a variety of benchmark datasets and show that our methods perform significantly better than existing label compression based methods and can scale up to very large datasets such as a Wikipedia dataset that has more than 200,000 labels.

S102 Dual Query: Practical Private Query Release for High Dimensional Data

Marco Gaboardi; Emilio Jesus Gallego Arias; Justin Hsu; Aaron Roth; Zhiwei Steven Wu

We present a practical, differentially private algorithm for answering a large number of queries on high dimensional datasets. Like all algorithms for this task, ours necessarily has worst-case complexity exponential in the dimension of the data. However, our algorithm packages the computationally hard step into a concisely defined integer program, which can be solved non-privately using standard solvers. We prove accuracy and privacy theorems for our algorithm, and then demonstrate experimentally that our algorithm performs well in practice. For example, our algorithm can efficiently and accurately answer millions of queries on the Netflix dataset, which has over 17,000 attributes; this is an improvement on the state of the art by multiple orders of magnitude.

S103 A Highly Scalable Parallel Algorithm for Isotropic Total Variation Models

Jie Wang; Qingyang Li; Sen Yang; Wei Fan; Peter Wonka; Jieping Ye

Total variation (TV) models are among the most popular and successful tools in signal processing. However, due to the complex nature of the TV term, it is challenging to efficiently compute a solution for large-scale problems. State-of-the-art algorithms that are based on the alternating direction method of multipliers (ADMM) often involve solving large-size linear systems. In this paper, we propose a highly scalable parallel algorithm for TV models that is based on a novel decomposition strategy of the problem domain. As a result, the TV models can be decoupled into a set of small and independent subproblems, which admit closed form solutions. This makes our approach particularly suitable for parallel implementation. Our algorithm is guaranteed to converge to its global minimum. With N variables and n_ p processes, the time complexity is O(N/(epsilon n_p)) to reach an epsilon-optimal solution. Extensive experiments demonstrate that our approach outperforms existing state-of-the-art algorithms, especially in dealing with high-resolution, mega-size images.

S104 Buffer k-d Trees: Processing Massive Nearest Neighbor Queries on GPUs

Fabian Gieseke; Justin Heinermann; Cosmin Oancea; Christian Igel

We present a new approach for combining k-d trees and graphics processing units for nearest neighbor search. It is well known that a direct combination of these tools leads to a non-satisfying performance due to conditional computations and suboptimal memory accesses. To alleviate these problems, we propose a variant of the classical k-d tree data structure, called buffer k-d tree, which can be used to reorganize the search. Our experiments show that we can take advantage of both the hierarchical subdivision induced by k-d trees and the huge computational resources provided by today's many-core devices. We demonstrate the potential of our approach in astronomy, where hundreds of million nearest neighbor queries have to be processed.

S105 Fast Multi-stage Submodular Maximization

Kai Wei; Rishabh Iyer; Jeff Bilmes

We introduce a new multi-stage algorithmic framework for submodular maximization. We are motivated by extremely large scale machine learning problems, where both storing the

whole data for function evaluation and running the standard accelerated greedy algorithm are prohibitive. We propose a multi-stage framework (called MultGreed), where at each stage we apply an approximate greedy procedure to maximize surrogate submodular functions. The surrogates serve as proxies for a target submodular function but require less memory and are easy to evaluate. We theoretically analyze the performance guarantee of the multi-stage framework, and give examples on how to design instances of MultGreed for a broad range of natural submodular functions. We show that MultGreed performs very close to the standard greedy algorithm, given appropriate surrogate functions, and argue how our framework can easily be integrated with distributive algorithms for optimization. We complement our theory by empirically evaluating on several real world problems, including data subset selection on millions of speech samples, where MultGreed yields at least a thousand times speedup and superior results over the state-of-the-art selection methods.

S106 Multi-label Classification via Feature-aware Implicit Label Space Encoding

Zijia Lin; Guiguang Ding; Mingqing Hu; Jianmin Wang

To tackle a multi-label classification problem with many classes, recently label space dimension reduction (LSDR) is proposed. It encodes the original label space to a low-dimensional latent space and uses a decoding process for recovery. In this paper, we propose a novel method termed FaIE to perform LSDR via Feature-aware Implicit label space Encoding. Unlike most previous work, the proposed FaIE makes no assumptions about the encoding process and directly learns a code matrix, i.e. the encoding result of some implicit encoding function, and a linear decoding matrix. To learn both matrices, FaIE jointly maximizes the recoverability of the original label space from the latent space, and the predictability of the latent space from the feature space, thus making itself feature-aware. FaIE can also be specified to learn an explicit encoding function, and extended with kernel tricks to handle non-linear correlations between the feature space and the latent space. Extensive experiments conducted on benchmark datasets well demonstrate its effectiveness.

MONDAY

June 23, 2014 at 8:30am

Title: Algorithmic Trading and Machine Learning
Convention Hall No.1

Keynote Speaker, Michael Kearns, University of Pennsylvania

Abstract:
Traditional financial markets have undergone rapid technological change due to increased automation and the introduction of new mechanisms. Such changes have brought with them challenging new problems in algorithmic trading, many of which invite a machine learning approach. In this talk I will examine several algorithmic trading problems, focusing on their novel ML aspects, including limiting market impact, dealing with censored data, and incorporating risk considerations.

Bio:
Michael Kearns is a professor in the Computer and Information Science department at the University of Pennsylvania, where he holds the National Center Chair and has joint appointments in the Wharton School. He is founder of Penn's Networked and Social Systems Engineering (NETS) program (www.nets.upenn.edu),and director of Penn's Warren Center for Network and Data Sciences (www.warrencenter.upenn.edu). His research interests include topics in machine learning, algorithmic game theory, social networks, and computational finance. He has consulted extensively in the technology and finance industries.

Monday June 23

10:30 - Track A - Latent Variable Models (Room 305A)

M1 A Discriminative Latent Variable Model for Online Clustering
Rajhans Samdani; Kai-Wei Chang; Dan Roth

M2 Exchangeable Variable Models
Mathias Niepert; Pedro Domingos

M3 Learning Latent Variable Gaussian Graphical Models
Zhaoshi Meng; Brian Eriksson; Al Hero

M4 Latent Variable Copula Inference for Bundle Pricing from Retail Transaction Data
Benjamin Letham; Wei Sun; Anshul Sheopuri

M5 Affinity Weighted Embedding
Jason Weston; Ron Weiss; Hector Yee

M6 Learning the Irreducible Representations of Commutative Lie Groups
Taco Cohen; Max Welling

Monday June 23

10:30 - Track B - Online Learning and Planning (Room 307)

M7 Covering Number for Efficient Heuristic-based POMDP Planning
Zongzhang Zhang; David Hsu; Wee Sun Lee

M8 Learning Complex Neural Network Policies with Trajectory Optimization
Sergey Levine; Vladlen Koltun

M9 A Physics-Based Model Prior for Object-Oriented MDPs
Jonathan Scholz; Martin Levihn; Charles Isbell

M10 Online Multi-Task Learning for Policy Gradient Methods
Haitham Bou Ammar; Eric Eaton; Paul Ruvolo; Matthew Taylor

M11 Pursuit-Evasion Without Regret, with an Application to Trading
Lili Dworkin; Michael Kearns; Yuriy Nevmyvaka

M12 Adaptivity and Optimism: An Improved Exponentiated Gradient Algorithm
Jacob Steinhardt; Percy Liang

Monday June 23

10:30 - Track C - Clustering (Room 201A)

M13 Demystifying Information-Theoretic Clustering
Greg Ver Steeg; Aram Galstyan; Fei **Sha;**

M14 Simon DeDeo Clustering in the Presence of Background Noise
Shai Ben-David; Nika Haghtalab

M15 Hierarchical Quasi-Clustering Methods for Asymmetric Networks
Gunnar Carlsson; Facundo Mémoli; Alejandro Ribeiro; Santiago Segarra

M16 Local algorithms for interactive clustering
Pranjal Awasthi; Maria Balcan; Konstantin Voevodski

M17 Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance
Simone Romano; James Bailey; Vinh Nguyen; Karin Verspoor

M18 A Single-Pass Algorithm for Efficiently Recovering Sparse Cluster Centers of High-dimensional Data
Jinfeng Yi; Lijun Zhang; Jun Wang; Rong Jin; Anil Jain

Monday June 23

10:30 - Track D - Metric Learning and Feature Selection (Room 201B)

M19 Large-Margin Metric Learning for Constrained Partitioning Problems
Rémi Lajugie; Francis Bach; Sylvain Arlot

M20 Robust Distance Metric Learning via Simultaneous L1-Norm Minimization and Maximization
Hua Wang; Feiping Nie; Heng Huang

M21Efficient Learning of Mahalanobis Metrics for Ranking
Daryl Lim; Gert Lanckriet

M22 Stochastic Neighbor Compression
Matt Kusner; Stephen Tyree; Kilian Weinberger; Kunal Agrawal

M23 Large-margin Weakly Supervised Dimensionality Reduction
Chang Xu; Dacheng Tao; Chao Xu; Yong Rui

M24 Sparse meta-Gaussian information bottleneck
Melanie Rey; Volker Roth; Thomas Fuchs

Monday June 23

10:30 - Track E - Optimization II (Room 201C)

M25 Fast Stochastic Alternating Direction Method of Multipliers
Wenliang Zhong; James Kwok

M26 Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization
Shai Shalev-Shwartz; Tong Zhang

M27 An Asynchronous Parallel Stochastic Coordinate Descent Algorithm
Ji Liu; Steve Wright; Christopher Re; Victor Bittorf; Srikrishna Sridhar

M28 Towards an optimal stochastic alternating direction method of multipliers
Samaneh Azadi; Suvrit Sra

M29 Stochastic Dual Coordinate Ascent with Alternating Direction Method of Multipliers
Taiji Suzuki

M30 Communication-Efficient Distributed Optimization using an Approximate Newton-type Method
Ohad Shamir; Nati Srebro; Tong Zhang

Monday June 23

10:30 - Track F - Neural Language and Speech (Room 305B)

M31 Multimodal Neural Language Models
Ryan Kiros; Ruslan Salakhutdinov; Rich Zemel

M32 Distributed Representations of Sentences and Documents
Quoc Le; Tomas Mikolov

M33 Learning Character-level Representations for Part-of-Speech Tagging
Cicero Dos Santos; Bianca Zadrozny

M34 Compositional Morphology for Word Representations and Language Modelling
Jan Botha; Phil Blunsom

M35 Towards End-To-End Speech Recognition with Recurrent Neural Networks
Alex Graves; Navdeep Jaitly

M36 A Clockwork RNN
Jan Koutnik; Klaus Greff; Faustino Gomez; Juergen Schmidhuber

Monday June 23

14:00 - Track A - Graphical Models and Approximate Inference (Room 305A)

M37 Probabilistic Partial Canonical Correlation Analysis
Yusuke Mukuta; Tatsuya Harada

M38 Min-Max Problems on Factor Graphs
Siamak Ravanbakhsh; Christopher Srinivasa; Brendan Frey; Russell Greiner

M39 Skip Context Tree Switching
Marc Bellemare; Joel Veness; Erik Talvitie

M40 Learning the Parameters of Determinantal Point Process Kernels
Raja Hafiz Affandi; Emily Fox; Ryan Adams; Ben Taskar

M41 Deterministic Anytime Inference for Stochastic Continuous-Time Markov Processes
E. Busra Celikkaya; Christian Shelton

M42 Doubly Stochastic Variational Bayes for non-Conjugate Inference
Michalis Titsias; Miguel Lázaro-Gredilla

Monday June 23

14:00 - Track B - Online Learning I (Room 307)

M43 On the convergence of no-regret learning in selfish routing
Walid Krichene; Benjamin Drighès; Alexandre Bayen

M44 Optimal PAC Multiple Arm Identification with Applications to Crowdsourcing
Yuan Zhou; Xi Chen; Jian Li

M45 Prediction with Limited Advice and Multiarmed Bandits with Paid Observations
Yevgeny Seldin; Peter Bartlett; Koby Crammer; Yasin Abbasi-Yadkori

M46 One Practical Algorithm for Both Stochastic and Adversarial Bandits
Yevgeny Seldin; Aleksandrs Slivkins

M47 A Bayesian Framework for Online Classifier Ensemble
Qinxun Bai; Henry Lam; Stan Sclaroff

M48 Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits
Alekh Agarwal; Daniel Hsu; Satyen Kale; John Langford; Lihong Li; Robert Schapire

Monday June 23

14:00 - Track C - Monte Carlo and Approximate Inference (Room 201A)

M49 Memory (and Time) Efficient Sequential Monte Carlo
Seong-Hwan Jun; Alexandre Bouchard-Côté

M50 Efficient Continuous-Time Markov Chain Estimation
Monir Hajiaghayi; Bonnie Kirkpatrick; Liangliang Wang; Alexandre Bouchard-Côté

M51 Filtering with Abstract Particles
Jacob Steinhardt; Percy Liang

M52 Spherical Hamiltonian Monte Carlo for Constrained Target Distributions
Shiwei Lan; Bo Zhou; Babak Shahbaba

M53 Hamiltonian Monte Carlo Without Detailed Balance
Jascha Sohl-Dickstein; Mayur Mudigonda; Michael DeWeese

M54 Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Ito Process
Issei Sato; Hiroshi Nakagawa

Monday June 23

14:00 - Track D - Method-Of-Moments and Spectral Methods (Room 201B)

M55 Computing Parametric Ranking Models via Rank-Breaking
Hossein Azari Soufiani; David Parkes; Lirong Xia

M56 Learning Mixtures of Linear Classifiers
Yuekai Sun; Stratis Ioannidis; Andrea Montanari

M57 Methods of Moments for Learning Stochastic Languages: Unified Presentation and Empirical Comparison
Borja Balle; William Hamilton; Joelle Pineau

M58 Estimating Latent-Variable Graphical Models using Moments and Likelihoods
Arun Tejasvi Chaganty; Percy Liang

M59 Alternating Minimization for Mixed Linear Regression
Xinyang Yi; Constantine Caramanis; Sujay Sanghavi

M60 Dimension-free Concentration Bounds on Hankel Matrices for Spectral Learning
François Denis; Mattias Gybels; Amaury Habrard

Monday June 23

14:00 - Track E - Boosting and Ensemble Methods (Room 201C)

M61 Boosting with Online Binary Learners for the Multiclass Bandit Problem
Shang-Tse Chen; Hsuan-Tien Lin; Chi-Jen Lu

M62 Narrowing the Gap: Random Forests In Theory and In Practice
Misha Denil; David Matheson; Nando De Freitas

M63 Ensemble Methods for Structured Prediction
Corinna Cortes; Vitaly Kuznetsov; Mehryar Mohri

M64 Deep Boosting
Corinna Cortes; Mehryar Mohri; Umar Syed

M65 Dynamic Programming Boosting for Discriminative Macro-Action Discovery
Leonidas Lefakis; Francois Fleuret

M66 A Convergence Rate Analysis for LogitBoost, MART and Their Variant
Peng Sun; Tong Zhang; Jie Zhou

Monday June 23

14:00 - Track F - Neural Networks and Deep Learning II (Room 305B)

M67 Learning to Disentangle Factors of Variation with Manifold Interaction
Scott Reed; Kihyuk Sohn; Yuting Zhang; Honglak Lee

M68 Marginalized Denoising Auto-encoders for Nonlinear Representations
Minmin Chen; Kilian Weinberger; Fei Sha; Yoshua Bengio

M69 Deep Generative Stochastic Networks Trainable by Backprop
Yoshua Bengio; Eric Laufer; Guillaume Alain; Jason Yosinski

M70 Learning Ordered Representations with Nested Dropout
Oren Rippel; Michael Gelbart; Ryan Adams

M71 Efficient Gradient-Based Inference through Transformations between Bayes Nets and Neural Nets
Diederik Kingma; Max Welling

M72 Signal recovery from Pooling Representations
Joan Bruna Estrach; Arthur Szlam; Yann LeCun

Monday June 23

16:20 - Track A - Matrix Factorization I (Room 305A)

M73 Stochastic Inference for Scalable Probabilistic Modeling of Binary Matrices
Jose Miguel Hernandez-Lobato; Neil Houlsby; Zoubin Ghahramani

M74 Cold-start Active Learning with Robust Ordinal Matrix Factorization
Neil Houlsby; Jose Miguel Hernandez-Lobato; Zoubin Ghahramani

M75 Probabilistic Matrix Factorization with Non-random Missing Data
Jose Miguel Hernandez-Lobato; Neil Houlsby; Zoubin Ghahramani

M76 A Deep Semi-NMF Model for Learning Hidden Representations
George Trigeorgis; Konstantinos Bousmalis; Stefanos Zafeiriou; Bjoern Schuller

M77 Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing
Benjamin Haeffele; Eric Young; Rene Vidal

Monday June 23

16:20 - Track B - Learning Theory II (Room 201A)

M78 Lower Bounds for the Gibbs Sampler over Mixtures of Gaussians
Christopher Tosh; Sanjoy Dasgupta

M79 (Near) Dimension Independent Risk Bounds for Differentially Private Learning
Prateek Jain; Abhradeep Guha Thakurta

M80 Learning Theory and Algorithms for revenue optimization in second price auctions with reserve
Mehryar Mohri; Andres Munoz Medina

M81 Multi-period Trading Prediction Markets with Connections to Machine Learning
Jinli Hu; Amos Storkeyl

M82 Towards Minimax Online Learning with Unknown Time Horizon
Haipeng Luo; Robert Schapire

Monday June 23

16:20 - Track C - Nonparametric Bayes I (Room 305B)

M83 Rectangular Tiling Process
Masahiro Nakano; Katsuhiko Ishiguro; Akisato Kimura; Takeshi Yamada; Naonori Ueda

M84 A reversible infinite HMM using normalised random measures
David Knowles; Zoubin Ghahramani; Konstantina Palla

M85 Scalable Bayesian Low-Rank Decomposition of Incomplete Multiway Tensors
Piyush Rai; Yingjian Wang; Shengbo Guo; Gary Chen; David Dunson; Lawrence Carin

M86 Input Warping for Bayesian Optimization of Non-Stationary Functions
Jasper Snoek; Kevin Swersky; Rich Zemel; Ryan Adams

M87 Beta Diffusion Trees
Creighton Heaukulani; David Knowles; Zoubin GhahramaniM

Monday June 23

16:20 - Track D – Manifolds (Room 201B)

M88 An Information Geometry of Statistical Manifold Learning
Ke Sun; Stéphane Marchand-Maillet
Geodesic Distance Function Learning

M89 via Heat Flow on Vector Fields
Binbin Lin; Ji Yang; Xiaofei He; Jieping Ye

M90 Two-Stage Metric Learning
Jun Wang; Ke Sun; Fei Sha; Stéphane Marchand-Maillet; Alexandros Kalousis

M91 Transductive Learning with Multi-class Volume Approximation
Gang Niu; Bo Dai; Christoffel du Plessis; Masashi Sugiyama

M92 Convergence rates for persistence diagram estimation in Topological Data Analysis
Frédéric Chazal; Marc Glisse; Catherine Labruère; Bertrand Michel

Monday June 23

16:20 - Track E - Kernel Methods I (Room 307)

M93 On p-norm Path Following in Multiple Kernel Learning for Non-linear Feature Selection
Pratik Jawanpuria; Manik Varma; Saketha Nath

M94 A Divide-and-Conquer Solver for Kernel Support Vector Machines
Cho-Jui Hsieh; Si Si; Inderjit Dhillon

M95 Memory Efficient Kernel Approximation
Si Si; Cho-Jui Hsieh; Inderjit Dhillon

M96 Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection
Arun Iyer; Saketha Nath; Sunita Sarawagi

M97 Robust and Efficient Kernel Hyperparameter Paths with Guarantees
Joachim Giesen; Soeren Laue; Patrick Wieschollek

M98 Nonparametric Estimation of Multi-View Latent Variable Models
Le Song; Animashree Anandkumar; Bo Dai; Bo Xie

Monday June 23

16:20 - Track F - Unsupervised Learning and Detection (Room 201C)

M99 Anomaly Ranking as Supervised Bipartite Ranking
Stephan Clémençon; Sylvain Robbiano

M100 On learning to localize objects with minimal supervision
Hyun Oh Song; Ross Girshick; Stefanie Jegelka; Julien Mairal; Zaid Harchaoui; Trevor Darrell

M101 Active Detection via Adaptive Submodularity
Yuxin Chen; Hiroaki Shioi; Cesar Fuentes Montesinos; Lian Pin Koh; Serge Wich; Andreas Krause

M102 Structured Generative Models of Natural Source Code
Chris Maddison; Daniel Tarlow

M103 Coordinate-descent for learning orthogonal matrices through Givens rotations
Uri Shalit; Gal Chechik

Monday June 23

10:30 - Track A - Latent Variable Models

M1 A Discriminative Latent Variable Model for Online Clustering

Rajhans Samdani; Kai-Wei Chang; Dan Roth

This paper presents a latent variable structured prediction model for discriminative supervised clustering of items called the Latent Left-linking Model (L3M). We present an online clustering algorithm for L3M based on a feature-based item similarity function. We provide a learning framework for estimating the similarity function and present a fast stochastic gradient-based learning technique. In our experiments on coreference resolution and document clustering, L3 M outperforms several existing online as well as batch supervised clustering techniques.

M2 Exchangeable Variable Models

Mathias Niepert; Pedro Domingos

A sequence of random variables is exchangeable if its joint distribution is invariant under variable permutations. We introduce exchangeable variable models (EVMs) as a novel class of probabilistic models whose basic building blocks are partially exchangeable sequences, a generalization of exchangeable sequences. We prove that a family of tractable EVMs is optimal under zero-one loss for a large class of functions, including parity and threshold functions, and strictly subsumes existing tractable independence-based model families. Extensive experiments show that EVMs outperform state of the art classifiers such as SVMs and probabilistic models which are solely based on independence assumptions.

M3 Learning Latent Variable Gaussian Graphical Models

Zhaoshi Meng; Brian Eriksson; Al Hero

Gaussian graphical models (GGM) have been widely used in many high-dimensional applications ranging from biological and financial data to recommender systems. Sparsity in GGM plays a central role both statistically and computationally. Unfortunately, real-world data often does not fit well to sparse graphical models. In this paper, we focus on a family of latent variable Gaussian graphical models (LVGGM), where the model is conditionally sparse given latent variables, but marginally non-sparse. In LVGGM, the inverse covariance matrix has a low-rank plus sparse structure, and can be learned in a regularized maximum likelihood framework. We derive novel parameter estimation error bounds for LVGGM under mild conditions in the high-dimensional setting. These results complement the existing theory on the structural learning, and open up new possibilities of using LVGGM for statistical inference.

M4 Latent Variable Copula Inference for Bundle Pricing from Retail Transaction Data

Benjamin Letham; Wei Sun; Anshul Sheopuri

M5 Affinity Weighted Embedding

Jason Weston; Ron Weiss; Hector Yee

Supervised linear embedding models like Wsabie (Weston et al., 2011) and supervised semantic indexing (Bai et al., 2010) have proven successful at ranking, recommendation and annotation tasks. However, despite being scalable to large datasets they do not take full advantage of the extra data due to their linear nature, and we believe they typically underfit. We propose a new class of models which aim to provide improved performance while retaining many of the benefits of the existing class of embedding models. Our approach works by reweighting each component of the embedding of features and labels with a potentially nonlinear affinity function. We describe several variants of the family, and show its usefulness on several datasets.

M6 Learning the Irreducible Representations of Commutative Lie Groups

Taco Cohen; Max Welling

We present a new probabilistic model of compact commutative Lie groups that produces invariant-equivariant and disentangled representations of data. To define the notion of disentangling, we borrow a fundamental principle from physics that is used to derive the elementary particles of a system from its symmetries. Our model employs a newfound Bayesian conjugacy relation that enables fully tractable probabilistic inference over compact commutative Lie groups -- a class that includes the groups that describe the rotation and cyclic translation of images. We train the model on pairs of transformed image patches, and show that the learned invariant representation is highly effective for classification.

10:30 - Track B - Online Learning and Planning

M7 Covering Number for Efficient Heuristic-based POMDP Planning

Zongzhang Zhang; David Hsu; Wee Sun Lee

The difficulty of POMDP planning depends on the size of the search space involved. Heuristics are often used to reduce the search space size and improve computational efficiency; however, there are few theoretical bounds on their effectiveness. In this paper, we use the covering number to characterize the size of the search space reachable under heuristics and connect the complexity of POMDP planning to the effectiveness of heuristics. With insights from the theoretical analysis, we have developed a practical POMDP algorithm, Packing-Guided Value Iteration (PGVI). Empirically, PGVI is competitive with the state-of-the-art point-based POMDP algorithms on 65 small benchmark problems and outperforms them on 4 larger problems.

M8 Learning Complex Neural Network Policies with Trajectory Optimization

Sergey Levine; Vladlen Koltun

Direct policy search methods offer the promise of automatically learning controllers for complex, high-dimensional tasks. However, prior applications of policy search often required specialized,

low-dimensional policy classes, limiting their generality. In this work, we introduce a policy search algorithm that can directly learn high-dimensional, general-purpose policies, represented by neural networks. We formulate the policy search problem as an optimization over trajectory distributions, alternating between optimizing the policy to match the trajectories, and optimizing the trajectories to match the policy and minimize expected cost. Our method can learn policies for complex tasks such as bipedal push recovery and walking on uneven terrain, while outperforming prior methods.

M9 A Physics-Based Model Prior for Object-Oriented MDPs

Jonathan Scholz; Martin Levihn; Charles Isbell

One of the key challenges in using reinforcement learning in robotics is the need for models that capture natural world structure. There are, methods that formalize multi-object dynamics using relational representations, but these methods are not sufficiently compact for real-world robotics. We present a physics-based approach that exploits modern simulation tools to efficiently parameterize physical dynamics. Our results show that this representation can result in much faster learning, by virtue of its strong but appropriate inductive bias in physical environments.

M10 Online Multi-Task Learning for Policy Gradient Methods

Haitham Bou Ammar; Eric Eaton; Paul Ruvolo; Matthew Taylor

Policy gradient algorithms have shown considerable recent success in solving high-dimensional sequential decision making tasks, particularly in robotics. However, these methods often require extensive experience in a domain to achieve high performance. To make agents more sample-efficient, we developed a multi-task policy gradient method to learn decision making tasks consecutively, transferring knowledge between tasks to accelerate learning. Our approach provides robust theoretical guarantees, and we show empirically that it dramatically accelerates learning on a variety of dynamical systems, including an application to quadrotor control.

M11 Pursuit-Evasion Without Regret, with an Application to Trading

Lili Dworkin; Michael Kearns; Yuriy Nevmyvaka

We propose a state-based variant of the classical online learning problem of tracking the best expert. In our setting, the actions of the algorithm and experts correspond to local moves through a continuous and bounded state space. At each step, Nature chooses payoffs as a function of each player's current position and action. Our model therefore integrates the problem of prediction with expert advice with the stateful formalisms of reinforcement learning. Traditional no-regret learning approaches no longer apply, but we propose a simple algorithm that provably achieves no-regret when the state space is any convex Euclidean region. Our algorithm combines techniques from online learning with results from the literature on pursuit-evasion games. We describe a quantitative trading application in which the convex region captures inventory risk constraints, and local moves limit market impact. Using historical market data, we show experimentally that our algorithm has a strong advantage over classic no-regret approaches.

M12 Adaptivity and Optimism: An Improved Exponentiated Gradient Algorithm

Jacob Steinhardt; Percy Liang

We present an adaptive variant of the exponentiated gradient algorithm. Leveraging the optimistic learning framework of Rakhlin & Sridharan (2012), we obtain regret bounds that in the learning from experts setting depend on the variance and path length of the best expert, improving on results by Hazan & Kale (2008) and Chiang et al. (2012), and resolving an open problem posed by Kale (2012). Our techniques naturally extend to matrix-valued loss functions, where we present an adaptive matrix exponentiated gradient algorithm. To obtain the optimal regret bound in the matrix case, we generalize the Follow-the-Regularized-Leader algorithm to vector-valued payoffs, which may be of independent interest.

10:30 - Track C - Clustering

M13 Demystifying Information-Theoretic Clustering

Greg Ver Steeg; Aram Galstyan; Fei Sha; Simon DeDeo

We propose a novel method for clustering data which is grounded in information-theoretic principles and requires no parametric assumptions. Previous attempts to use information theory to define clusters in an assumption-free way are based on maximizing mutual information between data and cluster labels. We demonstrate that this intuition suffers from a fundamental conceptual flaw that causes clustering performance to deteriorate as the amount of data increases. Instead, we return to the axiomatic foundations of information theory to define a meaningful clustering measure based on the notion of consistency under coarse-graining for finite data.

M14 Clustering in the Presence of Background Noise

Shai Ben-David; Nika Haghtalab

We address the problem of noise management in clustering algorithms. Namely, issues that arise when on top of some cluster structure the data also contains an unstructured set of points. We consider how clustering algorithms can be ``robustified" so that they recover the cluster structure in spite of the unstructured part of the input. We introduce some quantitative measures of such robustness that take into account the strength of the embedded cluster structure as well was the mildness of the noise subset. We propose a simple and efficient method to turn any centroid-based clustering algorithm into a noise-robust one, and prove robustness guarantees for our method with respect to these measures. We also prove that more straightforward ways of ``robustifying" clustering algorithms fail to achieve similar guarantees.

M15 Hierarchical Quasi-Clustering Methods for Asymmetric Networks

Gunnar Carlsson; Facundo Mémoli; Alejandro Ribeiro; Santiago Segarra

This paper introduces hierarchical quasi-clustering methods, a generalization of hierarchical clustering for asymmetric networks where the output structure preserves the asymmetry of the input data.

show that this output structure is equivalent to a finite quasi-ultrametric space and study admissibility with respect to two desirable properties. We prove that a modified version of single linkage is the only admissible quasi-clustering method. Moreover, we show stability of the proposed method and we establish invariance properties fulfilled by it. Algorithms are further developed and the value of quasi-clustering analysis is illustrated with a study of internal migration within United States.

M16 Local algorithms for interactive clustering

Pranjal Awasthi; Maria Balcan; Konstantin Voevodski

We study the design of interactive clustering algorithms for data sets satisfying natural stability assumptions. Our algorithms start with any initial clustering and only make local changes in each step; both are desirable features in many applications. We show that in this constrained setting one can still design provably efficient algorithms that produce accurate clusterings. We also show that our algorithms perform well on real-world data.

M17 Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance

Simone Romano; James Bailey; Vinh Nguyen; Karin Verspoor

Mutual information is a very popular measure for comparing clusterings. Previous work has shown that it is beneficial to make an adjustment for chance to this measure, by subtracting an expected value and normalizing via an upper bound. This yields the constant baseline property that enhances intuitiveness. In this paper, we argue that a further type of statistical adjustment for the mutual information is also beneficial - an adjustment to correct selection bias. This type of adjustment is useful when carrying out many clustering comparisons, to select one or more preferred clusterings. It reduces the tendency for the mutual information to choose clustering solutions i) with more clusters, or ii) induced on fewer data points, when compared to a reference one. We term our new adjusted measure the *standardized mutual information*. It requires computation of the variance of mutual information under a hypergeometric model of randomness, which is technically challenging. We derive an analytical formula for this variance and analyze its complexity. We then experimentally assess how our new measure can address selection bias and also increase interpretability. We recommend using the standardized mutual information when making multiple clustering comparisons in situations where the number of records is small compared to the number of clusters considered.

M18 A Single-Pass Algorithm for Efficiently Recovering Sparse Cluster Centers of High-dimensional Data

Jinfeng Yi; Lijun Zhang; Jun Wang; Rong Jin; Anil Jain

Learning a statistical model for high-dimensional data is an important topic in machine learning. Although this problem has been well studied in the supervised setting, little is known about its unsupervised counterpart. In this work, we focus on the problem of clustering high-dimensional data with sparse centers. In particular, we address the following open question in unsupervised learning: ``is it possible to reliably cluster high-dimensional data when the number of samples is smaller than the data dimensionality?" We develop an efficient clustering algorithm that is able to estimate sparse cluster centers with a single pass over the data. Our theoretical analysis shows that the proposed algorithm is able to accurately recover cluster centers with only $O(s\ log d)$ number of samples (data points), provided all the cluster centers are s-sparse vectors

in a d dimensional space. Experimental results verify both the effectiveness and efficiency of the proposed clustering algorithm compared to the state-of-the-art algorithms on several benchmark datasets.

10:30 - Track D - Metric Learning and Feature Selection

M19 Large-Margin Metric Learning for Constrained Partitioning Problems

Rémi Lajugie; Francis Bach; Sylvain Arlot

We consider unsupervised partitioning problems based explicitly or implicitly on the minimization of Euclidean distortions, such as clustering, image or video segmentation, and other change-point detection problems. We emphasize on cases with specific structure, which include many practical situations ranging from mean-based change-point detection to image segmentation problems. We aim at learning a Mahalanobis metric for these unsupervised problems, leading to feature weighting and/or selection. This is done in a supervised way by assuming the availability of several (partially) labeled datasets that share the same metric. We cast the metric learning problem as a large-margin structured prediction problem, with proper definition of regularizers and losses, leading to a convex optimization problem which can be solved efficiently. Our experiments show how learning the metric can significantly improve performance on bioinformatics, video or image segmentation problems.

M20 Robust Distance Metric Learning via Simultaneous L1-Norm Minimization and Maximization

Hua Wang; Feiping Nie; Heng Huang

Traditional distance metric learning with side information usually formulates the objectives using the covariance matrices of the data point pairs in the two constraint sets of must-links and cannot-links. Because the covariance matrix computes the sum of the squared L2-norm distances, it is prone to both outlier samples and outlier features. To develop a robust distance metric learning method, in this paper we propose a new objective for distance metric learning using the L1-norm distances. However, the resulted objective is very challenging to solve, because it simultaneously minimizes and maximizes (minmax) a number of non-smooth L1-norm terms. As an important theoretical contribution of this paper, we systematically derive an efficient iterative algorithm to solve the general L1-norm minmax problem, which is rarely studied in literature. We have performed extensive empirical evaluations, where our new distance metric learning method outperforms related state-of-the-art methods in a variety of experimental settings to cluster both noiseless and noisy data.

M21 Efficient Learning of Mahalanobis Metrics for Ranking

Daryl Lim; Gert Lanckriet

We develop an efficient algorithm to learn a Mahalanobis distance metric by directly optimizing a ranking loss. Our approach focuses on optimizing the top of the induced rankings, which is desirable in tasks such as visualization and nearest-neighbor retrieval. We further develop and justify a simple technique to reduce training time significantly with minimal impact on performance. Our proposed method significantly outperforms alternative methods on several real-world tasks, and can scale to large and high-dimensional data.

M22 Stochastic Neighbor Compression

Matt Kusner; Stephen Tyree; Kilian Weinberger; Kunal Agrawal

We present Stochastic Neighborhood Compression (SNC), an algorithm to compress a dataset for the purpose of k-nearest neighbor (kNN) classification. Given training data, SNC learns a much smaller synthetic data set, that minimizes the stochastic 1-nearest neighbor classification error on the training data. This approach has several appealing properties: due to its small size, the compressed set speeds up kNN testing drastically (up to several orders of magnitude, in our experiments); it makes the kNN classifier substantially more robust to label noise; on 4 of 7 data sets it yields lower test error than kNN on the entire training set, even at compression ratios as low as 2%; finally, the SNC compression leads to impressive speed ups over kNN even when kNN and SNC are both used with ball-tree data structures, hashing, and LMNN dimensionality reduction—demonstrating that it is complementary to existing state-of-the-art algorithms to speed up kNN classification and leads to substantial further improvements.

M23 Large-margin Weakly Supervised Dimensionality Reduction

Chang Xu; Dacheng Tao; Chao Xu; Yong Rui

This paper studies dimensionality reduction in a weakly supervised setting, in which the preference relationship between examples is indicated by weak cues. A novel framework is proposed that integrates two aspects of the large margin principle (angle and distance), which simultaneously encourage angle consistency between preference pairs and maximize the distance between examples in preference pairs. Two specific algorithms are developed: an alternating direction method to learn a linear transformation matrix and a gradient boosting technique to optimize a non-linear transformation directly in the function space. Theoretical analysis demonstrates that the proposed large margin optimization criteria can strengthen and improve the robustness and generalization performance of preference learning algorithms on the obtained low-dimensional subspace. Experimental results on real-world datasets demonstrate the significance of studying dimensionality reduction in the weakly supervised setting and the effectiveness of the proposed framework.

M24 Sparse meta-Gaussian information bottleneck

Melanie Rey; Volker Roth; Thomas Fuchs

We present a new sparse compression technique based on the information bottleneck (IB) principle, which takes into account side information. This is achieved by introducing a sparse variant of IB which preserves the information in only a few selected dimensions of the original data through compression. By assuming a Gaussian copula we can capture arbitrary non-Gaussian margins, continuous or discrete. We apply our model to select a sparse number of biomarkers relevant to the evolution of malignant melanoma and show that our sparse selection provides reliable predictors.

10:30 - Track E - Optimization II

M25 Fast Stochastic Alternating Direction Method of Multipliers

Wenliang Zhong; James Kwok

We propose a new stochastic alternating direction method of multipliers (ADMM) algorithm, which incrementally approximates the full gradient in the linearized ADMM formulation. Besides having a low per-iteration complexity as existing stochastic ADMM algorithms, it improves the convergence rate on convex problems from $\mO(1/\sqrt{T})$ to $\mO(1/T)$, where T is the number of iterations. This matches the convergence rate of the batch ADMM algorithm, but without the need to visit all the samples in each iteration. Experiments on the graph-guided fused lasso demonstrate that the new algorithm is significantly faster than state-of-the-art stochastic and batch ADMM algorithms.

M26 Accelerated Proximal Stochastic Dual Coordinate Ascent for Regularized Loss Minimization

Shai Shalev-Shwartz; Tong Zhang

We introduce a proximal version of the stochastic dual coordinate ascent method and show how to accelerate the method using an inner-outer iteration procedure. We analyze the runtime of the framework and obtain rates that improve state-of-the-art results for various key machine learning optimization problems including SVM, logistic regression, ridge regression, Lasso, and multiclass SVM. Experiments validate our theoretical findings.

M27 An Asynchronous Parallel Stochastic Coordinate Descent Algorithm

Ji Liu; Steve Wright; Christopher Re; Victor Bittorf; Srikrishna Sridhar

We describe an asynchronous parallel stochastic coordinate descent algorithm for minimizing smooth unconstrained or separably constrained functions. The method achieves a linear convergence rate on functions that satisfy an essential strong convexity property and a sublinear rate ($1/K$) on general convex functions. Near-linear speedup on a multicore system can be expected if the number of processors is $O(n^{1/2})$ in unconstrained optimization and $O(n^{1/4})$ in the separable-constrained case, where n is the number of variables. We describe results from implementation on 40-core processors.

M28 Towards an optimal stochastic alternating direction method of multipliers

Samaneh Azadi; Suvrit Sra

We study regularized stochastic convex optimization subject to linear equality constraints. This class of problems was recently also studied by Ouyang et al. (2013) and Suzuki (2013); both introduced similar stochastic alternating direction method of multipliers (SADMM) algorithms. However, the analysis of both papers led to suboptimal convergence rates. This paper presents two new SADMM methods: (i) the first attains the minimax optimal rate of O(1/k) for nonsmooth strongly-convex stochastic problems; while (ii) the second progresses towards an optimal rate by exhibiting an O(1/k^2) rate for the smooth part. We present several experiments with our new methods; the results indicate improved performance over competing ADMM methods.

M29 Stochastic Dual Coordinate Ascent with Alternating Direction Method of Multipliers

Taiji Suzuki

We propose a new stochastic dual coordinate ascent technique that can be applied to a wide range of regularized learning problems. Our method is based on alternating direction method of multipliers (ADMM) to deal with complex regularization functions such as structured regularizations. Although the original ADMM is a batch method, the proposed method offers a stochastic update rule where each iteration requires only one or few sample observations. Moreover, our method can naturally afford mini-batch update and it gives speed up of convergence. We show that, under mild assumptions, our method converges exponentially. The numerical experiments show that our method actually performs efficiently.

M30 Communication-Efficient Distributed Optimization using an Approximate Newton-type Method

Ohad Shamir; Nati Srebro; Tong Zhang

We present a novel Newton-type method for distributed optimization, which is particularly well suited for stochastic optimization and learning problems. For quadratic objectives, the method enjoys a linear rate of convergence which provably \emph{improves} with the data size, requiring an essentially constant number of iterations under reasonable assumptions. We provide theoretical and empirical evidence of the advantages of our method compared to other approaches, such as one-shot parameter averaging and ADMM.
Monday June 23, 10:30 - Track F - Neural Language and Speech

10:30 - Track F - Neural Language and Speech

M31 Multimodal Neural Language Models

Ryan Kiros; Ruslan Salakhutdinov; Rich Zemel

We introduce two multimodal neural language models: models of natural language that can be conditioned on other modalities. An image-text multimodal neural language model can be used to retrieve images given complex sentence queries, retrieve phrase descriptions given image queries, as well as generate text conditioned on images. We show that in the case of image-text modelling we can jointly learn word representations and image features by training our models together with a convolutional network. Unlike many of the existing methods, our approach can generate sentence descriptions for images without the use of templates, structured prediction, and/or syntactic trees. While we focus on image-text modelling, our algorithms can be easily applied to other modalities such as audio.

M32 Distributed Representations of Sentences and Documents

Quoc Le; Tomas Mikolov

Many machine learning algorithms require the input to be represented as a fixed length feature vector. When it comes to texts, one of the most common representations is bag-of-words. Despite their popularity, bag-of-words models have two major weaknesses: they lose the ordering of the words and they also ignore semantics of the words. For example, “powerful,” “strong” and “Paris” are equally distant. In this paper, we propose an unsupervised algorithm that learns vector representations of sentences and text doc- uments. T

his algorithm represents each docu- ment by a dense vector which is trained to predict words in the document. Its construction gives our algorithm the potential to overcome the weak- nesses of bag-of-words models. Empirical re- sults show that our technique outperforms bag- of-words models as well as other techniques for text representations. Finally, we achieve new state-of-the-art results on several text classifica- tion and sentiment analysis tasks.

M33 Learning Character-level Representations for Part-of-Speech Tagging

Cicero Dos Santos; Bianca Zadrozny

Distributed word representations have recently been proven to be an invaluable resource for NLP. These representations are normally learned using neural networks and capture syntactic and semantic information about words. Information about word morphology and shape is normally ignored when learning word representations. However, for tasks like part-of-speech tagging, intra-word information is extremely useful, specially when dealing with morphologically rich languages. In this paper, we propose a deep neural network that learns character-level representation of words and associate them with usual word representations to perform POS tagging. Using the proposed approach, while avoiding the use of any handcrafted feature, we produce state-of-the-art POS taggers for two languages: English, with 97.32% accuracy on the Penn Treebank WSJ corpus; and Portuguese, with 97.47% accuracy on the Mac-Morpho corpus, where the latter represents an error reduction of 12.2% on the best previous known result.

M34 Compositional Morphology for Word Representations and Language Modelling

Jan Botha; Phil Blunsom

This paper presents a scalable method for integrating compositional morphological representations into a vector-based probabilistic language model. Our approach is evaluated in the context of log-bilinear language models, rendered suitably efficient for implementation inside a machine translation decoder by factoring the vocabulary. We perform both intrinsic and extrinsic evaluations, presenting results on a range of languages which demonstrate that our model learns morphological representations that both perform well on word similarity tasks and lead to substantial reductions in perplexity. When used for translation into morphologically rich languages with large vocabularies, our models obtain improvements of up to 1.2 BLEU points relative to a baseline system using back-off n-gram models.

M35 Towards End-To-End Speech Recognition with Recurrent Neural Networks

Alex Graves; Navdeep Jaitly

This paper presents a speech recognition system that directly transcribes audio data with text, without requiring an intermediate phonetic representation. The system is based on a combination of the deep bidirectional LSTM recurrent neural network architecture and the Connectionist Temporal Classification objective function. A modification to the objective function is introduced that trains the network to minimise the expectation of an arbitrary transcription loss function. This allows a direct optimisation of the word error rate, even in the absence of a lexicon or language model. The system achieves a word error rate of 27.3% on the Wall Street Journal corpus with no prior linguistic information, 21.9% with only a lexicon of allowed words, and 8.2% with a trigram language model. Combining the network with a baseline system further reduces the error rate to 6.7%.

Networks (RNNs) have the ability, in theory, to cope with these temporal dependencies by virtue of the short-term memory implemented by their recurrent (feedback) connections. However, in practice they are difficult to train successfully when long-term memory is required. This paper introduces a simple, yet powerful modification to the simple RNN (SRN) architecture, the Clockwork RNN (CW-RNN), in which the hidden layer is partitioned into separate modules, each processing inputs at its own temporal granularity, making computations only at its prescribed clock rate. Rather than making the standard RNN models more complex, CW-RNN reduces the number of SRN parameters, improves the performance significantly in the tasks tested, and speeds up the network evaluation. The network is demonstrated in preliminary experiments involving three tasks: audio signal generation, TIMIT spoken word classification, where it outperforms both SRN and LSTM networks, and online handwriting recognition, where it outperforms SRNs.

14:00 - Track A - Graphical Models and Approximate Inference

M37 Probabilistic Partial Canonical Correlation Analysis

Yusuke Mukuta; Tatsuya Harada

Partial canonical correlation analysis (partial CCA) is a statistical method that estimates a pair of linear projections onto a low dimensional space, where the correlation between two multidimensional variables is maximized after eliminating the influence of a third variable. Partial CCA is known to be closely related to a causality measure between two time series. However, partial CCA requires the inverses of covariance matrices, so the calculation is not stable. This is particularly the case for high-dimensional data or small sample sizes. Additionally, we cannot estimate the optimal dimension of the subspace in the model. In this paper, we have addressed these problems by proposing a probabilistic interpretation of partial CCA and deriving a Bayesian estimation method based on the probabilistic model. Our numerical experiments demonstrated that our methods can stably estimate the model parameters, even in high dimensions or when there are a small number of samples.

M38 Min-Max Problems on Factor Graphs

Siamak Ravanbakhsh; Christopher Srinivasa; Brendan Frey; Russell Greiner

We study the min-max problem in factor graphs, which seeks the assignment that minimizes the maximum value over all factors. We reduce this problem to both min-sum and sum-product inference, and focus on the later. This approach reduces the min-max inference problem to a sequence of constraint satisfaction problems (CSPs) which allows us to sample from a uniform distribution over the set of solutions. We demonstrate how this scheme provides a message passing solution to several NP-hard combinatorial problems, such as min-max clustering (a.k.a. K-clustering), the asymmetric K-center problem, K-packing and the bottleneck traveling salesman problem. Furthermore we theoretically relate the min-max reductions to several NP hard decision problems, such as clique cover, set cover, maximum clique and Hamiltonian cycle, therefore also providing message passing solutions for these problems. Experimental results suggest that message passing often provides near optimal min-max solutions for moderate size instances.

M39 Skip Context Tree Switching

Marc Bellemare; Joel Veness; Erik Talvitie

Context Tree Weighting (CTW) is a powerful probabilistic sequence prediction technique that efficiently performs Bayesian model averaging over the class of all prediction suffix trees of bounded depth. In this paper we show how to generalize this technique to the class of K-skip prediction suffix trees. Contrary to regular prediction suffix trees, K-skip prediction suffix trees are permitted to ignore up to K contiguous portions of the context. This allows for significant improvements in predictive accuracy when irrelevant variables are present, a case which often occurs within record-aligned data and images. We provide a regret-based analysis of our approach, and empirically evaluate it on the Calgary corpus and a set of Atari 2600 screen prediction tasks.

M40 Learning the Parameters of Determinantal Point Process Kernels

Raja Hafiz Affandi; Emily Fox; Ryan Adams; Ben Taskar

Determinantal point processes (DPPs) are well-suited for modeling repulsion and have proven useful in applications where diversity is desired. While DPPs have many appealing properties, learning the parameters of a DPP is diffcult, as the likelihood is non-convex and is infeasible to compute in many scenarios. Here we propose Bayesian methods for learning the DPP kernel parameters. These methods are applicable in large-scale discrete and continuous DPP settings, even when the likelihood can only be bounded. We demonstrate the utility of our DPP learning methods in studying the progression of diabetic neuropathy based on the spatial distribution of nerve fibers, and in studying human perception of diversity in images.

M41 Deterministic Anytime Inference for Stochastic Continuous-Time Markov Processes

E. Busra Celikkaya; Christian Shelton

We describe a deterministic anytime method for calculating filtered and smoothed distributions in large variable-based continuous time Markov processes. Prior non-random algorithms do not converge to the true distribution in the limit of infinite computation time. Sampling algorithms give different results each time run, which can lead to instability when used inside expectation-maximization or other algorithms. Our method combines the anytime convergent properties of sampling with the non-random nature of variational approaches. It is built upon a sum of time-ordered products, an expansion of the matrix exponential. We demonstrate that our method performs as well as or better than the current best sampling approaches on benchmark problems.

M42 Doubly Stochastic Variational Bayes for non-Conjugate Inference

Michalis Titsias; Miguel Lázaro-Gredilla

We propose a simple and effective variational inference algorithm based on stochastic optimisation that can be widely applied for Bayesian non-conjugate inference in continuous parameter spaces. This algorithm is based on stochastic approximation and allows for efficient use of gradient information from the model joint density. We demonstrate these properties using illustrative examples as well as in challenging and diverse Bayesian inference problems such as variable selection in logistic regression and fully Bayesian inference over kernel hyperparameters in Gaussian process regression.

14:00 - Track B - Online Learning I

M43 On the convergence of no-regret learning in selfish routing

Walid Krichene; Benjamin Drighès; Alexandre Bayen

We study the repeated, non-atomic routing game, in which selfish players make a sequence of routing decisions. We consider a model in which players use regret-minimizing algorithms as the learning mechanism, and study the resulting dynamics. We are concerned in particular with the convergence to the set of Nash equilibria of the routing game. No-regret learning algorithms are known to guarantee convergence of a subsequence of population strategies. We are concerned with convergence of the actual sequence. We show that convergence holds for a large class of online learning algorithms, inspired from the continuous-time replicator dynamics. In particular, the discounted Hedge algorithm is proved to belong to this class, which guarantees its convergence.

M44 Optimal PAC Multiple Arm Identification with Applications to Crowdsourcing

Yuan Zhou; Xi Chen; Jian Li

We study the problem of selecting K arms with the highest expected rewards in a stochastic N-armed bandit game. Instead of using existing evaluation metrics (e.g., misidentification probability or the metric in EXPLORE-K), we propose to use the aggregate regret, which is defined as the gap between the average reward of the optimal solution and that of our solution. Besides being a natural metric by itself, we argue that in many applications, such as our motivating example from crowdsourcing, the aggregate regret bound is more suitable. We propose a new PAC algorithm, which, with probability at least $1-\delta$, identifies a set of K arms with regret at most ϵ. We provide the sample complexity bound of our algorithm. To complement, we establish the lower bound and show that the sample complexity of our algorithm matches the lower bound. Finally, we report experimental results on both synthetic and real data sets, which demonstrates the superior performance of the proposed algorithm.

M45 Prediction with Limited Advice and Multiarmed Bandits with Paid Observations

Yevgeny Seldin; Peter Bartlett; Koby Crammer; Yasin Abbasi-Yadkori

We study two problems of online learning under restricted information access. In the first problem, \emph{prediction with limited advice}, we consider a game of prediction with expert advice, where on each round of the game we query the advice of a subset of M out of N experts. We present an algorithm that achieves $O(\sqrt{(N/M)T\ln N})$ regret on T rounds of this game. The second problem, the \emph{multiarmed bandit with paid observations}, is a variant of the adversarial N-armed bandit game, where on round t of the game we can observe the reward of any number of arms, but each observation has a cost c. We present an algorithm that achieves $O((cN\ln N)^{1/3} T^{2/3} + \sqrt{T \ln N})$ regret on T rounds of this game in the worst case. Furthermore, we present a number of refinements that treat arm- and time-dependent observation costs and achieve lower regret under benign conditions. We present lower bounds that show that, apart from the logarithmic factors, the worst-case regret bounds cannot be improved.

M46 One Practical Algorithm for Both Stochastic and Adversarial Bandits

Yevgeny Seldin; Aleksandrs Slivkins

We present an algorithm for multiarmed bandits that achieves almost optimal performance in both stochastic and adversarial regimes without prior knowledge about the nature of the environment. Our

algorithm is based on augmentation of the EXP3 algorithm with a new control lever in the form of exploration parameters that are tailored individually for each arm. The algorithm simultaneously applies the ``old'' control lever, the learning rate, to control the regret in the adversarial regime and the new control lever to detect and exploit gaps between the arm losses. This secures problem-dependent ``logarithmic'' regret when gaps are present without compromising on the worst-case performance guarantee in the adversarial regime. We show that the algorithm can exploit both the usual expected gaps between the arm losses in the stochastic regime and deterministic gaps between the arm losses in the adversarial regime. The algorithm retains ``logarithmic'' regret guarantee in the stochastic regime even when some observations are contaminated by an adversary, as long as on average the contamination does not reduce the gap by more than a half. Our results for the stochastic regime are supported by experimental validation.

M47 A Bayesian Framework for Online Classifier Ensemble

Qinxun Bai; Henry Lam; Stan Sclaroff

We propose a Bayesian framework for recursively estimating the classifier weights in online learning of a classifier ensemble. In contrast with past methods, such as stochastic gradient descent or online boosting, our framework estimates the weights in terms of evolving posterior distributions. For a specified class of loss functions, we show that it is possible to formulate a suitably defined likelihood function and hence use the posterior distribution as an approximation to the global empirical loss minimizer. If the stream of training data is sampled from a stationary process, we can also show that our framework admits a superior rate of convergence to the expected loss minimizer than is possible with standard stochastic gradient descent. In experiments with real-world datasets, our formulation often performs better than online boosting algorithms.

M48 Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits

Alekh Agarwal; Daniel Hsu; Satyen Kale; John Langford; Lihong Li; Robert Schapire

We present a new algorithm for the contextual bandit learning problem, where the learner repeatedly takes one of K \emph{actions} in response to the observed \emph{context}, and observes the \emph{reward} only for that action. Our method assumes access to an oracle for solving fully supervised cost-sensitive classification problems and achieves the statistically optimal regret guarantee with only $\otil(\sqrt{KT})$ oracle calls across all T rounds. By doing so, we obtain the most practical contextual bandit learning algorithm amongst approaches that work for general policy classes. We conduct a proof-of-concept experiment which demonstrates the excellent computational and statistical performance of (an online variant of) our algorithm relative to several strong baselines.

14:00 - Track C - Monte Carlo and Approximate Inference

M49 Memory (and Time) Efficient Sequential Monte Carlo

Seong-Hwan Jun; Alexandre Bouchard-Côté

Memory efficiency is an important issue in Sequential Monte Carlo (SMC) algorithms, arising for example in inference of high-dimensional latent variables via Rao-Blackwellized SMC algorithms, where the size of individual particles combined with the required number of particles can stress the main

memory. Standard SMC methods have a memory requirement that scales linearly in the number of particles present at all stage of the algorithm. Our contribution is a simple scheme that makes the memory cost of SMC methods depends on the number of distinct particles that survive resampling. We show that this difference has a large empirical impact on the quality of the approximation in realistic scenarios, and also---since memory access is generally slow---on the running time. The method is based on a two pass generation of the particles, which are represented implicitly in the first pass. We parameterize the accuracy of our algorithm with a memory budget rather than with a fixed number of particles. Our algorithm adaptively selects an optimal number of particle to exploit this fixed memory budget. We show that this adaptation does not interfere with the usual consistency guarantees that come with SMC algorithms.

M50 Efficient Continuous-Time Markov Chain Estimation

Monir Hajiaghayi; Bonnie Kirkpatrick; Liangliang Wang; Alexandre Bouchard-Côté

Many problems of practical interest rely on Continuous-time Markov chains~(CTMCs) defined over combinatorial state spaces, rendering the computation of transition probabilities, and hence probabilistic inference, difficult or impossible with existing methods. For problems with countably infinite states, where classical methods such as matrix exponentiation are not applicable, the main alternative has been particle Markov chain Monte Carlo methods imputing both the holding times and sequences of visited states. We propose a particle-based Monte Carlo approach where the holding times are marginalized analytically. We demonstrate that in a range of realistic inferential setups, our scheme dramatically reduces the variance of the Monte Carlo approximation and yields more accurate parameter posterior approximations given a fixed computational budget. These experiments are performed on both synthetic and real datasets, drawing from two important examples of CTMCs having combinatorial state spaces: string-valued mutation models in phylogenetics and nucleic acid folding pathways.

M51 Filtering with Abstract Particles

Jacob Steinhardt; Percy Liang

Using particles, beam search and sequential Monte Carlo can approximate distributions in an extremely flexible manner. However, they can suffer from sparsity and inadequate coverage on large state spaces. We present a new filtering method that addresses this issue by using “abstract particles” that each represent an entire region of the state space. These abstract particles are combined into a hierarchical decomposition, yielding a representation that is both compact and flexible. Empirically, our method outperforms beam search and sequential Monte Carlo on both a text reconstruction task and a multiple object tracking task.

M52 Spherical Hamiltonian Monte Carlo for Constrained Target Distributions

Shiwei Lan; Bo Zhou; Babak Shahbaba

Statistical models with constrained probability distributions are abundant in machine learning. Some examples include regression models with norm constraints (e.g., Lasso), probit models, many copula models, and Latent Dirichlet Allocation (LDA) models. Bayesian inference involving probability distributions confined to constrained domains could be quite challenging for commonly used sampling algorithms. For such problems, we propose a novel Markov Chain Monte Carlo (MCMC) method that provides a general and computationally efficient framework for handling boundary conditions. Our method first maps the D-dimensional constrained domain of parameters to the unit ball ${\bf B}_0^D(1)$, then augments it to the D-dimensional

sphere ${\bf S}^D$ such that the original boundary corresponds to the equator of ${\bf S}^D$. This way, our method handles the constraints implicitly by moving freely on sphere generating proposals that remain within boundaries when mapped back to the original space. To improve the computational efficiency of our algorithm, we divide the dynamics into several parts such that the resulting split dynamics has a partial analytical solution as a geodesic flow on the sphere. We apply our method to several examples including truncated Gaussian, Bayesian Lasso, Bayesian bridge regression, and a copula model for identifying synchrony among multiple neurons. Our results show that the proposed method can provide a natural and efficient framework for handling several types of constraints on target distributions.

M53 Hamiltonian Monte Carlo Without Detailed Balance

Jascha Sohl-Dickstein; Mayur Mudigonda; Michael DeWeese

We present a method for performing Hamiltonian Monte Carlo that largely eliminates sample rejection. In situations that would normally lead to rejection, instead a longer trajectory is computed until a new state is reached that can be accepted. This is achieved using Markov chain transitions that satisfy the fixed point equation, but do not satisfy detailed balance. The resulting algorithm significantly suppresses the random walk behavior and wasted function evaluations that are typically the consequence of update rejection. We demonstrate a greater than factor of two improvement in mixing time on three test problems. We release the source code as Python and MATLAB packages.

M54 Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Ito Process

Issei Sato; Hiroshi Nakagawa

The stochastic gradient Langevin dynamics (SGLD) algorithm is appealing for large scale Bayesian learning. The SGLD algorithm seamlessly transit stochastic optimization and Bayesian posterior sampling. However, solid theories, such as convergence proof, have not been developed. We theoretically analyze the SGLD algorithm with constant stepsize in two ways. First, we show by using the Fokker-Planck equation that the probability distribution of random variables generated by the SGLD algorithm converges to the Bayesian posterior. Second, we analyze the convergence of the SGLD algorithm by using the Ito process, which reveals that the SGLD algorithm does not strongly but weakly converges. This result indicates that the SGLD algorithm can be an approximation method for posterior averaging.

14:00 - Track D - Method-Of-Moments and Spectral Methods

M55 Computing Parametric Ranking Models via Rank-Breaking

Hossein Azari Soufiani; David Parkes; Lirong Xia

Rank breaking is a methodology introduced by Azari Soufiani et al. (2013a) for applying a Generalized Method of Moments (GMM) algorithm to the estimation of parametric ranking models. Breaking takes full rankings and breaks, or splits them up, into counts for pairs of alternatives that occur in particular

positions (e.g., first place and second place, second place and third place). GMMs are of interest because they can achieve significant speed-up relative to maximum likelihood approaches and comparable statistical efficiency. We characterize the breakings for which the estimator is consistent for random utility models (RUMs) including Plackett-Luce and Normal-RUM, develop a general sufficient condition for a full breaking to be the only consistent breaking, and provide a trichotomy theorem in regard to single-edge breakings. Experimental results are presented to show the computational efficiency along with statistical performance of the proposed method.

M56 Learning Mixtures of Linear Classifiers

Yuekai Sun; Stratis Ioannidis; Andrea Montanari

We consider a discriminative learning (regression) problem, whereby the regression function is a convex combination of k linear classifiers. Existing approaches are based on the EM algorithm, or similar techniques, without provable guarantees. We develop a simple method based on spectral techniques and a `mirroring' trick, that discovers the subspace spanned by the classifiers' parameter vectors. Under a probabilistic assumption on the feature vector distribution, we prove that this approach has nearly optimal statistical efficiency.

M57 Methods of Moments for Learning Stochastic Languages: Unified Presentation and Empirical Comparison

Borja Balle; William Hamilton; Joelle Pineau

Probabilistic latent-variable models are a powerful tool for modelling structured data. However, traditional expectation-maximization methods of learning such models are both computationally expensive and prone to local-minima. In contrast to these traditional methods, recently developed learning algorithms based upon the method of moments are both computationally efficient and provide strong statistical guarantees. In this work, we provide a unified presentation and empirical comparison of three general moment-based methods in the context of modelling stochastic languages. By rephrasing these methods upon a common theoretical ground, introducing novel theoretical results where necessary, we provide a clear comparison, making explicit the statistical assumptions upon which each method relies. With this theoretical grounding, we then provide an in-depth empirical analysis of the methods on both real and synthetic data with the goal of elucidating performance trends and highlighting important implementation details.

M58 Estimating Latent-Variable Graphical Models using Moments and Likelihoods

Arun Tejasvi Chaganty; Percy Liang

Recent work in method of moments provide consistent estimates for latent-variable models, avoiding local optima issues, but these methods can only be applied to certain types of graphical models. In this work, we show that the method of moments in conjunction with a composite marginal likelihood objective yields consistent parameter estimates for a much broader class of directed and undirected graphical models, including loopy graphs with high treewidth. Specifically, we use tensor factorization to reveal partial information about the hidden variables, rendering the otherwise non-convex negative log-likelihood convex. Our approach gracefully extends to models outside our class by incorporating the partial information via posterior regulraization.

M59 Alternating Minimization for Mixed Linear Regression

Xinyang Yi; Constantine Caramanis; Sujay Sanghavi

Mixed linear regression involves the recovery of two (or more) unknown vectors from unlabeled linear measurements; that is, where each sample comes from exactly one of the vectors, but we do not know which one. It is a classic problem, and the natural and empirically most popular approach to its solution has been the EM algorithm. As in other settings, this is prone to bad local minima; however, each iteration is very fast (alternating between guessing labels, and solving with those labels). In this paper we provide a new initialization procedure for EM, based on finding the leading two eigenvectors of an appropriate matrix. We then show that with this, a re-sampled version of the EM algorithm provably converges to the correct vectors, under natural assumptions on the sampling distribution, and with nearly optimal (unimprovable) sample complexity. This provides not only the first characterization of EM's performance, but also much lower sample complexity as compared to both standard (randomly initialized) EM, and other methods for this problem.

M60 Dimension-free Concentration Bounds on Hankel Matrices for Spectral Learning

François Denis; Mattias Gybels; Amaury Habrard

Learning probabilistic models over strings is an important issue for many applications. Spectral methods propose elegant solutions to the problem of inferring weighted automata from finite samples of variable-length strings drawn from an unknown target distribution. These methods rely on a singular value decomposition of a matrix H_S, called the Hankel matrix, that records the frequencies of (some of) the observed strings. The accuracy of the learned distribution depends both on the quantity of information embedded in H_S and on the distance between H_S and its mean H_r. Existing concentration bounds seem to indicate that the concentration over H_r gets looser with its size, suggesting to make a trade-off between the quantity of used information and the size of H_r. We propose new dimension-free concentration bounds for several variants of Hankel matrices. Experiments demonstrate that these bounds are tight and that they significantly improve existing bounds. These results suggest that the concentration rate of the Hankel matrix around its mean does not constitute an argument for limiting its size.

14:00 - Track E - Boosting and Ensemble Methods

M61 Boosting with Online Binary Learners for the Multiclass Bandit Problem

Shang-Tse Chen; Hsuan-Tien Lin; Chi-Jen Lu

We consider the problem of online multiclass prediction in the bandit setting. Compared with the full-information setting, in which the learner can receive the true label as feedback after making each prediction, the bandit setting assumes that the learner can only know the correctness of the predicted label. Because the bandit setting is more restricted, it is difficult to design good bandit learners and currently there are not many bandit learners. In this paper, we propose an approach that systematically converts existing online binary classifiers to promising bandit learners with strong theoretical guarantee. The approach matches the idea of boosting, which has been shown to be powerful for batch learning as well as online learning. In particular, we establish the weak-learning condition on the online binary classifiers, and show that the condition allows automatically constructing a bandit learner with arbitrary

strength by combining several of those classifiers. Experimental results on several real-world data sets demonstrate the effectiveness of the proposed approach.

M62 Narrowing the Gap: Random Forests In Theory and In Practice

Misha Denil; David Matheson; Nando De Freitas

Despite widespread interest and practical use, the theoretical properties of random forests are still not well understood. In this paper we contribute to this understanding in two ways. We present a new theoreti- cally tractable variant of random regression forests and prove that our algorithm is con- sistent. We also provide an empirical eval- uation, comparing our algorithm and other theoretically tractable random forest models to the random forest algorithm used in prac- tice. Our experiments provide insight into the relative importance of different simplifi- cations that theoreticians have made to ob- tain tractable models for analysis.

M63 Ensemble Methods for Structured Prediction

Corinna Cortes; Vitaly Kuznetsov; Mehryar Mohri

We present a series of learning algorithms and theoretical guarantees for designing accurate ensembles of structured prediction tasks. This includes several randomized and deterministic algorithms devised by converting on-line learning algorithms to batch ones, and a boosting-style algorithm applicable in the context of structured prediction with a large number of labels. We give a detailed study of all these algorithms, including the description of new on-line-to-batch conversions and learning guarantees. We also report the results of extensive experiments with these algorithms in several structured prediction tasks.

M64 Deep Boosting

Corinna Cortes; Mehryar Mohri; Umar Syed

We present a new ensemble learning algorithm, DeepBoost, which can use as base classifiers a hypothesis set containing deep decision trees, or members of other rich or complex families, and succeed in achieving high accuracy without overfitting the data. The key to the success of the algorithm is a 'capacity-conscious' criterion for the selection of the hypotheses. We give new data-dependent learning bounds for convex ensembles expressed in terms of the Rademacher complexities of the sub-families composing the base classifier set, and the mixture weight assigned to each sub-family. Our algorithm directly benefits from these guarantees since it seeks to minimize the corresponding learning bound. We give a full description of our algorithm, including the details of its derivation, and report the results of several experiments showing that its performance compares favorably to that of AdaBoost and Logistic Regression and their L_1-regularized variants.

M65 Dynamic Programming Boosting for Discriminative Macro-Action Discovery

Leonidas Lefakis; Francois Fleuret

We consider the problem of automatic macro- action discovery in imitation learning, which we cast as one of change-point detection. Unlike prior work in change-point detection, the present work leverages discriminative learning algorithms. Our main contribution is a novel supervised learning algorithm which extends the classical Boosting framework by combining it with dynamic programming. The resulting

process alternatively improves the performance of individual strong predictors and the estimated change- points in the training sequence. Empirical evaluation is presented for the pro- posed method on tasks where change-points arise naturally as part of a classification problem. Finally we show the applicability of the algorithm to macro-action discovery in imitation learning and demonstrate it allows us to solve complex image-based goal-planning problems with thou- sands of features.

M66 A Convergence Rate Analysis for LogitBoost, MART and Their Variant

Peng Sun; Tong Zhang; Jie Zhou

LogitBoost, MART and their variant can be viewed as additive tree regression using logistic loss and boosting style optimization. We analyze their convergence rates based on a new weak learnability formulation. We show that it has $O(\frac{1}{T})$ rate when using gradient descent only, while a linear rate is achieved when using Newton descent. Moreover, introducing Newton descent when growing the trees, as LogitBoost does, leads to a faster linear rate. Empirical results on UCI datasets support our analysis.

14:00 - Track F - Neural Networks and Deep Learning II

M67 Learning to Disentangle Factors of Variation with Manifold Interaction

Scott Reed; Kihyuk Sohn; Yuting Zhang; Honglak Lee

Many latent factors of variation interact to generate sensory data; for example pose, morphology and expression in face images. We propose to learn manifold coordinates for the relevant factors of variation and to model their joint interaction. Most existing feature learning algorithms focus on a single task and extract features that are sensitive to the task-relevant factors and invariant to all others. However, models that just extract a single set of invariant features do not exploit the relationships among the latent factors. To address this we propose a higher-order Boltzmann machine that incorporates multiplicative interactions among groups of hidden units that each learn to encode a factor of variation. Furthermore, we propose a manifold-based training strategy that allows effective disentangling, meaning that units in each group encode a distinct type of variation. Our model achieves state-of-the-art emotion recognition and face verification performance on the Toronto Face Database, and we also demonstrate disentangled features learned on the CMU Multi-PIE dataset.

M68 Marginalized Denoising Auto-encoders for Nonlinear Representations

Minmin Chen; Kilian Weinberger; Fei Sha; Yoshua Bengio

Denoising auto-encoders (DAEs) have been successfully used to learn new representations for a wide range of machine learning tasks. During training, DAEs make many passes over the training dataset and reconstruct it from partial corruption generated from a pre-specified corrupting distribution. This process learns robust representation, though at the expense of requiring many training epochs, in which the data is explicitly corrupted. In this paper we present the marginalized Denoising Auto-encoder (mDAE), which (approximately) marginalizes out the corruption during training. Effectively, the mDAE takes into account infinitely many corrupted copies of the training data in every epoch, and therefore is able to match or outperform the DAE with much fewer training epochs. We analyze our proposed algorithm and show that it can be understood as a classic auto-encoder with a special form of

regularization. In empirical evaluations we show that it attains 1-2 order-of-magnitude speedup in training time over other competing approaches.

M69 Deep Generative Stochastic Networks Trainable by Backprop

Yoshua Bengio; Eric Laufer; Guillaume Alain; Jason Yosinski

We introduce a novel training principle for probabilistic models that is an alternative to maximum likelihood. The proposed Generative Stochastic Networks (GSN) framework is based on learning the transition operator of a Markov chain whose stationary distribution estimates the data distribution. Because the transition distribution is a conditional distribution generally involving a small move, it has fewer dominant modes, being unimodal in the limit of small moves. Thus, it is easier to learn, more like learning to perform supervised function approximation, with gradients that can be obtained by backprop. The theorems provided here generalize recent work on the probabilistic interpretation of denoising autoencoders and provide an interesting justification for dependency networks and generalized pseudolikelihood (along with defining an appropriate joint distribution and sampling mechanism, even when the conditionals are not consistent). GSNs can be used with missing inputs and can be used to sample subsets of variables given the rest. Successful experiments are conducted, validating these theoretical results, on two image datasets and with a particular architecture that mimics the Deep Boltzmann Machine Gibbs sampler but allows training to proceed with backprop, without the need for layerwise pretraining.

M70 Learning Ordered Representations with Nested Dropout

Oren Rippel; Michael Gelbart; Ryan Adams

In this paper, we present results on ordered representations of data in which different dimensions have different degrees of importance. To learn these representations we introduce nested dropout, a procedure for stochastically removing coherent nested sets of hidden units in a neural network. We first present a sequence of theoretical results in the simple case of a semi-linear autoencoder. We rigorously show that the application of nested dropout enforces identifiability of the units, which leads to an exact equivalence with PCA. We then extend the algorithm to deep models and demonstrate the relevance of ordered representations to a number of applications. Specifically, we use the ordered property of the learned codes to construct hash-based data structures that permit very fast retrieval, achieving retrieval in time logarithmic in the database size and independent of the dimensionality of the representation. This allows the use of codes that are hundreds of times longer than currently feasible for retrieval. We therefore avoid the diminished quality associated with short codes, while still performing retrieval that is competitive in speed with existing methods. We also show that ordered representations are a promising way to learn adaptive compression for efficient online data reconstruction.

M71 Efficient Gradient-Based Inference through Transformations between Bayes Nets and Neural Nets

Diederik Kingma; Max Welling

Pooling operators construct non-linear representations by cascading a redundant linear transform, followed by a point-wise nonlinearity and a local aggregation, typically implemented with a ℓ_p norm. Their efficiency in recognition architectures is based on their ability to locally contract the input space, but also on their capacity to retain as much stable information as possible. We address this latter

question by computing the upper and lower Lipschitz bounds of ℓ_p pooling operators for $p=1, 2, \infty$ as well as their half-rectified equivalents, which give sufficient conditions for the design of invertible pooling layers. Numerical experiments on MNIST and image patches confirm that pooling layers can be inverted with phase recovery algorithms. Moreover, the regularity of the inverse pooling, controlled by the lower Lipschitz constant, is empirically verified with a nearest neighbor regression.

16:20 - Track A - Matrix Factorization I

M73 Stochastic Inference for Scalable Probabilistic Modeling of Binary Matrices

Jose Miguel Hernandez-Lobato; Neil Houlsby; Zoubin Ghahramani

Fully observed large binary matrices appear in a wide variety of contexts. To model them, probabilistic matrix factorization (PMF) methods are an attractive solution. However, current batch algorithms for PMF can be inefficient because they need to analyze the entire data matrix before producing any parameter updates. We derive an efficient stochastic inference algorithm for PMF models of fully observed binary matrices. Our method exhibits faster convergence rates than more expensive batch approaches and has better predictive performance than scalable alternatives. The proposed method includes new data subsampling strategies which produce large gains over standard uniform subsampling. We also address the task of automatically selecting the size of the minibatches of data used by our method. For this, we derive an algorithm that adjusts this hyper-parameter online.

M74 Cold-start Active Learning with Robust Ordinal Matrix Factorization

Neil Houlsby; Jose Miguel Hernandez-Lobato; Zoubin Ghahramani

We present a new matrix factorization model for rating data and a corresponding active learning strategy to address the cold-start problem. Cold-start is one of the most challenging tasks for recommender systems: what to recommend with new users or items for which one has little or no data. An approach is to use active learning to collect the most useful initial ratings. However, the performance of active learning depends strongly upon having accurate estimates of i) the uncertainty in model parameters and ii) the intrinsic noisiness of the data. To achieve these estimates we propose a heteroskedastic Bayesian model for ordinal matrix factorization. We also present a computationally efficient framework for Bayesian active learning with this type of complex probabilistic model. This algorithm successfully distinguishes between informative and noisy data points. Our model yields state-of-the-art predictive performance and, coupled with our active learning strategy, enables us to gain useful information in the cold-start setting from the very first active sample.

M75 Probabilistic Matrix Factorization with Non-random Missing Data

Jose Miguel Hernandez-Lobato; Neil Houlsby; Zoubin Ghahramani

We propose a probabilistic matrix factorization model for collaborative filtering that learns from data that is missing not at random(MNAR). Matrix factorization models exhibit state-of-the-art predictive performance in collaborative filtering. However, these models usually assume that the data is missing at random (MAR), and this is rarely the case. For example, the data is not MAR if users rate items they like more than ones they dislike. When the MAR assumption is incorrect,

inferences are biased and predictive performance can suffer. Therefore, we model both the generative process for the data and the missing data mechanism. By learning these two models jointly we obtain improved performance over state-of-the-art methods when predicting the ratings and when modeling the data observation process. We present the first viable MF model for MNAR data. Our results are promising and we expect that further research on NMAR models will yield large gains in collaborative filtering.

M76 A Deep Semi-NMF Model for Learning Hidden Representations

George Trigeorgis; Konstantinos Bousmalis; Stefanos Zafeiriou; Bjoern Schuller

Semi-NMF is a matrix factorization technique that learns a low-dimensional representation of a dataset that lends itself to a clustering interpretation. It is possible that the mapping between this new representation and our original features contains rather complex hierarchical information with implicit lower-level hidden attributes, that classical one level clustering methodologies can not interpret. In this work we propose a novel model, Deep Semi-NMF, that is able to learn such hidden representations that allow themselves to an interpretation of clustering according to different, unknown attributes of a given dataset. We show that by doing so, our model is able to learn low-dimensional representations that are better suited for clustering, outperforming Semi-NMF, but also other NMF variants.

M77 Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing

Benjamin Haeffele; Eric Young; Rene Vidal

Recently, convex solutions to low-rank matrix factorization problems have received increasing attention in machine learning. However, in many applications the data can display other structures beyond simply being low-rank. For example, images and videos present complex spatio-temporal structures, which are largely ignored by current low-rank methods. In this paper we explore a matrix factorization technique suitable for large datasets that captures additional structure in the factors by using a projective tensor norm, which includes classical image regularizers such as total variation and the nuclear norm as particular cases. Although the resulting optimization problem is not convex, we show that under certain conditions on the factors, any local minimizer for the factors yields a global minimizer for their product. Examples in biomedical video segmentation and hyperspectral compressed recovery show the advantages of our approach on high-dimensional datasets.

16:20 - Track B - Learning Theory II

M78 Lower Bounds for the Gibbs Sampler over Mixtures of Gaussians

Christopher Tosh; Sanjoy Dasgupta

The mixing time of a Markov chain is the minimum time t necessary for the total variation distance between the distribution of the Markov chain's current state X_t and its stationary distribution to fall below some $\epsilon > 0$. In this paper, we present lower bounds for the mixing time of the Gibbs sampler over Gaussian mixture models with Dirichlet priors.

M79 (Near) Dimension Independent Risk Bounds for Differentially Private Learning

Prateek Jain; Abhradeep Guha Thakurta

In this paper, we study the problem of differentially private risk minimization where the goal is to provide differentially private algorithms that have small excess risk. In particular we address the following open problem: \emph{Is it possible to design computationally efficient differentially private risk minimizers with excess risk bounds that do not explicitly depend on dimensionality (p) and do not require structural assumptions like restricted strong convexity?} In this paper, we answer the question in the affirmative for a variant of the well-known \emph{output} and \emph{objective} perturbation algorithms [Chaudhuri et al., 2011]. In particular, we show that in generalized linear model, variants of both output and objective perturbation algorithms have no {\em explicit} dependence on p. Our results assume that the underlying loss function is a 1-Lipschitz convex function and we show that the excess risk depends only on L_2 norm of the true risk minimizer and that of training points. Next, we present a novel privacy preserving algorithm for risk minimization over simplex in the generalized linear model, where the loss function is a doubly differentiable convex function. Assuming that the training points have bounded L_∞-norm, our algorithm provides risk bound that has only {\em logarithmic} dependence on p. We also apply our technique to the online learning setting and obtain a regret bound with similar logarithmic dependence on p. In contrast, the existing differentially private online learning methods incur $O(\sqrt{p})$ dependence.

M80 Learning Theory and Algorithms for revenue optimization in second price auctions with reserve

Mehryar Mohri; Andres Munoz Medina

Second-price auctions with reserve play a critical role for modern search engine and popular online sites since the revenue of these companies often directly depends on the outcome of such auctions. The choice of the reserve price is the main mechanism through which the auction revenue can be influenced in these electronic markets. We cast the problem of selecting the reserve price to optimize revenue as a learning problem and present a full theoretical analysis dealing with the complex properties of the corresponding loss function (it is non-convex and discontinuous). We further give novel algorithms for solving this problem and report the results of encouraging experiments demonstrating their effectiveness.

M81 Multi-period Trading Prediction Markets with Connections to Machine Learning

Jinli Hu; Amos Storkey

We present a new model for prediction markets, in which we use risk measures to model agents and introduce a market maker to describe the trading process. This specific choice of modelling approach enables us to show that the whole market approaches a global objective, despite the fact that the market is designed such that each agent only cares about its own goal. In addition, the market dynamic provides a sensible algorithm for optimising the global objective. An intimate connection between machine learning and our markets is thus established, such that we could 1) analyse a market by applying machine learning methods to the global objective; and 2) solve machine learning problems by setting up and running certain markets.

M82 Towards Minimax Online Learning with Unknown Time Horizon

Haipeng Luo; Robert Schapire

We consider online learning when the time horizon is unknown. We apply a minimax analysis, beginning with the fixed horizon case, and then moving on to two unknown-horizon settings, one that assumes the horizon is chosen randomly according to some distribution, and the other which allows the adversary full control over the horizon. For the random horizon setting with restricted losses, we derive a fully optimal minimax algorithm. And for the adversarial horizon setting, we prove a nontrivial lower bound which shows that the adversary obtains strictly more power than when the horizon is fixed and known. Based on the minimax solution of the random horizon setting, we then propose a new adaptive algorithm which ``pretends'' that the horizon is drawn from a distribution from a special family, but no matter how the actual horizon is chosen, the worst-case regret is of the optimal rate. Furthermore, our algorithm can be combined and applied in many ways, for instance, to online convex optimization, follow the perturbed leader, exponential weights algorithm and first order bounds. Experiments show that our algorithm outperforms many other existing algorithms in an online linear optimization setting.

16:20 - Track C - Nonparametric Bayes I

M83 Rectangular Tiling Process

Masahiro Nakano; Katsuhiko Ishiguro; Akisato Kimura; Takeshi Yamada; Naonori Ueda

This paper proposes a novel stochastic process that represents the arbitrary rectangular partitioning of an infinite-dimensional matrix as the conditional projective limit. Rectangular partitioning is used in relational data analysis, and is classified into three types: regular grid, hierarchical, and arbitrary. Conventionally, a variety of probabilistic models have been advanced for the first two, including the product of Chinese restaurant processes and the Mondrian process. However, existing models for arbitrary partitioning are too complicated to permit the analysis of the statistical behaviors of models, which places very severe capability limits on relational data analysis. In this paper, we propose a new probabilistic model of arbitrary partitioning called the rectangular tiling process (RTP). Our model has a sound mathematical base in projective systems and infinite extension of conditional probabilities, and is capable of representing partitions of infinite elements as found in ordinary Bayesian nonparametric models.

M84 A reversible infinite HMM using normalised random measures

David Knowles; Zoubin Ghahramani; Konstantina Palla

We present a nonparametric prior over reversible Markov chains. We use completely random measures, specifically gamma processes, to construct a countably infinite graph with weighted edges. By enforcing symmetry to make the edges undirected we define a prior over random walks on graphs that results in a reversible Markov chain. The resulting prior over infinite transition matrices is closely related to the hierarchical Dirichlet process but enforces reversibility. A reinforcement scheme has recently been proposed with similar properties, but the de Finetti measure is not well characterised. We take the alternative approach of explicitly constructing the mixing measure, which allows more straightforward and efficient inference at the cost of no longer having a closed form predictive distribution. We use our process to construct a reversible infinite HMM which we apply to two real datasets, one from epigenomics and one ion channel recording.

M85 Scalable Bayesian Low-Rank Decomposition of Incomplete Multiway Tensors

Piyush Rai; Yingjian Wang; Shengbo Guo; Gary Chen; David Dunson; Lawrence Carin

We present a scalable Bayesian framework for low-rank decomposition of multiway tensor data with missing observations. The key issue of pre-specifying the rank of the decomposition is sidestepped in a principled manner using a multiplicative gamma process prior. Both continuous and binary data can be analyzed under the framework, in a coherent way using fully conjugate Bayesian analysis. In particular, the analysis in the non-conjugate binary case is facilitated via the use of the P\'olya-Gamma sampling strategy which elicits closed-form Gibbs sampling updates. The resulting samplers are efficient and enable us to apply our framework to large-scale problems, with time-complexity that is linear in the number of observed entries in the tensor. This is especially attractive in analyzing very large but sparsely observed tensors with very few known entries. Moreover, our method admits easy extension to the supervised setting where entities in one or more tensor modes have labels. Our method outperforms several state-of-the-art tensor decomposition methods on various synthetic and benchmark real-world datasets.

M86 Input Warping for Bayesian Optimization of Non-Stationary Functions

Jasper Snoek; Kevin Swersky; Rich Zemel; Ryan Adams

Bayesian optimization has proven to be a highly effective methodology for the global optimization of unknown, expensive and multimodal functions. The ability to accurately model distributions over functions is critical to the effectiveness of Bayesian optimization. Although Gaussian processes provide a flexible prior over functions, there are various classes of functions that remain difficult to model. One of the most frequently occurring of these is the class of non-stationary functions. The optimization of the hyperparameters of machine learning algorithms is a problem domain in which parameters are often manually transformed a priori, for example by optimizing in "log-space", to mitigate the effects of spatially-varying length scale. We develop a methodology for automatically learning a wide family of bijective transformations or warpings of the input space using the Beta cumulative distribution function. We further extend the warping framework to multi-task Bayesian optimization so that multiple tasks can be warped into a jointly stationary space. On a set of challenging benchmark optimization tasks, we observe that the inclusion of warping greatly improves on the state-of-the-art, producing better results faster and more reliably.

M87 Beta Diffusion Trees

Creighton Heaukulani; David Knowles; Zoubin Ghahramani

We define the beta diffusion tree, a random tree structure with a set of leaves that defines a collection of overlapping subsets of objects, known as a feature allocation. The generative process for the tree is defined in terms of particles (representing the objects) diffusing in some continuous space, analogously to the Dirichlet and Pitman–Yor diffusion trees (Neal, 2003b; Knowles & Ghahramani, 2011), both of which define tree structures over clusters of the particles. With the beta diffusion tree, however, multiple copies of a particle may exist and diffuse to multiple locations in the continuous space, resulting in (a random number of) possibly overlapping clusters of the objects. We demonstrate how to build a hierarchically-clustered factor analysis model with the beta diffusion tree and how to perform inference over the random tree structures with a Markov chain Monte Carlo algorithm. We conclude with several numerical experiments on missing data problems with data sets of gene expression arrays, international development statistics, and intranational socioeconomic measurements.

16:20 - Track D - Manifolds

M88 An Information Geometry of Statistical Manifold Learning

Ke Sun; Stéphane Marchand-Maillet

Manifold learning seeks low-dimensional representations of high-dimensional data. The main tactics have been exploring the geometry in an input data space and an output embedding space. We develop a manifold learning theory in a hypothesis space consisting of models. A model means a specific instance of a collection of points, e.g., the input data collectively or the output embedding collectively. The semi-Riemannian metric of this hypothesis space is uniquely derived in closed form based on the information geometry of probability distributions. There, manifold learning is interpreted as a trajectory of intermediate models. The volume of a continuous region reveals an amount of information. It can be measured to define model complexity and embedding quality. This provides deep unified perspectives of manifold learning theory.

M89 Geodesic Distance Function Learning via Heat Flow on Vector Fields

Binbin Lin; Ji Yang; Xiaofei He; Jieping Ye

Learning a distance function or metric on a given data manifold is of great importance in machine learning and pattern recognition. Many of the previous works first embed the manifold to Euclidean space and then learn the distance function. However, such a scheme might not faithfully preserve the distance function if the original manifold is not Euclidean. In this paper, we propose to learn the distance function directly on the manifold without embedding. We first provide a theoretical characterization of the distance function by its gradient field. Based on our theoretical analysis, we propose to first learn the gradient field of the distance function and then learn the distance function itself. Specifically, we set the gradient field of a local distance function as an initial vector field. Then we transport it to the whole manifold via heat flow on vector fields. Finally, the geodesic distance function can be obtained by requiring its gradient field to be close to the normalized vector field. Experimental results on both synthetic and real data demonstrate the effectiveness of our proposed algorithm.

M90 Two-Stage Metric Learning

Jun Wang; Ke Sun; Fei Sha; Stéphane Marchand-Maillet; Alexandros Kalousis

In this paper, we present a novel two-stage metric learning algorithm. We first map each learning instance to a probability distribution by computing its similarities to a set of fixed anchor points. Then, we define the distance in the input data space as the Fisher information distance on the associated statistical manifold. This induces in the input data space a new family of distance metric which presents unique properties. Unlike kernelized metric learning, we do not require the similarity measure to be positive semi-definite. Moreover, it can also be interpreted as a local metric learning algorithm with well defined distance approximation. We evaluate its performance on a number of datasets. It outperforms significantly other metric learning methods and SVM.

M91 Transductive Learning with Multi-class Volume Approximation

Gang Niu; Bo Dai; Christoffel du Plessis; Masashi Sugiyama

Given a hypothesis space, the large volume principle by Vladimir Vapnik prioritizes equivalence classes according to their volume in the hypothesis space. The volume approximation has hitherto been successfully applied to binary learning problems. In this paper, we propose a novel generalization to multiple classes, allowing applications of the large volume principle on more learning problems such as multi-class, multi-label and serendipitous learning in a transductive manner. Although the resultant learning method involves a non-convex optimization problem, the globally optimal solution is almost surely unique and can be obtained using O(n^3) time. Novel theoretical analyses are presented for the proposed method, and experimental results show it compares favorably with the one-vs-rest extension.

M92 Convergence rates for persistence diagram estimation in Topological Data Analysis

Frédéric Chazal; Marc Glisse; Catherine Labruère; Bertrand Michel

Computational topology has recently seen an important development toward data analysis, giving birth to Topological Data Analysis. Persistent homology appears as a fundamental tool in this field. We show that the use of persistent homology can be naturally considered in general statistical frameworks. We establish convergence rates of persistence diagrams associated to data randomly sampled from any compact metric space to a well defined limit diagram encoding the topological features of the support of the measure from which the data have been sampled. Our approach relies on a recent and deep stability result for persistence that allows to relate our problem to support estimation problems (with respect to the Gromov-Hausdorff distance). Some numerical experiments are performed in various contexts to illustrate our results.

16:20 - Track E - Kernel Methods I

M93 On p-norm Path Following in Multiple Kernel Learning for Non-linear Feature Selection

Pratik Jawanpuria; Manik Varma; Saketha Nath

Our objective is to develop formulations and algorithms for efficiently computing the feature selection path -- i.e. the variation in classification accuracy as the fraction of selected features is varied from null to unity. Multiple Kernel Learning subject to $l_{p\geq1}$ regularization (l_{p}-MKL) has been demonstrated to be one of the most effective techniques for non-linear feature selection. However, state-of-the-art l_p-MKL algorithms are too computationally expensive to be invoked thousands of times to determine the entire path. We propose a novel conjecture which states that, for certain l_p-MKL formulations, the number of features selected in the optimal solution monotonically decreases as p is decreased from an initial value to unity. We prove the conjecture, for a generic family of kernel target alignment based formulations, and show that the feature weights themselves decay (grow) monotonically once they are below (above) a certain threshold at optimality. This allows us to develop a path following algorithm that systematically generates optimal feature sets of decreasing size. The proposed algorithm sets certain feature weights directly to zero for potentially large intervals of p thereby reducing optimization costs while simultaneously providing approximation guarantees.

We empirically demonstrate that our formulation can lead to classification accuracies which are as much as 10\% higher on benchmark data sets not only as compared to other l_p-MKL formulations and uniform kernel baselines but also leading feature selection methods. We further demonstrate that our algorithm reduces training time significantly over other path following algorithms and state-of-the-art l_p-MKL optimizers such as SMO-MKL. In particular, we generate the entire feature selection path for data sets with a hundred thousand features in approximately half an hour on standard hardware.

M94 A Divide-and-Conquer Solver for Kernel Support Vector Machines

Cho-Jui Hsieh; Si Si; Inderjit Dhillon

The kernel support vector machine (SVM) is one of the most widely used classification methods; however, the amount of computation required becomes the bottleneck when facing millions of samples. In this paper, we propose and analyze a novel divide-and-conquer solver for kernel SVMs (DC-SVM). In the division step, we partition the kernel SVM problem into smaller subproblems by clustering the data, so that each subproblem can be solved independently and efficiently. We show theoretically that the support vectors identified by the subproblem solution are likely to be support vectors of the entire kernel SVM problem, provided that the problem is partitioned appropriately by kernel clustering. In the conquer step, the local solutions from the subproblems are used to initialize a global coordinate descent solver, which converges quickly as suggested by our analysis. By extending this idea, we develop a multilevel Divide-and-Conquer SVM algorithm with adaptive clustering and early prediction strategy, which outperforms state-of-the-art methods in terms of training speed, testing accuracy, and memory usage. As an example, on the covtype dataset with half-a-million samples, DC-SVM is 7 times faster than LIBSVM in obtaining the exact SVM solution (to within 10^{-6} relative error) which achieves 96.15% prediction accuracy. Moreover, with our proposed early prediction strategy, DC-SVM achieves about 96% accuracy in only 12 minutes, which is more than 100 times faster than LIBSVM.

M95 Memory Efficient Kernel Approximation

Si Si; Cho-Jui Hsieh; Inderjit Dhillon

The scalability of kernel machines is a big challenge when facing millions of samples due to storage and computation issues for large kernel matrices, that are usually dense. Recently, many papers have suggested tackling this problem by using a low rank approximation of the kernel matrix. In this paper, we first make the observation that the structure of shift-invariant kernels changes from low-rank to block-diagonal (without any low-rank structure) when varying the scale parameter. Based on this observation, we propose a new kernel approximation algorithm -- Memory Efficient Kernel Approximation (MEKA), which considers both low-rank and clustering structure of the kernel matrix. We show that the resulting algorithm outperforms state-of-the-art low-rank kernel approximation methods in terms of speed, approximation error, and memory usage. As an example, on the MNIST2M dataset with two-million samples, our method takes 550 seconds on a single machine using less than 500 MBytes memory to achieve 0.2313 test RMSE for kernel ridge regression, while standard Nystr\{o}m approximation takes more than 2700 seconds and uses more than 2 GBytes memory on the same problem to achieve 0.2318 test RMSE.

M96 Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection

Arun Iyer; Saketha Nath; Sunita Sarawagi

In recent times, many real world applications have emerged that require estimates of class ratios in an unlabeled instance collection as opposed to labels of individual instances in the collection. In this paper we investigate the use of maximum mean discrepancy (MMD) in a reproducing kernel Hilbert space (RKHS) for estimating such ratios. First, we theoretically analyze the MMD-based estimates. Our analysis establishes that, under some mild conditions, the estimate is statistically consistent. More importantly, it provides an upper bound on the error in the estimate in terms of intuitive geometric quantities like class separation and data spread. Next, we use the insights obtained from the theoretical analysis, to propose a novel convex formulation that automatically learns the kernel to be employed in the MMD-based estimation. We design an efficient cutting plane algorithm for solving this formulation. Finally, we empirically compare our estimator with several existing methods, and show significantly improved performance under varying datasets, class ratios, and training sizes.

M97 Robust and Efficient Kernel Hyperparameter Paths with Guarantees

Joachim Giesen; Soeren Laue; Patrick Wieschollek

Algorithmically, many machine learning tasks boil down to solving parameterized optimization problems. Finding good values for the parameters has significant influence on the statistical performance of these methods. Thus supporting the choice of parameter values algorithmically has received quite some attention recently, especially algorithms for computing the whole solution path of parameterized optimization problem. These algorithms can be used, for instance, to track the solution of a regularized learning problem along the regularization parameter path, or for tracking the solution of kernelized problems along a kernel hyperparameter path. Since exact path following algorithms can be numerically unstable, robust and efficient approximate path tracking algorithms became popular for regularized learning problems. By now algorithms with optimal path complexity are known for many regularized learning problems. That is not the case for kernel hyperparameter path tracking algorithms, where the exact path tracking algorithms can also suffer from numerical instabilities. The robust approximation algorithms for regularization path tracking can not be used directly for kernel hyperparameter path tracking problems since the latter fall into a different problem class. Here we address this problem by devising a robust and efficient path tracking algorithm that can also handle kernel hyperparameter paths and has asymptotically optimal complexity. We use this algorithm to compute approximate kernel hyperparamter solution paths for support vector machines and robust kernel regression. Experimental results for this problem applied to various data sets confirms the theoretical complexity analysis.

M98 Nonparametric Estimation of Multi-View Latent Variable Models

Le Song; Animashree Anandkumar; Bo Dai; Bo Xie

Spectral methods have greatly advanced the estimation of latent variable models, generating a sequence of novel and efficient algorithms with strong theoretical guarantees. However, current spectral algorithms are largely restricted to mixtures of discrete or Gaussian distributions. In this paper,

we propose a kernel method for learning multi-view latent variable models, allowing each mixture component to be nonparametric and learned from data in an unsupervised fashion. The key idea of our method is to embed the joint distribution of a multi-view latent variable model into a reproducing kernel Hilbert space, and then the latent parameters are recovered using a robust tensor power method. We establish that the sample complexity for the proposed method is quadratic in the number of latent components and is a low order polynomial in the other relevant parameters. Thus, our nonparametric tensor approach to learning latent variable models enjoys good sample and computational efficiencies. As a special case of our framework, we also obtain a first unsupervised conditional density estimator of the kind with provable guarantees. In both synthetic and real world datasets, the nonparametric tensor power method compares favorably to EM algorithm and other spectral algorithms.

16:20 - Track F - Unsupervised Learning and Detection

M99 Anomaly Ranking as Supervised Bipartite Ranking

Stephan Clémençon; Sylvain Robbiano

The Mass Volume (MV) curve is a visual tool to evaluate the performance of a scoring function with regard to its capacity to rank data in the same order as the underlying den- sity function. Anomaly ranking refers to the unsupervised learning task which consists in building a scoring function, based on unla- beled data, with a MV curve as low as pos- sible at any point. In this paper, it is proved that, in the case where the data generat- ing probability distribution has compact sup- port, anomaly ranking is equivalent to (su- pervised) bipartite ranking, where the goal is to discriminate between the underlying prob- ability distribution and the uniform distribu- tion with same support. In this situation, the MV curve can be then seen as a simple trans- form of the corresponding ROC curve. Ex- ploiting this view, we then show how to use bipartite ranking algorithms, possibly com- bined with random sampling, to solve the MV curve minimization problem. Numeri- cal experiments based on a variety of bipar- tite ranking algorithms well-documented in the literature are displayed in order to illus- trate the relevance of our approach.

M100 On learning to localize objects with minimal supervision

Hyun Oh Song; Ross Girshick; Stefanie Jegelka; Julien Mairal; Zaid Harchaoui; Trevor Darrell

Learning to localize objects with minimal supervision is an important problem in computer vision, since large fully annotated datasets are extremely costly to obtain. In this paper, we propose a new method that achieves this goal with only image-level labels of whether the objects are present or not. Our approach combines a discriminative submodular cover problem for automatically discovering a set of positive object windows with a smoothed latent SVM formulation. The latter allows us to leverage efficient quasi-Newton optimization techniques. Our experiments demonstrate that the proposed approach provides a 50% relative improvement in mean average precision over the current state-of-the-art on PASCAL VOC 2007 detection.

M101 Active Detection via Adaptive Submodularity

Yuxin Chen; Hiroaki Shioi; Cesar Fuentes Montesinos; Lian Pin Koh; Serge Wich; Andreas Krause

Efficient detection of multiple object instances is one of the fundamental challenges in computer vision. For certain object categories, even the best automatic systems are yet unable to produce high-quality detection results, and fully manual annotation would be an expensive process. How can detection algorithms interplay with human expert annotators? To make the best use of scarce (human) labeling resources, one needs to decide when to invoke the expert, such that the best possible performance can be achieved while requiring a minimum amount of supervision. In this paper, we propose a principled approach to active object detection, and show that for a rich class of base detectors algorithms, one can derive a natural sequential decision problem for deciding when to invoke expert supervision. We further show that the objective function satisfies adaptive submodularity, which allows us to derive strong performance guarantees for our algorithm. We demonstrate the proposed algorithm on three real-world tasks, including a problem for biodiversity monitoring from micro UAVs in the Sumatra rain forest. Our results show that active detection not only outperforms its passive counterpart; for certain tasks, it also works significantly better than straightforward application of existing active learning techniques. To the best of our knowledge, our approach is the first to rigorously address the active detection problem from both empirical and theoretical perspectives.

M102 Structured Generative Models of Natural Source Code

Chris Maddison; Daniel Tarlow

We study the problem of building generative models of natural source code (NSC); that is, source code written and understood by humans. Our primary contribution is to describe a family of generative models for NSC that have two key properties: First, they incorporate both sequential and hierarchical structure. Second, they are capable of integrating closely with a compiler, which allows leveraging compiler logic and abstractions when building structure into the model. We also develop an extension that includes more complex structure, refining how the model generates identifier tokens based on what variables are currently in scope. Our models can be learned efficiently, and we show empirically that including appropriate structure greatly improves the probability of generating test programs.

M103 Coordinate-descent for learning orthogonal matrices through Givens rotations

Uri Shalit; Gal Chechik

Optimizing over the set of orthogonal matrices is a central component in problems like sparse-PCA or tensor decomposition. Unfortunately, such optimization is hard since simple operations on orthogonal matrices easily break orthogonality, and correcting orthogonality usually costs a large amount of computation. Here we propose a framework for optimizing orthogonal matrices, that is the parallel of coordinate-descent in Euclidean spaces. It is based on {\em Givens-rotations}, a fast-to-compute operation that affects a small number of entries in the learned matrix, and preserves orthogonality. We show two applications of this approach: an algorithm for tensor decompositions used in learning mixture models, and an algorithm for sparse-PCA. We study the parameter regime where a Givens rotation approach converges faster and achieves a superior model on a genome-wide brain-wide mRNA expression dataset.

TUESDAY

June 24, 2014 at 8:30am

Title: On the Computational and Statistical Interface and "Big Data"

Convention Hall No.1

Keynote Speaker, Michael I. Jordan, University of California, Berkeley

Abstract:
The rapid growth in the size and scope of datasets in science and technology has created a need for novel foundational perspectives on data analysis that blend the statistical and computational sciences. That classical perspectives from these fields are not adequate to address emerging problems in "Big Data" is apparent from their sharply divergent nature at an elementary level---in computer science, the growth of the number of data points is a source of "complexity" that must be tamed via algorithms or hardware, whereas in statistics, the growth of the number of data points is a source of "simplicity" in that inferences are generally stronger and asymptotic results or concentration theorems can be invoked. We present several research vignettes on topics at the computation/statistics interface, an interface that we aim to characterize in terms of theoretical tradeoffs between statistical risk, amount of data and "externalities" such as computation, communication and privacy. [Joint work with VenkatChandrasekaran, John Duchi, Martin Wainwright and Yuchen Zhang.]

Bio:
Michael I. Jordan is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley., His research interests bridge the computational, statistical, cognitive and biological sciences, and have focused in recent years on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines and applications to problems in distributed computing systems, natural language processing, signal processing and statistical genetics. Prof. Jordan is a member of the National Academy of Sciences, a member of the National Academy of Engineering and a member of the American Academy of Arts and Sciences. He is a Fellow of the American Association for the Advancement of Science., He has been named a Neyman Lecturer and a Medallion Lecturer by the Institute of Mathematical Statistics, and has received the ACM/AAAI Allen Newell Award. He is a Fellow of the AAAI, ACM, ASA, CSS, IMS, IEEE and SIAM.

Tuesday June 24

10:50 - Track A - Matrix Factorization II (Room 305A)

T1 Rank-One Matrix Pursuit for Matrix Completion
Zheng Wang; Ming-Jun Lai; Zhaosong Lu; Wei Fan; Hasan Davulcu; Jieping Ye

T2 Convex Total Least Squares
Dmitry Malioutov; Nikolai Slavov

T3 Nuclear Norm Minimization via Active Subspace Selection
Cho-Jui Hsieh; Peder Olsen

T4 Riemannian Pursuit for Big Matrix Recovery
Mingkui Tan; Ivor W. Tsang; Li Wang; Bart Vandereycken; Sinno Jialin Pan

T5 Multiresolution Matrix Factorization
Risi Kondor; Nedelina Teneva; Vikas Garg

Tuesday June 24

10:50 - Track B - Bandits II (Room 201A)

T6 Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques
Jérémie Mary; Philippe Preux; Olivier Nicol

T7 Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem
Masrour Zoghi; Shimon Whiteson; Remi Munos; Maarten de Rijke

T8 Spectral Bandits for Smooth Graph Functions
Michal Valko; Remi Munos; Branislav Kveton; Tomáš Kocák

T9 Online Clustering of Bandits
Claudio Gentile; Shuai Li; Giovanni Zappella

T10 Latent Bandits
Odalric-Ambrym Maillard; Shie Mannor

Tuesday June 24

10:50 - Track C - Crowd-Sourcing (Room 307)

T11 Near-Optimally Teaching the Crowd to Classify
Adish Singla; Ilija Bogunovic; Gabor Bartok; Amin Karbasi; Andreas Krause

T12 Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy
Dengyong Zhou; Qiang Liu; John Platt; Christopher Meek

T13 Gaussian Process Classification and Active Learning with Multiple Annotators
Filipe Rodrigues; Francisco Pereira; Bernardete Ribeiro

T14 Ensemble-Based Tracking: Aggregating Crowdsourced Structured Time Series Data
Naiyan Wang; Dit-Yan Yeung

T15 Latent Confusion Analysis by Normalized Gamma Construction
Issei Sato; Hisashi Kashima; Hiroshi Nakagawa

Tuesday June 24

10:50 - Track D - Manifolds and Graphs (Room 201B)

T16 Graph-based Semi-supervised Learning: Realizing Pointwise Smoothness Probabilistically
Yuan Fang; Kevin Chang; Hady Lauw

T17 Wasserstein Propagation for Semi-Supervised Learning
Justin Solomon; Raif Rustamov; Leonidas Guibas; Adrian Butscher

T18 Optimization Equivalence of Divergences Improves Neighbor Embedding
Zhirong Yang; Jaakko Peltonen; Samuel Kaski

T19 Local Ordinal Embedding
Yoshikazu Terada; Ulrike von Luxburg

T20. The f-Adjusted Graph Laplacian: a Diagonal Modification with a Geometric Interpretation
Sven Kurras; Ulrike von Luxburg; Gilles Blanchard

Tuesday June 24

10:50 - Track E - Regularization and Lasso (Room 201C)

T21 A Unified Framework for Consistency of Regularized Loss Minimizers
Jean Honorio; Tommi Jaakkola

T22 Making the Most of Bag of Words: Sentence Regularization with Alternating Direction Method of Multipliers
Dani Yogatama; Noah Smith

T23 Coupled Group Lasso for Web-Scale CTR Prediction in Display Advertising
Ling Yan; Wu-Jun Li; Gui-Rong Xue; Dingyi Han

T24 Sample-based approximate regularization
Philip Bachman; Amir-Massoud Farahmand; Doina Precup

T25 Safe Screening with Variational Inequalities and Its Application to Lasso
Jun Liu; Zheng Zhao; Jie Wang; Jieping Ye

Tuesday June 24

10:50 - Track F - Nearest-Neighbors and Large-Scale Learning (Room 305B)

T26 Densifying One Permutation Hashing via Rotation for Fast Near Neighbor Search
Anshumali Shrivastava; Ping Li

T27 Coding for Random Projections
Ping Li; Michael Mitzenmacher; Anshumali Shrivastava

T28 Nearest Neighbors Using Compact Sparse Codes
Anoop Cherian

T29 Composite Quantization for Approximate Nearest Neighbor Search
Ting Zhang; Chao Du; Jingdong Wang

T30 Circulant Binary Embedding
Felix Yu; Sanjiv Kumar; Yunchao Gong; Shih-Fu Chang

Tuesday June 24

14:00 - Track A - Graphical Models II (Room 305A)

T31 Low-density Parity Constraints for Hashing-Based Discrete Integration
Stefano Ermon; Carla Gomes; Ashish Sabharwal; Bart Selman

T32 On Measure Concentration of Random Maximum A-Posteriori Perturbations
Francesco Orabona; Tamir Hazan; Anand Sarwate; Tommi Jaakkola

T33 Learning Sum-Product Networks with Direct and Indirect Variable Interactions
Amirmohammad Rooshenas; Daniel Lowd

T34 Multiple Testing under Dependence via Semiparametric Graphical Models
Jie Liu; Chunming Zhang; Elizabeth Burnside; David Page

T35 Discrete Chebyshev Classifiers
Elad Eban; Elad Mezuman; Amir Globerson

T36. Preserving Modes and Messages via Diverse Particle Selectio
Jason Pacheco; Silvia Zuffi; Michael Black; Erik Sudderth

Tuesday June 24

14:00 - Track B - Reinforcement Learning II (Room 201A)

T37 A new Q(lambda) with interim forward view and Monte Carlo equivalence
Rich Sutton; Ashique Rupam Mahmood; Doina Precup; Hado van Hasselt

T38 True Online TD(lambda)
Harm van Seijen; Rich Sutton

T39 Bias in Natural Actor-Critic Algorithms
Philip Thomas

T40 Deterministic Policy Gradient Algorithms
David Silver; Guy Lever; Nicolas Heess; Thomas Degris; Daan Wierstra; Martin Riedmiller

T41 Programming by Feedback
Riad Akrour; Marc Schoenauer; Jean-Christophe Souplet; Michele Sebag

T42 Active Learning of Parameterized Skills
Bruno Da Silva; George Konidaris; Andrew Barto

Tuesday June 24

14:00 - Track C - Topic Models (Room 305B)

T43 Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis
Jian Tang; Zhaoshi Meng; Xuanlong Nguyen; Qiaozhu Mei; Ming Zhang

T44 The Inverse Regression Topic Model
Maxim Rabinovich; David Blei

T45 On Modelling Non-linear Topical Dependencies
Zhixing Li; Siqiang Wen; Juanzi Li; Peng Zhang; Jie Tang

T46 Admixture of Poisson MRFs: A Topic Model with Word Dependencies
David Inouye; Pradeep Ravikumar; Inderjit Dhillon

T47 Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data
Zhiyuan Chen; Bing Liu

T48 Automated inference of point of view from user interactions in collective intelligence venues
Sanmay Das; Allen Lavoie

Tuesday June 24

14:00 - Track D - Sparsity (Room 201B)

T49 Gradient Hard Thresholding Pursuit for Sparsity-Constrained Optimization
Xiaotong Yuan; Ping Li; Tong Zhang

T50 Forward-Backward Greedy Algorithms for General Convex Smooth Functions over A Cardinality Constraint
Ji Liu; Jieping Ye; Ryohei Fujimaki

T51 Efficient Algorithms for Robust One-bit Compressive Sensing
Lijun Zhang; Jinfeng Yi; Rong Jin

T52 Nonlinear Information-Theoretic Compressive Measurement Design
Liming Wang; Abolfazl Razi; Miguel Rodrigues; Robert Calderbank; Lawrence Carin

T53 Elementary Estimators for High-Dimensional Linear Regression
Eunho Yang; Aurelie Lozano; Pradeep Ravikumar

T54 Statistical-Computational Phase Transitions in Planted Models: The High-Dimensional Setting
Yudong Chen; Jiaming Xu

Tuesday June 24

14:00 - Track E - Kernel Methods II (Room 201C)

T55 Kernel Mean Estimation and Stein Effect
Krikamol Muandet; Kenji Fukumizu; Bharath Sriperumbudur; Arthur Gretton; BernhardSchoelkopf

T56 A Kernel Independence Test for Random Processes
Kacper Chwialkowski; Arthur Gretton

T57 Quasi-Monte Carlo Feature Maps for Shift-Invariant Kernels
Jiyan Yang; Vikas Sindhwani; Haim Avron; Michael Mahoney

T58 A Unifying View of Representer Theorems
Andreas Argyriou; Francesco Dinuzzo

T59 Efficient Approximation of Cross-Validation for Kernel Methods using Bouligand Influence Function
Yong Liu; Shali Jiang; Shizhong Liao

Tuesday June 24

14:00 - Track F - Neural Theory and Spectral Methods (Room 307)

T60 Provable Bounds for Learning Some Deep Representations
Sanjeev Arora; Aditya Bhaskara; Rong Ge; Tengyu Ma

T61 K-means recovers ICA filters when independent components are sparse
Alon Vinnikov; Shai Shalev-Shwart

T62 Learning Polynomials with Neural Networks
Alexandr Andoni; Rina Panigrahy; Gregory Valiant; Li Zhang

T63 Anti-differentiating approximation algorithms:A case study with min-cuts, spectral, and flow
David Gleich; Michael Mahoney

T64 Nonnegative Sparse PCA with Provable Guarantees
Megasthenis Asteris; Dimitris Papailiopoulos; Alexandros Dimakis

T65 Finding Dense Subgraphs via Low-Rank Bilinear Optimization
Dimitris Papailiopoulos; Ioannis Mitliagkas; Alexandros Dimakis; Constantine Caramanis

Tuesday June 24

16:20 - Track A - Networks and Graph-Based Learning II (Room 305A)

T66 Learning Graphs with a Few Hubs
Rashish Tandon; Pradeep Ravikumar

T67 Global graph kernels using geometric embeddings
Fredrik Johansson; Vinay Jethava; Devdatt Dubhashi; Chiranjib Bhattacharyya

T68 Efficient Label Propagation
Yasuhiro Fujiwara; Go Irie

T69 Estimating Diffusion Network Structures: Recovery Conditions, Sample Complexity & Soft-thresholding Algorithm
Hadi Daneshmand; Manuel Gomez-Rodriguez; Le Song; Bernhard Schoelkopf

T70 Learning from Contagion (Without Timestamps)
Kareem Amin; Hoda Heidari; Michael Kearns

T71 Influence Function Learning in Information Diffusion Networks
Nan Du; Yingyu Liang; Maria Balcan; Le Song

Tuesday June 24

16:20 - Track B - Online Learning II (Room 307)

T72 Tracking Adversarial Targets
Yasin Abbasi-Yadkori; Peter Bartlett; Varun Kanade

T73 Sparse Reinforcement Learning via Convex Optimization
Zhiwei Qin; Weichang Li; Firdaus Janoos

T74 Online Learning in Markov Decision Processes with Changing Cost Sequences
Travis Dick; Andras Gyorgy; Csaba Szepesvari

T75 Linear Programming for Large-Scale Markov Decision Problems
Alan Malek; Yasin Abbasi-Yadkori; Peter Bartlett

T76 Statistical analysis of stochastic gradient methods for generalized linear models
Panagiotis Toulis; Edoardo Airoldi; Jason Rennie

T77 Preference-Based Rank Elicitation using Statistical Models: The Case of Mallows
Robert Busa-Fekete; Eyke Huellermeier; Balázs Szörényi

Tuesday June 24

16:20 - Track C - Nonparametric Bayes II (Room 305B)

T78 Bayesian Max-margin Multi-Task Learning with Data Augmentation
Chengtao Li; Jun Zhu; Jianfei Chen

T79 Variational Inference for Sequential Distance Dependent Chinese Restaurant Process
Sergey Bartunov; Dmitry Vetrov

T80 Pitfalls in the use of Parallel Inference for the Dirichlet Process
Yarin Gal; Zoubin Ghahramani

T81 Fast Allocation of Gaussian Process Experts
Trung Nguyen; Edwin Bonilla

T82 Scalable and Robust Bayesian Inference via the Median Posterior
Stanislav Minsker; Sanvesh Srivastava; Lizhen Lin; David Dunson

T83 Nonparametric Estimation of Renyi Divergence and Friends
Akshay Krishnamurthy; Kirthevasan Kandasamy; Barnabas Poczos; Larry Wasserman

Tuesday June 24

16:20 - Track D - Features and Feature Selection (Room 201A)

T84. Elementary Estimators for Sparse Covariance Matrices and other Structured Moments
Jun-Kun Wang; Shou-de Lin

T86. Making Fisher Discriminant Analysis Scalable
Bojun Tu; Zhihua Zhang; Shusen Wang; Hui Qian

T87. An Analysis of State-Relevance Weights and Sampling Distributions on L1-Regularized Approximate Linear Programming Approximation Accuracy
Gavin Taylor; Connor Geer; David Piekut

T88. Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball
Andrew Miller; Luke Bornn; Ryan Adams; Kirk Goldsberry

T89. Compact Random Feature Maps
Raffay Hamid; Ying Xiao; Alex Gittens; Dennis Decoste

Tuesday June 24

16:20 - Track E - Optimization III (Room 201B)

T90. New Primal SVM Solver with Linear Computational Cost for Big Data Classifications
Feiping Nie; Yizhen Huang; Xiaoqian Wang; Heng Huang

T91. Scaling SVM and Least Absolute Deviations via Exact Data Reduction
Jie Wang; Peter Wonka; Jieping Ye

T92. Margins, Kernels and Non-linear Smoothed Perceptrons
Aaditya Ramdas; Javier Peña

T93. Saddle Points and Accelerated Perceptron Algorithms
Adams Wei Yu; Fatma Kilinc-Karzan; Jaime Carbonell

T94. Outlier Path: A Homotopy Algorithm for Robust SVM
Shinya Suzumura; Kohei Ogawa; Masashi Sugiyama; Ichiro Takeuchi

T95. Optimal Budget Allocation: Theoretical Guarantee and Efficient Algorithm
Tasuku Soma; Naonori Kakimura; Kazuhiro Inaba; Ken-ichi Kawarabayashi

Tuesday June 24

16:20 - Track F - Time Series and Sequences (Room 201C)

T96. Boosting multi-step autoregressive forecasts
Souhaib Ben Taieb; Rob Hyndman

T97. Modeling Correlated Arrival Events with Latent Semi-Markov Processes
Wenzhao Lian; Vinayak Rao; Brian Eriksson; Lawrence Carin

T98. Asymptotically consistent estimation of the number of change points in highly dependent time series
Azadeh Khaleghi; Daniil Ryabko

T99. Effective Bayesian Modeling of Groups of Related Count Time Series
Nicolas Chapados

T100. Stochastic Variational Inference for Bayesian Time Series Models
Matthew Johnson; Alan Willsky

T101. Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models
Robert McGibbon; Bharath Ramsundar; Mohammad Sultan; Gert Kiss; Vijay Pande

M37	Probabilistic Partial Canonical Correlation Analysis	Yusuke Mukuta; Tatsuya Harada
M38	Min-Max Problems on Factor Graphs	Siamak Ravanbakhsh; Christopher Srinivasa; Brendan Frey; Russell Greiner
M39	Skip Context Tree Switching	Marc Bellemare; Joel Veness; Erik Talvitie
M40	Learning the Parameters of Determinantal Point Process Kernels	Raja Hafiz Affandi; Emily Fox; Ryan Adams; Ben Taskar
M41	Deterministic Anytime Inference for Stochastic Continuous-Time Markov Processes	E. Busra Celikkaya; Christian Shelton
M42	Doubly Stochastic Variational Bayes for non-Conjugate Inference	Michalis Titsias; Miguel Lázaro-Gredilla
M43	On the convergence of no-regret learning in selfish routing	Walid Krichene; Benjamin Drighès; Alexandre Bayen
M44	Optimal PAC Multiple Arm Identification with Applications to Crowdsourcing	Yuan Zhou; Xi Chen; Jian Li
M45	Prediction with Limited Advice and Multiarmed Bandits with Paid Observations	Yevgeny Seldin; Peter Bartlett; Koby Crammer; Yasin Abbasi-Yadkori
M46	One Practical Algorithm for Both Stochastic and Adversarial Bandits	Yevgeny Seldin; Aleksandrs Slivkins
M47	A Bayesian Framework for Online Classifier Ensemble	Qinxun Bai; Henry Lam; Stan Sclaroff
M48	Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits	Alekh Agarwal; Daniel Hsu; Satyen Kale; John Langford; Lihong Li; Robert Schapire
M49	Memory (and Time) Efficient Sequential Monte Carlo	Seong-Hwan Jun; Alexandre Bouchard-Côté
M50	Efficient Continuous-Time Markov Chain Estimation	Monir Hajiaghayi; Bonnie Kirkpatrick; Liangliang Wang; Alexandre Bouchard-Côté
M51	Filtering with Abstract Particles	Jacob Steinhardt; Percy Liang
M52	Spherical Hamiltonian Monte Carlo for Constrained Target Distributions	Shiwei Lan; Bo Zhou; Babak Shahbaba
M53	Hamiltonian Monte Carlo Without Detailed Balance	Jascha Sohl-Dickstein; Mayur Mudigonda; Michael DeWeese
M54	Approximation Analysis of Stochastic Gradient Langevin Dynamics by using Fokker-Planck Equation and Ito Process	Issei Sato; Hiroshi Nakagawa
M55	Computing Parametric Ranking Models via Rank-Breaking	Hossein Azari Soufiani; David Parkes; Lirong Xia
M56	Learning Mixtures of Linear Classifiers	Yuekai Sun; Stratis Ioannidis; Andrea Montanari
M57	Methods of Moments for Learning Stochastic Languages: Unified Presentation and Empirical Comparison	Borja Balle; William Hamilton; Joelle Pineau
M58	Estimating Latent-Variable Graphical Models using	Arun Tejasvi Chaganty; Percy Liang

	Moments and Likelihoods	
M59	Alternating Minimization for Mixed Linear Regression	Xinyang Yi; Constantine Caramanis; Sujay Sanghavi
M60	Dimension-free Concentration Bounds on Hankel Matrices for Spectral Learning	François Denis; Mattias Gybels; Amaury Habrard
M61	Boosting with Online Binary Learners for the Multiclass Bandit Problem	Shang-Tse Chen; Hsuan-Tien Lin; Chi-Jen Lu
M62	Narrowing the Gap: Random Forests In Theory and In Practice	Misha Denil; David Matheson; Nando De Freitas
M63	Ensemble Methods for Structured Prediction	Corinna Cortes; Vitaly Kuznetsov; Mehryar Mohri
M64	Deep Boosting	Corinna Cortes; Mehryar Mohri; Umar Syed
M65	Dynamic Programming Boosting for Discriminative Macro-Action Discovery	Leonidas Lefakis; Francois Fleuret
M66	A Convergence Rate Analysis for LogitBoost, MART and Their Variant	Peng Sun; Tong Zhang; Jie Zhou
M67	Learning to Disentangle Factors of Variation with Manifold Interaction	Scott Reed; Kihyuk Sohn; Yuting Zhang; Honglak Lee
M68	Marginalized Denoising Auto-encoders for Nonlinear Representations	Minmin Chen; Kilian Weinberger; Fei Sha; Yoshua Bengio
M69	Deep Generative Stochastic Networks Trainable by Backprop	Yoshua Bengio; Eric Laufer; Guillaume Alain; Jason Yosinski
M70	Learning Ordered Representations with Nested Dropout	Oren Rippel; Michael Gelbart; Ryan Adams
M71	Efficient Gradient-Based Inference through Transformations between Bayes Nets and Neural Nets	Diederik Kingma; Max Welling
M72	Signal recovery from Pooling Representations	Joan Bruna Estrach; Arthur Szlam; Yann LeCun
M73	Stochastic Inference for Scalable Probabilistic Modeling of Binary Matrices	Jose Miguel Hernandez-Lobato; Neil Houlsby; Zoubin Ghahramani
M74	Cold-start Active Learning with Robust Ordinal Matrix Factorization	Neil Houlsby; Jose Miguel Hernandez-Lobato; Zoubin Ghahramani
M75	Probabilistic Matrix Factorization with Non-random Missing Data	Jose Miguel Hernandez-Lobato; Neil Houlsby; Zoubin Ghahramani
M76	A Deep Semi-NMF Model for Learning Hidden Representations	George Trigeorgis; Konstantinos Bousmalis; Stefanos Zafeiriou; Bjoern Schuller
M77	Structured Low-Rank Matrix Factorization: Optimality, Algorithm, and Applications to Image Processing	Benjamin Haeffele; Eric Young; Rene Vidal
M78	Lower Bounds for the Gibbs Sampler over Mixtures of Gaussians	Christopher Tosh; Sanjoy Dasgupta
M79	(Near) Dimension Independent Risk Bounds for Differentially Private Learning	Prateek Jain; Abhradeep Guha Thakurta
M80	Learning Theory and Algorithms for revenue	Mehryar Mohri; Andres Munoz Medina

	optimization in second price auctions with reserve	
M81	Multi-period Trading Prediction Markets with Connections to Machine Learning	Jinli Hu; Amos Storkey
M82	Towards Minimax Online Learning with Unknown Time Horizon	Haipeng Luo; Robert Schapire
M83	Rectangular Tiling Process	Masahiro Nakano; Katsuhiko Ishiguro; Akisato Kimura; Takeshi Yamada; Naonori Ueda
M84	A reversible infinite HMM using normalised random measures	David Knowles; Zoubin Ghahramani; Konstantina Palla
M85	Scalable Bayesian Low-Rank Decomposition of Incomplete Multiway Tensors	Piyush Rai; Yingjian Wang; Shengbo Guo; Gary Chen; David Dunson; Lawrence Carin
M86	Input Warping for Bayesian Optimization of Non-Stationary Functions	Jasper Snoek; Kevin Swersky; Rich Zemel; Ryan Adams
M87	Beta Diffusion Trees	Creighton Heaukulani; David Knowles; Zoubin Ghahramani
M88	An Information Geometry of Statistical Manifold Learning	Ke Sun; Stéphane Marchand-Maillet
M89	Geodesic Distance Function Learning via Heat Flow on Vector Fields	Binbin Lin; Ji Yang; Xiaofei He; Jieping Ye
M90	Two-Stage Metric Learning	Jun Wang; Ke Sun; Fei Sha; Stéphane Marchand-Maillet; Alexandros Kalousis
M91	Transductive Learning with Multi-class Volume Approximation	Gang Niu; Bo Dai; Christoffel du Plessis; Masashi Sugiyama
M92	Convergence rates for persistence diagram estimation in Topological Data Analysis	Frédéric Chazal; Marc Glisse; Catherine Labruère; Bertrand Michel
M93	On p-norm Path Following in Multiple Kernel Learning for Non-linear Feature Selection	Pratik Jawanpuria; Manik Varma; Saketha Nath
M94	A Divide-and-Conquer Solver for Kernel Support Vector Machines	Cho-Jui Hsieh; Si Si; Inderjit Dhillon
M95	Memory Efficient Kernel Approximation	Si Si; Cho-Jui Hsieh; Inderjit Dhillon
M96	Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection	Arun Iyer; Saketha Nath; Sunita Sarawagi
M97	Robust and Efficient Kernel Hyperparameter Paths with Guarantees	Joachim Giesen; Soeren Laue; Patrick Wieschollek
M98	Nonparametric Estimation of Multi-View Latent Variable Models	Le Song; Animashree Anandkumar; Bo Dai; Bo Xie
M99	Anomaly Ranking as Supervised Bipartite Ranking	Stephan Clémençon; Sylvain Robbiano
M100	On learning to localize objects with minimal supervision	Hyun Oh Song; Ross Girshick; Stefanie Jegelka; Julien Mairal; Zaid Harchaoui; Trevor Darrell
M101	Active Detection via Adaptive Submodularity	Yuxin Chen; Hiroaki Shioi; Cesar Fuentes Montesinos; Lian Pin Koh; Serge Wich; Andreas Krause

M102	Structured Generative Models of Natural Source Code	Chris Maddison; Daniel Tarlow
M103	Coordinate-descent for learning orthogonal matrices through Givens rotations	Uri Shalit; Gal Chechik
T1	Rank-One Matrix Pursuit for Matrix Completion	Zheng Wang; Ming-Jun Lai; Zhaosong Lu; Wei Fan; Hasan Davulcu; Jieping Ye
T2	Convex Total Least Squares	Dmitry Malioutov; Nikolai Slavov
T3	Nuclear Norm Minimization via Active Subspace Selection	Cho-Jui Hsieh; Peder Olsen
T4	Riemannian Pursuit for Big Matrix Recovery	Mingkui Tan; Ivor W. Tsang; Li Wang; Bart Vandereycken; Sinno Jialin Pan
T5	Multiresolution Matrix Factorization	Risi Kondor; Nedelina Teneva; Vikas Garg
T6	Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques	Jérémie Mary; Philippe Preux; Olivier Nicol
T7	Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem	Masrour Zoghi; Shimon Whiteson; Remi Munos; Maarten de Rijke
T8	Spectral Bandits for Smooth Graph Functions	Michal Valko; Remi Munos; Branislav Kveton; Tomáš Kocák
T9	Online Clustering of Bandits	Claudio Gentile; Shuai Li; Giovanni Zappella
T10	Latent Bandits.	Odalric-Ambrym Maillard; Shie Mannor
T11	Near-Optimally Teaching the Crowd to Classify	Adish Singla; Ilija Bogunovic; Gabor Bartok; Amin Karbasi; Andreas Krause
T12	Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy	Dengyong Zhou; Qiang Liu; John Platt; Christopher Meek
T13	Gaussian Process Classification and Active Learning with Multiple Annotators	Filipe Rodrigues; Francisco Pereira; Bernardete Ribeiro
T14	Ensemble-Based Tracking: Aggregating Crowdsourced Structured Time Series Data	Naiyan Wang; Dit-Yan Yeung
T15	Latent Confusion Analysis by Normalized Gamma Construction	Issei Sato; Hisashi Kashima; Hiroshi Nakagawa
T16	Graph-based Semi-supervised Learning: Realizing Pointwise Smoothness Probabilistically	Yuan Fang; Kevin Chang; Hady Lauw
T17	Wasserstein Propagation for Semi-Supervised Learning	Justin Solomon; Raif Rustamov; Leonidas Guibas; Adrian Butscher
T18	Optimization Equivalence of Divergences Improves Neighbor Embedding	Zhirong Yang; Jaakko Peltonen; Samuel Kaski
T19	Local Ordinal Embedding	Yoshikazu Terada; Ulrike von Luxburg
T20	The f-Adjusted Graph Laplacian: a Diagonal Modification with a Geometric Interpretation	Sven Kurras; Ulrike von Luxburg; Gilles Blanchard
T21	A Unified Framework for Consistency of Regularized Loss Minimizers	Jean Honorio; Tommi Jaakkola
T22	Making the Most of Bag of Words: Sentence Regularization with Alternating Direction Method of Multipliers	Dani Yogatama; Noah Smith
T23	Coupled Group Lasso for Web-Scale CTR Prediction	Ling Yan; Wu-Jun Li; Gui-Rong Xue; Dingyi

	in Display Advertising	Han
T24	Sample-based approximate regularization	Philip Bachman; Amir-Massoud Farahmand; Doina Precup
T25	Safe Screening with Variational Inequalities and Its Application to Lasso	Jun Liu; Zheng Zhao; Jie Wang; Jieping Ye
T26	Densifying One Permutation Hashing via Rotation for Fast Near Neighbor Search	Anshumali Shrivastava; Ping Li
T27	Coding for Random Projections	Ping Li; Michael Mitzenmacher; Anshumali Shrivastava
T28	Nearest Neighbors Using Compact Sparse Codes	Anoop Cherian
T29	Composite Quantization for Approximate Nearest Neighbor Search	Ting Zhang; Chao Du; Jingdong Wang
T30	Circulant Binary Embedding	Felix Yu; Sanjiv Kumar; Yunchao Gong; Shih-Fu Chang
T31	Low-density Parity Constraints for Hashing-Based Discrete Integration	Stefano Ermon; Carla Gomes; Ashish Sabharwal; Bart Selman
T32	On Measure Concentration of Random Maximum A-Posteriori Perturbations	Francesco Orabona; Tamir Hazan; Anand Sarwate; Tommi Jaakkola
T33	Learning Sum-Product Networks with Direct and Indirect Variable Interactions	Amirmohammad Rooshenas; Daniel Lowd
T34	Multiple Testing under Dependence via Semiparametric Graphical Models	Jie Liu; Chunming Zhang; Elizabeth Burnside; David Page
T35	Discrete Chebyshev Classifiers	Elad Eban; Elad Mezuman; Amir Globerson
T36	Preserving Modes and Messages via Diverse Particle Selection	Jason Pacheco; Silvia Zuffi; Michael Black; Erik Sudderth
T37	A new Q(lambda) with interim forward view and Monte Carlo equivalence	Rich Sutton; Ashique Rupam Mahmood; Doina Precup; Hado van Hasselt
T38	True Online TD(lambda)	Harm van Seijen; Rich Sutton
T39	Bias in Natural Actor-Critic Algorithms	Philip Thomas
T40	Deterministic Policy Gradient Algorithms	David Silver; Guy Lever; Nicolas Heess; Thomas Degris; Daan Wierstra; Martin Riedmiller
T41	Programming by Feedback	Riad Akrour; Marc Schoenauer; Jean-Christophe Souplet; Michele Sebag
T42	Active Learning of Parameterized Skills	Bruno Da Silva; George Konidaris; Andrew Barto
T43	Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis	Jian Tang; Zhaoshi Meng; Xuanlong Nguyen; Qiaozhu Mei; Ming Zhang
T44	The Inverse Regression Topic Model	Maxim Rabinovich; David Blei
T45	On Modelling Non-linear Topical Dependencies	Zhixing Li; Siqiang Wen; Juanzi Li; Peng Zhang; Jie Tang
T46	Admixture of Poisson MRFs: A Topic Model with Word Dependencies	David Inouye; Pradeep Ravikumar; Inderjit Dhillon
T47	Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data	Zhiyuan Chen; Bing Liu

T48	Automated inference of point of view from user interactions in collective intelligence venues	Sanmay Das; Allen Lavoie
T49	Gradient Hard Thresholding Pursuit for Sparsity-Constrained Optimization	Xiaotong Yuan; Ping Li; Tong Zhang
T50	Forward-Backward Greedy Algorithms for General Convex Smooth Functions over A Cardinality Constraint	Ji Liu; Jieping Ye; Ryohei Fujimaki
T51	Efficient Algorithms for Robust One-bit Compressive Sensing	Lijun Zhang; Jinfeng Yi; Rong Jin
T52	Nonlinear Information-Theoretic Compressive Measurement Design	Liming Wang; Abolfazl Razi; Miguel Rodrigues; Robert Calderbank; Lawrence Carin
T53	Elementary Estimators for High-Dimensional Linear Regression	Eunho Yang; Aurelie Lozano; Pradeep Ravikumar
T54	Statistical-Computational Phase Transitions in Planted Models: The High-Dimensional Setting	Yudong Chen; Jiaming Xu
T55	Kernel Mean Estimation and Stein Effect	Krikamol Muandet; Kenji Fukumizu; Bharath Sriperumbudur; Arthur Gretton; Bernhard Schoelkopf
T56	A Kernel Independence Test for Random Processes	Kacper Chwialkowski; Arthur Gretton
T57	Quasi-Monte Carlo Feature Maps for Shift-Invariant Kernels	Jiyan Yang; Vikas Sindhwani; Haim Avron; Michael Mahoney
T58	A Unifying View of Representer Theorems	Andreas Argyriou; Francesco Dinuzzo
T59	Efficient Approximation of Cross-Validation for Kernel Methods using Bouligand Influence Function	Yong Liu; Shali Jiang; Shizhong Liao
T60	Provable Bounds for Learning Some Deep Representations	Sanjeev Arora; Aditya Bhaskara; Rong Ge; Tengyu Ma
T61	K-means recovers ICA filters when independent components are sparse	Alon Vinnikov; Shai Shalev-Shwartz
T62	Learning Polynomials with Neural Networks	Alexandr Andoni; Rina Panigrahy; Gregory Valiant; Li Zhang
T63	Anti-differentiating approximation algorithms:A case study with min-cuts, spectral, and flow	David Gleich; Michael Mahoney
T64	Nonnegative Sparse PCA with Provable Guarantees	Megasthenis Asteris; Dimitris Papailiopoulos; Alexandros Dimakis
T65	Finding Dense Subgraphs via Low-Rank Bilinear Optimization	Dimitris Papailiopoulos; Ioannis Mitliagkas; Alexandros Dimakis; Constantine Caramanis
T66	Learning Graphs with a Few Hubs	Rashish Tandon; Pradeep Ravikumar
T67	Global graph kernels using geometric embeddings	Fredrik Johansson; Vinay Jethava; Devdatt Dubhashi; Chiranjib Bhattacharyya
T68	Efficient Label Propagation	Yasuhiro Fujiwara; Go Irie
T69	Estimating Diffusion Network Structures: Recovery Conditions, Sample Complexity & Soft-thresholding Algorithm	Hadi Daneshmand; Manuel Gomez-Rodriguez; Le Song; Bernhard Schoelkopf

T70	Learning from Contagion (Without Timestamps)	Kareem Amin; Hoda Heidari; Michael Kearns
T71	Influence Function Learning in Information Diffusion Networks	Nan Du; Yingyu Liang; Maria Balcan; Le Song
T72	Tracking Adversarial Targets	Yasin Abbasi-Yadkori; Peter Bartlett; Varun Kanade
T73	Sparse Reinforcement Learning via Convex Optimization	Zhiwei Qin; Weichang Li; Firdaus Janoos
T74	Online Learning in Markov Decision Processes with Changing Cost Sequences	Travis Dick; Andras Gyorgy; Csaba Szepesvari
T75	Linear Programming for Large-Scale Markov Decision Problems	Alan Malek; Yasin Abbasi-Yadkori; Peter Bartlett
T76	Statistical analysis of stochastic gradient methods for generalized linear models	Panagiotis Toulis; Edoardo Airoldi; Jason Rennie
T77	Preference-Based Rank Elicitation using Statistical Models: The Case of Mallows	Robert Busa-Fekete; Eyke Huellermeier; Balázs Szörényi
T78	Bayesian Max-margin Multi-Task Learning with Data Augmentation	Chengtao Li; Jun Zhu; Jianfei Chen
T79	Variational Inference for Sequential Distance Dependent Chinese Restaurant Process	Sergey Bartunov; Dmitry Vetrov
T80	Pitfalls in the use of Parallel Inference for the Dirichlet Process	Yarin Gal; Zoubin Ghahramani
T81	Fast Allocation of Gaussian Process Experts	Trung Nguyen; Edwin Bonilla
T82	Scalable and Robust Bayesian Inference via the Median Posterior	Stanislav Minsker; Sanvesh Srivastava; Lizhen Lin; David Dunson
T83	Nonparametric Estimation of Renyi Divergence and Friends	Akshay Krishnamurthy; Kirthevasan Kandasamy; Barnabas Poczos; Larry Wasserman
T84	Elementary Estimators for Sparse Covariance Matrices and other Structured Moments	Eunho Yang; Aurelie Lozano; Pradeep Ravikumar
T85	Robust Inverse Covariance Estimation under Noisy Measurements	Jun-Kun Wang; Shou-de Lin
T86	Making Fisher Discriminant Analysis Scalable	Bojun Tu; Zhihua Zhang; Shusen Wang; Hui Qian
T87	An Analysis of State-Relevance Weights and Sampling Distributions on L1-Regularized Approximate Linear Programming Approximation Accuracy	Gavin Taylor; Connor Geer; David Piekut
T88	Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball	Andrew Miller; Luke Bornn; Ryan Adams; Kirk Goldsberry
T89	Compact Random Feature Maps	Raffay Hamid; Ying Xiao; Alex Gittens; Dennis Decoste
T90	New Primal SVM Solver with Linear Computational Cost for Big Data Classifications	Feiping Nie; Yizhen Huang; Xiaoqian Wang; Heng Huang
T91	Scaling SVM and Least Absolute Deviations via Exact Data Reduction	Jie Wang; Peter Wonka; Jieping Ye
T92	Margins, Kernels and Non-linear Smoothed	Aaditya Ramdas; Javier Peña

	Perceptrons	
T93	Saddle Points and Accelerated Perceptron Algorithms	Adams Wei Yu; Fatma Kilinc-Karzan; Jaime Carbonell
T94	Outlier Path: A Homotopy Algorithm for Robust SVM	Shinya Suzumura; Kohei Ogawa; Masashi Sugiyama; Ichiro Takeuchi
T95	Optimal Budget Allocation: Theoretical Guarantee and Efficient Algorithm	Tasuku Soma; Naonori Kakimura; Kazuhiro Inaba; Ken-ichi Kawarabayashi
T96	Boosting multi-step autoregressive forecasts	Souhaib Ben Taieb; Rob Hyndman
T97	Modeling Correlated Arrival Events with Latent Semi-Markov Processes	Wenzhao Lian; Vinayak Rao; Brian Eriksson; Lawrence Carin
T98	Asymptotically consistent estimation of the number of change points in highly dependent time series	Azadeh Khaleghi; Daniil Ryabko
T99	Effective Bayesian Modeling of Groups of Related Count Time Series	Nicolas Chapados
T100	Stochastic Variational Inference for Bayesian Time Series Models	Matthew Johnson; Alan Willsky
T101	Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models	Robert McGibbon; Bharath Ramsundar; Mohammad Sultan; Gert Kiss; Vijay Pande

Tuesday June 24

10:50 - Track A - Matrix Factorization II

T1 Rank-One Matrix Pursuit for Matrix Completion

Zheng Wang; Ming-Jun Lai; Zhaosong Lu; Wei Fan; Hasan Davulcu; Jieping Ye

Low rank matrix completion has been applied successfully in a wide range of machine learning applications, such as collaborative filtering, image inpainting and Microarray data imputation. However, many existing algorithms are not scalable to large-scale problems, as they involve computing singular value decomposition. In this paper, we present an efficient and scalable algorithm for matrix completion. The key idea is to extend the well-known orthogonal matching pursuit from the vector case to the matrix case. In each iteration, we pursue a rank-one matrix basis generated by the top singular vector pair of the current approximation residual and update the weights for all rank-one matrices obtained up to the current iteration. We further propose a novel weight updating rule to reduce the time and storage complexity, making the proposed algorithm scalable to large matrices. We establish the linear convergence of the proposed algorithm. The fast convergence is achieved due to the proposed construction of matrix bases and the estimation of the weights. We empirically evaluate the proposed algorithm on many real-world large scale datasets. Results show that our algorithm is much more efficient than state-of-the-art matrix completion algorithms while achieving similar or better prediction performance.

T2 Convex Total Least Squares

Dmitry Malioutov; Nikolai Slavov

We study the total least squares (TLS) problem that generalizes least squares regression by allowing measurement errors in both dependent and independent variables. TLS is widely used in applied fields including computer vision, system identifi cation and econometrics. The special case when all dependent and independent variables have the same level of uncorrelated Gaussian noise, known as ordinary TLS, can be solved by singular value decomposition (SVD). However, SVD cannot solve many important practical TLS problems with realistic noise structure, such as having varying measurement noise, known structure on the errors, or large outliers requiring robust error-norms. To solve such problems, we develop convex relaxation approaches for a general class of structured TLS (STLS). We show both theoretically and experimentally, that while the plain nuclear norm relaxation incurs large approximation errors for STLS, the re-weighted nuclear norm approach is very effective, and achieves better accuracy on challenging STLS problems than popular non-convex solvers. We describe a fast solution based on augmented Lagrangian formulation, and apply our approach to an important class of biological problems that use population average measurements to infer cell-type and physiological-state specific expression levels that are very hard to measure directly.

T3 Nuclear Norm Minimization via Active Subspace Selection

Cho-Jui Hsieh; Peder Olsen

We describe a novel approach to optimizing matrix problems involving nuclear norm regularization and apply it to the matrix completion problem. We combine methods from non-smooth and smooth optimization. At each step we use the proximal gradient to select an active subspace. We then find a smooth, convex relaxation of the smaller subspace problems and solve these using second order methods. We apply our methods to matrix completion problems including Netflix dataset, and show that they are more than 6 times faster than state-of-the-art nuclear norm solvers. Also, this is the first paper to scale nuclear norm solvers to the Yahoo-Music dataset, and the first time in the literature that the efficiency of nuclear norm solvers can be compared and even compete with non-convex solvers like Alternating Least Squares (ALS).

T4 Riemannian Pursuit for Big Matrix Recovery

Mingkui Tan; Ivor W. Tsang; Li Wang; Bart Vandereycken; Sinno Jialin Pan

Low rank matrix recovery is a fundamental task in many real-world applications. The performance of existing methods, however, deteriorates significantly when applied to ill-conditioned or large-scale matrices. In this paper, we therefore propose an efficient method, called Riemannian Pursuit (RP), that aims to address these two problems simultaneously. Our method consists of a sequence of fixed-rank optimization problems. Each subproblem, solved by a nonlinear Riemannian conjugate gradient method, aims to correct the solution in the most important subspace of increasing size. Theoretically, RP converges linearly under mild conditions and experimental results show that it substantially outperforms existing methods when applied to large-scale and ill-conditioned matrices.

T5 Multiresolution Matrix Factorization

Risi Kondor; Nedelina Teneva; Vikas Garg

The types of large matrices that appear in modern Machine Learning problems often have complex hierarchical structures that go beyond what can be found by traditional linear algebra tools, such as eigendecompositions. Inspired by ideas from multiresolution analysis, this paper introduces a new notion of matrix factorization that can capture structure in matrices at multiple different scales. The resulting Multiresolution Matrix Factorizations (MMFs) not only provide a wavelet basis for sparse approximation, but can also be used for matrix compression (similar to Nystrom approximations) and as a prior for matrix completion.

Tuesday June 24

10:50 - Track B - Bandits II

T6 Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques

Jérémie Mary; Philippe Preux; Olivier Nicol

In many recommendation applications such as news recommendation, the items that can be recommended come and go at a very fast pace. This is a challenge for recommender systems (RS) to face this setting. Online learning algorithms seem to be the most straight forward solution. The contextual bandit framework was introduced for that very purpose. In general the evaluation of a RS is a critical issue. Live evaluation is often avoided due to the potential loss of revenue, hence the need for offline evaluation methods. Two options are available. Model based methods are biased by nature and are thus difficult to trust when used alone. Data driven methods are therefore what we consider here. Evaluating online learning algorithms with past data is not simple but some methods exist in the literature. Nonetheless their accuracy is not satisfactory mainly due to their mechanism of data rejection that only allow the exploitation of a small fraction of the data. We precisely address this issue in this paper. After highlighting the limitations of the previous methods, we present a new method, based on bootstrapping techniques. This new method comes with two important improvements: it is much more accurate and it provides a measure of quality of its estimation. The latter is a highly desirable property in order to minimize the risks entailed by putting online a RS for the first time. We provide both theoretical and experimental proofs of its superiority compared to state-of-the-art methods, as well as an analysis of the convergence of the measure of quality.

T7 Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem

Masrour Zoghi; Shimon Whiteson; Remi Munos; Maarten de Rijke

This paper proposes a new method for the K-armed dueling bandit problem, a variation on the regular K-armed bandit problem that offers only relative feedback about pairs of arms. Our approach extends the Upper Confidence Bound algorithm to the relative setting by using estimates of the pairwise probabilities to select a promising arm and applying Upper Confidence Bound with the winner as a benchmark. We prove a sharp finite-time regret bound of order O(K log t) on a very general class of dueling bandit problems that matches a lower bound proven in (Yue et al., 2012). In addition, our empirical results using real data from an information retrieval application show that it greatly outperforms the state of the art.

T8 Spectral Bandits for Smooth Graph Functions

Michal Valko; Remi Munos; Branislav Kveton; Tomáš Kocák

Smooth functions on graphs have wide applications in manifold and semi-supervised learning. In this paper, we study a bandit problem where the payoffs of arms are smooth on a graph. This framework is suitable for solving online learning problems that involve graphs, such as content-based recommendation. In this problem, each item we can recommend is a node and its expected rating is similar to its neighbors. The goal is to recommend items that have high expected ratings. We aim for the algorithms where the cumulative regret with respect to the optimal policy would not scale poorly with the number of nodes. In particular, we introduce the notion of an effective dimension, which is small in

real-world graphs, and propose two algorithms for solving our problem that scale linearly and sublinearly in this dimension. Our experiments on real-world content recommendation problem show that a good estimator of user preferences for thousands of items can be learned from just tens of nodes evaluations.

T9 Online Clustering of Bandits

Claudio Gentile; Shuai Li; Giovanni Zappella

We introduce a novel algorithmic approach to content recommendation based on adaptive clustering of exploration-exploitation (``bandit'') strategies. We provide a sharp regret analysis of this algorithm in a standard stochastic noise setting, demonstrate its scalability properties, and prove its effectiveness on a number of artificial and real-world datasets. Our experiments show a significant increase in prediction performance over state-of-the-art methods for bandit problems.

T10 Latent Bandits.

Odalric-Ambrym Maillard; Shie Mannor

We consider a multi-armed bandit problem where the reward distributions are indexed by two sets -- one for arms, one for type-- and can be partitioned into a small number of clusters according to the type. First, we consider the setting where all reward distributions are known and all types have the same underlying cluster, the type's identity is, however, unknown. Second, we study the case where types may come from different classes, which is significantly more challenging. Finally, we tackle the case where the reward distributions are completely unknown. In each setting, we introduce specific algorithms and derive non-trivial regret performance. Numerical experiments show that, in the most challenging agnostic case, the proposed algorithm achieves excellent performance in several difficult scenarios.

10:50 - Track C - Crowd-Sourcing

T11 Near-Optimally Teaching the Crowd to Classify

Adish Singla; Ilija Bogunovic; Gabor Bartok; Amin Karbasi; Andreas Krause

How should we present training examples to learners to teach them classification rules? This is a natural problem when training workers for crowdsourcing labeling tasks, and is also motivated by challenges in data-driven online education. We propose a natural stochastic model of the learners, modeling them as randomly switching among hypotheses based on observed feedback. We then develop STRICT, an efficient algorithm for selecting examples to teach to workers. Our solution greedily maximizes a submodular surrogate objective function in order to select examples to show to the learners. We prove that our strategy is competitive with the optimal teaching policy. Moreover, for the special case of linear

separators, we prove that an exponential reduction in error probability can be achieved. Our experiments on simulated workers as well as three real image annotation tasks on Amazon Mechanical Turk show the effectiveness of our teaching algorithm.

T12 Aggregating Ordinal Labels from Crowds by Minimax Conditional Entropy

Dengyong Zhou; Qiang Liu; John Platt; Christopher Meek

We propose a method to aggregate noisy ordinal labels collected from a crowd of workers or annotators. Eliciting ordinal labels is important in tasks such as judging web search quality and consumer satisfaction. Our method is motivated by the observation that workers usually have difficulty distinguishing between two adjacent ordinal classes whereas distinguishing between two classes which are far away from each other is much easier. We develop the method through minimax conditional entropy subject to constraints which encode this observation. Empirical evaluations on real datasets demonstrate significant improvements over existing methods.

T13 Gaussian Process Classification and Active Learning with Multiple Annotators

Filipe Rodrigues; Francisco Pereira; Bernardete Ribeiro

Learning from multiple annotators took a valuable step towards modelling data that does not fit the usual single annotator setting. However, multiple annotators sometimes offer varying degrees of expertise. When disagreements arise, the establishment of the correct label through trivial solutions such as majority voting may not be adequate, since without considering heterogeneity in the annotators, we risk generating a flawed model. In this paper, we extend GP classification in order to account for multiple annotators with different levels expertise. By explicitly handling uncertainty, Gaussian processes (GPs) provide a natural framework to build proper multiple-annotator models. We empirically show that our model significantly outperforms other commonly used approaches, such as majority voting, without a significant increase in the computational cost of approximate Bayesian inference. Furthermore, an active learning methodology is proposed, which is able to reduce annotation cost even further.

T14 Ensemble-Based Tracking: Aggregating Crowdsourced Structured Time Series Data

Naiyan Wang; Dit-Yan Yeung

We study the problem of aggregating the contributions of multiple contributors in a crowdsourcing setting. The data involved is in a form not typically considered in most crowdsourcing tasks, in that the data is structured and has a temporal dimension. In particular, we study the visual tracking problem in which the unknown data to be estimated is in the form of a sequence of bounding boxes representing the trajectory of the target object being tracked. We propose a factorial hidden Markov model (FHMM) for ensemble-based tracking by learning jointly the unknown trajectory of the target and the reliability of each tracker in the ensemble. For efficient online inference of the FHMM, we devise a conditional particle filter algorithm by exploiting the structure of the joint posterior distribution of the hidden

variables. Using the largest open benchmark for visual tracking, we empirically compare two ensemble methods constructed from five state-of-the-art trackers with the individual trackers. The promising experimental results provide empirical evidence for our ensemble approach to "get the best of all worlds".

T15 Latent Confusion Analysis by Normalized Gamma Construction

Issei Sato; Hisashi Kashima; Hiroshi Nakagawa

We developed a flexible framework for modeling the annotation and judgment processes of humans, which we called ``normalized gamma construction of a confusion matrix.'' This framework enabled us to model three properties: (1) the abilities of humans, (2) a confusion matrix with labeling, and (3) the difficulty with which items are correctly annotated. We also provided the concept of ``latent confusion analysis (LCA),'' whose main purpose was to analyze the principal confusions behind human annotations and judgments. It is assumed in LCA that confusion matrices are shared between persons, which we called ``latent confusions'', in tribute to the ``latent topics'' of topic modeling. We aim at summarizing the workers' confusion matrices with the small number of latent principal confusion matrices because many personal confusion matrices is difficult to analyze. We used LCA to analyze latent confusions regarding the effects of radioactivity on fish and shellfish following the Fukushima Daiichi nuclear disaster in 2011.

Tuesday June 24

10:50 - Track D - Manifolds and Graphs

T16 Graph-based Semi-supervised Learning: Realizing Pointwise Smoothness Probabilistically

Yuan Fang; Kevin Chang; Hady Lauw

As the central notion in semi-supervised learning, smoothness is often realized on a graph representation of the data. In this paper, we study two complementary dimensions of smoothness: its pointwise nature and probabilistic modeling. While no existing graph-based work exploits them in conjunction, we encompass both in a novel framework of Probabilistic Graph-based Pointwise Smoothness (PGP), building upon two foundational models of data closeness and label coupling. This new form of smoothness axiomatizes a set of probability constraints, which ultimately enables class prediction. Theoretically, we provide an error and robustness analysis of PGP. Empirically, we conduct extensive experiments to show the advantages of PGP.

T17 Wasserstein Propagation for Semi-Supervised Learning

Justin Solomon; Raif Rustamov; Leonidas Guibas; Adrian Butscher

Probability distributions and histograms are natural representations for product ratings, traffic measurements, and other data considered in many machine learning applications. Thus, this paper

introduces a technique for graph-based semi-supervised learning of histograms, derived from the theory of optimal transportation. Our method has several properties making it suitable for this application; in particular, its behavior can be characterized by the moments and shapes of the histograms at the labeled nodes. In addition, it can be used for histograms on non-standard domains like circles, revealing a strategy for manifold-valued semi-supervised learning. We also extend this technique to related problems such as smoothing distributions on graph nodes.

T18 Optimization Equivalence of Divergences Improves Neighbor Embedding

Zhirong Yang; Jaakko Peltonen; Samuel Kaski

Visualization methods that arrange data objects in 2D or 3D layouts have followed two main schools, methods oriented for graph layout and methods oriented for vectorial embedding. We show the two previously separate approaches are tied by an optimization equivalence, making it possible to relate methods from the two approaches and to build new methods that take the best of both worlds. In detail, we prove a theorem of optimization equivalences between beta- and gamma-, as well as alpha- and Renyi-divergences through a connection scalar. Through the equivalences we represent several nonlinear dimensionality reduction and graph drawing methods in a generalized stochastic neighbor embedding setting, where information divergences are minimized between similarities in input and output spaces, and the optimal connection scalar provides a natural choice for the tradeoff between attractive and repulsive forces. We give two examples of developing new visualization methods through the equivalences: 1) We develop weighted symmetric stochastic neighbor embedding (ws-SNE) from Elastic Embedding and analyze its benefits, good performance for both vectorial and network data; in experiments ws-SNE has good performance across data sets of different types, whereas comparison methods fail for some of the data sets; 2) we develop a gamma-divergence version of a PolyLog layout method; the new method is scale invariant in the output space and makes it possible to efficiently use large-scale smoothed neighborhoods.

T19 Local Ordinal Embedding

Yoshikazu Terada; Ulrike von Luxburg

We study the problem of ordinal embedding: given a set of ordinal constraints of the form $distance(i,j) < distance(k,l)$ for some quadruples (i,j,k,l) of indices, the goal is to construct a point configuration $\hat{\bm{x}}_1, ..., \hat{\bm{x}}_n$ in $\R^p$ that preserves these constraints as well as possible. Our first contribution is to suggest a simple new algorithm for this problem, Soft Ordinal Embedding. The key feature of the algorithm is that it recovers not only the ordinal constraints, but even the density structure of the underlying data set. As our second contribution we prove that in the large sample limit it is enough to know ``local ordinal information'' in order to perfectly reconstruct a given point configuration. This leads to our Local Ordinal Embedding algorithm, which can also be used for graph drawing.

T20 The f-Adjusted Graph Laplacian: a Diagonal Modification with a Geometric Interpretation

Sven Kurras; Ulrike von Luxburg; Gilles Blanchard

Consider a neighborhood graph, for example a k-nearest neighbor graph, that is constructed on sample points drawn according to some density p. Our goal is to re-weight the graph's edges such that all cuts and volumes behave as if the graph was built on a different sample drawn from an alternative density q. We introduce the f-adjusted graph and prove that it provides the correct cuts and volumes as the sample size tends to infinity. From an algebraic perspective, we show that its normalized Laplacian, denoted as the f-adjusted Laplacian, represents a natural family of diagonal perturbations of the original normalized Laplacian. Our technique allows to apply any cut and volume based algorithm to the f-adjusted graph, for example spectral clustering, in order to study the given graph as if it were built on an unaccessible sample from a different density. We point out applications in sample bias correction, data uniformization, and multi-scale analysis of graphs.

10:50 - Track E - Regularization and Lasso

T21 A Unified Framework for Consistency of Regularized Loss Minimizers

Jean Honorio; Tommi Jaakkola

We characterize a family of regularized loss minimization problems that satisfy three properties: scaled uniform convergence, super-norm regularization, and norm-loss monotonicity. We show several theoretical guarantees within this framework, including loss consistency, norm consistency, sparsistency (i.e. support recovery) as well as sign consistency. A number of regularization problems can be shown to fall within our framework and we provide several examples. Our results can be seen as a concise summary of existing guarantees but we also extend them to new settings. Our formulation enables us to assume very little about the hypothesis class, data distribution, the loss, or the regularization. In particular, many of our results do not require a bounded hypothesis class, or identically distributed samples. Similarly, we do not assume boundedness, convexity or smoothness of the loss nor the regularizer. We only assume approximate optimality of the empirical minimizer. In terms of recovery, in contrast to existing results, our sparsistency and sign consistency results do not require knowledge of the sub-differential of the objective function.

T22 Making the Most of Bag of Words: Sentence Regularization with Alternating Direction Method of Multipliers

Dani Yogatama; Noah Smith

In many high-dimensional learning problems, only some parts of an observation are important to the prediction task; for example, the cues to correctly categorizing a document may lie in a handful of its sentences. We introduce a learning algorithm that exploits this intuition by encoding it in a regularizer. Specifically, we apply the sparse overlapping group lasso with one group for every bundle of features occurring together in a training-data sentence, leading to thousands to millions of overlapping groups. We show how to efficiently solve the resulting optimization challenge using the alternating directions

method of multipliers. We find that the resulting method significantly outperforms competitive baselines (standard ridge, lasso, and elastic net regularizers) on a suite of real-world text categorization problems.

T23 Coupled Group Lasso for Web-Scale CTR Prediction in Display Advertising

Ling Yan; Wu-Jun Li; Gui-Rong Xue; Dingyi Han

In display advertising, click through rate(CTR) prediction is the problem of estimating the probability that an advertisement (ad) is clicked when displayed to a user in a specific context. Due to its easy implementation and promising performance, logistic regression(LR) model has been widely used for CTR prediction, especially in industrial systems. However, it is not easy for LR to capture the nonlinear information, such as the conjunction information, from user features and ad features. In this paper, we propose a novel model, called coupled group lasso(CGL), for CTR prediction in display advertising. CGL can seamlessly integrate the conjunction information from user features and ad features for modeling. Furthermore, CGL can automatically eliminate useless features for both users and ads, which may facilitate fast online prediction. Scalability of CGL is ensured through feature hashing and distributed implementation. Experimental results on real-world data sets show that our CGL model can achieve state-of-the-art performance on web-scale CTR prediction tasks.

T24 Sample-based approximate regularization

Philip Bachman; Amir-Massoud Farahmand; Doina Precup

We introduce a method for regularizing linearly parameterized functions using general derivative-based penalties, which relies on sampling as well as finite-difference approximations of the relevant derivatives. We call this approach sample-based approximate regularization (SAR). We provide theoretical guarantees on the fidelity of such regularizers, compared to those they approximate, and prove that the approximations converge efficiently. We also examine the empirical performance of SAR on several datasets.

T25 Safe Screening with Variational Inequalities and Its Application to Lasso

Jun Liu; Zheng Zhao; Jie Wang; Jieping Ye

Sparse learning techniques have been routinely used for feature selection as the resulting model usually has a small number of non-zero entries. Safe screening, which eliminates the features that are guaranteed to have zero coefficients for a certain value of the regularization parameter, is a technique for improving the computational efficiency. Safe screening is gaining increasing attention since 1) solving sparse learning formulations usually has a high computational cost especially when the number of features is large and 2) one needs to try several regularization parameters to select a suitable model. In this paper, we propose an approach called ``Sasvi" (Safe screening with variational inequalities). Sasvi makes use of the variational inequality that provides the sufficient and necessary optimality condition for the dual problem. Several existing approaches for Lasso screening can be casted as relaxed versions

of the proposed Sasvi, thus Sasvi provides a stronger safe screening rule. We further study the monotone properties of Sasvi for Lasso, based on which a sure removal regularization parameter can be identified for each feature. Experimental results on both synthetic and real data sets are reported to demonstrate the effectiveness of the proposed Sasvi for Lasso screening.

Tuesday June 24

10:50 - Track F - Nearest-Neighbors and Large-Scale Learning

T26 Densifying One Permutation Hashing via Rotation for Fast Near Neighbor Search

Anshumali Shrivastava; Ping Li

The query complexity of {\em locality sensitive hashing (LSH)} based similarity search is dominated by the number of hash evaluations, and this number grows with the data size~\cite{Proc:Indyk_STOC98}. In industrial applications such as search where the data are often high-dimensional and binary (e.g., text n-grams), {\em minwise hashing} is widely adopted, which requires applying a large number of permutations on the data. This is costly in computation and energy-consumption. In this paper, we propose a hashing technique which generates all the necessary hash evaluations needed for similarity search, using one single permutation. The heart of the proposed hash function is a ``rotation'' scheme which densifies the sparse sketches of {\em one permutation hashing}~\cite{Proc:Li_Owen_Zhang_NIPS12} in an unbiased fashion thereby maintaining the LSH property. This makes the obtained sketches suitable for hash table construction. This idea of rotation presented in this paper could be of independent interest for densifying other types of sparse sketches. Using our proposed hashing method, the query time of a (K,L)-parameterized LSH is reduced from the typical $O(dKL)$ complexity to merely $O(KL+dL)$, where d is the number of nonzeros of the data vector, K is the number of hashes in each hash table, and L is the number of hash tables. Our experimental evaluation on real data confirms that the proposed scheme significantly reduces the query processing time over minwise hashing without loss in retrieval accuracies.

T27 Coding for Random Projections

Ping Li; Michael Mitzenmacher; Anshumali Shrivastava

The method of random projections has become popular for large-scale applications in statistical learning, information retrieval, bio-informatics and other applications. Using a well-designed \textbf{coding} scheme for the projected data, which determines the number of bits needed for each projected value and how to allocate these bits, can significantly improve the effectiveness of the algorithm, in storage cost as well as computational speed. In this paper, we study a number of simple coding schemes, focusing on the task of similarity estimation and on an application to training linear classifiers. We demonstrate that \textbf{uniform quantization} outperforms the standard and influential method~\cite{Proc:Datar_SCG04}, which used a {\em window-and-random offset} scheme. Indeed, we argue that in many cases coding with just a small number of bits suffices. Furthermore, we also develop a \textbf{non-uniform 2-bit} coding scheme that generally performs well in practice, as confirmed by our

experiments on training linear support vector machines (SVM). Proofs and additional experiments are available at {\em arXiv:1308.2218}. In the context of using coded random projections for \textbf{approximate near neighbor search} by building hash tables ({\em arXiv:1403.8144})~\cite{Report:RPCodeLSH2014}, we show that the step of random offset in~\cite{Proc:Datar_SCG04} is again not needed and may hurt the performance. Furthermore, we show that, unless the target similarity level is high, it usually suffices to use only 1 or 2 bits to code each hashed value for this task. Section~\ref{sec_LSH} presents some experimental results for LSH.

T28 Nearest Neighbors Using Compact Sparse Codes

Anoop Cherian

In this paper, we propose a novel scheme for approximate nearest neighbor (ANN) retrieval based on dictionary learning and sparse coding. Our key innovation is to build compact codes, dubbed SpANN codes, using the active set of sparse coded data. These codes are then used to index an inverted file table for fast retrieval. The active sets are often found to be sensitive to small differences among data points, resulting in only near duplicate retrieval. We show that this sensitivity is related to the coherence of the dictionary; small coherence resulting in better retrieval. To this end, we propose a novel dictionary learning formulation with incoherence constraints and an efficient method to solve it. Experiments are conducted on two state-of-the-art computer vision datasets with 1M data points and show an order of magnitude improvement in retrieval accuracy without sacrificing memory and query time compared to the state-of-the-art methods.

T29 Composite Quantization for Approximate Nearest Neighbor Search

Ting Zhang; Chao Du; Jingdong Wang

This paper presents a novel compact coding approach, composite quantization, for approximate nearest neighbor search. The idea is to use the composition of several elements selected from the dictionaries to accurately approximate a vector and to represent the vector by a short code composed of the indices of the selected elements. To efficiently compute the approximate distance of a query to a database vector using the short code, we introduce an extra constraint, constant inter-dictionary-element-product, resulting in that approximating the distance only using the distance of the query to each selected element is enough for nearest neighbor search. Experimental comparison with state-of-the-art algorithms over several benchmark datasets demonstrates the efficacy of the proposed approach.

T30 Circulant Binary Embedding

Felix Yu; Sanjiv Kumar; Yunchao Gong; Shih-Fu Chang

Binary embedding of high-dimensional data requires long codes to preserve the discriminative power of the input space. Traditional binary coding methods often suffer from very high computation and storage costs in such a scenario. To address this problem, we propose Circulant Binary Embedding (CBE) which generates binary codes by projecting the data with a circulant matrix. The circulant structure enables

the use of Fast Fourier Transformation to speed up the computation. Compared to methods that use unstructured matrices, the proposed method improves the time complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(d\log{d})$, and the space complexity from $\mathcal{O}(d^2)$ to $\mathcal{O}(d)$ where d is the input dimensionality. We also propose a novel time-frequency alternating optimization to learn data-dependent circulant projections, which alternatively minimizes the objective in original and Fourier domains. We show by extensive experiments that the proposed approach gives much better performance than the state-of-the-art approaches for fixed time, and provides much faster computation with no performance degradation for fixed number of bits.

Tuesday June 24

14:00 - Track A - Graphical Models II

T31 Low-density Parity Constraints for Hashing-Based Discrete Integration

Stefano Ermon; Carla Gomes; Ashish Sabharwal; Bart Selman

In recent years, a number of probabilistic inference and counting techniques have been proposed that exploit pairwise independent hash functions to infer properties of succinctly defined high-dimensional sets. While providing desirable statistical guarantees, typical constructions of such hash functions are themselves not amenable to efficient inference. Inspired by the success of LDPC codes, we propose the use of low-density parity constraints to make inference more tractable in practice. While not strongly universal, we show that such sparse constraints belong to a new class of hash functions that we call Average Universal. These weaker hash functions retain the desirable statistical guarantees needed by most such probabilistic inference methods. Thus, they continue to provide provable accuracy guarantees while at the same time making a number of algorithms significantly more scalable in practice. Using this technique, we provide new, tighter bounds for challenging discrete integration and model counting problems.

T32 On Measure Concentration of Random Maximum A-Posteriori Perturbations

Francesco Orabona; Tamir Hazan; Anand Sarwate; Tommi Jaakkola

The maximum a-posteriori (MAP) perturbation framework has emerged as a useful approach for inference and learning in high dimensional complex models. By maximizing a randomly perturbed potential function, MAP perturbations generate unbiased samples from the Gibbs distribution. Unfortunately, the computational cost of generating so many high-dimensional random variables can be prohibitive. More efficient algorithms use sequential sampling strategies based on the expected value of low dimensional MAP perturbations. This paper develops new measure concentration inequalities that bound the number of samples needed to estimate such expected values. Applying the general result to MAP perturbations can yield a more efficient algorithm to approximate sampling from the Gibbs distribution. The measure concentration result is of general interest and may be applicable to other areas involving Monte Carlo estimation of expectations.

T33 Learning Sum-Product Networks with Direct and Indirect Variable Interactions

Amirmohammad Rooshenas; Daniel Lowd

Sum-product networks (SPNs) are a deep probabilistic representation that allows for efficient, exact inference. SPNs generalize many other tractable models, including thin junction trees, latent tree models, and many types of mixtures. Previous work on learning SPN structure has mainly focused on using top-down or bottom-up clustering to find mixtures, which capture variable interactions indirectly through implicit latent variables. In contrast, most work on learning graphical models, thin junction trees, and arithmetic circuits has focused on finding direct interactions among variables. In this paper, we present ID-SPN, a new algorithm for learning SPN structure that unifies the two approaches. In experiments on 20 benchmark datasets, we find that the combination of direct and indirect interactions leads to significantly better accuracy than several state-of-the-art algorithms for learning SPNs and other tractable models.

T34 Multiple Testing under Dependence via Semiparametric Graphical Models

Jie Liu; Chunming Zhang; Elizabeth Burnside; David Page

It has been shown that graphical models can be used to leverage the dependence in large-scale multiple testing problems with significantly improved performance (Sun & Cai, 2009; Liu et al., 2012). These graphical models are fully parametric and require that we know the parameterization of f1 — the density function of the test statistic under the alternative hypothesis. However in practice, f1 is often heterogeneous, and cannot be estimated with a simple parametric distribution. We propose a novel semiparametric approach for multiple testing under dependence, which estimates f1 adaptively. This semiparametric approach exactly generalizes the local FDR procedure (Efron et al., 2001) and connects with the BH procedure (Benjamini & Hochberg, 1995). A variety of simulations show that our semiparametric approach outperforms classical procedures which assume independence and the parametric approaches which capture dependence.

T35 Discrete Chebyshev Classifiers

Elad Eban; Elad Mezuman; Amir Globerson

In large scale learning problems it is often easy to collect simple statistics of the data, but hard or impractical to store all the original data. A key question in this setting is how to construct classifiers based on such partial information. One traditional approach to the problem has been to use maximum entropy arguments to induce a complete distribution on variables from statistics. However, this approach essentially makes conditional independence assumptions about the distribution, and furthermore does not optimize prediction loss. Here we present a framework for discriminative learning given a set of statistics. Specifically, we address the case where all variables are discrete and we have access to various marginals. Our approach minimizes the worst case hinge loss in this case, which upper bounds the generalization error. We show that for certain sets of statistics the problem is tractable, and in the general case can be approximated using MAP LP relaxations. Empirical results show that the method is competitive with other approaches that use the same input.

T36 Preserving Modes and Messages via Diverse Particle Selection

Jason Pacheco; Silvia Zuffi; Michael Black; Erik Sudderth

In applications of graphical models arising in domains such as computer vision and signal processing, we often seek the most likely configurations of high-dimensional, continuous variables. We develop a particle-based max-product algorithm which maintains a diverse set of posterior mode hypotheses, and is robust to initialization. At each iteration, the set of hypotheses at each node is augmented via stochastic proposals, and then reduced via an efficient selection algorithm. The integer program underlying our optimization-based particle selection minimizes errors in subsequent max-product message updates. This objective automatically encourages diversity in the maintained hypotheses, without requiring tuning of application-specific distances among hypotheses. By avoiding the stochastic resampling steps underlying particle sum-product algorithms, we also avoid common degeneracies where particles collapse onto a single hypothesis. Our approach significantly outperforms previous particle-based algorithms in experiments focusing on the estimation of human pose from single images.

Tuesday June 24

14:00 - Track B - Reinforcement Learning II

T37 A new Q(lambda) with interim forward view and Monte Carlo equivalence

Rich Sutton; Ashique Rupam Mahmood; Doina Precup; Hado van Hasselt

Q-learning, the most popular of reinforcement learning algorithms, has always included an extension to eligibility traces to enable more rapid learning and improved asymptotic performance on non-Markov problems. The lambda parameter smoothly shifts on-policy algorithms such as TD(lambda) and Sarsa(lambda) from a pure bootstrapping form (lambda=0) to a pure Monte Carlo form (lambda=1). In off-policy algorithms, including Q(lambda), GQ(lambda), and off-policy LSTD(lambda), the lambda parameter is intended to play the same role, but does not; on every exploratory action these algorithms bootstrap regardless of the value of lambda, and as a result they fail to approximate Monte Carlo learning when lambda=1. It may seem that this is inevitable for any online off-policy algorithm; if updates are made on each step on which the target policy is followed, then how could just the right updates be `un-made' upon deviation from the target policy? In this paper, we introduce a new version of Q(lambda) that does exactly that, without significantly increased algorithmic complexity. En route to our new Q(lambda), we introduce a new derivation technique based on the forward-view/backward-view analysis familiar from TD(lambda) but extended to apply at every time step rather than only at the end of episodes. We apply this technique to derive first a new off-policy version of TD(lambda), called PTD(lambda), and then our new Q(lambda), called PQ(lambda).

T38 True Online TD(lambda)

Harm van Seijen; Rich Sutton

TD(lambda) is a core algorithm of modern reinforcement learning. Its appeal comes from its equivalence to a clear and conceptually simple forward view, and the fact that it can be implemented online in an inexpensive manner. However, the equivalence between TD(lambda) and the forward view is exact only for the off-line version of the algorithm (in which updates are made only at the end of each episode). In the online version of TD(lambda) (in which updates are made at each step, which generally performs better and is always used in applications) the match to the forward view is only approximate. In a sense this is unavoidable for the conventional forward view, as it itself presumes that the estimates are unchanging during an episode. In this paper we introduce a new forward view that takes into account the possibility of changing estimates and a new variant of TD(lambda) that exactly achieves it. Our algorithm uses a new form of eligibility trace similar to but different from conventional accumulating and replacing traces. The overall computational complexity is the same as TD(lambda), even when using function approximation. In our empirical comparisons, our algorithm outperformed TD(lambda) in all of its variations. It seems, by adhering more truly to the original goal of TD(lambda)---matching an intuitively clear forward view even in the online case---that we have found a new algorithm that simply improves on classical TD(lambda).

T39 Bias in Natural Actor-Critic Algorithms

Philip Thomas

We show that several popular discounted reward natural actor-critics, including the popular NAC-LSTD and eNAC algorithms, do not generate unbiased estimates of the natural policy gradient as claimed. We derive the first unbiased discounted reward natural actor-critics using batch and iterative approaches to gradient estimation. We argue that the bias makes the existing algorithms more appropriate for the average reward setting. We also show that, when Sarsa(lambda) is guaranteed to converge to an optimal policy, the objective function used by natural actor-critics is concave, so policy gradient methods are guaranteed to converge to globally optimal policies as well.

T40 Deterministic Policy Gradient Algorithms

David Silver; Guy Lever; Nicolas Heess; Thomas Degris; Daan Wierstra; Martin Riedmiller

In this paper we consider deterministic policy gradient algorithms for reinforcement learning with continuous actions. The deterministic policy gradient has a particularly appealing form: it is the expected gradient of the action-value function. This simple form means that the deterministic policy gradient can be estimated much more efficiently than the usual stochastic policy gradient. To ensure adequate exploration, we introduce an off-policy actor-critic algorithm that learns a deterministic target policy from an exploratory behaviour policy. Deterministic policy gradient algorithms outperformed their stochastic counterparts in several benchmark problems, particularly in high-dimensional action spaces.

T41 Programming by Feedback

Riad Akrour; Marc Schoenauer; Jean-Christophe Souplet; Michele Sebag

This paper advocates a new ML-based programming framework, called Programming by Feedback (PF), which involves a sequence of interactions between the active computer and the user. The latter only provides preference judgments on pairs of solutions supplied by the active computer. The active computer involves two components: the learning component estimates the user's utility function and accounts for the user's (possibly limited) competence; the optimization component explores the search space and returns the most appropriate candidate solution. A proof of principle of the approach is proposed, showing that PF requires a handful of interactions in order to solve some discrete and continuous benchmark problems.

T42 Active Learning of Parameterized Skills

Bruno Da Silva; George Konidaris; Andrew Barto

We introduce a method for actively learning parameterized skills. Parameterized skills are flexible behaviors that can solve any task drawn from a distribution of parameterized reinforcement learning problems. Approaches to learning such skills have been proposed, but limited attention has been given to identifying which training tasks allow for rapid skill acquisition. We construct a non-parametric Bayesian model of skill performance and derive analytical expressions for a novel acquisition criterion capable of identifying tasks that maximize expected improvement in skill performance. We also introduce a spatiotemporal kernel tailored for non-stationary skill performance models. The proposed method is agnostic to policy and skill representation and scales independently of task dimensionality. We evaluate it on a non-linear simulated catapult control problem over arbitrarily mountainous terrains.

14:00 - Track C - Topic Models

T43 Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis

Jian Tang; Zhaoshi Meng; Xuanlong Nguyen; Qiaozhu Mei; Ming Zhang

Topic models such as the latent Dirichlet allocation (LDA) have become a standard staple in the modeling toolbox of machine learning. They have been applied to a vast variety of data sets, contexts, and tasks to varying degrees of success. However, to date there is almost no formal theory explicating the LDA's behavior, and despite its familiarity there is very little systematic analysis of and guidance on the properties of the data that affect the inferential performance of the model. This paper seeks to address this gap, by providing a systematic analysis of factors which characterize the LDA's performance. We present theorems elucidating the posterior contraction rates of the topics as the amount of data increases, and a thorough supporting empirical study using synthetic and real data sets, including news and web-based articles and tweet messages. Based on these results we provide practical guidance on how to identify suitable data sets for topic models, and how to specify particular model parameters.

T44 The Inverse Regression Topic Model

Maxim Rabinovich; David Blei

\citet{taddy13mnir} proposed multinomial inverse regression (MNIR) as a new model of annotated text based on the influence of metadata and response variables on the distribution of words in a document. While effective, MNIR has no way to exploit structure in the corpus to improve its predictions or facilitate exploratory data analysis. On the other hand, traditional probabilistic topic models (like latent Dirichlet allocation) capture natural heterogeneity in a collection but do not account for external variables. In this paper, we introduce the inverse regression topic model (IRTM), a mixed-membership extension of MNIR that combines the strengths of both methodologies. We present two inference algorithms for the IRTM: an efficient batch estimation algorithm and an online variant, which is suitable for large corpora. We apply these methods to a corpus of 73K Congressional press releases and another of 150K Yelp reviews, demonstrating that the IRTM outperforms both MNIR and supervised topic models on the prediction task. Further, we give examples showing that the IRTM enables systematic discovery of in-topic lexical variation, which is not possible with previous supervised topic models.

T45 On Modelling Non-linear Topical Dependencies

Zhixing Li; Siqiang Wen; Juanzi Li; Peng Zhang; Jie Tang

Probabilistic topic models such as Latent Dirichlet Allocation (LDA) discover latent topics from large corpora by exploiting words' co-occurring relation. By observing the topical similarity between words, we find that some other relations, such as semantic or syntax relation between words, lead to strong dependence between their topics. In this paper, sentences are represented as dependency trees and a Global Topic Random Field (GTRF) is presented to model the non-linear dependencies between words. To infer our model, a new global factor is defined over all edges and the normalization factor of GRF is proven to be a constant. As a result, no independent assumption is needed when inferring our model. Based on it, we develop an efficient expectation-maximization (EM) procedure for parameter estimation. Experimental results on four data sets show that GTRF achieves much lower perplexity than LDA and linear dependency topic models and produces better topic coherence.

T46 Admixture of Poisson MRFs: A Topic Model with Word Dependencies

David Inouye; Pradeep Ravikumar; Inderjit Dhillon

This paper introduces a new topic model based on an admixture of Poisson Markov Random Fields (APM), which can model dependencies between words as opposed to previous independent topic models such as PLSA (Hofmann, 1999), LDA (Blei et al., 2003) or SAM (Reisinger et al., 2010). We propose a class of admixture models that generalizes previous topic models and show an equivalence between the conditional distribution of LDA and independent Poissons—suggesting that APM subsumes the modeling power of LDA. We present a tractable method for estimating the parameters of an APM based on the pseudo log-likelihood and demonstrate the benefits of APM over previous models by preliminary qualitative and quantitative experiments.

T47 Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data

Zhiyuan Chen; Bing Liu

Topic modeling has been commonly used to discover topics from document collections. However, unsupervised models can generate many incoherent topics. To address this problem, several knowledge-based topic models have been proposed to incorporate prior domain knowledge from the user. This work advances this research much further and shows that without any user input, we can mine the prior knowledge automatically and dynamically from topics already found from a large number of domains. This paper first proposes a novel method to mine such prior knowledge dynamically in the modeling process, and then a new topic model to use the knowledge to guide the model inference. What is also interesting is that this approach offers a novel lifelong learning algorithm for topic discovery, which exploits the big (past) data and knowledge gained from such data for subsequent modeling. Our experimental results using product reviews from 50 domains demonstrate the effectiveness of the proposed approach.

T48 Automated inference of point of view from user interactions in collective intelligence venues

Sanmay Das; Allen Lavoie

Empirical evaluation of trust and manipulation in large-scale collective intelligence processes is challenging. The datasets involved are too large for thorough manual study, and current automated options are limited. We introduce a statistical framework which classifies point of view based on user interactions. The framework works on Web-scale datasets and is applicable to a wide variety of collective intelligence processes. It enables principled study of such issues as manipulation, trustworthiness of information, and potential bias. We demonstrate the model's effectiveness in determining point of view on both synthetic data and a dataset of Wikipedia user interactions. We build a combined model of topics and points-of-view on the entire history of English Wikipedia, and show how it can be used to find potentially biased articles and visualize user interactions at a high level.

14:00 - Track D - Sparsity

T49 Gradient Hard Thresholding Pursuit for Sparsity-Constrained Optimization

Xiaotong Yuan; Ping Li; Tong Zhang

Hard Thresholding Pursuit (HTP) is an iterative greedy selection procedure for finding sparse solutions of underdetermined linear systems. This method has been shown to have strong theoretical guarantees and impressive numerical performance. In this paper, we generalize HTP from compressed sensing to a generic problem setup of sparsity-constrained convex optimization. The proposed algorithm iterates between a standard gradient descent step and a hard truncation step with or without debiasing. We prove that our method enjoys the strong guarantees analogous to HTP in terms of rate of convergence and parameter estimation accuracy. Numerical evidences show that our method is superior to the state-

of-the-art greedy selection methods when applied to learning tasks of sparse logistic regression and sparse support vector machines.

T50 Forward-Backward Greedy Algorithms for General Convex Smooth Functions over A Cardinality Constraint

Ji Liu; Jieping Ye; Ryohei Fujimaki

We consider forward-backward greedy algorithms for solving sparse feature selection problems with general convex smooth functions. A state-of-the-art greedy method, the Forward-Backward greedy algorithm (FoBa-obj) requires to solve a large number of optimization problems, thus it is not scalable for large-size problems. The FoBa-gdt algorithm, which uses the gradient information for feature selection at each forward iteration, significantly improves the efficiency of FoBa-obj. In this paper, we systematically analyze the theoretical properties of both algorithms. Our main contributions are: 1) We derive better theoretical bounds than existing analyses regarding FoBa-obj for general smooth convex functions; 2) We show that FoBa-gdt achieves the same theoretical performance as FoBa-obj under the same condition: restricted strong convexity condition. Our new bounds are consistent with the bounds of a special case (least squares) and fills a previously existing theoretical gap for general convex smooth functions; 3) We show that the restricted strong convexity condition is satisfied if the number of independent samples is more than $\bar{k}\log d$ where $\bar{k}$ is the sparsity number and d is the dimension of the variable; 4) We apply FoBa-gdt (with the conditional random field objective) to the sensor selection problem for human indoor activity recognition and our results show that FoBa-gdt outperforms other methods based on forward greedy selection and L1-regularization.

T51 Efficient Algorithms for Robust One-bit Compressive Sensing

Lijun Zhang; Jinfeng Yi; Rong Jin

While the conventional compressive sensing assumes measurements of infinite precision, one-bit compressive sensing considers an extreme setting where each measurement is quantized to just a single bit. In this paper, we study the vector recovery problem from noisy one-bit measurements, and develop two novel algorithms with formal theoretical guarantees. First, we propose a passive algorithm, which is very efficient in the sense it only needs to solve a convex optimization problem that has a closed-form solution. Despite the apparent simplicity, our theoretical analysis reveals that the proposed algorithm can recover both the exactly sparse and the approximately sparse vectors. In particular, for a sparse vector with s nonzero elements, the sample complexity is $O(s \log n/\epsilon^2)$, where n is the dimensionality and ϵ is the recovery error. This result improves significantly over the previously best known sample complexity in the noisy setting, which is $O(s \log n/\epsilon^4)$. Second, in the case that the noise model is known, we develop an adaptive algorithm based on the principle of active learning. The key idea is to solicit the sign information only when it cannot be inferred from the current estimator. Compared with the passive algorithm, the adaptive one has a lower sample complexity if a high-precision solution is desired.

T52 Nonlinear Information-Theoretic Compressive Measurement Design

Liming Wang; Abolfazl Razi; Miguel Rodrigues; Robert Calderbank; Lawrence Carin

We investigate design of general nonlinear functions for mapping high-dimensional data into a lower-dimensional (compressive) space. The nonlinear measurements are assumed contaminated by additive Gaussian noise. Depending on the application, we are either interested in recovering the high-dimensional data from the nonlinear compressive measurements, or performing classification directly based on these measurements. The latter case corresponds to classification based on nonlinearly constituted and noisy features. The nonlinear measurement functions are designed based on constrained mutual-information optimization. New analytic results are developed for the gradient of mutual information in this setting, for arbitrary input-signal statistics. We make connections to kernel-based methods, such as the support vector machine. Encouraging results are presented on multiple datasets, for both signal recovery and classification. The nonlinear approach is shown to be particularly valuable in high-noise scenarios.

T53 Elementary Estimators for High-Dimensional Linear Regression

Eunho Yang; Aurelie Lozano; Pradeep Ravikumar

We consider the problem of structurally constrained high-dimensional linear regression. This has attracted considerable attention over the last decade, with state of the art statistical estimators based on solving regularized convex programs. While these typically non-smooth convex programs can be solved in polynomial time, scaling the state of the art optimization methods to very large-scale problems is an ongoing and rich area of research. In this paper, we attempt to address this scaling issue at the source, by asking whether one can build \emph{simpler} possibly closed-form estimators, that yet come with statistical guarantees that are nonetheless comparable to regularized likelihood estimators! We answer this question in the affirmative, with variants of the classical ridge and OLS (ordinary least squares estimators) for linear regression. We analyze our estimators in the high-dimensional setting, and moreover provide empirical corroboration of its performance on simulated as well as real world microarray data.

T54 Statistical-Computational Phase Transitions in Planted Models: The High-Dimensional Setting

Yudong Chen; Jiaming Xu

The planted models assume that a graph is generated from some unknown clusters by randomly placing edges between nodes according to their cluster memberships; the task is to recover the clusters given the graph. Special cases include planted clique, planted partition, planted densest subgraph and planted coloring. Of particular interest is the High-Dimensional setting where the number of clusters is allowed to grow with the number of nodes. We show that the space of model parameters can be partitioned into four disjoint regions corresponding to decreasing statistical and computational complexities: (1) the impossible regime, where all algorithms fail; (2) the hard regime, where the exponential-time Maximum Likelihood Estimator (MLE) succeeds, and no polynomial-time method is known; (3) the easy regime, where the polynomial-time convexified MLE succeeds; (4) the simple regime, where a simple

counting/thresholding procedure succeeds. Moreover, each of these algorithms provably fails in the previous harder regimes. Our theorems establish the first minimax recovery results for the high-dimensional setting, and provide the best known guarantees for polynomial-time algorithms. Our results extend to the related problem of submatrix localization, a.k.a. bi-clustering. These results demonstrate the tradeoffs between statistical and computational considerations.

Tuesday June 24

14:00 - Track E - Kernel Methods II

T55 Kernel Mean Estimation and Stein Effect

Krikamol Muandet; Kenji Fukumizu; Bharath Sriperumbudur; Arthur Gretton; Bernhard Schoelkopf

A mean function in reproducing kernel Hilbert space (RKHS), or a kernel mean, is an important part of many algorithms ranging from kernel principal component analysis to Hilbert-space embedding of distributions. Given a finite sample, an empirical average is the standard estimate for the true kernel mean. We show that this estimator can be improved due to a well-known phenomenon in statistics called Stein phenomenon. After consideration, our theoretical analysis reveals the existence of a wide class of estimators that are better than the standard one. Focusing on a subset of this class, we propose efficient shrinkage estimators for the kernel mean. Empirical evaluations on several applications clearly demonstrate that the proposed estimators outperform the standard kernel mean estimator.

T56 A Kernel Independence Test for Random Processes

Kacper Chwialkowski; Arthur Gretton

A non-parametric approach to the problem of testing the independence of two random processes is developed. The test statistic is the Hilbert-Schmidt Independence Criterion (HSIC), which was used previously in testing independence for i.i.d. pairs of variables. The asymptotic behaviour of HSIC is established when computed from samples drawn from random processes. It is shown that earlier bootstrap procedures which worked in the i.i.d. case will fail for random processes, and an alternative consistent estimate of the p-values is proposed. Tests on artificial data and real-world forex data indicate that the new test procedure discovers dependence which is missed by linear approaches, while the earlier bootstrap procedure returns an elevated number of false positives.

T57 Quasi-Monte Carlo Feature Maps for Shift-Invariant Kernels

Jiyan Yang; Vikas Sindhwani; Haim Avron; Michael Mahoney

We consider the problem of improving the efficiency of randomized Fourier feature maps to accelerate training and testing speed of kernel methods on large datasets. These approximate feature maps arise as Monte Carlo approximations to integral representations of shift-invariant kernel functions (e.g., Gaussian kernel). In this paper, we propose to use Quasi-Monte Carlo (QMC) approximations instead

where the relevant integrands are evaluated on a low-discrepancy sequence of points as opposed to random point sets as in the Monte Carlo approach. We derive a new discrepancy measure called box discrepancy based on theoretical characterizations of the integration error with respect to a given sequence. We then propose to learn QMC sequences adapted to our setting based on explicit box discrepancy minimization. Our theoretical analyses are complemented with empirical results that demonstrate the effectiveness of classical and adaptive QMC techniques for this problem.

T58 A Unifying View of Representer Theorems

Andreas Argyriou; Francesco Dinuzzo

It is known that the solution of regularization and interpolation problems with Hilbertian penalties can be expressed as a linear combination of the data. This very useful property, called the representer theorem, has been widely studied and applied to machine learning problems. Analogous optimality conditions have appeared in other contexts, notably in matrix regularization. In this paper we propose a unified view, which generalizes the concept of representer theorems and extends necessary and sufficient conditions for such theorems to hold. Our main result shows a close connection between representer theorems and certain classes of regularization penalties, which we call orthomonotone functions. This result not only subsumes previous representer theorems as special cases but also yields a new class of optimality conditions, which goes beyond the classical linear combination of the data. Moreover, orthomonotonicity provides a useful criterion for testing whether a representer theorem holds for a specific regularization problem.

T59 Efficient Approximation of Cross-Validation for Kernel Methods using Bouligand Influence Function

Yong Liu; Shali Jiang; Shizhong Liao

Model selection is one of the key issues both in recent research and application of kernel methods. Cross-validation is a commonly employed and widely accepted model selection criterion. However, it requires multiple times of training the algorithm under consideration, which is computationally intensive. In this paper, we present a novel strategy for approximating the cross-validation based on the Bouligand influence function (BIF), which only requires the solution of the algorithm once. The BIF measures the impact of an infinitesimal small amount of contamination of the original distribution. We first establish the link between the concept of BIF and the concept of cross-validation. The BIF is related to the first order term of a Taylor expansion. Then, we calculate the BIF and higher order BIFs, and apply these theoretical results to approximate the cross-validation error in practice. Experimental results demonstrate that our approximate cross-validation criterion is sound and efficient.

Tuesday June 24

14:00 - Track F - Neural Theory and Spectral Methods

T60 Provable Bounds for Learning Some Deep Representations

Sanjeev Arora; Aditya Bhaskara; Rong Ge; Tengyu Ma

We give algorithms with provable guarantees that learn a class of deep nets in the generative model view popularized by Hinton and others. Our generative model is an n node multilayer neural net that has degree at most n^{γ} for some $\gamma < 1$ and each edge has a random edge weight in [-1,1]. Our algorithm learns almost all networks in this class with polynomial running time. The sample complexity is quadratic or cubic depending upon the details of the model. The algorithm uses layerwise learning. It is based upon a novel idea of observing correlations among features and using these to infer the underlying edge structure via a global graph recovery procedure. The analysis of the algorithm reveals interesting structure of neural nets with random edge weights.

T61 K-means recovers ICA filters when independent components are sparse

Alon Vinnikov; Shai Shalev-Shwartz

Unsupervised feature learning is the task of using unlabeled examples for building a representation of objects as vectors. This task has been extensively studied in recent years, mainly in the context of unsupervised pre-training of neural networks. Recently, (Coates et al., 2011) conducted extensive experiments, comparing the accuracy of a linear classifier that has been trained using features learnt by several unsupervised feature learning methods. Surprisingly, the best performing method was the simplest feature learning approach that was based on applying the K-means clustering algorithm after a whitening of the data. The goal of this work is to shed light on the success of K-means with whitening for the task of unsupervised feature learning. Our main result is a close connection between K-means and ICA (Independent Component Analysis). Specifically, we show that K-means and similar clustering algorithms can be used to recover the ICA mixing matrix or its inverse, the ICA filters. It is well known that the independent components found by ICA form useful features for classification (Le et al., 2012; 2011; 2010), hence the connection between K-mean and ICA explains the empirical success of K-means as a feature learner. Moreover, our analysis underscores the significance of the whitening operation, as was also observed in the experiments reported in (Coates et al., 2011). Finally, our analysis leads to a better initialization of K-means for the task of feature learning.

T62 Learning Polynomials with Neural Networks

Alexandr Andoni; Rina Panigrahy; Gregory Valiant; Li Zhang

We study the effectiveness of learning low degree polynomials using neural networks by the gradient descent method. While neural networks have been shown to have great expressive power, and gradient descent has been widely used in practice for learning neural networks, few theoretical guarantees are known for such methods. In particular, it is well known that gradient descent can get stuck at local minima, even for simple classes of target functions. In this paper, we present several positive theoretical results to support the effectiveness of neural networks. We focus on two-layer neural networks (i.e. one hidden layer) where the top layer node is a linear function, similar to~\cite{barron93}. First we show that for a randomly initialized neural network with sufficiently many hidden units, the gradient descent

method can learn any low degree polynomial. Secondly, we show that if we use complex-valued weights (the target function can still be real), then under suitable conditions, there are no ``robust local minima'': the neural network can always escape a local minimum by performing a random perturbation. This property does not hold for real-valued weights. Thirdly, we discuss whether sparse polynomials can be learned with \emph{small} neural networks, where the size is dependent on the sparsity of the target function.

T63 Anti-differentiating approximation algorithms:A case study with min-cuts, spectral, and flow

David Gleich; Michael Mahoney

We formalize and illustrate the general concept of algorithmic anti-differentiation: given an algorithmic procedure, e.g., an approximation algorithm for which worst-case approximation guarantees are available or a heuristic that has been engineered to be practically-useful but for which a precise theoretical understanding is lacking, an algorithmic anti-derivative is a precise statement of an optimization problem that is exactly solved by that procedure. We explore this concept with a case study of approximation algorithms for finding locally-biased partitions in data graphs, demonstrating connections between min-cut objectives, a personalized version of the popular PageRank vector, and the highly effective "push" procedure for computing an approximation to personalized PageRank. We show, for example, that this latter algorithm solves (exactly, but implicitly) an l1-regularized l2-regression problem, a fact that helps to explain its excellent performance in practice. We expect that, when available, these implicit optimization problems will be critical for rationalizing and predicting the performance of many approximation algorithms on realistic data.

T64 Nonnegative Sparse PCA with Provable Guarantees

Megasthenis Asteris; Dimitris Papailiopoulos; Alexandros Dimakis

We introduce a novel algorithm to compute nonnegative sparse principal components of positive semidefinite (PSD) matrices. Our algorithm comes with approximation guarantees contingent on the spectral profile of the input matrix A: the sharper the eigenvalue decay, the better the approximation quality. If the eigenvalues decay like any asymptotically vanishing function, we can approximate nonnegative sparse PCA within any accuracy ϵ in time polynomial in the matrix size n and desired sparsity k, but not in $1/\epsilon$. Further, we obtain a data-dependent bound that is computed by executing an algorithm on a given data set. This bound is significantly tighter than a-priori bounds and can be used to show that for all tested datasets our algorithm is provably within 40%-90% from the unknown optimum. Our algorithm is combinatorial and explores a subspace defined by the leading eigenvectors of A. We test our scheme on several data sets, showing that it matches or outperforms the previous state of the art.

T65 Finding Dense Subgraphs via Low-Rank Bilinear Optimization

Dimitris Papailiopoulos; Ioannis Mitliagkas; Alexandros Dimakis; Constantine Caramanis

Given a graph, the Densest k-Subgraph (DkS) problem asks for the subgraph on k vertices that contains the largest number of edges. In this work, we develop a new algorithm for DkS that searches a low-dimensional space for provably good solutions. Our algorithm comes with novel performance bounds that depend on the graph spectrum. Our graph-dependent bounds are in practice significantly tighter than worst case a priori bounds: for most tested real-world graphs we find subgraphs with density provably within 70\% of the optimum. Our algorithm runs in nearly linear time, under spectral assumptions satisfied by most graphs found in applications. Moreover, it is highly scalable and parallelizable. We demonstrate this by implementing it in MapReduce and executing numerous experiments on massive real-world graphs that have up to billions of edges. We empirically show that our algorithm can find subgraphs of significantly higher density compared to the previous state of the art.

Tuesday June 24

16:20 - Track A - Networks and Graph-Based Learning II

T66 Learning Graphs with a Few Hubs

Rashish Tandon; Pradeep Ravikumar

We consider the problem of recovering the graph structure of a ``hub-networked'' Ising model given iid samples, under high-dimensional settings, where number of nodes $p$ could be potentially larger than the number of samples $n$. By a ``hub-networked'' graph, we mean a graph with a few ``hub nodes'' with very large degrees. State of the art estimators for Ising models have a sample complexity that scales polynomially with the maximum node-degree, and are thus ill-suited to recovering such graphs with a few hub nodes. Some recent proposals for specifically recovering hub graphical models do not come with theoretical guarantees, and even empirically provide limited improvements over vanilla Ising model estimators. Here, we show that under such low sample settings, instead of estimating ``difficult'' components such as hub-neighborhoods, we can use quantitative indicators of our inability to do so, and thereby identify hub-nodes. This simple procedure allows us to recover hub-networked graphs with very strong statistical guarantees even under very low sample settings.

T67 Global graph kernels using geometric embeddings

Fredrik Johansson; Vinay Jethava; Devdatt Dubhashi; Chiranjib Bhattacharyya

Applications of machine learning methods increasingly deal with graph structured data through kernels. Most existing graph kernels compare graphs in terms of features defined on small subgraphs such as walks, paths or graphlets, adopting an inherently local perspective. However, several interesting properties such as girth or chromatic number are global properties of the graph, and are not captured in local substructures. This paper presents two graph kernels defined on unlabeled graphs which capture global properties of graphs using the celebrated Lovász number and its associated orthonormal representation. We make progress towards theoretical results aiding kernel choice, proving a result about the separation margin of our kernel for classes of graphs. We give empirical results on

classification of synthesized graphs with important global properties as well as established benchmark graph datasets, showing that the accuracy of our kernels is better than or competitive to existing graph kernels.

T68 Efficient Label Propagation

Yasuhiro Fujiwara; Go Irie

Label propagation is a popular graph-based semi-supervised learning framework. So as to obtain the optimal labeling scores, the label propagation algorithm requires an inverse matrix which incurs the high computational cost of O(n^3+cn^2), where n and c are the numbers of data points and labels, respectively. This paper proposes an efficient label propagation algorithm that guarantees exactly the same labeling results as those yielded by optimal labeling scores. The key to our approach is to iteratively compute lower and upper bounds of labeling scores to prune unnecessary score computations. This idea significantly reduces the computational cost to O(cnt) where t is the average number of iterations for each label and t << n in practice. Experiments demonstrate the significant superiority of our algorithm over existing label propagation methods.

T69 Estimating Diffusion Network Structures: Recovery Conditions, Sample Complexity & Soft-thresholding Algorithm

Hadi Daneshmand; Manuel Gomez-Rodriguez; Le Song; Bernhard Schoelkopf

Information spreads across social and technological networks, but often the network structures are hidden from us and we only observe the traces left by the diffusion processes, called cascades. Can we recover the hidden network structures from these observed cascades? What kind of cascades and how many cascades do we need? Are there some network structures which are more difficult than others to recover? Can we design efficient inference algorithms with provable guarantees? Despite the increasing availability of cascade data and methods for inferring networks from these data, a thorough theoretical understanding of the above questions remains largely unexplored in the literature. In this paper, we investigate the network structure inference problem for a general family of continuous-time diffusion models using an l1-regularized likelihood maximization framework. We show that, as long as the cascade sampling process satisfies a natural incoherence condition, our framework can recover the correct network structure with high probability if we observe O(d^3 log N) cascades, where d is the maximum number of parents of a node and N is the total number of nodes. Moreover, we develop a simple and efficient soft-thresholding inference algorithm, which we use to illustrate the consequences of our theoretical results, and show that our framework outperforms other alternatives in practice.

T70 Learning from Contagion (Without Timestamps)

Kareem Amin; Hoda Heidari; Michael Kearns

We introduce and study new models for learning from contagion processes in a network. A learning algorithm is allowed to either choose or passively observe an initial set of seed infections. This seed set

then induces a final set of infections resulting from the underlying stochastic contagion dynamics. Our models differ from prior work in that detailed vertex-by-vertex timestamps for the spread of the contagion are not observed. The goal of learning is to infer the unknown network structure. Our main theoretical results are efficient and provably correct algorithms for exactly learning trees. We provide empirical evidence that our algorithm performs well more generally on realistic sparse graphs.

T71 Influence Function Learning in Information Diffusion Networks

Nan Du; Yingyu Liang; Maria Balcan; Le Song

Can we learn the influence of a set of people in a social network from cascades of information diffusion? This question is often addressed by a two-stage approach: first learn a diffusion model, and then calculate the influence based on the learned model. Thus, the success of this approach relies heavily on the correctness of the diffusion model which is hard to verify for real world data. In this paper, we exploit the insight that the influence functions in many diffusion models are coverage functions, and propose a novel parameterization of such functions using a convex combination of random basis functions. Moreover, we propose an efficient maximum likelihood based algorithm to learn such functions directly from cascade data, and hence bypass the need to specify a particular diffusion model in advance. We provide both theoretical and empirical analysis for our approach, showing that the proposed approach can provably learn the influence function with low sample complexity, be robust to the unknown diffusion models, and significantly outperform existing approaches in both synthetic and real world data.

Tuesday June 24

16:20 - Track B - Online Learning II

T72 Tracking Adversarial Targets

Yasin Abbasi-Yadkori; Peter Bartlett; Varun Kanade

We study linear control problems with quadratic losses and adversarially chosen tracking targets. We present an efficient algorithm for this problem and show that, under standard conditions on the linear system, its regret with respect to an optimal linear policy grows as $O(\log^2 T)$, where T is the number of rounds of the game. We also study a problem with adversarially chosen transition dynamics; we present an exponentially-weighted average algorithm for this problem, and we give regret bounds that grow as $O(\sqrt T)$.

T73 Sparse Reinforcement Learning via Convex Optimization

Zhiwei Qin; Weichang Li; Firdaus Janoos

We propose two new algorithms for the sparse reinforcement learning problem based on different formulations. The first algorithm is an off-line method based on the alternating direction method of

multipliers for solving a constrained formulation that explicitly controls the projected Bellman residual. The second algorithm is an online stochastic approximation algorithm that employs the regularized dual averaging technique, using the Lagrangian formulation. The convergence of both algorithms are established. We demonstrate the performance of these algorithms through two classical examples.

T74 Online Learning in Markov Decision Processes with Changing Cost Sequences

Travis Dick; Andras Gyorgy; Csaba Szepesvari

In this paper we consider online learning in finite Markov decision processes (MDPs) with changing cost sequences under full and bandit-information. We propose to view this problem as an instance of online linear optimization. We propose two methods for this problem: MD^2 (mirror descent with approximate projections) and the continuous exponential weights algorithm with Dikin walks. We provide a rigorous complexity analysis of these techniques, while providing near-optimal regret-bounds (in particular, we take into account the computational costs of performing approximate projections in MD^2). In the case of full-information feedback, our results complement existing ones. In the case of bandit-information feedback we consider the online stochastic shortest path problem, a special case of the above MDP problems, and manage to improve the existing results by removing the previous restrictive assumption that the state-visitation probabilities are uniformly bounded away from zero under all policies.

T75 Linear Programming for Large-Scale Markov Decision Problems

Alan Malek; Yasin Abbasi-Yadkori; Peter Bartlett

We consider the problem of controlling a Markov decision process (MDP) with a large state space, so as to minimize average cost. Since it is intractable to compete with the optimal policy for large scale problems, we pursue the more modest goal of competing with a low-dimensional family of policies. We use the dual linear programming formulation of the MDP average cost problem, in which the variable is a stationary distribution over state-action pairs, and we consider a neighborhood of a low-dimensional subset of the set of stationary distributions (defined in terms of state-action features) as the comparison class. We propose two techniques, one based on stochastic convex optimization, and one based on constraint sampling. In both cases, we give bounds that show that the performance of our algorithms approaches the best achievable by any policy in the comparison class. Most importantly, these results depend on the size of the comparison class, but not on the size of the state space. Preliminary experiments show the effectiveness of the proposed algorithms in a queuing application.

T76 Statistical analysis of stochastic gradient methods for generalized linear models

Panagiotis Toulis; Edoardo Airoldi; Jason Rennie

We study the statistical properties of stochastic gradient descent (SGD) using explicit and implicit updates for fitting generalized linear models (GLMs). Initially, we develop a computationally efficient algorithm to implement implicit SGD learning of GLMs. Next, we obtain exact formulas for the bias and variance of both updates which leads to two important observations on their comparative statistical

properties. First, in small samples, the estimates from the implicit procedure are more biased than the estimates from the explicit one, but their empirical variance is smaller and they are more robust to learning rate misspecification. Second, the two procedures are statistically identical in the limit: they are both unbiased, converge at the same rate and have the same asymptotic variance. Our set of experiments confirm our theory and more broadly suggest that the implicit procedure can be a competitive choice for fitting large-scale models, especially when robustness is a concern.

T77 Preference-Based Rank Elicitation using Statistical Models: The Case of Mallows

Robert Busa-Fekete; Eyke Huellermeier; Balázs Szörényi

We address the problem of rank elicitation assuming that the underlying data generating process is characterized by a probability distribution on the set of all rankings (total orders) of a given set of items. Instead of asking for complete rankings, however, our learner is only allowed to query pairwise preferences. Using information of that kind, the goal of the learner is to reliably predict properties of the distribution, such as the most probable top-item, the most probable ranking, or the distribution itself. More specifically, learning is done in an online manner, and the goal is to minimize sample complexity while guaranteeing a certain level of confidence.

16:20 - Track C - Nonparametric Bayes II

T78 Bayesian Max-margin Multi-Task Learning with Data Augmentation

Chengtao Li; Jun Zhu; Jianfei Chen

Both max-margin and Bayesian methods have been extensively studied in multi-task learning, but have rarely been considered together. We present Bayesian max-margin multi-task learning, which conjoins the two schools of methods, thus allowing the discriminative max-margin methods to enjoy the great flexibility of Bayesian methods on incorporating rich prior information as well as performing nonparametric Bayesian feature learning with the latent dimensionality resolved from data. We develop Gibbs sampling algorithms by exploring data augmentation to deal with the non-smooth hinge loss. For nonparametric models, our algorithms do not need to make mean-field assumptions or truncated approximation. Empirical results demonstrate superior performance than competitors in both multi-task classification and regression.

T79 Variational Inference for Sequential Distance Dependent Chinese Restaurant Process

Sergey Bartunov; Dmitry Vetrov

Recently proposed distance dependent Chinese Restaurant Process (ddCRP) generalizes extensively used Chinese Restaurant Process (CRP) by accounting for dependencies between data points. Its posterior is intractable and so far only MCMC methods were used for inference. Because of very different nature of

ddCRP no prior developments in variational methods for Bayesian nonparametrics are appliable. In this paper we propose novel variational inference for important sequential case of ddCRP (seqddCRP) by revealing its connection with Laplacian of random graph constructed by the process. We develop efficient algorithm for optimizing variational lower bound and demonstrate its efficiency comparing to Gibbs sampler. We also apply our variational approximation to CRP-equivalent seqddCRP-mixture model, where it could be considered as alternative to one based on truncated stick-breaking representation. This allowed us to achieve significantly better variational lower bound than variational approximation based on truncated stick breaking for Dirichlet process.

T80 Pitfalls in the use of Parallel Inference for the Dirichlet Process

Yarin Gal; Zoubin Ghahramani

Recent work done by Lovell, Adams, and Mansingka (2012) and Williamson, Dubey, and Xing (2013) has suggested an alternative parametrisation for the Dirichlet process in order to derive non-approximate parallel MCMC inference for it – work which has been picked-up and implemented in several different fields. In this paper we show that the approach suggested is impractical due to an extremely unbalanced distribution of the data. We characterise the requirements of efficient parallel inference for the Dirichlet process and show that the proposed inference fails most of these requirements (while approximate approaches often satisfy most of them). We present both theoretical and experimental evidence, analysing the load balance for the inference and showing that it is independent of the size of the dataset and the number of nodes available in the parallel implementation. We end with suggestions of alternative paths of research for efficient non-approximate parallel inference for the Dirichlet process.

T81 Fast Allocation of Gaussian Process Experts

Trung Nguyen; Edwin Bonilla

We propose a scalable nonparametric Bayesian regression model based on a mixture of Gaussian process (GP) experts and the inducing points formalism underpinning sparse GP approximations. Each expert is augmented with a set of inducing points, and the allocation of data points to experts is defined probabilistically based on their proximity to the experts. This allocation mechanism enables a fast variational inference procedure for learning of the inducing inputs and hyperparameters of the experts. When using K experts, our method can run K^2 times faster and use K^2 times less memory than popular sparse methods such as the FITC approximation. Furthermore, it is easy to parallelize and handles non-stationarity straightforwardly. Our experiments show that on medium-sized datasets (of around 10^4 training points) it trains up to 5 times faster than FITC while achieving comparable accuracy. On a large dataset of 10^5 training points, our method significantly outperforms six competitive baselines while requiring only a few hours of training.

T82 Scalable and Robust Bayesian Inference via the Median Posterior

Stanislav Minsker; Sanvesh Srivastava; Lizhen Lin; David Dunson

Many Bayesian learning methods for massive data benefit from working with small subsets of observations. In particular, significant progress has been made in scalable Bayesian learning via stochastic approximation. However, Bayesian learning methods in distributed computing environments are often problem- or distribution-specific and use ad hoc techniques. We propose a novel general approach to Bayesian inference that is scalable and robust to corruption in the data. Our technique is based on the idea of splitting the data into several non-overlapping subgroups, evaluating the posterior distribution given each independent subgroup, and then combining the results. The main novelty is the proposed aggregation step which is based on finding the geometric median of posterior distributions. We present both theoretical and numerical results illustrating the advantages of our approach.

T83 Nonparametric Estimation of Renyi Divergence and Friends

Akshay Krishnamurthy; Kirthevasan Kandasamy; Barnabas Poczos; Larry Wasserman

We consider nonparametric estimation of L_2, Renyi-α and Tsallis-α divergences between continuous distributions. Our approach is to construct estimators for particular integral functionals of two densities and translate them into divergence estimators. For the integral functionals, our estimators are based on corrections of a preliminary plug-in estimator. We show that these estimators achieve the parametric convergence rate of $n^{-1/2}$ when the densities' smoothness, s, are both at least $d/4$ where d is the dimension. We also derive minimax lower bounds for this problem which confirm that $s > d/4$ is necessary to achieve the $n^{-1/2}$ rate of convergence. We validate our theoretical guarantees with a number of simulations.

16:20 - Track D - Features and Feature Selection

T84 Elementary Estimators for Sparse Covariance Matrices and other Structured Moments

Eunho Yang; Aurelie Lozano; Pradeep Ravikumar

We consider the problem of estimating distributional parameters that are expected values of given feature functions. We are interested in recovery under high-dimensional regimes, where the number of variables p is potentially larger than the number of samples n, and where we need to impose structural constraints upon the parameters. In a natural distributional setting for this problem, the feature functions comprise the sufficient statistics of an exponential family, so that the problem would entail estimating structured moments of exponential family distributions. A special case of the above involves estimating the covariance matrix of a random vector, and where the natural distributional setting would correspond to the multivariate Gaussian distribution. Unlike the inverse covariance estimation case, we show that the regularized MLEs for covariance estimation, as well as natural Dantzig variants, are \emph{non-convex}, even when the regularization functions themselves are convex; with the same holding for the general structured moment case. We propose a class of elementary convex estimators, that in many cases are available in \emph{closed-form}, for estimating general structured

moments. We then provide a unified statistical analysis of our class of estimators. Finally, we demonstrate the applicability of our class of estimators on real-world climatology and biology datasets.

T85 Robust Inverse Covariance Estimation under Noisy Measurements

Jun-Kun Wang; Shou-de Lin

This paper proposes a robust method to estimate the inverse covariance under noisy measurements. The method is based on the estimation of each column in the inverse covariance matrix independently via robust regression, which enables parallelization. Different from previous linear programming based methods that cannot guarantee a positive semi-definite covariance matrix, our method adjusts the learned matrix to satisfy this condition, which further facilitates the tasks of forecasting future values. Experiments on time series prediction and classification under noisy condition demonstrate the effectiveness of the approach.

T86 Making Fisher Discriminant Analysis Scalable

Bojun Tu; Zhihua Zhang; Shusen Wang; Hui Qian

The Fisher linear discriminant analysis (LDA) is a classical method for classification and dimension reduction jointly. A major limitation of the conventional LDA is a so-called singularity issue. Many LDA variants, especially two-stage methods such as PCA+LDA and LDA/QR, were proposed to solve this issue. In the two-stage methods, an intermediate stage for dimension reduction is developed before the actual LDA method works. These two-stage methods are scalable because they are an approximate alternative of the LDA method. However, there is no theoretical analysis on how well they approximate the conventional LDA problem. In this paper we present theoretical analysis on the approximation error of a two-stage algorithm. Accordingly, we develop a new two-stage algorithm. Furthermore, we resort to a random projection approach, making our algorithm scalable. We also provide an implemention on distributed system to handle large scale problems. Our algorithm takes LDA/QR as its special case, and outperforms PCA+LDA while having a similar scalability. We also generalize our algorithm to kernel discriminant analysis, a nonlinear version of the classical LDA. Extensive experiments show that our algorithms outperform PCA+LDA and have a similar scalability with it.

T87 An Analysis of State-Relevance Weights and Sampling Distributions on L1-Regularized Approximate Linear Programming Approximation Accuracy

Gavin Taylor; Connor Geer; David Piekut

Recent interest in the use of L_1 regularization in the use of value function approximation includes Petrik et al.'s introduction of L_1-Regularized Approximate Linear Programming (RALP). RALP is unique among L_1-regularized approaches in that it approximates the optimal value function using off-policy samples. Additionally, it produces policies which outperform those of previous methods, such as LSPI. RALP's value function approximation quality is affected heavily by the choice of state-relevance weights in the objective function of the linear program, and by the distribution from which samples are

drawn; however, there has been no discussion of these considerations in the previous literature. In this paper, we discuss and explain the effects of choices in the state-relevance weights and sampling distribution on approximation quality, using both theoretical and experimental illustrations. The results provide insight not only onto these effects, but also provide intuition into the types of MDPs which are especially well suited for approximation with RALP.

T88 Factorized Point Process Intensities: A Spatial Analysis of Professional Basketball

Andrew Miller; Luke Bornn; Ryan Adams; Kirk Goldsberry

We develop a machine learning approach to represent and analyze the underlying spatial structure that governs shot selection among professional basketball players in the NBA. Typically, NBA players are discussed and compared in an heuristic, imprecise manner that relies on unmeasured intuitions about player behavior. This makes it difficult to draw comparisons between players and make accurate player specific predictions. Modeling shot attempt data as a point process, we create a low dimensional representation of offensive player types in the NBA. Using non-negative matrix factorization (NMF), an unsupervised dimensionality reduction technique, we show that a low-rank spatial decomposition summarizes the shooting habits of NBA players. The spatial representations discovered by the algorithm correspond to intuitive descriptions of NBA player types, and can be used to model other spatial effects, such as shooting accuracy.

T89 Compact Random Feature Maps

Raffay Hamid; Ying Xiao; Alex Gittens; Dennis Decoste

Kernel approximation using randomized feature maps has recently gained a lot of interest. In this work, we identify that previous approaches for polynomial kernel approximation create maps that are rank deficient, and therefore do not utilize the capacity of the projected feature space effectively. To address this challenge, we propose compact random feature maps (CRAFTMaps) to approximate polynomial kernels more concisely and accurately. We prove the error bounds of CRAFTMaps demonstrating their superior kernel reconstruction performance compared to the previous approximation schemes. We show how structured random matrices can be used to efficiently generate CRAFTMaps, and present a single-pass algorithm using CRAFTMaps to learn non-linear multi-class classifiers. We present experiments on multiple standard data-sets with performance competitive with state-of-the-art results.

16:20 - Track E - Optimization III

T90 New Primal SVM Solver with Linear Computational Cost for Big Data Classifications

Feiping Nie; Yizhen Huang; Xiaoqian Wang; Heng Huang

Support Vector Machines (SVM) is among the most popular classification techniques in machine learning, hence designing fast primal SVM algorithms for large-scale datasets is a hot topic in recent years. This paper presents a new L2-norm regularized primal SVM solver using Augmented Lagrange Multipliers, with linear-time computational cost for Lp-norm loss functions. The most computationally intensive steps (that determine the algorithmic complexity) of the proposed algorithm is purely and simply matrix-by-vector multiplication, which can be easily parallelized on a multi-core server for parallel computing. We implement and integrate our algorithm into the interfaces and framework of the well-known LibLinear software toolbox. Experiments show that our algorithm is with stable performance and on average faster than the state-of-the-art solvers such as SVMperf , Pegasos and the LibLinear that integrates the TRON, PCD and DCD algorithms.

T91 Scaling SVM and Least Absolute Deviations via Exact Data Reduction

Jie Wang; Peter Wonka; Jieping Ye

The support vector machine (SVM) is a widely used method for classification. Although many efforts have been devoted to develop efficient solvers, it remains challenging to apply SVM to large-scale problems. A nice property of SVM is that the non-support vectors have no effect on the resulting classifier. Motivated by this observation, we present fast and efficient screening rules to discard non-support vectors by analyzing the dual problem of SVM via variational inequalities (DVI). As a result, the number of data instances to be entered into the optimization can be substantially reduced. Some appealing features of our screening method are: (1) DVI is safe in the sense that the vectors discarded by DVI are guaranteed to be non-support vectors; (2) the data set needs to be scanned only once to run the screening, and its computational cost is negligible compared to that of solving the SVM problem; (3) DVI is independent of the solvers and can be integrated with any existing efficient solver. We also show that the DVI technique can be extended to detect non-support vectors in the least absolute deviations regression (LAD). To the best of our knowledge, there are currently no screening methods for LAD. We have evaluated DVI on both synthetic and real data sets. Experiments indicate that DVI significantly outperforms the existing state-of-the-art screening rules for SVM, and it is very effective in discarding non-support vectors for LAD. The speedup gained by DVI rules can be up to two orders of magnitude.

T92 Margins, Kernels and Non-linear Smoothed Perceptrons

Aaditya Ramdas; Javier Peña

We focus on the problem of finding a non-linear classification function that lies in a Reproducing Kernel Hilbert Space (RKHS) both from the primal point of view (finding a perfect separator when one exists) and the dual point of view (giving a certificate of non-existence), with special focus on generalizations of two classical schemes - the Perceptron (primal) and Von-Neumann (dual) algorithms. We cast our problem as one of maximizing the regularized normalized hard-margin (ρ) in an RKHS and %use the Representer Theorem to rephrase it in terms of a Mahalanobis dot-product/semi-norm associated with the kernel's (normalized and signed) Gram matrix. We derive an accelerated smoothed algorithm with a convergence rate of $\tfrac{\sqrt {\log n}}{\rho}$ given n separable points, which is strikingly similar to the classical kernelized Perceptron algorithm whose rate is $\tfrac1{\rho^2}$. When no such classifier

exists, we prove a version of Gordan's separation theorem for RKHSs, and give a reinterpretation of negative margins. This allows us to give guarantees for a primal-dual algorithm that halts in $\min\{\tfrac{\sqrt n}{|\rho|}, \tfrac{\sqrt n}{\epsilon}\}$ iterations with a perfect separator in the RKHS if the primal is feasible or a dual ϵ-certificate of near-infeasibility.

T93 Saddle Points and Accelerated Perceptron Algorithms

Adams Wei Yu; Fatma Kilinc-Karzan; Jaime Carbonell

In this paper, we consider the problem of finding a linear (binary) classifier or providing a near-infeasibility certificate if there is none. We bring a new perspective to addressing these two problems simultaneously in a single efficient process, by investigating a related Bilinear Saddle Point Problem (BSPP). More specifically, we show that a BSPP-based approach provides either a linear classifier or an ϵ-infeasibility certificate. We show that the accelerated primal-dual algorithm, Mirror Prox, can be used for this purpose and achieves the best known convergence rate of $O({\sqrt{\log n}\over\rho(A)})$ ($O({\sqrt{\log n}\over\epsilon})$), which is \emph{almost independent of the problem size}, n. Our framework also solves kernelized and conic versions of the problem, with the same rate of convergence. We support our theoretical findings with an empirical study on synthetic and real data, highlighting the efficiency and numerical stability of our algorithms, especially on large-scale instances.

T94 Outlier Path: A Homotopy Algorithm for Robust SVM

Shinya Suzumura; Kohei Ogawa; Masashi Sugiyama; Ichiro Takeuchi

In recent applications with massive but less reliable data (e.g., labels obtained by a semi-supervised learning method or crowdsourcing), non-robustness of the support vector machine (SVM) often causes considerable performance deterioration. Although improving the robustness of SVM has been investigated for long time, robust SVM (RSVM) learning still poses two major challenges: obtaining a good (local) solution from a non-convex optimization problem and optimally controlling the robustness-efficiency trade-off. In this paper, we address these two issues simultaneously in an integrated way by introducing a novel homotopy approach to RSVM learning. Based on theoretical investigation of the geometry of RSVM solutions, we show that a path of local RSVM solutions can be computed efficiently when the influence of outliers is gradually suppressed as simulated annealing. We experimentally demonstrate that our algorithm tends to produce better local solutions than the alternative approach based on the concave-convex procedure, with the ability of stable and efficient model selection for controlling the influence of outliers.

T95 Optimal Budget Allocation: Theoretical Guarantee and Efficient Algorithm

Tasuku Soma; Naonori Kakimura; Kazuhiro Inaba; Ken-ichi Kawarabayashi

We consider the budget allocation problem over bipartite influence model proposed by Alon et al. This problem can be viewed as the well-known influence maximization problem with budget constraints. We

first show that this problem and its much more general form fall into a general setting; namely the monotone submodular function maximization over integer lattice subject to a knapsack constraint. Our framework includes Alon et al.'s model, even with a competitor and with cost. We then give a (1-1/e)-approximation algorithm for this more general problem. Furthermore, when influence probabilities are nonincreasing, we obtain a faster (1-1/e)-approximation algorithm, which runs essentially in linear time in the number of nodes. This allows us to implement our algorithm up to almost 10M edges (indeed, our experiments tell us that we can implement our algorithm up to 1 billion edges. It would approximately take us only 500 seconds.).

Tuesday June 24

16:20 - Track F - Time Series and Sequences

T96 Boosting multi-step autoregressive forecasts

Souhaib Ben Taieb; Rob Hyndman

Multi-step forecasts can be produced recursively by iterating a one-step model, or directly using a specific model for each horizon. Choosing between these two strategies is not an easy task since it involves a trade-off between bias and estimation variance over the forecast horizon. Using a nonlinear machine learning model makes the tradeoff even more difficult. To address this issue, we propose a new forecasting strategy which boosts traditional recursive linear forecasts with a direct strategy using a boosting autoregression procedure at each horizon. First, we investigate the performance of the proposed strategy in terms of bias and variance decomposition of the error using simulated time series. Then, we evaluate the proposed strategy on real-world time series from two forecasting competitions. Overall, we obtain excellent performance with respect to the standard forecasting strategies.

T97 Modeling Correlated Arrival Events with Latent Semi-Markov Processes

Wenzhao Lian; Vinayak Rao; Brian Eriksson; Lawrence Carin

The analysis and characterization of correlated point process data has wide applications, ranging from biomedical research to network analysis. In this work, we model such data as generated by a latent collection of continuous-time binary semi-Markov processes, corresponding to external events appearing and disappearing. A continuous-time modeling framework is more appropriate for multichannel point process data than a binning approach requiring time discretization, and we show connections between our model and recent ideas from the discrete-time literature. We describe an efficient MCMC algorithm for posterior inference, and apply our ideas to both synthetic data and a real-world biometrics application.

T98 Asymptotically consistent estimation of the number of change points in highly dependent time series

Azadeh Khaleghi; Daniil Ryabko

The problem of change point estimation is considered in a general framework where the data are generated by arbitrary unknown stationary ergodic process distributions. This means that the data may have long-range dependencies of an arbitrary form. In this context the consistent estimation of the number of change points is provably impossible. A formulation is proposed which overcomes this obstacle: it is possible to find the correct number of change points at the expense of introducing the additional constraint that the correct number of process distributions that generate the data is provided. This additional parameter has a natural interpretation in many real-world applications. It turns out that in this formulation change point estimation can be reduced to time series clustering. Based on this reduction, an algorithm is proposed that finds the number of change points and locates the changes. This algorithm is shown to be asymptotically consistent. The theoretical results are complemented with empirical evaluations.

T99 Effective Bayesian Modeling of Groups of Related Count Time Series

Nicolas Chapados

Time series of counts arise in a variety of forecasting applications, for which traditional models are generally inappropriate. This paper introduces a hierarchical Bayesian formulation applicable to count time series that can easily account for explanatory variables and share statistical strength across groups of related time series. We derive an efficient approximate inference technique, and illustrate its performance on a number of datasets from supply chain planning.

T100 Stochastic Variational Inference for Bayesian Time Series Models

Matthew Johnson; Alan Willsky

Bayesian models provide powerful tools for analyzing complex time series data, but performing inference with large datasets is a challenge. Stochastic variational inference (SVI) provides a new framework for approximating model posteriors with only a small number of passes through the data, enabling such models to be fit at scale. However, its application to time series models has not been studied. In this paper we develop SVI algorithms for several common Bayesian time series models, namely the hidden Markov model (HMM), hidden semi-Markov model (HSMM), and the nonparametric HDP-HMM and HDP-HSMM. In addition, because HSMM inference can be expensive even in the minibatch setting of SVI, we develop fast approximate updates for HSMMs with durations distributions that are negative binomials or mixtures of negative binomials.

T101 Understanding Protein Dynamics with L1-Regularized Reversible Hidden Markov Models

Robert McGibbon; Bharath Ramsundar; Mohammad Sultan; Gert Kiss; Vijay Pande

We present a machine learning framework for modeling protein dynamics. Our approach uses L1-regularized, reversible hidden Markov models to understand large protein datasets generated via molecular dynamics simulations. Our model is motivated by three design principles: (1) the requirement of massive scalability; (2) the need to adhere to relevant physical law; and (3) the necessity of providing

accessible interpretations, critical for rational protein engineering and drug design. We present an EM algorithm for learning and introduce a model selection criteria based on the physical notion of relaxation timescales. We contrast our model with standard methods in biophysics and demonstrate improved robustness. We implement our algorithm on GPUs and apply the method to two large protein simulation datasets generated respectively on the NCSA Bluewaters supercomputer and the Folding@Home distributed computing network. Our analysis identifies the conformational dynamics of the ubiquitin protein responsible for signaling, and elucidates the stepwise activation mechanism of the c-Src kinase protein.

WORKSHOP

Wednesday, June 25th

Topological Methods for Machine Learning
Convention Hall No.4A
Jerry Zhu, Yuan Yao, Lek-Heng Lim, Jun Zhu

Workshop on Crowdsourcing and Human Computing
Convention Hall No.4B
Adish Singla, Xi Chen, Gagan Goel, Nihar Shah, Dengyong Zhou

The 3rd Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)
Convention Hall No.4C
Yunqing Xia, Erik Cambria, Yongzheng Zhang, Newton Howard

Learning, Security & Privacy
Convention Hall No.4D
Daniel Lowd, Pavel Laskov, Christos Dimitrakakis, Benjamin Rubinstein, Elaine Shi

Designing Machine Learning Platforms for Big Data
Room 201A
Xiangxiang Meng, Wayne Thompson, Xiaodong Lin

New Learning Frameworks and Models for Big Data
Room 201B
Massih-Reza Amini, Eric Gaussier, James Kwok, Yiming Yang

Deep Learning Models for Emerging Big Data Applications
Room 201C
Shan Suthaharan, Jinzhu Jia

Workshop on the Method of Moments and Spectral Learning
Convention Hall No.4E
Daniel Hsu, David Sontag, Percy Liang, Borja Balle Pigem, Byron Boots, Yoni Halpern

Causal Modeling & Machine Learning
Convention Hall No.4F
Kun Zhang, Bernhard Schölkopf, Elias Bareinboim, Jiji Zhang

†Machine Learning in China
Room 308
Min-Ling Zhang, Wu-Jun Li

‡(Joint with APSYS) ML Meets Systems
Room 307
Lorenzo Alvisi, Haibo Chen

Thursday, June 26th

The AutoML
Convention Hall No.4A
Frank Hutter, Rich Caruana, Remi Bardenet, Misha Bilenko, Isabelle Guyon, Balazs Kegl, Hugo Larochelle

Unsupervised Learning for Bioacoustic Big Data
Convention Hall No.4B
Hervé Glotin, Peter Dugan, Faicel Chamroukhi, Christopher Clark, Yann LeCun

Knowledge-Powered Deep Learning for Text Mining
Convention Hall No.4C
Bin Gao, Scott Yih, Richard Socher, Jiang Bian

Covariance Selection & Graphical Model Structure Learning
Convention Hall No.4D
Cho-Jui Hsieh, Peder Olsen, Inderjit S. Dhillon, Pradeep Ravikumar, Jorge Nocedal

Optimizing Customer Lifetime Value in Online Marketing
Convention Hall No.4E
Georgios Theocharous, Mohammad Ghavamzadeh, Shie Mannor

Learning Tractable Probabilistic Models
Convention Hall No.4F
Mathias Niepert, Pedro Domingos, Daniel Lowd

Divergence Methods for Probabilistic Inference
Room 308
Oluwasanmi Koyejo, Jun Zhu, Mark Reid, Eric Xing

‡(Joint with APSYS) ML Meets Systems
Room 307
Lorenzo Alvisi, Haibo Chen

REVIEWER LIST

Yasin Abbasi-Yadkori
Naoki Abe
Margareta Ackerman
Ryan Adams
Alekh Agarwal
Arvind Agarwal
John Agosta
Nir Ailon
Morteza Alamgir
Genevera Allen
Jose Alvarez
Chris Amato
Kareem Amin
Massih-Reza Amini
Hyrum Anderson
David Andrzejewski
Ery Arias-Castro
Raman Arora
Hossein Azari Soufiani
Stephen Bach
Krishnakumar Balasubramanian
Luca Baldassarre
David Balduzzi
Borja Balle
Akshay Balsubramani
Remi Bardenet
Elias Bareinboim
Andre Barreto
Gabor Bartok
Aviad Barzilay
Dhruv Batra
Stephen Becker
Ron Bekkerman
Kedar Bellare
Djalel Benbouzid
Shai Ben-David
James Bergstra
Michael Betancourt
Alina Beygelzimer
Misha Bilenko
Mustafa Bilgic
Charles Blundell
Sagan Bolliger
Edwin Bonilla
Byron Boots
Antoine Bordes
Luke Bornn
Alexandre Bouchard-Cote
Y-Lan Boureau
Jordan Boyd-Graber
Gavin Brown
Marcus Brubaker
Michael Brueckner
Sebastian Bubeck
Chris Burges
Trevor Campbell
Stephane Canu
Francois Caron
Nicolo Cesa-Bianchi
Volkan Cevher
Brahim Chaib-draa
Program Chairs
Antoni Chan
Jonathan Chang
Kai-Wei Chang
Ming-Wei Chang
Nicolas Chapados
Sanjay Chawla
Gal Chechik
Minmin Chen
Ning Chen
Songcan Chen
Tian Chen
Xi Chen
Yixin Chen
Yudong Chen
Yutian Chen
Alexey Chernov
David Chiang
Arthur Choi
Jaesik Choi
Anna Choromanska
Adam Coates
Shay Cohen
Ronan Collobert
Julien Cornebise
Aaron Courville
Aaron Courville
Jacob Crandall
Antonio Criminisi
James Cussens
Marco Cuturi
Alfredo Cuzzocrea
Bo Dai
Gal Dalal
Florence D'Alche-Buc
Anirban Dasgupta
Yann Dauphin
Ian Davidson
Jesse Davis
Cassio de Campos
Marc Deisenroth
Krzysztof Dembczynski
Thomas Deselaers
Debadeepta Dey
Paramveer Dhillon
Francesco Dinuzzo
Carlos Diuk
Nemanja Djuric
Justin Domke
Janardhan Rao (Jana) Doppa
Doug Downey
Kurt Driessens
Greg Druck
Kevin Duh
Haimonti Dutta
David Duvenaud
Chris Dyer
Jacob Eisenstein
Khalid El-Arini
Tal El-Hay
Lloyd Elliott
Dumitru Erhan
Xingyuan Fan
Amir-Massoud Farahmand
Sergey Feldman
Aasa Feragen
Xiaoli Fern
Dean Foster
Nicholas Foti
Rina Foygel
Vojtech Franc

REVIEWER LIST

Peter Frazier
Dayne Freitag
Arik Friedman
Johannes Fuernkranz
Kenji Fukumizu
Kuzman Ganchev
Jing Gao
Roman Garnett
Romaric Gaudel
Rong Ge
Peter Gehler
Alborz Geramifard
Sebastien Gerchinovitz
Pierre Geurts
Mohammad Ghavamzadeh
Mohammad Gheshlaghi Azar
Ran Gilad-Bachrach
Lee Giles
Inmar Givoni
Sharad Goel
Robby Goetschalckx
Jacob Goldberger
Ya'ara Goldschmidt
Daniel Golovin
Manuel Gomez-Rodriguez
Alon Gonen
Mehmet G?nen
Yunchao Gong
Joseph Gonzalez
Ian Goodfellow
Aditya Gopalan
Amit Goyal
Thore Graepel
David Grangier
Mihajlo Grbovic
Roger Grosse
Alex Grubb
Steffen Grunewalder
Marek Grzes
Sergio Guadarrama
Vincent Guigue
Yuhong Guo
Michael Gutmann
Andras Gyorgy
Hannaneh Hajishirzi
Yoni Halpern
Greg Hamerly
Zaid Harchaoui
Anne-Claire Haury
Jingrui He
Xuming He
Matthias Hein
David Helmbold
Lisa Hendricks
Philipp Hennig
Jose Miguel Hernandez-Lobato
Qirong Ho
Shen-Shyang Ho
Julia Hockenmaier
Jesse Hoey
Matt Hoffman
Jake Hofman
Steven Hoi
Cho-Jui Hsieh
Chun-Nan Hsu
Daniel Hsu
Jun Huan
Fei Huang
Furong Huang
Gao Huang
Jonathan Huang
Junzhou Huang
Qixing Huang
Shuai Huang
Eyke Huellermeier
Michael Hughes
Rebecca Hutchinson
Tsuyoshi Ide
Christian Igel
Alex Ihler
Rishabh Iyer
Nori Jacoby
Martin Jaggi
Prateek Jain
Ragesh Jaiswal
Michael James
Aditya Jami
Majid Janzamin
Shervin Javdani
Tony Jebara
David Jensen
Yacine Jernite
Yangqing Jia
Yun Jiang
Rong Jin
Vladimir Jojic
Armand Joulin
Tobias Jung
Adam Kalai
Hetunandan Kamisetty
Varun Kanade
Dimitri Kanevsky
Tapas Kanungo
Ashish Kapoor
Nikos Karampatziakis
Amin Karbasi
Hisashi Kashima
Koray Kavukcuoglu
Yoshinobu Kawahara
Balazs Kegl
Roni Khardon
Kee-Eung Kim
Myunghwan Kim
Sungwoong Kim
Jyrki Kivinen
Marius Kloft
Jens Kober
Mikko Koivisto
Stanley Kok
Alek Kolcz
Vladimir Kolmogorov
Nikos Komodakis
George Konidaris
Christian Konig
Aryeh Kontorovich
Hema Koppula
Anoop Korattikara
Wojciech Kotlowski
Alex Kulesza
Abhishek Kumar
Sanjiv Kumar
Sesh Kumar

REVIEWER LIST

James Kwok
Alexandre Lacoste
Simon Lacoste-Julien
Sebastien Lahaie
Remi Lajugie
Christoph Lampert
Niels Landwehr
Terran Lane
John Langford
Hugo Larochelle
Alessandro Lazaric
Quoc Le
Hoai An Le Thi
Guillaum Lecue
Jason Lee
Wee Sun Lee
Augustin Lefevre
Fuxin Li
Lei Li
Limin Li
Ping Li
Wu-Jun Li
Yu-Feng Li
Xuejun Liao
Chih-Jen Lin
Hui Lin
Nan Lin
Yuanqing Lin
Zhouchen Lin
Christoph Lippert
Michael Littman
Hanzhong Liu
Jun Liu
MeiZhu Liu
Qiang Liu
Tie-Yan Liu
Yi-Kai Liu
Ying Liu
Roi Livni
Daniel Lizotte
Ashley Llorens
Po-Ling Loh
Ben London
Mingsheng Long
Yin Lou
Yucheng Low
Daniel Lowd
Aurelie Lozano
Heng Luo
Andrew Maas
Dougal Maclaurin
Rich Maclin
Sridhar Mahadevan
Ashique Rupam Mahmood
Odalric-Ambrym Maillard
Julien Mairal
Michael Mandel
Vikash Mansinghka
Ben Marlin
James Martens
Andre Martins
Hamed Masnadi-Shirazi
Shin Matsushima
Stan Matwin
Julian McAuley
Jon McAuliffe
Andrew McCallum
Brian McFee
Scott McQuade
Roland Memisevic
Aditya Menon
Srujana Merugu
Ofer Meshi
Tomas Mikolov
Mahdi Milani Fard
David Mimno
Paul Mineiro
Roni Mittelman
Joseph Modayil
Shakir Mohamed
Karthik Mohan
Gregoire Montavon
Guido MontufarCuartas
Juston Moore
Alessandro Moschitti
Remi Munos
Daniel Munoz
Enrique Munoz de Cote
Andres Munoz Medina
Fionn Murtagh
Shinichi Nakajima
Atsuyoshi Nakamura
Karthik Narayan
Saketha Nath
Willie Neiswanger
Praneeth Netrapalli
Gergely Neu
Marion Neumann
Eric Nichols
Alexandru Niculescu-Mizil
Jordi Nin
Yang Ning
William Noble
Richard Nock
Ann Nowe
Sebastian Nowozin
Guillaume Obozinski
Brendan O'Connor
Sylvie Ong
Francesco Orabona
Michael Osborne
Simon Osindero
Takayuki Osogami
Hua Ouyang
Hua Ouyang
Diane Oyen
Alain Pagani
John Paisley
Konstantina Palla
Ankur Parikh
Neal Parikh
Nathan Parrish
Razvan Pascanu
Andrea Passerini
Dmitry Pechyony
Jian Peng
Fernando Perez-Cruz
Franz Pernkopf
Jonas Peters
Marek Petrik
Daniel Polani
David Poole

REVIEWER LIST

Pascal Poupart
Philippe Preux
Kriti Puniyani
Novi Quadrianto
Qichao Que
Michael Rabbat
Filip Radlinski
Piyush Rai
Tapani Raiko
Barbara Rakitsch
Alain Rakotomamonjy
Liva Ralaivola
Deepak Ramachandran
Daniel Ramage
Karthik Raman
Aaditya Ramdas
Jan Ramon
Fabio Ramos
bharath Ramsundar
Vinayak Rao
Nathan Ratliff
Sujith Ravi
Soumya Ray
Narges Razavian
Mark Reid
Lev Reyzin
Oren Rippel
Irene Rodriguez
Stephane Ross
Fabrice Rossi
Volker Roth
Daniel Roy
Avraham Ruderman
Nicholas Ruozzi
Alexander Rush
Daniil Ryabko
Sivan Sabato
Ankan Saha
Avishek Saha
Tara Sainath
Ruslan Salakhutdinov
Mathieu Salzmann
Rajhans Samdani
Ted Sandler
Scott Sanner
James Saunderson
Ivan Savov
Christoph Sawade
Mark Schmidt
Tobias Schnabel
D. Sculley
Michele Sebag
Hanie Sedghi
Yevgeny Seldin
Parikshit Shah
James Sharpnack
Daniel Sheldon
Christian Shelton
Bin Shen
WEINING Shen
Xiaotong Shen
Shohei Shimizu
Pannagadatta Shivaswamy
Lavi Shpigelman
Ozgur Simsek
Vikas Sindhwani
Sameer Singh
Kaushik Sinha
Mathieu Sinn
Martin Slawski
Jasper Snoek
Richard Socher
Jascha Sohl-Dickstein
Kihyuk Sohn
Kihyuk Sohn
KyungAh Sohn
Daria Sorokina
Vivek Srikumar
Bharath Sriperumbudur
Jacob Steinhardt
Veselin Stoyanov
Jiang Su
Amarnag Subramanya
Mahito Sugiyama
Masashi Sugiyama
Liang Sun
Tingni Sun
Yijun Sun
Ilya Sutskever
Taiji Suzuki
Kevin Swersky
Istvan Szita
Arthur Szlam
Partha Talukdar
Cheng Tang
Yichuan Tang
Daniel Tarlow
Graham Taylor
Matthew Taylor
matus Telgarsky
Nedelina Teneva
Gerald Tesauro
Ambuj Tewari
Ivan Titov
Ryota Tomioka
Hanghang Tong
Marc Toussaint
Long Tran-Thanh
Grigorios Tsoumakas
Zhuowen Tu
Stephen Tyree
Niranjan U N
Ruth Urner
Daniel Vaisencher
Laurens van der Maaten
Tim Van Erven
Harm van Seijen
Bart Vandereycken
Vincent Vanhoucke
Kush Varshney
Nakul Verma
Paul Vernaza
Jean-Philippe Vert
Bernardo Bernardo ávila Pires
Pascal Vincent
Oriol Vinyals
S V N Vishwanathan
Fabio Vitale
Nikos Vlassis
Maksims Volkovs
Jan Vondrak
Slobodan Vucetic

REVIEWER LIST

Yoav Wald
Thomas Walsh
Chong Wang
Fei Wang
Hongning Wang
Huan Wang
Jun Wang
Jun Wang
Lie Wang
Wei Wang
Fabian Wauthier
Adrian Weller
Jason Weston
Adam White
Shimon Whiteson
Marco Wiering
Becca Willett
Aaron Wilson
Anthony Wirth
Frank Wood
Lin Xiao
Lexing Xie
Pengtao Xie
Eddie Xu
Huan Xu
Jinbo Xu
Linli Xu
Min Xu
Minjie Xu
Zhao Xu
Feng Yan
Liu Yang
Chen Yanover
Guibo Ye
Scott Yih
Emine Yilmaz
Junming Yin
Wotao Yin
Yiming Ying
Elad Yom-Tov
Chun-Nam Yu
Jian Yu
Kai Yu
Shipeng Yu
Yang Yu
Yaoliang Yu
Lei Yuan
Xiaotong Yuan
Yisong Yue
Zhang Yuhang
Junping Zhang
Min-Ling Zhang
Ning Zhang
Teng Zhang
Xinhua Zhang
Ya Zhang
Yu Zhang
Yuchen Zhang
Bin Zhao
Peilin Zhao
Tuo Zhao
Shandian Zhe
Dengyong Zhou
Mingyuan Zhou
Qiang Zhou
Shenghuo Zhu
Brian Ziebart
Matt Zucker
Alon Zweig

AUTHOR INDEX

Abbasi-Yadkori, Yasin: M45, T72, T75
Abrahao, Bruno: S46
Adams, Ryan: S6, M40, M70, M86, T88
Affandi, Raja Hafiz: M40
Agarwal, Alekh: S99, M48
Agarwal, Arpit: S29
Agarwal, Shivani: S28, S29
Agrawal, Kunal: M22
Ahmed, Bilal: S69
Ahn, Sungjin: S51
Ailon, Nir: S45
Airoldi, Edoardo: S3, S77, T76
Akrour, Riad: T41
Alain, Guillaume: M69
Amin, Kareem: T70
Ammar, Haitham Bou: M10
Anandkumar, Animashree: M98
Anaraki, Farhad Pourkamali: S23
Andoni, Alexandr: T62
Andrieu, Christophe: S52
Argyriou, Andreas: T58
Arias, Emilio Jesus Gallego: S102
Arlot, Sylvain: M19
Arora, Sanjeev: T60
Asteris, Megasthenis: T64
Avron, Haim: T57
Awasthi, Pranjal: M16
Azadi, Samaneh: M28
Azar, Mohammad Gheshlaghi: S47
Azizi, Elham: S3
Böhm, Klemens: S20
Bach, Francis: M19
Bachman, Philip: T24
Bai, Qinxun: M47
Bailey, James: M17
Balcan, Maria: M16, T71
Balle, Borja: S63, M57
Barber, David: S18
Bardenet, Rémi: S50
Bartlett, Peter: M45, T72, T75
Barto, Andrew: T42
Bartok, Gabor: T11
Bartunov, Sergey: T79
Bayen, Alexandre: M43
Beijbom, Oscar: S30
Bellemare, Marc: M39
Ben-David, Shai: M14
Benavoli, Alessio: S59
Bengio, Yoshua: M68, M69
Bhaskara, Aditya: T60
Bhattacharyya, Chiranjib: T67
Bhojanapalli, Srinadh: S74, S75
Bilmes, Jeff: S105
Bittorf, Victor: M27
Black, Michael: T36
Blackmon, Karen: S69
Blanchard, Gilles: T20
Blei, David: T44
Blundell, Charles: S34
Blunsom, Phil: M34
Bogunovic, Ilija: T11
Bonilla, Edwin: T81
Bornn, Luke: T88
Botha, Jan: M34
Bouchard-Côté, Alexandre: M49, M50
Bousmalis, Konstantinos: M76
Bratieres, Sebastien: S61
Brodley, Carla: S69
Brunskill, Emma: S7, S47
Bui, Hung: S86
Burnside, Elizabeth: T34
Busa-Fekete, Robert: T77
Butscher, Adrian: T17
Caetano, Tiberio: S96
Calderbank, Robert: T52
Caramanis, Constantine: M59, T65
Carbonell, Jaime: T93
Carin, Lawrence: M85, T52, T97
Carlsson, Gunnar: M15
Carreras, Xavier: S63
Celik, Safiye: S5
Celikkaya, E. Busra: M41
Chaganty, Arun Tejasvi: M58
Chakrabarti, Deepayan: S1
Chan, Stanley: S77
Chang, Jonathan: S1
Chang, Kai-Wei: M1
Chang, Kevin: T16
Chang, Shih-Fu: T30
Chapados, Nicolas: T99
Chazal, Frédéric: M92
Chechik, Gal: M103
Chen, Gary: M85
Chen, Jianfei: T78
Chen, Minmin: M68
Chen, Shang-Tse: M61
Chen, Tianqi: S53
Chen, Wei: S46
Chen, Xi: M44
Chen, Yudong: S4, S74, T54
Chen, Yutian: S49
Chen, Yuxin: S40, S73, M101
Chen, Zhiyuan: T47
Cherian, Anoop: T28
Chien, Hao-Heng: S2
Chwialkowski, Kacper: T56
Cicalese, Ferdinando: S90
Clémençon, Stephan: M99
Cohen, Taco: M6
Collobert, Ronan: S66
Combes, Richard: S44
Contal, Emile: S94
Corani, Giorgio: S59
Cortes, Corinna: M63, M64
Crammer, Koby: S58, M45
Cunningham, John: S16
Cuturi, Marco: S88
Dai, Bo: M91, M98
Daneshmand, Hadi: T69
Danihelka, Ivo: S34
Darrell, Trevor: S68, M100
Das, Sanmay: T48
Dasgupta, Sanjoy: M78
Davulcu, Hasan: T1
DeDeo, Simon: M13
DeWeese, Michael: M53
Decoste, Dennis: T89
Defazio, Aaron: S96
Degris, Thomas: T40
Denil, Misha: S37, M62
Denis, François: M60
Devinsky, Orrin: S69
Dhillon, Inderjit: S101, M94, M95, T46
Dick, Travis: T74
Dietterich, Thomas: S39, S80
Dimakis, Alexandros: T64, T65
Ding, Guiguang: S106
Dinuzzo, Francesco: T58
Domingos, Pedro: M2
Domke, Justin: S96
Donahue, Jeff: S68
Doucet, Arnaud: S50, S88
Drighès, Benjamin: M43
Du, Chao: T29
Du, Nan: T71
Dubhashi, Devdatt: T67
Dukkipati, Ambedkar: S71
Dunson, David: M85, T82
Dworkin, Lili: M11
Eaton, Eric: M10
Eban, Elad: T35
Efros, Pavel: S20
El-Yaniv, Ran: S58
Eriksson, Brian: M3, T97
Ermon, Stefano: T31
Estrach, Joan Bruna: M72

AUTHOR INDEX

Fan, Wei: S103, T1
Fang, Yuan: T16
Farahmand, Amir-Massoud: T24
Fleuret, Francois: M65
Fox, Emily: S53, M40
Freitas, Nando De: S37, M62
Frey, Brendan: M38
Fuchs, Thomas: M24
Fujimaki, Ryohei: T50
Fujiwara, Yasuhiro: T68
Fukumizu, Kenji: T55
Funiak, Stanislav: S1
Gaboardi, Marco: S102
Gal, Yarin: T80
Galagan, James: S3
Galstyan, Aram: M13
Ganguli, Surya: S98
Garcia, Maria Lomeli: S52
Gardner, Jacob: S16
Garg, Vikas: T5
Ge, Rong: T60
Geer, Connor: T87
Gelbart, Michael: M70
Gentile, Claudio: T9
Ghahramani, Zoubin: S22, S61, M73, M74, M75, M84, M87, T80
Ghosh, Joydeep: S76
Gieseke, Fabian: S104
Giesen, Joachim: M97
Gionis, Aristides: S65
Girshick, Ross: M100
Gittens, Alex: T89
Gleich, David: T63
Glisse, Marc: M92
Globerson, Amir: S42, S63, T35
Goldfarb, Donald: S72
Goldsberry, Kirk: T88
Gomes, Carla: T31
Gomez, Faustino: M36
Gomez-Rodriguez, Manuel: T69
Gong, Yunchao: T30
Gopal, Siddharth: S84
Gopalan, Aditya: S43
Grande, Robert: S82
Graves, Alex: M35
Greff, Klaus: M36
Gregor, Karol: S34, S36
Greiner, Russell: S27, M38
Gretton, Arthur: S52, T55, T56
Guestrin, Carlos: S53
Guibas, Leonidas: S40, S73, T17
Gunasekar, Suriya: S76
Guo, Shengbo: M85
Guo, Yuhong: S67
Gybels, Mattias: M60
Gyorgy, Andras: S48, T74
Habrard, Amaury: M60
Haeffele, Benjamin: M77
Haghtalab, Nika: M14
Hajiaghayi, Monir: M50
Hamid, Raffay: T89
Hamilton, William: M57
Han, Dingyi: T23
Harada, Tatsuya: M37
Harchaoui, Zaid: M100
Harel, Maayan: S58
Hasselt, Hado van: T37
Hazan, Elad: S92
Hazan, Tamir: S41, T32
He, Xiaofei: M89
Heaukulani, Creighton: M87
Heess, Nicolas: T40
Heidari, Hoda: T70
Heinemann, Uri: S42
Heinermann, Justin: S104
Hernandez-Lobato, Jose Miguel: M73, M74, M75
Hero, Al: M3
Hoang, Trong Nghia: S91
Hoffman, Judy: S68
Holmes, Chris: S50
Honorio, Jean: T21
Hoos, Holger: S15
Houlsby, Neil: M73, M74, M75
How, Jonathan: S82
Hsieh, Cho-Jui: M94, M95, T3
Hsu, Daniel: S79, M48
Hsu, David: M7
Hsu, Justin: S102
Hu, Jinli: M81
Hu, Mingqing: S106
Huang, Bo: S72
Huang, Heng: S24, M20, T90
Huang, Qixing: S40, S73
Huang, Tzu-Kuo: S93
Huang, Yizhen: T90
Huellermeier, Eyke: T77
Hughes, Shannon: S23
Hutter, Frank: S15
Hyndman, Rob: T96
Igel, Christian: S104
Ihler, Alex: S60
Inaba, Kazuhiro: T95
Inouye, David: T46
Ioannidis, Stratis: M56
Irie, Go: T68
Isbell, Charles: M9
Ishiguro, Katsuhiko: M83
Iyer, Arun: M96
Iyer, Rishabh: S105
Jaakkola, Tommi: T21, T32
Jaillet, Patrick: S91
Jain, Anil: M18
Jain, Prateek: S75, S101, M79
Jaitly, Navdeep: M35
Janoos, Firdaus: T73
Janzing, Dominik: S56
Jawanpuria, Pratik: M93
Jegelka, Stefanie: M100
Jethava, Vinay: T67
Jia, Yangqing: S68
Jiang, Shali: T59
Jin, Rong: M18, T51
Joachims, Thorsten: S45
Johansson, Fredrik: T67
Johnson, Matthew: T100
Jun, Seong-Hwan: M49
Kakade, Sham: S99
Kakimura, Naonori: T95
Kale, Satyen: M48
Kalousis, Alexandros: M90
Kalyanakrishnan, Shivaram: S29
Kanade, Varun: T72
Kandasamy, Kirthevasan: T83
Kankanhalli, Mohan: S91
Kar, Purushottam: S101
Karampatziakis, Nikos: S21, S99
Karbasi, Amin: T11
Karnin, Zohar: S45, S92
Kashima, Hisashi: T15
Kaski, Samuel: T18
Kawarabayashi, Ken-ichi: T95
Kearns, Michael: M11, T70
Khaleghi, Azadeh: T98
Kilinc-Karzan, Fatma: T93
Kim, Dongwoo: S87
Kimura, Akisato: M83
Kingma, Diederik: M71
Kirkpatrick, Bonnie: M50
Kiros, Ryan: M31
Kiss, Gert: T101
Kleinberg, Robert: S46
Knowles, David: M84, M87
Kocák, Tomáš: T8
Koh, Lian Pin: M101

AUTHOR INDEX

Koltun, Vladlen: M8
Kondor, Risi: T5
Konidaris, George: T42
Kontorovich, Aryeh: S78, S81
Korattikara, Anoop: S49
Koutnik, Jan: M36
Kpotufe, Samory: S56
Krause, Andreas: M101, T11
Krichene, Walid: M43
Kriegman, David: S30
Krishnamurthy, Akshay: T83
Kuipers, Benjamin: S31
Kumar, Sanjiv: T30
Kung, H. T.: S70
Kurras, Sven: T20
Kusner, Matt: S16, M22
Kuznetsov, Vitaly: M63
Kuzniecky, Ruben: S69
Kveton, Branislav: T8
Kwok, James: S97, M25
Laber, Eduardo: S90
Labruère, Catherine: M92
Lacoste, Alexandre: S13
Lai, Ming-Jun: T1
Lajugie, Rémi: M19
Lam, Henry: M47
Lampert, Christoph: S17
Lan, Shiwei: M52
Lanckriet, Gert: M21
Langford, John: M48
Larochelle, Hugo: S13, S32
Laue, Soeren: M97
Laufer, Eric: M69
Lauw, Hady: T16
Laviolette, Francois: S13
Lavoie, Allen: T48
Lazaric, Alessandro: S47
Le, Quoc: M32
LeCun, Yann: M72
Lee, Honglak: S31, M67
Lee, Su-In: S5
Lee, Wee Sun: M7
Lefakis, Leonidas: M65
Letham, Benjamin: M4
Lever, Guy: T40
Levihn, Martin: M9
Levine, Sergey: M8
Leyton-Brown, Kevin: S15
Li, Chengtao: T78
Li, Chun-Liang: S26
Li, Jian: M44
Li, Juanzi: T45
Li, Lihong: S7, M48
Li, Ping: T26, T27, T49
Li, Qingyang: S103
Li, Shuai: T9
Li, Weichang: T73
Li, Wu-Jun: T23
Li, Xin: S67
Li, Yujia: S62
Li, Zhixing: T45
Lian, Wenzhao: T97
Liang, Percy: M12, M51, M58
Liang, Yingyu: T71
Liao, Shizhong: T59
Lim, Daryl: M21
Lim, Shiau Hong: S4
Lin, Binbin: M89
Lin, Hsuan-Tien: S26, M61
Lin, Kuan-Hua: S2
Lin, Lizhen: T82
Lin, Qihang: S95
Lin, Shou-de: T85
Lin, Tian: S46
Lin, Tsung-Han: S70
Lin, Zijia: S106
Linderman, Scott: S6
Liu, Bing: T47
Liu, Ji: M27, T50
Liu, Jie: T34
Liu, Jun: T25
Liu, Liping: S39, S80
Liu, Qiang: S60, T12
Liu, Yong: T59
Logsdon, Benjamin: S5
Lopez-Paz, David: S22
Low, Bryan Kian Hsiang: S91
Lowd, Daniel: S64, T33
Lozano, Aurelie: T53, T84
Lu, Chi-Jen: M61
Lu, Zhaosong: T1
Lui, John: S46
Luo, Haipeng: M82
Luxburg, Ulrike von: T19, T20
Lázaro-Gredilla, Miguel: M42
Ma, Ping: S100
Ma, Tengyu: T60
Macskassy, Sofus: S1
Maddison, Chris: M102
Mahmood, Ashique Rupam: T37
Mahoney, Michael: S100, T57, T63
Maillard, Odalric-Ambrym: T10
Mairal, Julien: M100
Malek, Alan: T75
Malioutov, Dmitry: T2
Mangili, Francesca: S59
Mankowitz, Daniel: S8
Mann, Timothy: S8, S83
Mannor, Shie: S8, S12, S43, S58, S83, T10
Mansour, Yishay: S43
Marchand, Mario: S13
Marchand-Maillet, Stéphane: M88, M90
Mary, Jérémie: T6
Matheson, David: M62
McGibbon, Robert: T101
Medina, Andres Munoz: M80
Meek, Christopher: T12
Mei, Qiaozhu: T43
Mei, Shike: S14
Meng, Deyu: S19
Meng, Zhaoshi: M3, T43
Mezuman, Elad: T35
Michel, Bertrand: M92
Mikolov, Tomas: M32
Miller, Andrew: T88
Mineiro, Paul: S21
Minsker, Stanislav: T82
Mitliagkas, Ioannis: T65
Mittelman, Roni: S31
Mitzenmacher, Michael: T27
Mizrahi, Yariv: S37
Mnih, Andriy: S34, S36
Mohamed, Shakir: S35
Mohri, Mehryar: M63, M64, M80
Montanari, Andrea: M56
Montesinos, Cesar Fuentes: M101
Mu, Cun: S72
Muandet, Krikamol: T55
Mudigonda, Mayur: M53
Mukuta, Yusuke: M37
Munos, Remi: T7, T8
Murray, Iain: S32
Mémoli, Facundo: M15
Müller, Emmanuel: S20
Nakagawa, Hiroshi: M54, T15
Nakano, Masahiro: M83
Narasimhan, Harikrishna: S29
Nath, Saketha: M93, M96
Neufeld, James: S48
Nevmyvaka, Yuriy: M11
Ngo, Vien: S10
Nguyen, Hoang Vu: S20
Nguyen, Tien Vu: S86
Nguyen, Trung: T81
Nguyen, Vinh: M17

AUTHOR INDEX

Nguyen, Xuanlong: S86, T43
Nicol, Olivier: T6
Nie, Feiping: S24, M20, T90
Niepert, Mathias: M2
Niu, Gang: M91
Novikov, Alexander: S38
Nowozin, Sebastian: S61
Oancea, Cosmin: S104
Ogawa, Kohei: T94
Oh, Alice: S87
Olsen, Peder: T3
Orabona, Francesco: T32
Osokin, Anton: S38
Pacheco, Jason: T36
Page, David: T34
Paige, Brooks: S54
Palla, Konstantina: M84
Pan, Sinno Jialin: T4
Pande, Vijay: T101
Pandey, Gaurav: S71
Panigrahy, Rina: T62
Papailiopoulos, Dimitris: T64, T65
Parkes, David: M55
Peña, Javier: T92
Peltonen, Jaakko: T18
Pentina, Anastasia: S17
Perchet, Vianney: S94
Pereira, Francisco: T13
Phung, Dinh: S86
Piekut, David: T87
Pineau, Joelle: M57
Ping, Wei: S60
Pinheiro, Pedro: S66
Platt, John: T12
Plessis, Christoffel du: M91
Poczos, Barnabas: S55, T83
Pollefeys, Marc: S41
Poole, Ben: S98
Precup, Doina: T24, T37
Preux, Philippe: T6
Proutiere, Alexandre: S44
Qian, Hui: T86
Qin, Zhiwei: T73
Quadrianto, Novi: S61
Quattoni, Ariadna: S63
Rabinovich, Maxim: T44
Rai, Piyush: M85
Rajkumar, Arun: S28
Ramdas, Aaditya: T92
Ramsundar, Bharath: T101
Rao, Vinayak: T97
Ravanbakhsh, Siamak: M38
Ravikumar, Pradeep: S76, T46, T53, T66, T84
Razi, Abolfazl: T52
Re, Christopher: M27
Reed, Scott: M67
Rennie, Jason: T76
Rey, Melanie: M24
Rezende, Danilo Jimenez: S35
Ribeiro, Alejandro: M15
Ribeiro, Bernardete: T13
Riedmiller, Martin: T40
Rijke, Maarten de: T7
Rippel, Oren: M70
Robbiano, Sylvain: M99
Rodomanov, Anton: S38
Rodrigues, Filipe: T13
Rodrigues, Miguel: T52
Romano, Simone: M17
Rooshenas, Amirmohammad: T33
Roth, Aaron: S102
Roth, Dan: M1
Roth, Volker: M24
Rousu, Juho: S65
Ruggeri, Fabrizio: S59
Rui, Yong: M23
Rustamov, Raif: T17
Ruvolo, Paul: M10
Ryabko, Daniil: T98
Sabato, Sivan: S79
Saberian, Mohammad: S30
Sabharwal, Ashish: T31
Saettler, Aline Medeiros: S90
Salakhutdinov, Ruslan: M31
Samdani, Rajhans: M1
Sanghavi, Sujay: S74, M59
Santos, Cicero Dos: M33
Sarawagi, Sunita: M96
Sarwate, Anand: T32
Sato, Issei: M54, T15
Savarese, Silvio: S31
Schapire, Robert: M48, M82
Scherrer, Bruno: S9
Schmidhuber, Juergen: M36
Schneider, Jeff: S93
Schoelkopf, Bernhard: S22, S56, T55, T69
Schoenauer, Marc: T41
Scholz, Jonathan: M9
Schuller, Bjoern: M76
Schuurmans, Dale: S48
Schwing, Alexander: S41
Sclaroff, Stan: M47
Sebag, Michele: T41
Segarra, Santiago: M15
Seijen, Harm van: T38
Sejdinovic, Dino: S52
Seldin, Yevgeny: M45, M46
Selman, Bart: T31
Sgouritsa, Eleni: S56
Sha, Fei: M13, M68, M90
Shahbaba, Babak: S51, M52
Shalev-Shwartz, Shai: M26, T61
Shalit, Uri: M103
Shamir, Ohad: M30
Sheldon, Daniel: S39
Shelton, Christian: M41
Sheopuri, Anshul: M4
Shi, Tianlin: S85
Shioi, Hiroaki: M101
Shrivastava, Anshumali: T26, T27
Si, Si: M94, M95
Silva, Bruno Da: T42
Silver, David: T40
Sim, Melvyn: S25
Sindhwani, Vikas: T57
Singh, Shashank: S55
Singla, Adish: T11
Slavov, Nikolai: T2
Slivkins, Aleksandrs: M46
Smith, Noah: T22
Smola, Alex: S22, S57
Snoek, Jasper: M86
Sohl-Dickstein, Jascha: S98, M53
Sohn, Kihyuk: M67
Solomon, Justin: T17
Soma, Tasuku: T95
Song, Hyun Oh: M100
Song, Le: S99, M98, T69, T71
Soufiani, Hossein Azari: M55
Souplet, Jean-Christophe: T41
Sra, Suvrit: S22, M28
Srebro, Nati: M30
Sridhar, Srikrishna: M27
Srinivasa, Christopher: M38
Sriperumbudur, Bharath: T55
Srivastava, Sanvesh: T82
Steeg, Greg Ver: M13
Steinhardt, Jacob: M12, M51
Storkey, Amos: M81
Strathmann, Heiko: S52
Su, Hongyu: S65
Sudderth, Erik: T36
Sugiyama, Masashi: M91, T94
Sultan, Mohammad: T101

Sun, Ke: M88, M90
Sun, Peng: M66
Sun, Wei: M4
Sun, Yuekai: M56
Sutton, Rich: T37, T38
Suzuki, Taiji: M29
Suzumura, Shinya: T94
Swersky, Kevin: M86
Syed, Umar: M64
Szörényi, Balázs: T77
Szepesvari, Csaba: S48, T74
Szlam, Arthur: M72
Taieb, Souhaib Ben: T96
Takeuchi, Ichiro: T94
Talvitie, Erik: M39
Tamar, Aviv: S12
Tan, Mingkui: T4
Tandon, Rashish: T66
Tang, Jian: T43
Tang, Jie: T45
Tao, Dacheng: M23
Tarlow, Daniel: M102
Taskar, Ben: M40
Taylor, Gavin: T87
Taylor, Matthew: M10
Teneva, Nedelina: T5
Terada, Yoshikazu: T19
Thakurta, Abhradeep Guha: M79
Thesen, Thomas: S69
Thomas, Philip: S11, T39
Tibshirani, Ryan: S57
Titsias, Michalis: M42
Torkamani, Mohamad Ali: S64
Tosh, Christopher: M78
Toulis, Panagiotis: T76
Toussaint, Marc: S10
Trigeorgis, George: M76
Troyanskaya, Olga: S33
Tsang, Ivor W.: T4
Tu, Bojun: T86
Tyree, Stephen: M22
Tzeng, Eric: S68
Ueda, Naonori: M83
Uria, Benigno: S32
Urtasun, Raquel: S41
Valiant, Gregory: S99, T62
Valko, Michal: T8
Vandereycken, Bart: T4
Varma, Manik: M93
Vasconcelos, Nuno: S30
Vayatis, Nicolas: S94
Veness, Joel: M39
Venkatesh, Swetha: S86
Verspoor, Karin: M17
Vetrov, Dmitry: S38, T79
Vidal, Rene: M77
Vinnikov, Alon: T61
Vinyals, Oriol: S68
Voevodski, Konstantin: M16
Vreeken, Jilles: S20
Walsh, Thomas: S82
Wang, Hua: M20
Wang, Jianmin: S106
Wang, Jie: S103, T25, T91
Wang, Jingdong: T29
Wang, Jun: M18, M90
Wang, Jun-Kun: T85
Wang, Li: T4
Wang, Liangliang: M50
Wang, Liming: T52
Wang, Naiyan: T14
Wang, Shusen: T86
Wang, Xiaoqian: T90
Wang, Xuezhi: S93
Wang, Yali: S18
Wang, Yingjian: M85
Wang, Yu-Xiang: S57
Wang, Zheng: T1
Ward, Rachel: S74
Wasserman, Larry: T83
Wei, Kai: S105
Weinberger, Kilian: S16, M22, M68
Weiss, Roi: S81
Weiss, Ron: M5
Welling, Max: S49, S51, M6, M71
Wen, Junfeng: S27
Wen, Siqiang: T45
Weston, Jason: M5
Whiteson, Shimon: T7
Wich, Serge: M101
Wierstra, Daan: S34, S35, T40
Wieschollek, Patrick: M97
Willsky, Alan: T100
Wonka, Peter: S103, T91
Wood, Frank: S54
Wright, John: S72
Wright, Steve: M27
Wu, Shan-Hung: S2
Wu, Zhiwei Steven: S102
Xia, Lirong: M55
Xiao, Lin: S95
Xiao, Ying: T89
Xie, Bo: M98
Xu, Zhixiang: S16
Xu, Chang: M23
Xu, Chao: M23
Xu, Huan: S4, S12, S25
Xu, Jiaming: T54
Xu, Zongben: S19
Xue, Gui-Rong: T23
Yamada, Takeshi: M83
Yan, Ling: T23
Yang, Eunho: T53, T84
Yang, Ji: M89
Yang, Jiyan: T57
Yang, Sen: S103
Yang, Wenzhuo: S25
Yang, Yiming: S84
Yang, Zhirong: T18
Ye, Jieping: S103, M89, T1, T25, T50, T91
Yee, Hector: M5
Yeung, Dit-Yan: T14
Yi, Jinfeng: M18, T51
Yi, Xinyang: M59
Yogatama, Dani: T22
Yosinski, Jason: M69
Young, Eric: M77
Yu, Adams Wei: T93
Yu, Bin: S100
Yu, Chun-Nam: S27
Yu, Felix: T30
Yu, Hsiang-Fu: S101
Yu, Philip: S2
Yuan, Jianjun: S24
Yuan, Xiaotong: T49
Zadrozny, Bianca: M33
Zafeiriou, Stefanos: M76
Zaffalon, Marco: S59
Zappella, Giovanni: T9
Zemel, Rich: S62, M31, M86
Zhang, Aonan: S89
Zhang, Bo: S89
Zhang, Chunming: T34
Zhang, Lei: S19
Zhang, Li: T62
Zhang, Lijun: M18, T51
Zhang, Ming: T43
Zhang, Ning: S68
Zhang, Peng: T45
Zhang, Ruiliang: S97
Zhang, Ting: T29
Zhang, Tong: M26, M30, M66, T49
Zhang, Yuting: M67
Zhang, Zhihua: T86

AUTHOR INDEX

Zhang, Zongzhang: M7
Zhao, Qian: S19
Zhao, Yijun: S69
Zhao, Zheng: T25
Zhong, Wenliang: M25
Zhou, Bo: M52
Zhou, Dengyong: T12
Zhou, Jian: S33
Zhou, Jie: M66
Zhou, Yuan: M44
Zhu, Jerry: S14
Zhu, Jun: S14, S85, S89, T78
Zoghi, Masrour: T7
Zuffi, Silvia: T36
Zuo, Wangmeng: S19


ICML 2014
BEIJING CHINA