Report from Dagstuhl Seminar 15152

Machine Learning with Interdependent and Non-identically Distributed Data

Edited by

Trevor Darrell[1], Marius Kloft[2], Massimiliano Pontil[3], Gunnar Rätsch[4], and Erik Rodner[5]

1 **University of California - Berkeley, US,** `trevor@eecs.berkeley.edu`
2 **HU Berlin, DE,** `kloft@hu-berlin.de`
3 **University College London, GB,** `m.pontil@cs.ucl.ac.uk`
4 **Memorial Sloan-Kettering Cancer Center – New York, US**
5 **Friedrich Schiller University Jena, DE,** `erik.rodner@uni-jena.de`

Abstract

One of the most common assumptions in many machine learning and data analysis tasks is that the given data points are realizations of independent and identically distributed (IID) random variables. However, this assumption is often violated, e.g., when training and test data come from different distributions (dataset bias or domain shift) or the data points are highly interdependent (e.g., when the data exhibits temporal or spatial correlations). Both scenarios are typical situations in visual recognition and computational biology. For instance, computer vision and image analysis models can be learned from object-centric internet resources, but are often rather applied to real-world scenes. In computational biology and personalized medicine, training data may be recorded at a particular hospital, but the model is applied to make predictions on data from different hospitals, where patients exhibit a different population structure. In the seminar report, we discuss, present, and explore new machine learning methods that can deal with non-i.i.d. data as well as new application scenarios.

Seminar April 7–10, 2015 – http://www.dagstuhl.de/15152
1998 ACM Subject Classification G.3 Probability and Statistics, I.4.8. Scene Analysis, J.3 Biology and Genetics
Keywords and phrases machine learning, computer vision, computational biology, transfer learning, domain adaptation
Digital Object Identifier 10.4230/DagRep.5.4.18

1 Executive Summary

Erik Rodner
Trevor Darrell
Marius Kloft
Massimiliano Pontil
Gunnar Rätsch

The seminar broadly dealt with *machine learning*, the area of computer science that concerns developing computational methods using data to make accurate predictions. The classical machine learning theory is built upon the assumption of independent and identically distributed random variables. In practical applications, however, this assumption is often violated,

Machine Learning with Interdependent and Non-identically Distributed Data, *Dagstuhl Reports*, Vol. 5, Issue 4, pp. 18–55
Editors: Trevor Darrell, Marius Kloft, Massimiliano Pontil, Gunnar Rätsch, and Erik Rodner

Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

for instance, when training and test data come from different distributions (dataset bias or domain shift) or when the data exhibits temporal or spatial correlations. In general, there are three major reasons why the assumption of independent and identically distributed data can be violated:

1. The draw of a data point influences the outcome of a subsequent draw (inter-dependencies).
2. The distribution changes at some point (non-stationarity).
3. The data is not generated by a distribution at all (adversarial).

The seminar focused on the scenarios (a) and (b). This general research direction comprises several subfields of machine learning: transfer and multi-task learning, learning with inter-dependent data, and two application fields, that is, visual recognition and computational biology. Both application areas are not only two of the main application areas for machine learning algorithms in general, but their recognition tasks are often characterized by multiple related learning problems that require transfer and multitask learning approaches. For example, in visual recognition tasks, object categories are often visually related or hierarchically organized, and tasks in computational biology are often characterized by different but related organisms and phenotypes. The problems and techniques discussed during the seminar are also important for other more general application areas, such as scientific data analysis or data-oriented decision making.

Results of the Seminar and Topics Discussed

In the following, the important research fields related to the seminar topic are introduced and we also give a short list of corresponding research questions discussed at the seminar. In contrast to other workshops and seminars often associated with larger conferences, the aim of the Dagstuhl seminar was to reflect on open issues in each of the individual research areas.

Foundations of Transfer Learning

Transfer Learning (TL) [2, 18] refers to the problem of retaining and applying the knowledge available for one or more source tasks, in order to efficiently develop an hypothesis for a new target task. Each task may contain common (domain adaptation [25, 10]) or different label sets (across category transfer). Most of the effort has been devoted to binary classification [23], while interesting practical transfer problems are often intrinsically multi-class and the number of classes can increase in time [17, 22]. Accordingly the following research questions arise:

- How to formalize knowledge transfer across multi-class tasks and provide theoretical guarantees on this setting?
- Moreover, can inter-class transfer and incremental class learning be properly integrated?
- Can learning guarantees be provided when the adaptation relies only on pre-trained source hypotheses without explicit access to the source samples, as it is often the case in real world scenarios?

Foundations of Multi-task Learning

Learning over multiple related tasks can outperform learning each task in isolation. This is the principal assertion of Multi-task learning (MTL) [3, 7, 1] and implies that the learning process may benefit from common information shared across the tasks. In the simplest case,

the transfer process is symmetric and all the tasks are considered as equally related and appropriate for joint training. Open questions in this area are:

- What happens when the condition of equally related tasks does not hold, e.g., how to avoid negative transfer?
- Moreover, can non-parametric statistics [27] be adequately integrated into the learning process to estimate and compare the distributions underlying the multiple tasks in order to learn the task similarity measure?
- Can recent semi-automatic methods, like deep learning [9] or multiple kernel learning [13, 12, 11, 4], help to get a step closer towards the complete automatization of multi-task learning, e.g., by learning the task similarity measure?
- How can insights and views of researcher be shared across domains (e.g., regarding the notation of *source task selection* in reinforcement learning)?

Foundations of Learning with Inter-dependent Data

Dependent data arises whenever there are inherent correlations in between observations. For example, this is to be expected for time series, where we would intuitively expect that instances with similar time stamps have stronger dependencies than ones that are far away in time. Another domain where dependent data occurs are spatially-indexed sequences, such as windows taken from DNA sequences. Most of the body of work on machine learning theory is on learning with i.i.d. data. Even the few analyses (e.g., [28]) allowing for "slight" violations of the assumption (mixing processes) analyze the same algorithms as in the i.i.d. case, while it should be clear that also novel algorithms are needed to most effectively adapt to rich dependency structures in the data. The following aspects have been discussed during the seminar:

- Can we develop algorithms that exploit rich dependency structures in the data?
- Do such algorithms enjoy theoretical generalization guarantees?
- Can such algorithms be phrased in a general framework in order to jointly analyze them?
- How can we appropriately measure the degree of inter-dependencies (theoretically) such that it can be also empirically estimated from data (overcoming the so-called *mixing* assumption)?
- Can theoretical bounds be obtained for more practical dependency measures than mixing?

Visual Transfer and Adaptation

Visual recognition tasks are one of the main applications for knowledge transfer and adaptation techniques. For instance, transfer learning can put to good use in the presence of visual categories with only a few number of labels, while across category transfer can help to exploit training data available for related categories to improve the recognition performance [14, 21, 20, 22]. Multi-task learning can be applied for learning multiple object detectors [30] or binary image classifiers [19] jointly, which is beneficial because visual features can be shared among categories and tasks. Another important topic is domain adaptation, which is very effective in object recognition applications [24], where the image distribution used for training (source domain) is different from the image distribution encountered during testing (target domain). This distribution shift is typically caused by a data collection bias. Sophisticated methods are needed as in general the visual domains can differ in a combination of (often unknown) factors including scene, object location and pose, viewing angle, resolution, motion blur, scene illumination, background clutter, camera characteristics, etc. Recent studies have demonstrated a significant degradation in the performance of state-of-the-art image

classifiers due to domain shift from pose changes [8], a shift from commercial to consumer video [5, 6, 10], and, more generally, training datasets biased by the way in which they were collected [29].

The following open questions have been discussed during the seminar:

- Which types of representations are suitable for transfer learning?
- How can we extend and update representations to avoid negative transfer?
- Are current adaptation and transfer learning methods efficient enough to allow for large-scale continuous visual learning and recognition?
- How can we exploit huge amounts of unlabeled data with certain dependencies to minimize supervision during learning and adaptation?
- Are deep learning methods already compensating for common domain changes in visual recognition applications?

Application Scenarios in Computational Biology

Non-i.i.d. data arises in biology, e.g., when transferring information from one organism to another or when learning from multiple organisms simultaneously [31]. A scenario where dependent data occurs is when extracting local features from genomic DNA by running a sliding window over a DNA sequence, which is a common approach to detect transcription start sites (TSS) [26]. Windows close by on the DNA strand – or even overlapping – show stronger dependencies than those far away. Another application scenario comes from statistical genetics. Many efforts in recent years focused on models to correct for population structure [16], which can arise from inter dependencies in the population under investigation. Correcting for such rich dependency structures is also a challenge in prediction problems in machine learning [15]. The seminar brought ideas together from the different fields of machine learning, statistical genetics, Bayesian probabilistic modeling, and frequentist statistics. In particular, we discussed the following open research questions:

- How can we empirically measure the degree of inter-dependencies, e.g., from a kinship matrix of patients?
- Do theoretical guarantees of algorithms (see above) break down for realistic values of "the degree of dependency"?
- What are effective prediction and learning algorithms correcting for population structure and inter-dependencies in general and can they be phrased in a general framework?
- What are adequate benchmarks to evaluate learning with non-i.i.d. data?
- How can information be transferred between organisms, taking into account the varying noise level and experimental conditions from which data are derived?
- How can non-stationarity be exploited in biological applications?
- What are promising applications of non-i.i.d. learning in the domains of bioinformatics and personalized medicine?

Conclusion

The idea of the seminar bringing together people from theory, algorithms, computer vision, and computational biology, was very successful, since many discussions and joint research questions came up that have not been anticipated in the beginning. These aspects were not completely limited to non-i.i.d. learning and also touched ubiquitous topics like learning with deeper architectures. It was the agreement of all participants that the seminar should be the beginning of an ongoing series of longer Dagstuhl seminars focused on non-i.i.d. learning.

References

1 Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

2 Jonathan Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

3 Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, July 1997.

4 C. Cortes, Marius Kloft, and M. Mohri. Learning kernels using local rademacher complexity. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, 2013. in press.

5 L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

6 L. Duan, D. Xu, I. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

7 Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117. ACM, 2004.

8 Ali Farhadi and Mostafa Kamali Tabrizi. Learning to recognize activities from the wrong view point. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.

9 Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, July 2006.

10 Judy Hoffman, Erik Rodner, Jeff Donahue, Trevor Darrell, and Kate Saenko. Efficient learning of domain-invariant image representations. In *International Conference on Learning Representations (ICLR)*, 2013.

11 Marius Kloft and Gilles Blanchard. On the convergence rate of ℓ_p-norm multiple kernel learning. *Journal of Machine Learning Research*, 13:2465–2502, Aug 2012.

12 Marius Kloft, U. Brefeld, S. Sonnenburg, and A. Zien. Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12:953–997, Mar 2011.

13 G. Lanckriet, N. Cristianini, L. E. Ghaoui, P. Bartlett, and M. I. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

14 Fei-Fei Li, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006.

15 Limin Li, Barbara Rakitsch, and Karsten M. Borgwardt. ccsvm: correcting support vector machines for confounding factors in biological data classification. *Bioinformatics [ISMB/ECCB]*, 27(13):342–348, 2011.

16 Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. FaST linear mixed models for genome-wide association studies. *Nat Meth*, 8(10):833–835, October 2011.

17 Jie Luo, Tatiana Tommasi, and Barbara Caputo. Multiclass transfer learning from unconstrained priors. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1863–1870, 2011.

18 Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

19 Ariadna Quattoni, Michael Collins, and Trevor Darrell. Transfer learning for image classification with sparse prototype representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8. IEEE, 2008.

20 Erik Rodner and Joachim Denzler. Learning with few examples by transferring feature relevance. In *Proceedings of the 31st Annual Symposium of the German Association for Pattern Recognition (DAGM)*, pages 252–261, 2009.

21 Erik Rodner and Joachim Denzler. One-shot learning of object categories using dependent gaussian processes. In *Proceedings of the 32nd Annual Symposium of the German Association for Pattern Recognition (DAGM)*, pages 232–241, 2010.

22 Erik Rodner and Joachim Denzler. Learning with few examples for binary and multiclass classification using regularization of randomized trees. *Pattern Recognition Letters*, 32(2):244–251, 2011.

23 Ulrich Rückert and Marius Kloft. Transfer learning with adaptive regularizers. In *ECML/PKDD (3)*, pages 65–80, 2011.

24 Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, pages 213–226, 2010.

25 Gabriele Schweikert, Christian Widmer, Bernhard Schölkopf, and Gunnar Rätsch. An empirical analysis of domain adaptation algorithms for genomic sequence analysis. In *Advances in Neural Information Processing Systems 21*, pages 1433–1440, 2009.

26 S. Sonnenburg, A. Zien, and G. Rätsch. Arts: Accurate recognition of transcription starts in human. *Bioinformatics*, 22(14):e472–e480, 2006.

27 Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

28 Ingo Steinwart, Don R. Hush, and Clint Scovel. Learning from dependent observations. *J. Multivariate Analysis*, 100(1):175–194, 2009.

29 Antonio Torralba and Alyosha Efros. Unbiased look at dataset bias. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

30 Antonio Torralba, Kevin P Murphy, and William T Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages II–762. IEEE, 2004.

31 C. Widmer, M. Kloft, and G. Rätsch. Multi-task learning for computational biology: Overview and outlook. In *B. Schoelkopf, Z. Luo, and V. Vovk, editors, Empirical Inference – Festschrift in Honor of Vladimir N. Vapnik*, 2013.

2 Table of Contents

3 Overview of Talks

3.1 Transfer Learning using Marginal Distribution Information

Gilles Blanchard (University of Potsdam, DE)

Consider a setting where a large number N of labeled training samples $S_i := (X_{ij}, Y_{ij})_{1 \leq j \leq n_i}$ $(i = 1, \ldots, N)$ on $\mathcal{X} \times \mathcal{Y}$ are available. The primary goal is not to find an adequate classification (or regression) function for each of these samples, but rather to find an appropriate prediction function $f : \mathcal{X} \times \mathcal{Y}$ for a *new, unlabeled* test sample $S^T := (X_j^T)_{1 \leq j \leq n^T}$. Such a situation occurs, for instance, for the *automatic gating* problem for flow cytometry data, a high-throughput measurement platform that is an important clinical tool for the diagnosis of many blood-related pathologies. The index i indicates a particular patient; for each patient a blood sample is taken, and measured by the device. This blood sample contains n_i invidual cells – potentially several dozens of thousands – each of which is separately analyzed by the device, giving rise to a feature vector X_{ij} of attributes related to physical and chemical properties of the individual cell. The label Y_{ij}, input manually by an expert, gives the type of each cell (blood cell, white cell, etc.). The goal is to make this last labelling (or "gating") step automatic, using the available labeled data. Note that in this case, for a new test patient zero label information is available, only the feature vectors of the cells present in the blood sample.

This problem belongs to the vast landscape of transfer learning. A classical approach to the problem (the covariate shift setting) assumes that the marginal distribution $P_X^{(i)}$ differs between samples, but that the conditional $P_{Y|X}$ stays the same. This is a very strong assumption that we want to avoid. We propose the following alternative Ansatz: there is a relationship common to all samples from the marginal $P_X^{(i)}$ to the conditional $P_{Y|X}^{(i)}$. Thus, we posit that there is some pattern making it possible to learn a mapping from marginal distributions to labels. We call this setting *marginal predictor learning*. In other words, we want to learn a mapping

$$f : \mathfrak{P}_{\mathcal{X}} \times \mathcal{X} \to \mathbb{R}\,,$$

(where $\mathfrak{P}_{\mathcal{X}}$ denotes the set of marginal distributions on $\mathcal{X}$) which, for a new unlabeled sample with corresponding empirical marginal distribution $\widehat{P}_X^T$, will predict the label $f(\widehat{P}_X^T, x)$ for a specific feature vector x belonging to that sample.

We show that this setting is amenable to a reproducing kernel learning method. The gist of our approach is to combine recent developments about kernels on distributions (Christmann and Steinwart 2010, Sriperumbudur et al. 2010) with ideas of kernel multitask learning (Evgeniou and Pontil 2005). In a nutshell, the abstract "kernel task similarity matrix" present in kernel multitask learning is replaced by a task similarity matrix determined by the similarity between empirical marginal distributions, as measured by a distribution kernel. We show in particular that this approach is universally consistent under weak assumptions, is practically applicable and can outperform other approaches.

References

1 G. Blanchard, G. Lee, C. Scott. Generalizing from Several Related Classification Tasks to a New Unlabeled Sample. In *NIPS*, 2178-2186, 2011.
2 A. Christmann and I. Steinwart. Universal kernels on non-standard input spaces. In *NIPS*, 406–414, 2010.

3 T. Evgeniou and M. Pontil. Learning multiple tasks with kernel methods. In *JMLR* (6), 615–637, 2005.
4 B. Sriperumbudur, A. Gretton, K. Fukumizu, B. Schölkopf, and G. Lanckriet, Hilbert space embeddings and metrics on probability measures. In *JMLR* (11), 1517–1561, 2010.

3.2 Non-i.i.d. Deep Learning

Trevor Darrell (UC Berkeley, US), Kate Saenko (UMass Lowell, US), Judy Hoffman (UC Berkeley, US)

LSDA: Detection as Domain Adaptation

One of the fundamental challenges in training object detection systems is the need to collect a large of amount of images with bounding box annotations. The introduction of detection challenge datasets, such as PASCAL VOC [9], have propelled progress by providing the research community a dataset with enough fully annotated images to train competitive models although only for 20 classes. Even though the more recent ImageNet detection challenge dataset [3] has extended the set of annotated images, it only contains data for 200 categories. As we look forward towards the goal of scaling our systems to human-level category detection, it becomes impractical to collect a large quantity of bounding box labels for tens or hundreds of thousands of categories.

We ask, is there something generic in the transformation from classification to detection that can be learned on a subset of categories and then transferred to other classifiers? We cast this task as a domain adaptation problem, considering the data used to train classifiers (images with category labels) as our source domain, and the data used to train detectors (images with bounding boxes and category labels) as our target domain. We then seek to find a general transformation from the source domain to the target domain, that can be applied to any future classifier to adapt it into a detector.

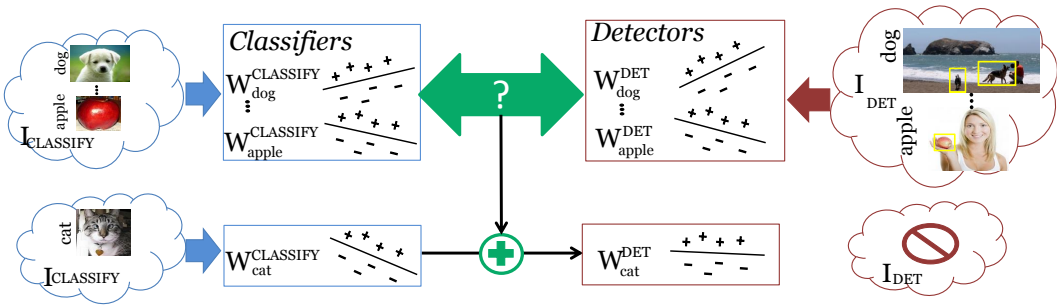

Figure 1 The core idea is that we can learn detectors (weights) from labeled classification data (left), for a wide range of classes. For some of these classes (top) we also have detection labels (right), and can learn detectors. But what can we do about the classes with classification data but no detection data (bottom)? Can we learn something from the paired relationships for the classes for which we have both classifiers and detectors, and transfer that to the classifier at the bottom to make it into a detector?

We have already released a 7.6K visually grounded lexicon comprised of detectors adapted from ImageNet classifiers, available at https://github.com/jhoffman/lsda. Our model is based

on a technique we call Large Scale Detection through Adaptation (LSDA), an algorithm that learns to transform an image classifier into an object detector [15]. To accomplish this goal, we use supervised convolutional neural networks (CNNs), which have recently been shown to perform well both for image classification [18] and object detection [10, 21]. We have recently extended this model to also solve a latent variable task to identify inlier visual regions, further improving learning from images of complex scenes [16].

In the future, we will extend this model beyond its present formulation based on WordNet to include similar concepts which can be learned from static imagery, including adjectives, and to incorporate motion representations for learning verbs. E.g., we hope to provide groundings similar to that in the Columbia "adjective noun pairs" dataset of [6], but integrated into the LSDA detector framework.

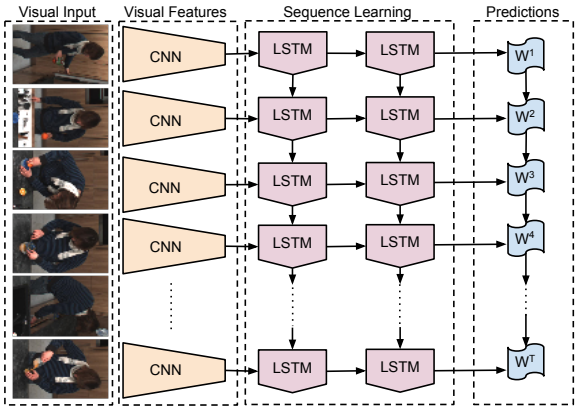

Figure 2 We introduced *Long-term Recurrent Convolutional Networks* (LRCNs), a class of architectures leveraging the strengths of rapid progress in CNNs for visual recognition problem, and the growing desire to apply such models to time-varying inputs and outputs. This enables learning from images and videos with only weak labels in the form of tags or captions. LRCN processes the (possibly) variable-length visual input (left) with a CNN (middle-left), whose outputs are fed into a stack of recurrent sequence models (*LSTMs*, middle-right), which finally produce a variable-length prediction (right). Please see goo.gl/cZRM4U for example output sentences.

LCRN: Weak learning from images, videos, and captions

Image data collection for individual concepts may have reached a plateau in productivity, and we predict stronger models will result from models which leverage images and text in context, with only indirect labeling. Learning models from images or videos and associated captions or descriptive text is an especially appealing method for grounding elementary units in perceptual experience, as the system learns how to align image and textual content without explicit supervision.

Recognition and description of images and videos is a fundamental challenge of computer vision. Dramatic progress has been achieved by supervised convolutional models on image recognition tasks, and a number of extensions to process video have been recently proposed. Ideally, a video model should allow processing of variable length input sequences, and also provide for variable length outputs, including generation of full-length sentence descriptions that go beyond conventional one-versus-all prediction tasks. We have produced *long-term recurrent convolutional networks* (LRCNs), a novel architecture for visual recognition and

description which combines convolutional layers and long-range temporal recursion and is end-to-end trainable (see Figure 2).

We have instantiated our architecture for specific video activity recognition, image caption generation, and video description tasks. We have shown that long-term recurrent convolutional models are generally applicable to visual time-series modeling and that these models improve generation of descriptions from intermediate visual representations derived from conventional visual models. We instantiate our proposed architecture in three experimental settings. First, we show that directly connecting a visual convolutional model to deep LSTM networks, we are able to train video recognition models that capture complex temporal state dependencies. While existing labeled video activity datasets may not have actions or activities with extremely complex time dynamics, we nonetheless see improvements on the order of 4% on conventional benchmarks. and importantly enable direct end-to-end trainable image-to-sentence mappings. Strong results for machine translation tasks have recently been reported [22, 7]; such models are encoder/decoder pairs based on LSTM networks. Our multimodal architecture consists of a visual CNN to encode a deep state vector and an LSTM to decode the vector into an natural language string. This model can be trained end-to-end on large-scale image and text datasets, and even with modest training provides competitive generation results compared to existing methods.

To date, there has only been limited investigation of what has been learned in these models, and little systematic exploration of how such knowledge can be extracted and leveraged in related tasks. Anecdotal results suggest that the LCRN model does learn how to localize specific noun phrases and can learn to ground complex and/or idiosyncratic terms.

We propose to combine the variable input weak learning model with our large scale detection through adaptation approach to create models that not only produce captions and descriptions for novel videos/images, but are also able to localize the salient nouns and verbs. This will enable interactive applications and provide an intuitive medium through which to communicate with users.

Towards Deep Confusion

The methods proposed above presume a (possibly weakly labeled) supervised learning regime, with test and training data coming from the same domain. It is a widely recognized phenomenon that models trained in one environment, even with large data sources, suffer from degraded performance when deployed in a new or specialized environment. For example, a model trained on web search images may not perform very well for recognition on a robot mounted camera in a warehouse or office environment. In order for our large scale models to be widely applicable, we will develop algorithms that quickly adapt to new scenarios without the expensive overhead of collecting new labeled data and retraining a model from scratch.

Dataset bias is a well known and theoretically understood problem with traditional supervised approaches to image recognition [23]. A number of recent theoretical and empirical results have shown that supervised methods' test error increases in proportion to the difference between the test and training input distribution [2, 4, 20, 23]. In the last few years, several methods for visual domain adaptation have been suggested to overcome this issue [8, 24, 1, 20, 19, 17, 12, 11, 13, 14], but were limited to shallow models. The traditional approach to adapting deep models has been fine-tuning; see [10] for a recent example.

We propose a new CNN architecture, outlined in Figure 3, which uses an adaptation layer along with a domain confusion loss based on maximum mean discrepancy (MMD) [5] to automatically learn a representation jointly trained to optimize for classification and domain invariance. Our domain confusion metric can be used both to select the dimension

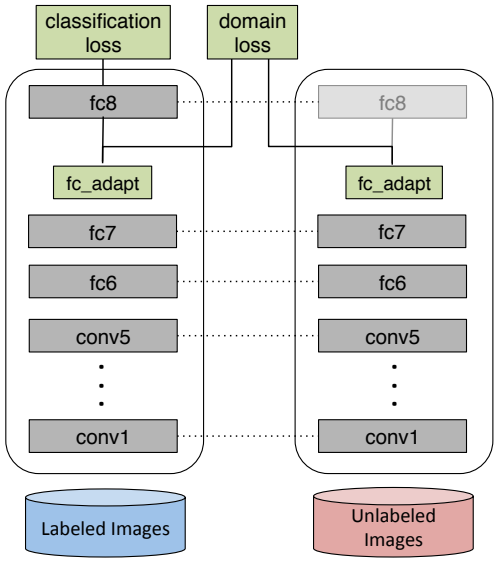

Figure 3 Our architecture optimizes a deep CNN for both classification performance and domain invariance. The model can be trained for *supervised* adaptation, when there is a small number of target labels available, or *unsupervised* adaptation, when no target labels are available. We introduce domain invariance through *domain confusion* guided selection of the depth and width of the adaptation layer, as well as an additional loss term during fine-tuning that directly minimizes the distance between source and target representations.

of the adaptation layers, choose an effective placement for a new adaptation layer within a pre-trained CNN architecture, and fine-tune the representation. Our architecture can be used to solve both *supervised adaptation*, when a small amount of target labeled data is available, and *unsupervised adaptation*, when no labeled target training data is available.

References

1 Y. Aytar and A. Zisserman. Tabula rasa: Model transfer for object category detection. In *Proc. ICCV*, 2011.
2 Shai Ben-David, John Blitzer, Koby Crammer, Fernando Pereira, et al. Analysis of representations for domain adaptation. *Proc. NIPS*, 2007.
3 A. Berg, J. Deng, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. 2012.
4 John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Proc. NIPS*, 2007.
5 Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *Bioinformatics*, 2006.
6 Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232. ACM, 2013.
7 Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.
8 H. Daumé III. Frustratingly easy domain adaptation. In *ACL*, 2007.

9 M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

10 R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *In Proc. CVPR*, 2014.

11 B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Proc. CVPR*, 2012.

12 R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proc. ICCV*, 2011.

13 J. Hoffman, B. Kulis, T. Darrell, and K. Saenko. Discovering latent domains for multisource domain adaptation. In *Proc. ECCV*, 2012.

14 J. Hoffman, E. Rodner, J. Donahue, K. Saenko, and T. Darrell. Efficient learning of domain-invariant image representations. In *Proc. ICLR*, 2013.

15 Judy Hoffman, Sergio Guadarrama, Eric S Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. Lsda: Large scale detection through adaptation. In *Advances in Neural Information Processing Systems*, pages 3536–3544, 2014.

16 Judy Hoffman, Deepak Pathak, Trevor Darrell, and Kate Saenko. Detector discovery in the wild: Joint multiple instance and representation learning, 2015.

17 A. Khosla, T. Zhou, T. Malisiewicz, A. Efros, and A. Torralba. Undoing the damage of dataset bias. In *Proc. ECCV*, 2012.

18 A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proc. NIPS*, 2012.

19 B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011.

20 K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010.

21 P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

22 Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

23 A. Torralba and A. Efros. Unbiased look at dataset bias. In *Proc. CVPR*, 2011.

24 J. Yang, R. Yan, and A. Hauptmann. Adapting SVM classifiers to data with shifted distributions. In *ICDM Workshops*, 2007.

3.3 Computer Vision to Support Decision Making in Ecology

Joachim Denzler (Friedrich Schiller University Jena, DE)

Ecology is the study of life and its interaction with the physical environment. Scientists are interested in quantifying relations between atmospheric, oceanic, and terrestrial processes. For a long time, analysis has been done locally both with respect to the region of investigation as well as with respect to the field in which phenomena are studied. Due to the possibilities to record data all over the world, the increase in resolution, the quality of recordings from satellites, distributions of data sets over the world wide web, and computing in the cloud new opportunities arise. Such heterogenous and globally collected data may make it possible to answer questions that are of fundamental importance for the future of our planet.

In this research domain computer vision can play an important role in the future. Today, most work by researchers in ecology is done by analyzing data manually. For example, the number of butterflies in a certain region is determined by visual inspection of traps installed in the environment.

Over the last years, computer vision research already tackled problems that are of high relevance for ecology as well. One example is the automatic analysis of remote sensing data. A second example is the identification of animals from images and videos. Birds, dogs, mushrooms, flowers build databases for object recognition benchmarks, since those objects not just offer very challenging problems but also call for new methods, that lead to the area of fine-grained recognition. Works directly related to ecology are for example, the classification of insects [1], or computer vision methods for coral reef assessment [2].

One hypothesis of our research is that computer vision methods can only be accepted and successful in ecology, if we are able to exploit all knowledge (labeled data from similar domains, common feature representations, etc.) already available, to incrementally improve performance, and to keep the human in the loop, for example, to check of correct automatic decision. This allows to build automatic systems with minimal user efforts – a preliminary, if researchers from other disciplines shall accept modern techniques from computer vision for their research. Domain adaptation and transfer learning will play one key role to success.

When working together with people from ecology and biodiversity research, specific problems arise that must be solved from the computer vision and machine learning perspective:

1. can we configure initial classifiers for ecology applications using already existing data bases or images from the internet?
2. can we adapt existing classifiers using a minimal set of training data from a specific application scenario, to reduce the effort by researchers from ecology?
3. can the process of domain adaptation be supported by the human in the loop, for example, to embed it into a life-long learning scenario?
4. can we exploit data from additional modalities besides visual data to support transfer learning in the visual domain?
5. are there common principles in transfer learning that can also be applied to analyse dynamic processes, for example, the interactions between animals – with special focus on behaviour changing over time
6. can we benefit from the huge amount of data that will be collected in the future, and are existing methods from machine learning already capable to deal with streams of input data for model update

The Computer Vision Group Jena aims at life-long learning scenarios, including large scale visual learning and recognition [3], active learning [4, 5], novelty detection [6, 7, 8], incremental learning [9], and fine-grained recognition [10, 11]. For dynamic scene analysis, computer vision in sensor networks has been one goal during the past years as well, with the focus on supervised and unsupervised activity recognition [12, 13]. Applications so far came from biology (unsupervised mytosis detection [14]) and medicine (classification of facial paralysis [15]). In the later case we investigated domain adaptation for active appearnance models [16].

The Computer Vision Group, headed by Joachim Denzler, consists of two senior researchers, Erik Rodner and Wolfgang Ortmann, and currently 12 PhD students. Joachim is also faculty member of the Abbe-School of Photonics and the International Max-Planck-Research School for Global Biochemical Cycles. He is co-founder of the Michael Stifel Center for Data-Driven and Simulation Science Jena.

References

1 N. Larios, B. Soran: Haar random forest features and svm spatial matching kernel for stonefly species identification. In: International Conference on Pattern Recognition. (2010)
2 Beijbom, O., Edmunds, P.J., Kline, D.I., Mitchell, B.G., Kriegman, D.: Automated annotation of coral reef survey images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, Rhode Island (2012)
3 Fröhlich, B., Rodner, E., Kemmler, M., Denzler, J.: Large-scale gaussian process multi-class classification for semantic segmentation and facade recognition. Machine Vision and Applications **24**(5) (2013) 1043–1053
4 Käding, C., Freytag, A., Rodner, E., Bodesheim, P., Denzler, J.: Active learning and discovery of object categories in the presence of unnameable instances. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2015)
5 Freytag, A., Rodner, E., Denzler, J.: Selecting influential examples: Active learning with expected model output changes. In: European Conference on Computer Vision (ECCV). Volume 8692. (2014) 562–577
6 Bodesheim, P., Rodner, E., Freytag, A., Denzler, J.: Divergence-based one-class classification using gaussian processes. In: British Machine Vision Conference (BMVC). (2012) 50.1–50.11
7 Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M., Denzler, J.: Kernel null space methods for novelty detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2013)
8 Bodesheim, P., Freytag, A., Rodner, E., Denzler, J.: Local novelty detection in multi-class recognition problems. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2015) 813–820
9 Lütz, A., Rodner, E., Denzler, J.: I want to know more: Efficient multi-class incremental learning using gaussian processes. Pattern Recognition and Image Analysis **23**(3) (2013) 402–407
10 Simon, M., Rodner, E., Denzler, J.: Fine-grained classification of identity document types with only one example. In: Machine Vision Applications (MVA). (2015)
11 Göring, C., Rodner, E., Freytag, A., Denzler, J.: Nonparametric part transfer for fine-grained recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 2489–2496
12 Krishna, M.V., Denzler, J.: A combination of generative and discriminative models for fast unsupervised activity recognition from traffic scene videos. In: IEEE Winter Conference on Applications of Computer Vision (WACV). (2014) 640–645
13 Körner, M., Denzler, J.: Temporal self-similarity for appearance-based action recognition in multi-view setups. In: Computer Analysis of Images and Patterns. Volume 8047. (2013) 163–171
14 Krishna, M.V., Denzler, J.: A hierarchical bayesian approach for unsupervised cell phenotype clustering. In: German Conference on Pattern Recognition (GCPR). (2014) 69- -80
15 Haase, D., Kemmler, M., Guntinas-Lichius, O., Denzler, J.: Efficient measuring of facial action unit activation intensities using active appearance models. In: IAPR International Conference on Machine Vision Applications (MVA). (2013) 141–144
16 Haase, D., Rodner, E., Denzler, J.: Instance-weighted transfer learning of active appearance models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 1426–1433

3.4 Reproducing Kernel Hilbert Space Embeddings in Computational Biology

Philipp Drewe (Max-Delbrück-Centrum, DE)

Joint work of Drewe, Philipp; Stegle, Oliver; Hartmann, Lisa; Kahles, André; Bohnert, Regina, Wachter, Andreas, Borgwardt, Karsten; Rätsch, Gunnar
Main reference P. Drewe, O. Stegle, L. Hartmann, A. Kahles, R. Bohnert, A. Wachter, K. Borgwardt, G. Rätsch, "Accurate detection of differential RNA processing," Nucleic Acids Research, 41(10):5189–5198, 2013.
URL http://dx.doi.org/10.1093/nar/gkt211

A fundamental problem in computational biology is identifying genes in a cell that are processed differently upon perturbation of the cell. However, this is challenging as the processing of the genes cannot be directly measured, but has to be inferred from a set of incomplete observations (reads) of the genes. These reads are high-dimensional, structured and typically non-iid distributed. Therefore, classical statistical test, such as the Kolmogorov-Smirnov test, cannot be applied in this setting. In this work, we show that Reproducing Kernel Hilbert Space (RKHS) embeddings allow a suitable representation of read-data. Furthermore, we present RKHS-embedding-based approaches to test for homogeneity of two sets of observations, in order to accurately identify genes whose processing has changed.

3.5 Bridging the Gap Between Synthetic and Real Data

Mario Fritz (MPI für Informatik – Saarbrücken, DE)

There is a long tradition of using generative models in combination with discriminative classifiers [5, 6, 7]. Equally the recently successful deep learning technique [3] use jittering techniques [1, 2] that imply sampling from an underlying distribution. Although in both cases the the model is postulated and all parameters are in our control, we rarely achieve an accurate representation of the true underlying distribution. Yet, these techniques have shown improved performance as learning is guided by prior knowledge encoded in such generative models.

Learning and Prediction from Rendered/Synthesized Data

Many applications greatly benefit by means of synthesizing additional training data. For visual recognition this often involves a rendering process for creating new images. The employed model represents prior knowledge about the target domain. In this section, several examples are listed where we have directly used the rendered data – assuming that the domain mismatch between real and virtual examples is negligible.

Detection by Rendering. In early work, we have captured a light-field of an object and rendered new views of the object on demand in order to evaluate the posterior in a particle filter tracking framework [8].

New View Synthesis. Human generalize easily from a single view of an object to novel view-points. Today's computer vision algorithms are mostly learning and example based and therefore have to be shown variations across style and viewpoints in order to succeed. We

have presented an approach that uses a 3D model to guide novel view synthesis, that is able to fill in disocculsion areas truthfully [9]. The object models trained on such augmented data show a greatly improved view point generalization.

Differentiable Vision Pipeline. Most recently, we have established a fully differentiable vision pipeline [10] that builds on top of an approximately differentiable renderer [4] and a differentiated HOG image representation. This allows us to estimate object poses by exploiting the prescribed image synthesis procedure in the gradient computation.

Adaptation to Rendered/Synthesized Data

Although significant progress has been achieved by solely relying on realistic rendering and synthesis, quite often the domain shift between the virtual and the real world introduces a distribution mismatch that should be treated separately.

Visual Domain Adaptation via Metric Learning. We have proposed to reduce the effects of domain shifts by a metric learning formulation [11]. Hereby we have improved recognition across different data sources such a webcam, dslr or data from the web.

Recognition from Virtual Examples. We have employed the concept of metric learning for domain adaptation to the problem of visual material recognition [12]. The approach helps to bridge the gap between rendered and real data.

Prediction under changing prior distribution. Most recently, we have have shown how to perform gaze estimation in the wild [13]. Considering the change in the prior distribution of head pose and eye fixation distribution has been critical when training across datasets.

Unsupervised Adaptation

Future challenges include scenarios where no training data for adaptation is available. Less work has been performed in this direction. We have proposed to adapt to new conditions in a road segmentation task by assuming a stationary, structured prior over the label space, which allows us to successfully adapt a semantic labeler to unseen weather conditions [14]. Beyond the traditional recognition scenarios, we have also attempted to bring the required adaptivity to learning settings. E.g. we have adapted active learning strategies via reinforcement learning to different training distributions [15]. We hypothesize that non-parametric learning techniques for visual recognition and grouping [16] can be well suited to transfer structural relations across domains, while being less affected by changes in individual appearances.

References

1 P. Simard, B. Victorri, Y. LeCun, J. Denker. Tangent prop-a formalism for specifying selected invariances in an adaptive network. In *Advances in neural information processing systems (NIPS)*, 1992
2 D. Decoste, B. Schölkopf. Training invariant support vector machines. In *Journal of Machine Learning*, 2002
3 A. Krizhevsky, I. Sutskever, G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.
4 M. Loper, M. Black. Opendr: An approximate differentiable renderer. In *European Conference on Computer Vision (ECCV)*, 2014
5 T. Jaakkola, D. Haussler. Exploiting generative models in discriminative classifiers. In *Advances in neural information processing systems (NIPS)*, 1999

6 M. Fritz, B. Leibe, B. Caputo, B. Schiele. Integrating representative and discriminant models for object category detection. In *IEEE International Conference on Computer Vision (ICCV)*, 2005

7 A. Holub, M. Welling, P. Perona. Combining generative models and fisher kernels for object recognition. In *IEEE International Conference on Computer Vision (CVPR)*, 2005

8 M. Zobel, M. Fritz, and I. Scholz. Object tracking and pose estimation using light-field object models. In *Vision, Modeling, and Visualization Conference (VMV)*, 2002.

9 K. Rematas, T. Ritschel, M. Fritz, and T. Tuytelaars. Image-based synthesis and re-synthesis of viewpoints guided by 3d models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

10 W.-C. Chiu and M. Fritz. See the difference: Direct pre-image reconstruction and pose estimation by differentiating hog. *arXiv:1505.00663 [cs.CV]*, 2015.

11 K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *European Conference on Computer Vision (ECCV)*, 2010.

12 W. Li and M. Fritz. Recognizing materials from virtual examples. In *European Conference on Computer Vision (ECCV)*, 2012.

13 X. Zhang, Y. Sugano, M. Fritz, and A. Bulling. Appearance-based gaze estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

14 E. Levinkov and M. Fritz. Sequential bayesian model update under structured scene prior for semantic road scenes labeling. In *IEEE International Conference on Computer Vision (ICCV)*, 2013.

15 S. Ebert, M. Fritz, B. Schiele. Ralf: A reinforced active learning formulation for object class recognition In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

16 W.-C. Chiu, M. Fritz. Multi-class video co-segmentation with a generative multi-video model. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013

3.6 On the Need of Theory and Algorithms Correcting for Confouding Factors

Marius Kloft (HU Berlin, DE)

A classic assumption in machine learning states that the data is independently realized from an unknown distribution. This assumptions greatly simplifies theory [1] and algorithms [2]. However, it is common in several applications that the data exhibit dependencies and inherent correlations between observations. Clearly, this occurs especially for time series, for instance, in network security (e.g., HTTP requests) and computer vision (video streams). Under the assumption of time-structured dependencies, several algorithms and theory have been proposed [3]. But few theory and algorithms have been developed for complexer dependencies, in particular for confounding ones.

For instance in statistical genetics, it is one of the central challenges to detect – among ten thousands of genes – the ones that are strong predictors of complex diseases or other binary outcomes [4, 5], as it is a first step in identifying regulatory components controlling heritability. However, for various diseases such as type 2 diabetes [6], these sparse signals are yet largely undetected, which is why these missing associations have been entitled the *The Dark Matter of Genomic Associations* [7]. Central problems include that these signals are

often very weak, and the found signals can be spurious due to confounding. Confounding can stem from varying experimental conditions and demographics such as age, ethnicity, gender [8], and – crucially – population structure, which is due to the relatedness between the samples [9, 8, 10]. Ignoring such confounders can often lead to spurious false positive findings that cannot be replicated on independent data [11]. Correcting for such confounding dependencies is considered one of the greatest challenges in statistical genetics [12]. Another example is content- and anomaly-based network intrusion detection and malware detection, where attacks are recorded within sandboxes [13]. Thus attributes that are specific to sandboxes help in discriminating attacks from benign data so that these attributes may be falsely promoted by the learning algorithm.

In the present Dagstuhl workshop, we found that there is a lack of research in the above respect. Which is why we advocate to develop theory and algorithms learning and estimation in the presence of confounding, the basic aim of which would be to understand and create statistical machine learning from confounded data. In particular, the following open problems arose at the workshop:

- How can we quantify "confoundedness" in learning settings?
- Can we develop theory similar to uniform convergence kind of analyses [1] under the assumption of confounders? And in order for this to work which assumptions do we need to state?
- How to design effective learning algorithms in presence of confounding and dependent labels?
- How to address feature selection under confounders?
- How to automatically learn the confounders?

Addressing the above stated open questions will subject to interesting future work. A good starting point to this end will be previous theoretical analyses regarding time series [3, 14] and probabilistic models such as the probit regression model [15, 16], and its extentions to GP classification [17, 18] and generalized linear mixed models [19].

Acknowledgments. MK acknowledges support by the German Research Foundation (DFG) under KL 2698/2-1.

References

1 V. N. Vapnik and A. Y. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities," *Theory of Probability and its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

2 C. Cortes and V. Vapnik, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

3 M. Mohri and A. Rostamizadeh, "Rademacher complexity bounds for non-i.i.d. processes," in *NIPS*, pp. 1097–1104, 2008.

4 T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, *et al.*, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–753, 2009.

5 S. Vattikuti, J. J. Lee, C. C. Chang, S. D. Hsu, and C. C. Chow, "Applying compressed sensing to genome-wide association studies," *GigaScience*, vol. 3, no. 1, p. 10, 2014.

6 N. Craddock, M. E. Hurles, N. Cardin, *et al.*, "Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls," *Nature*, vol. 464, no. 7289, pp. 713–720, 2010.

7 T. N. H. G. R. Institute, "Proceedings of the workshop on the dark matter of genomic associations with complex diseases: Explaining the unexplained heritability from genome-wide association studies," 2009.

8 L. Li, B. Rakitsch, and K. M. Borgwardt, "ccsvm: correcting support vector machines for confounding factors in biological data classification," *Bioinformatics*, vol. 27, no. 13, pp. 342–348, 2011.

9 C. Lippert, J. Listgarten, Y. Liu, C. Kadie, R. Davidson, and D. Heckerman, "Fast linear mixed models for genome-wide association studies," *Nature Methods*, vol. 8, pp. 833–835, October 2011.

10 N. Fusi, O. Stegle, and N. D. Lawrence, "Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical studies," *PLoS comp. bio.*, vol. 8, no. 1, 2012.

11 P. Kraft, E. Zeggini, and J. P. Ioannidis, "Replication in genome-wide association studies," *Statistical Science: A review journal of the Institute of Mathematical Statistics*, vol. 24, no. 4, p. 561, 2009.

12 B. J. Vilhjálmsson and M. Nordborg, "The nature of confounding in genome-wide association studies," *Nature Reviews Genetics*, vol. 14, no. 1, pp. 1–2, 2013.

13 D. Arp, M. Spreitzenbarth, M. Hübner, H. Gascon, K. Rieck, and C. Siemens, "Drebin: Effective and explainable detection of android malware in your pocket," in *Proc. of NDSS*, 2014.

14 I. Steinwart, D. R. Hush, and C. Scovel, "Learning from dependent observations," *J. Multivariate Analysis*, vol. 100, no. 1, pp. 175–194, 2009.

15 C. I. Bliss, "The method of probits," *Science*, vol. 79, no. 2037, pp. 38–39, 1934.

16 L. Fahrmeir, T. Kneib, S. Lang, and B. Marx, *Regression*. Springer, 2013.

17 C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. Cambridge, MA, USA: MIT Press, 2006.

18 J. P. Cunningham, P. Hennig, and S. Lacoste-Julien, "Gaussian probabilities and expectation propagation," *arXiv preprint arXiv:1111.6832*, 2011.

19 N. E. Breslow and D. G. Clayton, "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 9–25, 1993.

3.7 Transfer Learning in Computer Vision

Christoph H. Lampert (IST Austria – Klosterneuburg, AT)

Joint work of Lampert, Christoph H.; Pentina, Anastatia; Sharmanska, Viktoriia
Main reference A. Pentina, C. H. Lampert, "A PAC-Bayesian Bound for Lifelong Learning," in Proc. of the 31th Int'l Conf. on Machine Learning (ICML'14), pp. 991–999, JMLR.org, 2014.
URL http://jmlr.org/proceedings/papers/v32/pentina14.html
URL http://pub.ist.ac.at/~chl/erc

Computer Vision offers a wide range of problems where transfer learning techniques, such as domain adaptation and multi-task learning, can be applied. Several such techniques have proven useful in practice, but a solid theoretical understanding of when and how transfer learning offer benefits for computer vision tasks is still lacking. In my research group at IST Austria, we are particularly interested in the problem of lifelong learning. A lifelong learner continuously and autonomously learns from a stream of data, potentially for years or decades [1, 2]. During this time the learner should build an ever-improving base of generic information, and use this as background knowledge and context for solving different tasks. Using PAC-Bayesian learning theory, we have developed theoretic foundations that allow us to study different lifelong learning situations [3]. The generalization bounds that we obtain consist only of computable quantities and can therefore be used to analyze existing

lifelong learning algorithms and derive new ones. Similar techniques also allow the analysis of algorithms for sequential multi-task learning [4].

Acknowledgments. The described work was funded by the European Research Council under the European Unions Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no 308036.

References

1 S. Thrun and T. M. Mitchell. Lifelong robot learning. *Robotics and autonomous systems*, 15(1):25–46, 1995.
2 J. Baxter. A model of inductive bias learning. *Journal of Artificial Intelligence Research* , 12:149–198, 2000.
3 A. Pentina and C. H. Lampert. A PAC-Bayesian bound for lifelong learning. In *International Conference on Machine Learing (ICML)*, 2014.
4 A. Pentina, V. Sharmanska and C. H. Lampert. Curriculum Learning of Multiple Tasks. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

3.8 Optimization for Machine Learning – Made Easy yet Efficient

Soeren Laue (Friedrich Schiller University Jena, DE)

Many machine learning problems are cast as continuous optimization problems. A non-exhaustive list of such problems includes support vector machines [2], elastic nets [8], dimension reduction [1], and sparse PCA [9]. Moreover, for a given machine learning problem there is typically not only a single formulation as an optimization problem but different formulations that, for example, take previous knowledge or constraints into account. In the case of support vector machines the original formulation uses an ℓ_2-regularization term combined with the hinge loss. Different variants include the use of different loss functions, e.g., an ℓ_2-loss term for adapting to Gaussian noise, ℓ_1-regularization to obtain sparse predictors [7], or a combination of ℓ_1- and ℓ_2-regularization. Adding to this already large variety is the use of kernels in many of the problem formulations. However, up to this day, efficient solutions to any of these formulations still require the implementation of specialized, and highly-tuned solvers, not only in the case of support vector machines but for almost any machine learning problem that has been formulated as an optimization problem. This of course poses a problem when dealing with data sets whose size is well beyond the reach of easy to use modeling languages combined with a generic solver.

We present a novel approach to mitigate this problem by tightly coupling the modeling language and the generic solver. This results in code that is a few orders of magnitude more efficient than state-of-the-art modeling language/generic solver combinations like CVX/Gurobi [3, 4, 5] and CVX/Mosek [3, 4, 6]. The tight coupling is achieved by a generative programming approach that generates an individual solver for each problem as an instance of a generic solver. The generic optimizer is able to solve almost any continuous optimization problem with constraints over $\mathbb{R}^n$ that has been proposed for machine learning tasks. It combines the ease of use of commonly used modeling languages with the efficiency of highly-tuned, specialized state-of-the-art solvers for the individual machine learning problems. In the end, the automatically generated solver can be either deployed as a callable library or as a stand-alone solver.

References

1 Christopher J. C. Burges. Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning*, 2(4), 2010.
2 Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.
3 CVX Research, Inc. CVX: Matlab software for disciplined convex programming, version 2.0. http://cvxr.com/cvx, August 2012.
4 Michael Grant and Stephen Boyd. Graph Implementations for Nonsmooth Convex Programs. In *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag, 2008.
5 Gurobi Optimization, Inc. Gurobi Optimizer Reference Manual, 2013.
6 MOSEK ApS. The MOSEK Optimization Software, 2013.
7 Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, pages 267–288, 1996.
8 Hui Zou and Trevor Hastie. Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, pages 301–320, 2005.
9 Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15(2):265–286, 2006.

3.9 Transfer and Multi-Task Learning in Reinforcement Learning

Alessandro Lazaric (INRIA Lille, FR)

The Context

Reinforcement learning's (RL) [5, 1] challenging objective is to develop autonomous agents able to learn how to act optimally in an unknown and uncertain environment by trial-and-error and with limited level of supervision (i.e., a reinforcement signal). RL is mostly applied in domains where a precise formalization of the environment and/or the efficient computation of the optimal control policy is particularly difficult (e.g., robotics, human-computer interaction, recommendation systems). An RL problem is formalized as a Markov decision process (MDP) $\mathcal{M}$ characterized by a state space $\mathcal{X}$, an action space $\mathcal{A}$, a (stochastic) dynamics $p : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$ that determines the transition from states to states depending on the action, a reward function $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ that determines the value of a transition x, a, x'. An MDP defines a control **task**. The solution to an MDP/task is an optimal policy $\pi^* : \mathcal{X} \to \mathcal{A}$ that prescribes the actions to take in each state to maximize the (discounted) sum of rewards measured by the optimal value function $V^* = \max_\pi \mathbb{E}[\sum_t \gamma^t r_t]$ with $\gamma \in (0, 1)$ and $r_t = r(x_t, \pi(x_t), x_{t+1})$. Two of the most difficult challenges in RL are:

1. How to explore the unknown environment so as to maximize the cumulative reward. This requires solving the **exploration-exploitation** problem, well formalized and studied at its core by the multi-armed bandit framework [2].
2. How to effectively represent the policy and/or the value function. This requires defining an **approximation space** which is well-suited for the specific MDP at hand.

Both previous aspects may greatly benefit from techniques able to define suitable exploration strategies and approximation spaces from past experience or joint experience from other tasks (e.g., designing a intelligent tutoring system for a student and reuse the teaching strategy to

other students). The **objective** of my research is to study the problems of transfer learning, multi-task learning, and domain adaptation in the RL (and related) field.

The Past

Unlike in supervised learning, transfer learning faces challenges which are specific to field of RL:

- many different things can be transferred (e.g., the MDP parameters, policies, value functions, samples, features),
- the definition of "unsupervised" samples is not clear and thus, domain adaptation methods exploiting target unsupervised samples cannot be easily applied,
- samples are often non-i.i.d. because they are obtained from policies
- tasks may be similar in terms of policies but neither MDPs nor value functions or viceversa.

For this reason, borrowing techniques from "supervised" transfer/multi-task learning is not always trivial or even possible. Early research focused on studying transfer of different kind of solutions from a source to a target task[1]. Later, more sophisticated transfer/multi-task scenarios and algorithms have been developed (e.g., using hierarchical Bayesian solutions to learn "priors" from multiple tasks) to improve the accuracy of the approximation of optimal policies/value functions. The results obtained in the past show a significant sample complexity reduction and an improvement in asymptotic accuracy when transfer/multi-task is applied.

The Future

My main interest in the short-term is to study the problem of how transfer/multi-task learning can actually improve exploration-exploitation strategies in multi-armed bandit (MAB) and RL. While the problem of approximation is common in supervised learning as well, the active collection of information is very much specific to RL and MAB.

So far, I have investigated a sequential transfer scenario and investigated two approaches in the linear MAB framework: *(i)* transfer of samples *(under review)*, *(ii)* use of transferred samples to identify the set of possible MAB problems and speed-up the problem identification phase [3]. In both cases, we proved that the cumulative reward (i.e., reduce the regret) of exploration-exploitation strategies in MAB can be actually improved and that negative transfer can be avoid. Nonetheless, a number of very important questions remain unanswered:

- Is it possible to incrementally and efficiently estimate the potential bias due to transfer from different tasks? Under which assumptions? *In specific cases, this can be done in supervised learning.*
- What is the measure of similarity between two MDPs that determines the difference in performance of an exploration-exploitation strategy when applied to the two MDPs?
- Is it worth it to explore more in earlier tasks to "unveil" the generative process of the sequence of tasks and exploit it to enhance the transfer? In which scenarios?
- MDPs with different state-action spaces may still be very much similar. Is it possible to map different MDP to an "underlying" common MDP structure in which similar exploration-exploitation solutions can be identified and transferred?

As motivating fields of application, I will focus on *intelligent tutoring systems, recommendation systems, and computer games.*

[1] See [6, 4] for a survey.

References

1 Bertsekas, D. and Tsitsiklis, J. *Neuro-Dynamic Programming.* Athena Scientific, 1996.
2 Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.
3 Gheshlaghi-Azar, M., Lazaric, A., and Brunskill, E. Sequential transfer in multi-arm bandit with finite set of models. In *Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems (NIPS'13)*, 2013.
4 Lazaric, A. Transfer in reinforcement learning: a framework and a survey. In Wiering, M. and van Otterlo, M. (eds.), *Reinforcement Learning: State of the Art.* Springer, 2011.
5 Sutton, R. and Barto, A. *Reinforcement Learning, An introduction.* BradFord Book. The MIT Press, 1998.
6 Taylor, M. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10:1633–1685, 2000.

3.10 Deep unsupervised domain adaptation by backpropagation

Victor Lempitsky (Skoltech – Scolkovo, RU)

Joint work of Ganin, Yaroslav; Lempitsky, Victor
Main reference Y. Ganin, V. S. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in Proc. of the 32nd Int'l Conf. on Machine Learning (ICML'15), pp. 1180–1189, JMLR.org, 2015.
URL http://jmlr.org/proceedings/papers/v37/ganin15.html

The method

Consider the problem of learning a deep feedforward classifier in the presence of domain shift. Assume that a large number of labeled source examples and a large number of unlabeled target examples are present (e.g. train on synthetic images, test on real one). Our approach [1] to this unsupervised domain adaptation problem is to combine deep learning and domain adaptation into a single optimization process driven by simple backpropagation updates. The goal of the optimization is to obtain a deep model that has domain-invariant feature representations in its higher layers, while providing good predictions on the source data.

Let $\mathbf{x}$ be the input sample and y be the output of a network. Consider feature representation f that emerge after a certain layer L in the middle of the network. Let $\mathbf{f} = G_f(\mathbf{x}; \theta_f)$, $y = G_y(\mathbf{x}; \theta_y)$, where G_f and G_y are parts of the network before and after the layer L, while θ_f and θ_y are their parameters. Our goal is then to train a deep model where the features f are domain-invariant, i.e. have similar distribution in the source and the target domains. We denote these distributions as $S(\mathbf{f})$ and $T(\mathbf{f})$. While trying to match these distributions, one still needs to minimize the loss of the label prediction $y = G_y(G_f(\mathbf{x}; \theta_f); \theta_y)$ for source-domain data.

To measure the (dis)similarity of distributions $S(\mathbf{f})$ and $T(\mathbf{f})$, we augment our deep model with a domain classifier $d = G_d(\mathbf{f}; \theta_d)$. Given a feature vector $\mathbf{f}$ this multi-layer classifier tries to predict whether it corresponds to the source or to the target example (i.e. whether it comes from $S(\mathbf{f})$ or $T(\mathbf{f})$). The lower is the loss of this classifier, the larger is the gap between $S(\mathbf{f})$ and $T(\mathbf{f})$. In the ideal case ($S(\mathbf{f})$ is the same as $T(\mathbf{f})$) this classifier would perform no better than chance and have a high loss. The resulting three-part network has a fork shape (forward pass through the network works as: $\mathbf{x} \to \mathbf{f}$, $\mathbf{f} \to y$, $\mathbf{f} \to d$).

The learning process trains all three parts of the network simultaneously using backpropagation. The training incorporates both labeled source examples and unlabeled target

examples. The parameters θ_y and θ_d are optimized by an SGD, with each update minimizing the losses of the respective classifiers G_y (that only looks at labeled source data) and G_d (that looks both at source and target data). The updates of the parameters θ_f of the feature mapping are driven by the minimization of the loss of the label predictor G_y and the *maximization* of the loss of the domain classifier G_d (as we want features to be predictive of y and domain-invariant).

We can achieve this behavior within standard deep learning packages based on SGD using a simple trick. We reverse (multiply by a negative constant) the gradient that comes out of the domain classifier G_y during backpropagation and pass it further back into the feature extractor. This can be implemented as a simple *gradient reversal* layer. When this layer is inserted between the feature extractor G_f and the domain classifier G_d, SGD moves the parameters θ_f against the direction suggested by the minimization of the domain classifier' loss (thus maximizing it). This reverse direction is combined with the direction suggested by the minimization of the label predictor' loss (as G_f and G_y are connected sequentially in a standard way). Overall, SGD training makes the features $\mathbf{f}$ discriminative (good for predicting y), while trying to mix the the distributions $S(\mathbf{f})$ and $T(\mathbf{f})$ as much as possible. The resulting stochastic process can be seen as an example of *adversarial learning* and is reminiscent of adversarial generative networks [2].

Further outlook

Supervised deep learning methods are highly-successful across many applications. Yet training such models require lots of labeled data. Training them on surrogate data will therefore remain an important avenue for research. Unsupervised deep domain adaptation is becoming of particular interest for computer vision, since we almost always have some source of surrogate labeled data (the two most notable sources being Internet images and computer graphics).

The initial hope was that deep architectures will turn out to be invariant to domain shifts, yet this has not proven to be the case. On the one hand the networks show impressive ability to build invariance to some nuisance parameters towards higher level layers and thus mitigate the domain shift. On the other hand, the sheer number of parameters within modern deep architectures means that it is easier for deep models to overfit the peculiarities of a certain domain.

It is no wonder that several groups including ours started working in parallel on the unsupervised deep domain adaptation, i.e. training on labeled surrogate data and unlabeled target domain data (e.g. [1, 3, 4, 5]).

Overall, the goal seems to be to learn deep architectures where bottom layers are domain/modality specific with a gradually reducing specificity, middle layers are domain-invariant and task-unspecific, and then top layers are task specific (and class-specific). Parameters of the bottom layers of such networks can be either shared between domains or be different across domains.

References

1 Yaroslav Ganin and Victor Lempitsky, Unsupervised Domain Adaptation by Backpropagation, CoRR abs/1409.7495 (2014)

2 Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, Yoshua Bengio, Generative Adversarial Nets, NIPS 2014: 2672-2680

3 Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, Trevor Darrell: Deep Domain Confusion: Maximizing for Domain Invariance. CoRR abs/1412.3474 (2014)

4 Qiang Chen et al., Deep Domain Adaptation for Prediction of Fine-Grained Clothing Attributes. CVPR 2015
5 Mingsheng Long, Jianmin Wang: Learning Transferable Features with Deep Adaptation Networks. CoRR abs/1502.02791 (2015)

3.11 Feature Learning in a Probit Model with Correlated Noise

Stephan Mandt (Institute for Data Sciences and Engineering, Columbia University, US)

A large class of problems in statistical genetics amounts to finding a sparse linear effect in a binary classification setup, such as finding a small set of genes that most strongly predict a disease. Very often, these signals are spurious and obfuscated by confounders such as age, ethnicity or population structure. Beyond statistical genetics, sparse estimation is a general problem in binary classification, and has wide applications in science and technology, including, among many others, neuroscience, medicine, text classification, credit scoring, and computer malware detection. In all of these applications, confounding of the sparse signal can have dramatic consequences such as false medical diagnoses or violations of financial regulations. There is a need for statistical methods for feature selection that are robust to these confounding influences.

The model

In my talk I showed that by generalizing the probit model in a way that it captures correlated label noise is a way to eliminating confounders in the linear effect. Consider the following model:

$$Y_i = \text{sign}\left(X_i^\top w + \epsilon_i\right), \qquad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top \sim \mathcal{N}(0, \Sigma).$$

This is just the probit regression model with the addition of a covariance matrix for the label noises. By making the simplifying assumptions that all obeserved labels are 1 (this can be achieved by a linear transformation on the noise covariance and data matrix), the central computational problem amounts to optimizing the following objective function:

$$\mathcal{L}(w) = -\log \int_{\mathbb{R}^n_+} \mathcal{N}(\epsilon; X^\top w, \Sigma)\, d^n\epsilon \; + \; \lambda_0 ||w||_1^1.$$

Here, the ℓ_1 regularizer enforces sparsity in w, which is what we want in feature learning. In the uncorrelated case, the above integral decomposes into a sum of one-dimensional integrals that can be efficiently computed, but in the presence of correlations, the integral is intractable. In my talk, I derived an approximate inference algorithm for this task.

Why correlated label noises?

The correlated probit model delivers two alternative explanations of the observed labels Y_i: one in terms of a sparse linear effect (this is what we are interested in), and another explanation in terms of correlated label noise. The correlated label noise says, roughly speaking, that data points X_i that are similar, will also have similar labels Y_i. Similarity is

expressed in terms of a set of known kernels K_i (e.g., based on side information) that are the building blocks of the covariance matrix

$$\Sigma = \lambda_1 \mathbf{I} + \sum_{i=2}^{m} \lambda_i K_i.$$

The coefficients λ_i are determined by cross-validation. Now, by conditioning on the labels, the linear effect and the noise distribution will become correlated; in other words, thinking Bayesian, the correlated noise will *explain away* parts of the observed labels. Therefore the sparse linear effect will try to fit only those labels that are hard to fit with a correlated noise distribution, but better to fit with a sparse linear effect. Including a noise covariance matrix is therefore a possible way to include effects into our model that we do *not* want to have an effect on the sparse signal of interest.

Summary

Removing confounders in classification and regression task is an active and highly relevant field of research. A challenge is to make these more complex models computationally tractable. Variational methods offer a promising path.

References

1 Stephan Mandt, Florian Wenzel, Shinichi Nakajima, John P. Cunningham, Christoph Lippert, and Marius Kloft. Sparse Estimation in a Correlated Probit Model. arXiv preprint, arxiv:1507.04777.

3.12 A Resampling Method for Importance Weight Estimation

Shinichi Nakajima (TU Berlin, DE)

Joint work of Panknin, Danny; Braun, Mikio; Müller, Klaus-Robert;

Under the covariate shift setting, accurate estimation of importance weight is a key step, and several methods have been proposed for this purpose. We consider a new resampling method for density ratio estimation between two distributions, and introduce our plan to show its usefulness in theory and experiment.

3.13 Not IID Data in Advertising

Francesco Orabona (Yahoo! Labs – New York, US)

Main reference F. Orabona, "Simultaneous Model Selection and Optimization through Parameter-free Stochastic Learning," in Proc. of the 2014 Annual Conf. on Neural Information Processing Systems (NIPS'14), pp. 1116-1124, 2014; pre-print available as arXiv:1406.3816v1 [cs.LG].
URL http://papers.nips.cc/paper/5503-simultaneous-model-selection-and-optimization-through-parameter-free-stochastic-learning
URL http://arxiv.org/abs/1406.3816v1

We present the problem of click prediction and show what is the most common solution employed in industry to not-IID training data. Latest achievements in automatic parameter tuning for stochastic gradient descent are also shown.

3.14 The Benefit of Multitask Representation Learning

Massimiliano Pontil (University College London, GB)

Main reference A. Maurer, M. Pontil, B. Romera-Paredes, "The Benefit of Multitask Representation Learning," arXiv:1505.06279v1 [stat.ML], 2015.
URL http://arxiv.org/abs/1505.06279v1

We discuss a general method to learn data representations from multiple tasks. We provide a justification for this method in both settings of multitask learning and learning-to-learn. The method is illustrated in detail in the special case of linear feature learning. Conditions on the theoretical advantage offered by multitask representation learning over independent tasks learning are established. In particular, focusing on the important example of halfspace learning, we derive the regime in which multitask representation learning is beneficial over independent task learning, as a function of the sample size, the number of tasks and the intrinsic data dimensionality. Other potential applications of our results include multitask feature learning in reproducing kernel Hilbert spaces and multilayer, deep networks.

3.15 Adaptive Lifelong Learning for Visual Recognition and Data Analysis

Erik Rodner (Friedrich Schiller University Jena, DE)

Current and previous work

Whereas my studies focused on transfer learning with Gaussian process models [8] and random decision forests [9], my current main research topic is lifelong learning and adaptive scientific data analysis. In particular, I have worked on aspects of adaptation [6, 10] (model sharing, learning from different but related datasets), active learning [3, 7] (selecting unlabeled examples which are likely beneficial when being labeled by an annotator), novelty detection [1] (determining whether an example belongs to an unknown category), and fine-grained recognition [11, 4, 2] (discriminating very similar categories). Learning with non-iid. data has been always part of my research on domain adaptation, where I search for handy solutions applicable to some of the large-scale learning problems we have in vision and scientific data analysis. One example is the MMDT (max-margin domain transforms) method presented in [6], which jointly learns classifier parameters as well as a linear transformation that maps labeled examples of one dataset to the feature space of another but related labeled dataset. The method itself is a straightforward extension of standard one-vs-all SVMs and can be used in large-scale scenarios [10].

Recently, people have boosted the performance on nearly all vision datasets and tasks by using a feature representation learned with high complexity models (*e.g.,* CNNs) on large-scale datasets, such as ImageNet. This strategy can be seen as non-iid. learning with two related but different datasets (ImageNet and another vision data set). In a recent publication, we brought this concept to an extreme by using pre-trained CNN models for object part discovery [11].

Adaptive lifelong learning

I am currently developing an approach which allows for adapting to new input data and especially new tasks (set of categories) in a semi-supervised learning setting. First of all, think about the scenario where we have ImageNet $\tilde{\mathcal{D}} = \left(\tilde{\mathbf{X}}, \tilde{\mathbf{y}}\right)$ (labels $\tilde{\mathbf{y}}$, input examples $\tilde{\mathbf{X}}$), from which we can learn quite a lot of object categories, and an unlabeled set of images $\mathcal{D} = \mathbf{X}$ acquired in a new environment/domain (*e.g.,* video sequence of your office). The goal is now to learn an object classifier for the new domain by exploiting the fact that the input examples are related but different and the set of categories for the new domain might also contain new categories not part of ImageNet (have you ever searched for *toothpaste* in ImageNet?), *i.e.,* the label space changed.

In particular, $\tilde{\mathcal{D}}$ is sampled from $p(y, \tilde{\mathbf{x}}|\tilde{\mathbf{q}})$ and the unlabeled set $\mathcal{D}$ is sampled from $p(\mathbf{x}|\mathbf{q})$, where $\tilde{\mathbf{q}}$ and $\mathbf{q}$ are parameters of the distributions and are assumed to be sampled from a world model $p(\tilde{\mathbf{q}}|\mathbf{Q})$ and $p(\mathbf{q}|\mathbf{Q})$. The goal is to find a model for $p(y|\mathbf{x}, \mathbf{q})$ by using both datasets $\mathcal{D}$ and $\tilde{\mathcal{D}}$ and carefully coupling of the distributions through the world model. In summary, this is a learning framework that allows for adaptation of the label and the input space jointly. Furthermore, it can be extended to learning over time by assuming continuously changing distributions parameterized by $\mathbf{q}_t$.

Further challenges

In general, I am also interested in studying the effects current fine-tuning strategies have for adaptation. In contrast to vision research a few years ago, people make indirectly use of domain adaptation principles when fine-tuning is performed on models initially learned on other datasets. How can we control the degree of adaptation performed? Are there any theoretical results that might help us to select the parameters that should be fine-tuned and the ones that should be fixed to their initial value?

Furthermore, adapting to the right output space, a user might need and expect, will be extremely important in future in my opinion, especially for scientific data analysis where the goal is not always defined in advance. In vision, object detection methods can now detect thousands of categories and without focusing and re-focusing on the subset and the granularity of semantic information the user needs, we are likely not be able to make use of the results at all.

References

1 Paul Bodesheim, Alexander Freytag, Erik Rodner, Michael Kemmler, and Joachim Denzler. Kernel null space methods for novelty detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3374–3381, 2013.

2 Alexander Freytag, Erik Rodner, Trevor Darrell, and Joachim Denzler. Exemplar-specific patch features for fine-grained recognition. In *German Conference on Pattern Recognition (GCPR)*, pages 144–156, 2014.

3 Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *European Conference on Computer Vision (ECCV)*, volume 8692 of *Lecture Notes in Computer Science*, pages 562–577, 2014.

4 Christoph Göring, Erik Rodner, Alexander Freytag, and Joachim Denzler. Nonparametric part transfer for fine-grained recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2489–2496, 2014.

5 Judy Hoffman, Sergio Guadarrama, Eric Tzeng, Ronghang Hu, Jeff Donahue, Ross Girshick, Trevor Darrell, and Kate Saenko. LSDA: Large scale detection through adaptation. In *Neural Information Processing Systems (NIPS)*, 2014.

6 Judy Hoffman, Erik Rodner, Jeff Donahue, Brian Kulis, and Kate Saenko. Asymmetric and category invariant feature transformations for domain adaptation. *International Journal of Computer Vision (IJCV)*, 109(1-2):28–41, 2014.

7 Christoph Käding, Alexander Freytag, Erik Rodner, Paul Bodesheim, and Joachim Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

8 Erik Rodner and Joachim Denzler. One-shot learning of object categories using dependent gaussian processes. In *Annual Symposium of the German Association for Pattern Recognition (DAGM)*, pages 232–241. Springer, 2010.

9 Erik Rodner and Joachim Denzler. Learning with few examples for binary and multi-class classification using regularization of randomized trees. *Pattern Recognition Letters*, 32(2):244–251, January 2011.

10 Erik Rodner, Judy Hoffman, Jeff Donahue, Trevor Darrell, and Kate Saenko. Transform-based domain adaptation for big data. In *NIPS Workshop on New Directions in Transfer and Multi-Task Learning*, 2013. abstract version of arXiv:1308.4200.

11 Marcel Simon, Erik Rodner, and Joachim Denzler. Part detector discovery in deep convolutional neural networks. In *Asian Conference on Computer Vision (ACCV)*, 2014.

3.16 Covariate Shift and Varying-Coefficient Models

Tobias Scheffer (University of Potsdam, DE)

Joint work of Niels Landwehr, Matthias Bussas, Christoph Sawade, and Tobias Scheffer

The Past: Discriminative Learning of Importance Weights for Covariate Shift

Consider a data generation process in which there is a source variable $\sigma \in \{\text{train}, \text{test}\}$. Training instances are governed by $p(\mathbf{x}|\sigma = \text{train})$ whereas test instances are governed by a potentially different $p(\mathbf{x}|\sigma = \text{train})$. In either case, labels are created according to $p(y|\mathbf{x})$.

In order to minimize the regularizes risk under the test distribution, one has to minimize

$$\sum_{i=1}^{n} \frac{p(\mathbf{x}|\sigma = \text{test})}{p(\mathbf{x}|\sigma = \text{train})} \ell(f_{\mathbf{w}}(\mathbf{x}_i), y_i) + \Omega(\mathbf{w}).$$

Estimating the training and test density functions [5] is unnecessarily difficult, because those are high-dimensional density functions and really only a scalar factor is needed for each instance. However, observe that, by simple arithmetics [1]:

$$\frac{p(\mathbf{x}|\sigma = \text{test})}{p(\mathbf{x}|\sigma = \text{train})} = \frac{p(\sigma = \text{train})}{p(\sigma = \text{test})} \left(\frac{1}{p(\sigma = \text{train}|\mathbf{x})} - 1 \right).$$

The density ratio can be written in terms of $p(\sigma = \text{train}|\mathbf{x})$ which can be estimated with a logistic regression model

$$p(\sigma = \text{train}|\mathbf{x}, \mathbf{v}) = \frac{1}{1 + exp(\mathbf{v}^\mathsf{T}\mathbf{x})}.$$

This model is trained using the training data as positive, and the test data as negative examples.

Over KLIEP [6], this method has the advantage that the optimization problems are more directly linked to minimizing the risk under the test distribution. Over kernel mean matching

[4] it has the advantage that the regularization parameter for model $f_{\mathbf{v}}$ can be tuned easily. Since it is trained on labeled data (with label σ), it can simply be tuned on held-out data.

The Future: Varying-Coefficient Models with Isotropic GP Priors

Consider problems with continuous task variables $\mathbf{t}$ (*e.g.,* time and space), regular attributes $\mathbf{x}$, and outputs y. Assume that $p_{\mathbf{t}}(y|\mathbf{x})$ changes smoothly in $\mathbf{t}$. For standard learning problems, parameters $\mathbf{w}$ of a model $p(y|\mathbf{x},\mathbf{w})$ are usually assumed to be governed by an isotropic Gaussian prior (hence ℓ_2 regularization of $\mathbf{w}$). Instead, let us assume that a function $\boldsymbol{\omega} : \mathbf{t} \mapsto \mathbf{w}$ that generates task-specific parameters $\boldsymbol{\omega}(\mathbf{t})$ of a model $p(y|\mathbf{x},\boldsymbol{\omega}(\mathbf{t}))$ is governed by an isotropic Gaussian Process prior.

The Gaussian Process couples $p(y|\mathbf{x},\boldsymbol{\omega}(\mathbf{t}))$ for different values of $\mathbf{t}$. A constant $\boldsymbol{\omega}(\mathbf{t})$ corresponds to an *iid* model; generally, $\boldsymbol{\omega}$ allows the model to change smoothly in $\mathbf{t}$.

"Theorem". Let $\mathbf{X},\mathbf{T},\mathbf{y}$ be the training data and $\mathbf{x}^*,\mathbf{t}^*$ a test instance for which y^* has to be inferred. The predictive distribution $p(y^*|\mathbf{X},\mathbf{y},\mathbf{T},\mathbf{x}^*,\mathbf{t}^*)$ of the above model is equal to the predictive distribution of a standard Gaussian process that uses concatenated attribute vectors $(\mathbf{x},\mathbf{t})$ and product kernel $k((\mathbf{x}_i,\mathbf{t}_i),(\mathbf{x}_j,\mathbf{t}_j)) = k(\mathbf{x}_i,\mathbf{x}_j)k(\mathbf{t}_i,\mathbf{t}_j)$.

The theorem shows that Bayesian inference for varying-coefficient models can be done in $O(n^3 + dn)$ in the dual instead of in $O(n^3d^3)$ [3] for n observations and d attributes. It also makes assumptions explicit that justify the use of products of task and instance kernels [2]. The model works great for geospatial problems such as predicting rents or real estate prices.

Acknowledgment. This is joint work with Niels Landwehr, Matthias Bussas, and Christoph Sawade.

References

1 S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *ICML*, 2007.

2 E. V. Bonilla, F. V. Agakov, and C. K. I. Williams. Kernel multi-task learning using task-specific features. In *AISTATS*, 2007.

3 A. Gelfand, H. Kim, C. Sirmans, and S. Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462), 2003.

4 J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Sample sel. bias by unlabeled data. In *NIPS*, 2007.

5 Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

6 M. Sugiyama, T. Suzuki, S. Nakajima, P. von Bünau, and M. Kawanabe. Direct importance estimation for covariate shift adaptation. *Ann. of the Inst. of Stat. Math.*, 60(4), 2008.

3.17 Kernel Hypothesis Tests on Dependent Data

Dino Sejdinovic (University of Oxford, GB)

Joint work of Chwialkowski, Kacper; Sejdinovic, Dino; Gretton, Arthur
Main reference K. P. Chwialkowski, D. Sejdinovic, A. Gretton, "A wild bootstrap for degenerate kernel tests," in Proc. of the 2014 Annual Conf. on Neural Information Processing Systems (NIPS'14), pp. 3608–3616, 2014.
URL http://papers.nips.cc/paper/5452-a-wild-bootstrap-for-degenerate-kernel-tests

Statistical tests based on embeddings of probability distributions into reproducing kernel Hilbert spaces have been applied in many contexts, including two sample testing [6], tests of independence [5, 1], tests of conditional independence [4, 10], and tests for higher order (Lancaster) interactions [8].

For these tests, consistency is guaranteed if and only if the observations are independent and identically distributed. Much real-world data fails to satisfy the i.i.d. assumption: audio signals, EEG recordings, text documents, financial time series, and samples obtained when running Markov Chain Monte Carlo (MCMC), all show significant temporal dependence patterns. The asymptotic behaviour of kernel test statistics becomes quite different when temporal dependencies exist – the difference in their asymptotic null distributions has important implications in practice: the permutation-based tests return an elevated number of false positives.

An alternative estimate of the null distribution for the problem of independence testing was proposed in [2] (where one signal is repeatedly *shifted* relative to the other). There is, however, no obvious way to generalise this approach to other testing contexts. For instance, we might have two time series, with the goal of comparing their marginal distribution. In [3], it was shown that an external randomization with *wild bootstrap* [7] may be applied to simulate from the null distribution for *all* kernel hypothesis tests for which V-statistics are employed, and not just for independence tests. This result has a potential to lead to a powerful set of model checking and MCMC diagnostic tools – where a nonparametric test can be constructed whether a Markov chain has reached its stationary distribution using Maximum Mean Discrepancy (MMD) [6] as a test statistic, similarly as in [9]. While a permutation-based test of whether the sampler has converged leads to too many rejections of the null hypothesis due to chain dependence (implying that one requires heavily thinned chains, which is wasteful of samples and computationally burdensome), the wild bootstrap approach can be applied directly on chains and is demonstrated to attain a desired number of false positives in [3].

Future Work

Consistency of the above procedures requires strong mixing conditions on the time series at hand. Moreover, the wild bootstrap procedure has a tuning parameter which requires some knowledge of the mixing properties in order to be properly calibrated. Finally, the interplay between the kernel choice and the test performance in the case of dependent data is not well understood. What are the inherent tradeoffs when trying to learn such tuning parameters on a held out portion of the data before performing a test? Moreover, many outstanding practical considerations arise in the application of tests to MCMC diagnostics. When to perform a test? Can tuning parameters be learned on the fly?

Acknowledgments. This is joint work with Kacper Chwialkowski and Arthur Gretton.

References

1. Besserve, M., Logothetis, N.K., and Schölkopf, B. Statistical analysis of coupled time series with kernel cross-spectral density operators. In *NIPS*. 2013.
2. Chwialkowski, K. and Gretton, A. A kernel independence test for random processes. In *ICML*, 2014.
3. Chwialkowski, K., Sejdinovic, D., and Gretton, A. A wild bootstrap for degenerate kernel tests. In *NIPS*, 2014.
4. Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS*, volume 20, pp. 489–496, 2007.
5. Gretton, A., Fukumizu, K., Teo, C, Song, L., Schölkopf, B., and Smola, A. A kernel statistical test of independence. In *NIPS*, volume 20, pp. 585–592, 2007.
6. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., and Smola, A. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.
7. Leucht, A. and Neumann, M.H. Dependent wild bootstrap for degenerate U- and V-statistics. *J. Multivar. Anal.*, 117:257–280, 2013.
8. Sejdinovic, D., Gretton, A., and Bergsma, W. A kernel test for three-variable interactions. In *NIPS*, pp. 1124–1132, 2013.
9. Sejdinovic, D., Strathmann, H., Lomeli Garcia, M., Andrieu, C., and Gretton, A. Kernel Adaptive Metropolis-Hastings. In *ICML*, 2014.
10. Zhang, K., Peters, J., Janzing, D., B., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *UAI*, 2011.

3.18 Zero-shot learning via synthesized classifiers

Fei Sha (University of Southern California – Los Angeles, US)

Joint work of Sha, Fei; Chao, Weilun; Changpinyo, Soravit; Gong, Boqing

Real-world objects have a long-tailed distribution, making it difficult to collect labeled images of rare objects for visual object recognition. One appealing way to address this problem is zero-shot learning. We propose a unified framework based on the key insight that the classifiers of semantically similar objects can be constructed from a set of *base* classifiers of "phantom" classes. In sharp contrast to previous work, the classifiers of both seen and unseen objects are synthesized from the base classifiers, enabling us to effectively learn the bases using the labeled data of the seen classes and then readily apply them to synthesizing the classifiers of unseen classes. We further consider a *generalized* zero-shot learning setting, in which the test phase is a multi-way classification problem over both seen and unseen classes. This generalized case reflects more closely how test data are distributed in real applications, leading to a more challenging task. We demonstrate superior performance of our approach over the state of the art for (generalized) zero-shot learning on two benchmark datasets.

I would like to acknowledge the beneficial discussions with Prof. Christoph Lampert (IST, Austria) at the Dagstuhl Seminar, in particular, pointers to his earlier work on generalized zero-shot learning.

3.19 A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning

Ingo Steinwart (Universität Stuttgart, DE)

Joint work of Hang, Hanyuan; Steinwart, Ingo
Main reference H. Hang, I. Steinwart, "A Bernstein-type Inequality for Some Mixing Processes and Dynamical Systems with an Application to Learning," arXiv:1501.03059v1 [math.PR], 2015.
URL http://arxiv.org/abs/1501.03059v1

We establish a Bernstein-type inequality for a class of stochastic processes that include the classical geometrically ϕ-mixing processes, Rio's generalization of these processes, as well as many time-discrete dynamical systems. Modulo a logarithmic factor and some constants, our Bernstein-type inequality coincides with the classical Bernstein inequality for i.i.d. data. We further use this new Bernstein-type inequality to derive an oracle inequality for generic regularized empirical risk minimization algorithms and data generated by such processes. Applying this oracle inequality to support vector machines using the Gaussian kernels for both least squares and quantile regression, it turns out that the resulting learning rates match, up to some arbitrarily small extra term in the exponent, the optimal rates for i.i.d. processes.

3.20 Sampling without replacement: direct approach vs. reduction to i.i.d.

Ilya Tolstikhin (MPI for Intelligent Systems – Tübingen, DE)

Joint work of Tolstikhin, Ilya; Blanchard, Gilles; Kloft, Marius
Main reference I. O. Tolstikhin, G. Blanchard, M. Kloft, "Localized complexities for transductive learning," in Proc. of the 27th Conf. on Learning Theory (COLT'14), pp.857–884, JMLR.org, 2014.
URL http://jmlr.org/proceedings/papers/v35/tolstikhin14.html

We consider two closely related questions: (1) general properties of random variables sampled *without* replacement from arbitrary finite domains and (2) risk bounds in transductive learning, which is a particular setting of statistical learning theory introduced by V. Vapnik.

Formally, let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be some fixed finite *population*. Let $Z_1, \ldots, Z_n$ be sampled uniformly *without replacement* from $\mathcal{C}$ for $n \leq N$. independent. which may be more useful depending on situations: n of them and then take the first subset. Random variables sampled without replacement naturally appear in many modern applications of statistics, probability, and machine learning. First example which comes to mind is cross-validation, where sample is randomly partitioned into training and validation subsets. Other examples include matrix completion problems, various iterative stochastic algorithms like stochastic gradient descent, low-rank matrix factorization problems, and many others.

Arguably, one of the most useful tools when it comes to analysis of stochastic procedures are *concentration inequalities*, which control a deviation of random variables from their expected values with high probability. Generally one would like to upper bound tail probabilities $\mathbb{P}\{\xi - E[\xi] > t\}$ or $\mathbb{P}\{E[\xi] - \xi > t\}$ for $t > 0$ and $\xi := f(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ are random variables taking values in domain X and $f\colon X^n \to \mathbb{R}$. The case when $X_1, \ldots, X_n$ are independent is very well studied and many useful results are available, including Hoeffding's and Bernstein's inequalities for sums of independent real-valued random variables and McDiarmid's inequality for functions f with bounded differences. However,

when random variables are sampled without replacement $\xi := f(Z_1, \ldots, Z_n)$ new techniques are needed.

First results in this direction were derived by Hoeffding, who showed that classic inequalities for sums mentioned above also hold for $\xi := \sum_{i=1}^n Z_i$. This result was based on the elegant *reduction* of the sampling without replacement scheme to the i.i.d. setting. Later results showed that a direct approach can be tighter than the reduction: using a *martingale technique* Serfling derived an improved version of Hoeffding's inequality for $\xi := \sum_{i=1}^n Z_i$, containing additional factor $\frac{N-n+1}{N}$ which decreases as $n \to N$. The same technique was later used to derive versions of Bernstein's and McDiarmid's inequalities for sampling without replacement, which improve upon the i.i.d. counterparts in the similar way.

In trunsductive learning, a learner observes n labeled training points together with u unlabeled test points with the final goal of giving correct answers for the test points. This process can be modeled using sampling without replacement described above, with fixed population of N input-output pairs $\mathcal{C} := \{(X_i, Y_i)\}_{i=1}^N$, random labeled training sample $S_n := \{Z_1, \ldots, Z_n\}$, and unlabeled test sample X_u containing $u = N - n$ inputs of remaining elements $S_u := \mathcal{C} \setminus S_n$. Usually the learner fixes a class of predictors $\mathcal{H}$ and a bounded loss function ℓ and seeks for an optimal predictor h_u^* minimizing an average test loss $\mathrm{err}(h, S_u)$ over $\mathcal{H}$. However, labels of the test objects are unknown, and the learner resorts to $\hat{h}_n$ which minimizes an empirical loss $\mathrm{err}(h, S_n)$ over $\mathcal{H}$. The main question is: how large can be the *excess risk* $\mathrm{err}(\hat{h}_n, S_u) - \mathrm{err}(h_u^*, S_u)$? The excess risk can be upper bounded in a standard way by uniform deviations of risks computed on two disjoint finite samples $Q_n := \sup_{h \in \mathcal{H}} |\mathrm{err}(h, S_u) - \mathrm{err}(h, S_n)|$. Note that this construction naturally appears as a middle step in proofs of standard i.i.d. risk bounds as a result of symmetrization or the so-called *double-sample trick*. Since Q_n is a function of the random training set S_n, we can apply concentration inequalities for sampling without replacement in order to upper bound it using $E[Q_n]$. This can be done using a version of McDiarmid's inequality or more powerful versions of Talagrand's inequality for sampling without replacement, which were recently derived in [1].

It was also shown in [1] using Hoeffding's reduction trick that $E[Q_n]$ is upper bounded by $E[\tilde{Q}_n]$, where $\tilde{Q}_n$ is a supremum of the standard i.i.d. empirical process. Using well-known *symmetrization inequalities* one can further upper bound $E[\tilde{Q}_n]$ (and thus $E[Q_n]$) with *Rademcher complexity* of the class $\mathcal{H}$. Together with concentration argument this shows that most of the i.i.d. risk bounds also hold in the transductive learning setting. However, we would like to argue that this reduction to i.i.d. setting can give suboptimal results compared to direct analysis of $E[Q_n]$ (in the same way as Hoeffding's reduction trick leads to suboptimal inequalities compared to the direct martingale technique).

We introduce a new complexity measure for transductive learning called *permutational Rademacher complexity* (PRC), which is similar to the standard Rademacher complexity. The only difference is that in PRC ± 1 signs are obtained using random permutation of a sequence containing equal number of "-1" and "$+1$", while in the Rademacher complexity signs are sampled i.i.d. We provide the preliminary results on PRC, including a novel symmetrization inequality, which shows that $E[Q_n]$ is upper bounded by PRC.

References

1 Tolstikhin, I., Blanchard, G., Kloft, M.: Localized complexities for transductive learning. In: COLT 2014, pp. 857–884 (2014)

3.21 Active Learning for Domain Adaptation

Ruth Urner (MPI for Intelligent Systems – Tübingen, DE)

Main reference C. Berlind, R. Urner, "Active Nearest Neighbors in Changing Environments," in Proc. of the 32nd Int'l Conf. on Machine Learning (ICML'15), pp. 1870–1879, JMLR.org, 2015.
URL http://jmlr.org/proceedings/papers/v37/berlind15.html

While classic machine learning paradigms assume training and test data are generated from the same process, domain adaptation addresses the more realistic setting in which the learner has large quantities of labeled data from some *source* task but limited or no labeled data from the *target* task it is attempting to learn.

In the paper, we give the first formal analysis showing that using *active learning for domain adaptation* yields a way to address the challenges inherent in this scenario. As is common, we assume that the learner receives *labeled data* from the source task and *unlabeled data* from the target task. In our model, the learner can make a small number of queries for labels of target examples. Now the goal is to accurately learn a classifier for the target task while making as few label requests as possible.

We propose a simple nonparametric algorithm, ANDA, that combines an active nearest neighbor querying strategy with nearest neighbor prediction. ANDA receives a labeled sample from the source distribution and an unlabeled sample from the target task. It first actively selects a subset of the target data to be labeled based on the amount of source data among the k' nearest neighbors of each target example. Then it outputs a k-nearest neighbor classifier on the combined source and target labeled data.

We prove that ANDA enjoys strong performance guarantees on both the risk of the resulting classifier and the number of queries ANDA will make. Simply put, ANDA is guaranteed to make enough queries to be consistent but will not make unnecessary ones.

4 Working Groups, Presentations, and Panel Discussion

Working groups were an essential part of the seminar and have been integrated in the schedule in two versions: (1) discussion groups always directly after a presentation session and (2) working groups on Thursday during a longer time slot with topics voted for by the participants in a pseudo-random fashion. Especially the discussion groups directly after presentations led to interesting questions and comments by all participants. Although time was limited, results from the groups were summarized and supported a very interactive atmosphere of the seminar. The talks of the seminar had three different lengths: (1) longer keynotes for vision, algorithms, and computational biology for 25 minutes, (2) ongoing research talks for 12 minutes, and (3) quick presentations for just 3 minutes. This mix allowed a presentation for every participant and the quick presentations often led to interesting discussions in the evening.

The seminar ended with a panel discussion in the garden with Fei Sha, Shai-Ben David, and Oliver Stegle on the topic of open problems and upcoming research challenges in the area of non-i.i.d. learning. The topic quickly shifted towards recent advances in deep learning and how they are currently affecting the methodology used for non-i.i.d. learning. Especially for computational biology topics, the lack of large-scale training data was mentioned as the main obstacle for using these techniques. The panel ended with a summary of the seminar and a feedback to the organizers about its structure.

Participants

- Shai Ben-David
University of Waterloo, CA
- Gilles Blanchard
Universität Potsdam, DE
- Trevor Darrell
University of California – Berkeley, US
- Joachim Denzler
Universität Jena, DE
- Philipp Drewe
Max-Delbrück-Centrum, DE
- Mario Fritz
MPI für Informatik – Saarbrücken, DE
- Judy Hoffman
University of California – Berkeley, US
- Josef Kittler
University of Surrey, GB
- Marius Kloft
HU Berlin, DE
- Brian Kulis
Ohio State University – Columbus, US
- Christoph H. Lampert
IST Austria – Klosterneuburg, AT
- Soeren Laue
Universität Jena, DE
- Alessandro Lazaric
INRIA – University of Lille 1, FR
- Victor Lempitsky
Skoltech – Scolkovo, RU
- Christoph Lippert
Los Angeles, US
- Stephan Mandt
Columbia University, US
- Shin Nakajima
TU Berlin, DE
- Francesco Orabona
Yahoo! Labs – New York, US
- Massimiliano Pontil
University College London, GB
- Gunnar Rätsch
Memorial Sloan-Kettering Cancer Center – New York, US
- Erik Rodner
Universität Jena, DE
- Kate Saenko
University of Massachusetts – Lowell, US
- Tobias Scheffer
Universität Potsdam, DE
- Dino Sejdinovic
University of Oxford, GB
- Fei Sha
University of Southern California – Los Angeles, US
- Oliver Stegle
European Bioinformatics Institute – Cambridge, GB
- Ingo Steinwart
Universität Stuttgart, DE
- Ilya Tolstikhin
MPI für Intelligente Systeme – Tübingen, DE
- Ruth Urner
MPI für Intelligente Systeme – Tübingen, DE
- Alexander Zimin
IST Austria – Klosterneuburg, AT