

Foundations of Unsupervised Learning

Edited by

Maria-Florina Balcan¹, Shai Ben-David², Ruth Urner³, and
Ulrike von Luxburg⁴

- 1 Carnegie Mellon University, US, ninamf@cs.cmu.edu
- 2 University of Waterloo, CA, shai@uwaterloo.ca
- 3 MPI für Intelligente Systeme – Tübingen, DE, ruth.urner@tuebingen.mpg.de
- 4 Universität Tübingen, DE, luxburg@informatik.uni-tuebingen.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 16382 “Foundations of Unsupervised Learning”. Unsupervised learning techniques are frequently used in practice of data analysis. However, there is currently little formal guidance as to how, when and to what effect to use which unsupervised learning method. The goal of the seminar was to initiate a broader and more systematic research on the foundations of unsupervised learning with the ultimate aim to provide more support to practitioners. The seminar brought together academic researchers from the fields of theoretical computer science and statistics as well as some researchers from industry.

Seminar September 18–23, 2016 – <http://www.dagstuhl.de/16382>

1998 ACM Subject Classification I.2.6 Learning, H.3.3 Information Search and Retrieval


Keywords and phrases Machine learning, theory of computing, unsupervised learning, representation learning

Digital Object Identifier 10.4230/DagRep.6.9.94

1 Executive summary

Ruth Urner

Shai Ben-David

License  Creative Commons BY 3.0 Unported license
© Ruth Urner and Shai Ben-David

The success of Machine Learning methods for prediction crucially depends on data preprocessing such as building a suitable feature representation. With the recent explosion of data availability, there is a growing tendency to “let the data speak itself”. Thus, unsupervised learning is often employed as a first step in data analysis to build a good feature representation, but also, more generally, to detect patterns and regularities independently of any specific prediction task. There is a wide range of tasks frequently performed for these purposes such as representation learning, feature extraction, outlier detection, dimensionality reduction, manifold learning, clustering and latent variable models.

The outcome of such an unsupervised learning step has far reaching effects. The quality of a feature representation will affect the quality of a predictor learned based on this representation, a learned model of the data generating process may lead to conclusions about causal relations, a data mining method applied to a database of people may identify certain groups of individuals as “suspects” (for example of being prone to developing a specific disease or of being likely to commit certain crimes).



Except where otherwise noted, content of this report is licensed
under a Creative Commons BY 3.0 Unported license

Foundations of Unsupervised Learning, *Dagstuhl Reports*, Vol. 6, Issue 9, pp. 94–109

Editors: Maria-Florina Balcan, Shai Ben-David, Ruth Urner, and Ulrike von Luxburg



DAGSTUHL
REPORTS

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

However, in contrast to the well-developed theory of supervised learning, currently systematic analysis of unsupervised learning tasks is scarce and our understanding of the subject is rather meager. It is therefore more than timely to put effort into developing solid foundations for unsupervised learning methods. It is important to understand and be able to analyze the validity of conclusions being drawn from them. The goal of this Dagstuhl Seminar was to foster the development of a solid and useful theoretical foundation for unsupervised machine learning tasks.

The seminar hosted academic researchers from the fields of theoretical computer science and statistics as well as some researchers from industry. Bringing together experts from a variety of backgrounds, highlighted the many facets of unsupervised learning. The seminar included a number of technical presentations and discussions about the state of the art of research on statistical and computational analysis of unsupervised learning tasks.

We have held lively discussions concerning the development of objective criteria for the evaluation of unsupervised learning tasks, such as clustering. These converged to a consensus that such universal criteria cannot exist and that there is need to incorporate specific domain expertise to develop different objectives for different intended uses of the clusterings. Consequently, there was a debate concerning ways in which theoretical research could build useful tools for practitioners to assist them in choosing suitable methods for their tasks. One promising direction for progress towards better alignment of algorithmic objectives with application needs is the development of paradigms for interactive algorithms for such unsupervised learning tasks, that is, learning algorithms that incorporate adaptive “queries” to a domain expert. The seminar included presentations and discussions of various frameworks for the development of such active algorithms as well as tools for analysis of their benefits.

We believe, the seminar was a significant step towards further collaborations between different research groups with related but different views on the topic. A very active interchange of ideas took place and participants expressed their satisfactions of having gained new insights into directions of research relevant to their own. As a group, we developed a higher level perspective of the important challenges that research of unsupervised learning is currently facing.

2 Table of Contents

Executive summary

<i>Ruth Urner and Shai Ben-David</i>	94
--	----

Overview of Talks

Linear Algebraic Structure of Word Meanings <i>Sanjeev Arora</i>	98
Interactive Clustering <i>Pranjal Awasthi</i>	98
Two recent clustering paradigms <i>Shai Ben-David</i>	98
Questions in Representation Learning <i>Olivier Bousquet</i>	99
Active Learning Beyond Label Feedback <i>Kamalika Chaudhuri</i>	99
A cost function for similarity-based hierarchical clustering <i>Sanjoy Dasgupta</i>	100
Two sample tests for large random graphs <i>Debarghya Ghoshdastidar</i>	100
Globally Optimal Training of Generalized Polynomial Neural Networks with Non-linear Spectral Methods <i>Matthias Hein</i>	100
Multicriterion cluster validation <i>Christian Hennig</i>	101
What are the true clusters? <i>Christian Hennig</i>	101
Meta-unsupervised-learning: a supervised approach to unsupervised learning <i>Adam Tauman Kalai</i>	102
Planted Gaussian Problem: Beating the Spectral Bound <i>Ravindran Kannan</i>	102
Recent work on clustering and mode estimation with kNN graphs <i>Samory Kpotufe</i>	103
Proving clusterability <i>Marina Meila</i>	103
On Resilience in Graph Coloring and Boolean Satisfiability <i>Lev Reyzin</i>	103
Active Nearest-Neighbor Learning in Metric Spaces <i>Sivan Sabato</i>	104
Aversion k-clustering: How constraints make clustering harder <i>Melanie Schmidt</i>	104
Gradient descent for sequential analysis operator learning <i>Karin Schnass</i>	104

Towards an Axiomatic Approach to Hierarchical Clustering of Measures
Ingo Steinwart 105

On some properties of MMD and its relation to other distances
Ilya Tolstikhin 105

Lifelong Learning with Weighted Majority Votes
Ruth Urner 105

A Modular Theory of Feature Learning
Robert C. Williamson 106

Open problems

Valid cost functions for nonlinear dimensionality reduction
Barbara Hammer 106

Scaling up Spectral Clustering: The Case of Sparse Data Graphs
Claire Monteleoni 107

Participants 109

3 Overview of Talks

3.1 Linear Algebraic Structure of Word Meanings

Sanjeev Arora (Princeton University, US)

License © Creative Commons BY 3.0 Unported license
© Sanjeev Arora

Joint work of Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, Andrej Risteski, Kiran Vodrahalli

What does a word – or more generally, a piece of text – mean? While a precise answer is difficult, many approaches involve a distributional view of semantics. I will give a 30-min survey of this area focusing on use of word embeddings. Our papers give theoretical explanations of why word embeddings exhibit linear algebraic structure even though they are derived from nonlinear methods. A more recent discovery of ours shows that different senses of a polysemous words reside in linear superposition inside the word embedding, which has implications for use of word embeddings in linguistics tasks as well as fMRI studies of the brain, as I'll sketch.

Based upon joint works with Yuanzhi Li, Yingyu Liang, Tengyu Ma, Andrej Risteski, Kiran Vodrahalli.

3.2 Interactive Clustering

Pranjal Awasthi (Rutgers University – New Brunswick, US)

License © Creative Commons BY 3.0 Unported license
© Pranjal Awasthi

Joint work of Pranjal Awasthi, Maria-Florina Balcan, Konstantin Voevodski

Main reference P. Awasthi, M.-F. Balcan, K. Voevodski, “Local algorithms for interactive clustering”, arXiv:1312.6724 [cs.DS], 2014.

URL <https://128.84.21.199/abs/1312.6724v2>

Clustering is typically studied in the unsupervised learning setting. But in many applications, such as personalized recommendations, one cannot reach the optimal clustering without interacting with the end user. In this talk, I will describe a recent framework for interactive clustering with human in the loop. The algorithm can interact with the human in stages and receive limited, potentially noisy feedback to improve the clustering. I will present our preliminary results in this model and mention open questions.

3.3 Two recent clustering paradigms

Shai Ben-David (University of Waterloo, CA)

License © Creative Commons BY 3.0 Unported license
© Shai Ben-David

Joint work of Hassan Ashtiani, Shrinu Kushagra, Shai Ben-David

We consider two paradigms for semi supervised clustering. In the first, [1] the learner is allowed to interact with a domain expert, asking whether two given instances belong to the same cluster or not. We study the query and computational complexity of clustering in this framework. We consider a setting where the expert conforms to a center-based clustering with a notion of margin. We show that there is a trade off between computational complexity

and query complexity; We prove that for the case of k-means clustering (i.e., when the expert conforms to a solution of k-means), having access to relatively few such queries allows efficient solutions to otherwise NP-hard problems. In the second framework, [2], we ask the domain expert to cluster a small subset of the input data and use it to learn a metric over which k-means clustering conforms with that sample clustering. We analyze the sample complexity of that paradigm.

References

- 1 Hassan Ashtiani, Shrinu Kushagra and Shai Ben-David. *Clustering with Same-Cluster Queries*. Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS'16) 2016.
- 2 Hassan Ashtiani and Shai Ben-David. *Representation Learning for Clustering: A Statistical Framework*. Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence (UAI) 2015.

3.4 Questions in Representation Learning

Olivier Bousquet (Google Switzerland – Zürich, CH)

License © Creative Commons BY 3.0 Unported license
© Olivier Bousquet

Recent successes of Deep Learning seem to rely on the ability to automatically extract and exploit structure in the data. But this process is not well understood and often ignored in theoretical analyses where the input data is treated as points in a space with some given similarity measure (which may not fully capture the internal structure of these points). However, by taking a generative point of view one can try and uncover some of the input data structure. This has led to many surprising results in image and text processing. This talk attempts to frame several recent algorithms as conditional generative density estimation and present some theoretical questions that can lead to a better understanding of representation learning.

3.5 Active Learning Beyond Label Feedback

Kamalika Chaudhuri (University of California – San Diego, US)

License © Creative Commons BY 3.0 Unported license
© Kamalika Chaudhuri
Joint work of Chicheng Zhang
Main reference C. Zhang, K. Chaudhuri, “Active Learning from Weak and Strong Labelers”, arXiv:1510.02847v2 [cs.LG], 2015.
URL <https://arxiv.org/abs/1510.02847v2>

An active learner is given a hypothesis class, a large set of unlabeled examples and the ability to interactively query labels of a subset of them; the learner’s goal is to learn a hypothesis in the class that fits the data well by making as few label queries as possible. While active learning can yield considerable label savings in the realizable case – when there is a perfect hypothesis in the class that fits the data – the savings are not always as substantial when labels provided by the annotator may be noisy or biased. Thus an open question is whether more complex feedback can help active learning in the presence of noise.

In this talk, I will present a feedback mechanism – when the active learner has access to a weak and a strong labeler – and talk about when it can help reduce the label complexity of active learning. If time permits, I will also discuss active learning when the annotator can say “I don’t know” instead of providing an incorrect label.

3.6 A cost function for similarity-based hierarchical clustering

Sanjoy Dasgupta (University of California – San Diego, US)

License © Creative Commons BY 3.0 Unported license
© Sanjoy Dasgupta

Main reference S. Dasgupta, “A cost function for similarity-based hierarchical clustering”, in Proc. of the 48th Annual ACM Symp. on Theory of Computing (STOC 2016), pp. 118–127, ACM, 2016.

URL <https://doi.org/10.1145/2897518.2897527>

The development of algorithms for hierarchical clustering has been hampered by a shortage of precise objective functions. To help address this situation, we introduce a simple cost function on hierarchies over a set of points, given pairwise similarities between those points. We show that this criterion behaves sensibly in canonical instances and that it admits a top-down construction procedure with a provably good approximation ratio.

We show, moreover, that this procedure lends itself naturally to an interactive setting in which the user is repeatedly shown snapshots of the hierarchy and makes corrections to these.

3.7 Two sample tests for large random graphs

Debarghya Ghoshdastidar (Universität Tübingen, DE)

License © Creative Commons BY 3.0 Unported license
© Debarghya Ghoshdastidar

Joint work of Debarghya Ghoshdastidar, Ulrike von Luxburg

Standard two-sample tests can achieve a high test power in the presence of large number of samples, but little is known about their performance in the small sample regime. On the other hand, it is well known that a large random graph usually concentrates about its expected (population) version. One can exploit this fact to devise two sample tests for large (inhomogeneous Erdos-Renyi) random graphs, for which a high test power can be achieved with a small population of graphs. In this talk, we will look into different variations of the problem, and present some simple tests based on matrix concentration inequalities.

3.8 Globally Optimal Training of Generalized Polynomial Neural Networks with Nonlinear Spectral Methods

Matthias Hein (Universität des Saarlandes, DE)

License © Creative Commons BY 3.0 Unported license
© Matthias Hein

Joint work of Antoine Gautier, Matthias Hein, Quynh Nguyen Ngoc

We show that a particular class of non-standard feedforward neural networks can be trained globally optimal under relatively mild conditions on the data. The nonlinear spectral method

has a linear convergence rate and the conditions for global optimality can be easily checked before running the algorithm. While the algorithm can in principle be applied to neural networks of arbitrary depth, we present in the talk for simplicity the results for a one hidden layer network. The proof is based on a novel kind of Perron-Frobenius-type theorem for nonlinear eigenproblems. First experimental results show that the resulting classifiers are competitive with standard methods.

References

- 1 A. Gautier, Q. Nguyen Ngoc, M. Hein. *Globally Optimal Training of Generalized Polynomial Neural Networks with Nonlinear Spectral Methods*. Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS'16) 2016.

3.9 Multicriterion cluster validation

Christian Hennig (University College London, GB)

License © Creative Commons BY 3.0 Unported license
© Christian Hennig

Main reference C. Hennig, “Clustering strategy and method selection”, in Handbook of Cluster Analysis, pp. 703–730, Chapman & Hall/CRC, 2015.

Cluster validity measurement is the evaluation of the quality of a clustering, which is often used for comparing different clusterings on a dataset, stemming from different methods or with different parameters such as the number of clusters.

There are various measurements for cluster validity. Often these are used in such a way that the validity of the whole clustering is measured by a single number such as the Average Silhouette Width. But the quality of a clustering has various aspects such as within-cluster homogeneity, between-cluster separation, representation of cluster members by a centroid object, stability or within-cluster normal distribution shape, and what is most important depends on the aim of clustering. Furthermore, in many clusterings, various aspects of cluster validity differ between clusters.

In this presentation I will discuss a number of measurements of different aspects of cluster validity, partly to be evaluated for every single cluster, including some plots to summarize the measurements. A key aspect is calibration, i.e., making different measurements comparable, so that they can be used, for example, to compare different numbers of clusters. The proposed approach is to explore the variation of the index over several clusterings of the same dataset that can be generated by random clustering methods called “stupid k-means” (i.e., assigning points to a random set of centroids) or “stupid nearest neighbor” (i.e., adding nearest neighbors starting from random points).

3.10 What are the true clusters?

Christian Hennig (University College London, GB)

License © Creative Commons BY 3.0 Unported license
© Christian Hennig

In much of the literature on cluster analysis there is the implicit assumption that in any situation in which cluster analysis is applied, there are some “true” clusters at which the analysis aims; and usually the “true” clustering is assumed to be unique. Benchmarking of

clustering algorithms usually is based on datasets with some assumed truth, so that it can be seen how well this truth is recovered by the algorithms.

I will argue that there are several legitimate clusterings on the same data and that defining “true” clusters is highly problematic.

I will discuss a number of related issues: philosophical background, constructive and realist aims of clustering, and various ways to define “true clusters”, namely based on the data alone, on an underlying true class variable, or on probability models. Implications for cluster benchmarking and variable selection in clustering are also mentioned.

3.11 Meta-unsupervised-learning: a supervised approach to unsupervised learning

Adam Tauman Kalai (Microsoft New England R&D Center – Cambridge, US)

License  Creative Commons BY 3.0 Unported license
© Adam Tauman Kalai

Joint work of Adam Tauman Kalai, Vikas Garg

Unsupervised Learning (UL) and exploratory data analysis remain one of the murkiest areas within machine learning. Theorists debate the objective of UL, and for many practical UL problems, humans dramatically outperform ML systems using prior experience in UL and prior domain knowledge or common sense acquired from prior ML tasks.

We introduce the problem of meta-unsupervised-learning from a distribution of related or unrelated learning problems. We present simple agnostic models and algorithms illustrating how the meta approach circumvents impossibility results for novel “meta” problems such as meta-clustering, meta-outlier-removal, meta-feature-selection, and meta-embedding. We also present empirical results showing how the meta approach improves over standard UL techniques for these problems of outlier removal, choosing the number of clusters and a UL neural network that learns from prior supervised classification problems drawn from the openml collection of learning problems.

3.12 Planted Gaussian Problem: Beating the Spectral Bound

Ravindran Kannan (Microsoft Research India – Bangalore, IN)

License  Creative Commons BY 3.0 Unported license
© Ravindran Kannan

Joint work of Ravi Kannan, Santosh Vempala

Main reference R. Kannan, S. Vempala, “Chi-squared Amplification: Identifying Hidden Hubs”, arXiv:1608.03643v2 [cs.LG], 2016.

URL <https://arxiv.org/abs/1608.03643v2>

Spectral methods can find a planted clique of size $c\sqrt{n}$ in a random graph. In spite of some effort, this is the best we know so far. Here, for a different natural problem (of a similar flavor), we show that we can do better than spectral methods.

Given an n times n matrix with i.i.d. $N(0, 1)$ entries everywhere except a planted k by k submatrix which has $N(0, \sigma^2)$ entries, we show that if $\sigma^2 > 2$, then we can find a planted clique of size $o(\sqrt{n})$. We also show that if $\sigma^2 \leq 2$, no poly time Statistical algorithm can find the planted part if it is $o(\sqrt{n})$ sized. The algorithm as well as the lower bound are based on the chi-squared distance between the planted and ground densities. Some extensions will be discussed.

3.13 Recent work on clustering and mode estimation with kNN graphs

Samory Kpotufe (Princeton University, US)

License © Creative Commons BY 3.0 Unported license
© Samory Kpotufe

Joint work of Heinrich Jiang, Samory Kpotufe

Main reference H. Jiang, S. Kpotufe, “Modal-set estimation with an application to clustering”, arXiv:1606.04166v1 [stat.ML], 2016.

URL <https://arxiv.org/abs/1606.04166v1>

We present a first procedure that can estimate – with statistical consistency guarantees – any local-maxima of a density, under benign distributional conditions. The procedure estimates all such local maxima, or *modal-sets*, of any bounded shape or dimension, including usual point-modes. In practice, modal-sets can arise as dense low-dimensional structures in noisy data, and more generally serve to better model the rich variety of locally-high-density structures in data. The procedure is then shown to be competitive on clustering applications, and moreover is quite stable to a wide range of settings of its tuning parameter.

3.14 Proving clusterability

Marina Meila (University of Washington, US)

License © Creative Commons BY 3.0 Unported license
© Marina Meila

Joint work of Marina Meila, Yali Wan

Main reference M. Meila, Y. Wan, “Graph Clustering: Block-models and model free results”, in Proc. of the 30th Annual Conf. on Neural Information Processing Systems (NIPS’16), 2016.

URL <http://papers.nips.cc/paper/6140-graph-clustering-block-models-and-model-free-results>

Main reference M. Meila, “The stability of a good clustering”, Technical Report, 2011.

URL <http://www.stat.washington.edu/research/reports/2014/tr624.pdf>

Clustering graphs under the Stochastic Block Model (SBM) and extensions are well studied. Guarantees of correctness exist under the assumption that the data is sampled from a model. In this paper, we propose a framework, in which we obtain “correctness” guarantees without assuming the data comes from a model. The guarantees we obtain depend instead on the statistics of the data that can be checked. We also show that this framework ties in with the existing model-based framework, and that we can exploit results in model-based recovery, as well as strengthen the results existing in that area of research.

3.15 On Resilience in Graph Coloring and Boolean Satisfiability

Lev Reyzin (University of Illinois at Chicago, US)

License © Creative Commons BY 3.0 Unported license
© Lev Reyzin

Joint work of Jeremy Kun, Lev Reyzin

Main reference J. Kun, L. Reyzin, “On Coloring Resilient Graphs”, arXiv:1402.4376v2 [cs.CC], 2016.

URL <https://arxiv.org/abs/1402.4376v2>

Inspired by notions of stability arising in the clustering literature, I will introduce a new definition of resilience for constraint satisfaction problems, with the goal of more precisely determining the boundary between NP-hardness and the existence of efficient algorithms for resilient instances. In particular, I will examine r -resiliently k -colorable graphs, which are those k -colorable graphs that remain k -colorable even after the addition of any r new edges. I will also discuss the corresponding notion of resilience for k -SAT. This notion of resilience suggests an array of open questions for graph coloring and other combinatorial problems.

3.16 Active Nearest-Neighbor Learning in Metric Spaces

Sivan Sabato (Ben Gurion University – Beer Sheva, IL)

License © Creative Commons BY 3.0 Unported license
© Sivan Sabato

Joint work of Aryeh Kontorovich, Sivan Sabato, Ruth Uner

Main reference A. Kontorovich, S. Sabato, R. Uner, “Active Nearest-Neighbor Learning in Metric Spaces”, in Proc. of the 30th Annual Conf. on Neural Information Processing Systems (NIPS’16); pre-print available at arXiv:1605.06792v2 [cs.LG], 2016.

URL <https://papers.nips.cc/paper/6100-active-nearest-neighbor-learning-in-metric-spaces>

URL <https://arxiv.org/abs/1605.06792v2>

We propose a pool-based non-parametric active learning algorithm for general metric spaces, which outputs a nearest-neighbor classifier. We give prediction error guarantees that depend on the noisy-margin properties of the input sample, and are competitive with those obtained by previously proposed passive learners. We prove that the label complexity of the new algorithm is significantly lower than that of any passive learner with similar error guarantees. Our algorithm is based on a generalized sample compression scheme and a new label-efficient active model-selection procedure.

Sivan Sabato is supported by the Lynne and William Frankel Center for Computer Science.

3.17 Aversion k-clustering: How constraints make clustering harder

Melanie Schmidt (Universität Bonn, DE)

License © Creative Commons BY 3.0 Unported license
© Melanie Schmidt

Joint work of Melanie Schmidt, Anupam Gupta, Guru Guruganesh

Main reference A. Gupta, G. Guruganesh, M. Schmidt, “Approximation Algorithms for Aversion k-Clustering via Local k-Median”, in Proc. of the 43rd Int’l Colloquium on Automata, Languages, and Programming (ICALP 2016), LIPIcs, Vol. 55, pp. 66:1–66:13, Schloss Dagstuhl – Leibniz-Zentrum fuer Informatik, 2016.

URL <http://dx.doi.org/10.4230/LIPIcs.ICALP.2016.66>

There is a huge body of work on approximating clustering problems like k-median or k-means in their standard form. Less is known about the approximability of these problems once we constraint the possible solutions by, e.g., adding lower or upper bounds on the capacities of the facilities. This talk is about a side constraint that we name locality. It assumes that facilities have radii and demands that a client can only connect to a facility if it is within the facility’s radius. We see how a clustering problem from game theory inspires a k-median problem with this type of constraint. This local k-median problem turns out to be surprisingly hard to approximate.

3.18 Gradient descent for sequential analysis operator learning

Karin Schnass (Universität Innsbruck, AT)

License © Creative Commons BY 3.0 Unported license
© Karin Schnass

Joint work of Michael Sandbichler, Karin Schnass

We will shortly present ongoing work on analysis operator learning. We will describe the concept of co-sparsity in an analysis operator as dual concept to sparsity in a dictionary.

Based on this duality we will then propose optimization principles and associated algorithms for learning such an operator. We will show some recent results and discuss the difficulties that arise with a theoretical treatment and practical applications.

3.19 Towards an Axiomatic Approach to Hierarchical Clustering of Measures

Ingo Steinwart (Universität Stuttgart, DE)

License © Creative Commons BY 3.0 Unported license
© Ingo Steinwart

Joint work of Philipp Thomann, Ingo Steinwart, Nico Schmid

Main reference P. Thomann, I. Steinwart, N. Schmid, “Towards an axiomatic approach to hierarchical clustering of measures”, *J. of Machine Learning Research*, Vol. 16, pp. 1949–2002, 2015.

URL <http://www.jmlr.org/papers/volume16/thomann15a/thomann15a.pdf>

We propose some axioms for hierarchical clustering of probability measures and investigate their ramifications. The basic idea is to let the user stipulate the clusters for some elementary measures. This is done without the need of any notion of metric, similarity or dissimilarity. Our main results then show that for each suitable choice of user-defined clustering on elementary measures we obtain a unique notion of clustering on a large set of distributions satisfying a set of additivity and continuity axioms.

3.20 On some properties of MMD and its relation to other distances

Ilya Tolstikhin (MPI für Intelligente Systeme – Tübingen, DE)

License © Creative Commons BY 3.0 Unported license
© Ilya Tolstikhin

Joint work of Carl-Johann Simon-Gabriel, Ilya Tolstikhin

Maximum Mean Discrepancy (MMD) is a metric defined on the class of probability measures and induced by a positive-definite reproducing kernel. In the recent years MMD was getting more and more attention in the ML community. In this short talk I will discuss several results on MMD, including its relation to other stronger distances like Hellinger and Total-Variation, and try to outline some of important questions for the future research.

3.21 Lifelong Learning with Weighted Majority Votes

Ruth Urner (MPI für Intelligente Systeme – Tübingen, DE)

License © Creative Commons BY 3.0 Unported license
© Ruth Urner

Joint work of Anastasia Pentina, Ruth Urner

Main reference A. Pentina, R. Urner, “Lifelong Learning with Weighted Majority Votes”, in *Proc. of the 30th Annual Conf. on Neural Information Processing Systems (NIPS’16)*, 2016.

URL <https://papers.nips.cc/paper/6095-lifelong-learning-with-weighted-majority-votes>

Better understanding of the potential benefits of information transfer and representation learning is an important step towards the goal of building intelligent systems that are able to persist in the world and learn over time. In this talk, we discuss possible directions

for evaluating representation learning within the framework of statistical learning theory. We then focus on learning a representation from a sequence of tasks in a lifelong learning framework. We consider a setting where the learner encounters a stream of tasks but is able to retain only limited information from each encountered task, such as a learned predictor. In contrast to most previous works analyzing this scenario, we do not make any distributional assumptions on the task generating process. Instead, we formulate a complexity measure that captures the diversity of the observed tasks. We provide a lifelong learning algorithm with error guarantees for every observed task (rather than on average). We show sample complexity reductions in comparison to solving every task in isolation in terms of our task complexity measure. Further, our algorithmic framework can naturally be viewed as learning a representation from encountered tasks with a neural network.

3.22 A Modular Theory of Feature Learning

Robert C. Williamson (Australian National University)

License © Creative Commons BY 3.0 Unported license
© Robert C. Williamson

Joint work of Daniel McNamara, Cheng Soon Ong, Robert C. Williamson

Main reference D. McNamara, C. S. Ong, R. C. Williamson, “A Modular Theory of Feature Learning”, arXiv:1611.03125v1 [cs.LG], 2016.

URL <https://arxiv.org/abs/1611.03125v1>

Learning representations of data, and in particular learning features for a subsequent prediction task, has been a fruitful area of research delivering impressive empirical results in recent years. However, relatively little is understood about what makes a representation ‘good’. We propose the idea of a risk gap induced by representation learning for a given prediction context, which measures the difference in the risk of some learner using the learned features as compared to the original inputs. We describe a set of sufficient conditions for unsupervised representation learning to provide a benefit, as measured by this risk gap. These conditions decompose the problem of when representation learning works into its constituent parts, which can be separately evaluated using an unlabeled sample, suitable domain-specific assumptions about the joint distribution, and analysis of the feature learner and subsequent supervised learner. We provide two examples of such conditions in the context of specific properties of the unlabeled distribution, namely when the data lies close to a low-dimensional manifold and when it forms clusters. We compare our approach to a recently proposed analysis of semi-supervised learning.

4 Open problems

4.1 Valid cost functions for nonlinear dimensionality reduction

Barbara Hammer (Universität Bielefeld, DE)

License © Creative Commons BY 3.0 Unported license
© Barbara Hammer

Nonlinear dimensionality reduction techniques have made great strides in recent years [1], and ready-to-use techniques such as the popular t-distributed stochastic neighbor embedding and efficient approximations enable a fast inspection of structure which is inherent in big

data sets [2]. These methods are not only used for interactive data inspection in striking applications e.g. from bioinformatics [3], but they have also proved valuable as a preprocessing step for high dimensional data clustering [4]. In the presentation, we will demonstrate its use for the automated contamination detection in single-cell-sequencing, an important first step in the automated analysis of data as occur in this extremely promising biotechnology [5]. Despite their popularity, however, means of their formal quantitative evaluation are yetr lacking. One of the probably most popular quantitative evaluation methods of nonlinear dimensionality reduction is offered by the quality framework, which quantifies the degree of neighborhood preservation of a nonlinear dimensionality reduction method in terms of a single number [6]. We will formally introduce this measure, and we will argue why it is not suited as a loss function for the evaluation of nonlinear dimensionality reduction in a learning-theoretical sense. Hence, up to our knowledge, it is open how to define a cost term for nonlinear nonparametric dimensionality reduction based on a finite set of data in such a way that it extends to a natural generalization if the number of data points is not fixed.

References

- 1 Andrej Gisbrecht, Barbara Hammer: Data visualization by nonlinear dimensionality reduction. *Wiley Interdisc. Rev.: Data Mining and Knowledge Discovery* 5(2):51–73 (2015)
- 2 Laurens van der Maaten: Barnes-Hut-SNE. *CoRR* abs/1301.3342 (2013)
- 3 Laczny CC, Pinel N, Vlassis N, Wilmes P. Alignment-free visualization of metagenomic data by nonlinear dimension reduction. *Sci Rep.* 2014; 4:4516.
- 4 Automated Contamination Detection in Single-Cell Sequencing, Markus Lux, Barbara Hammer, Alexander Sczyrba *bioRxiv* 020859; <http://dx.doi.org/10.1101/020859>
- 5 Single-cell genome sequencing: current state of the science, Charles Gawad, Winston Koch, and Stephen R. Quake, *Nature Reviews Genetics* 17, 175–188 (2016)
- 6 John Aldo Lee, Michel Verleysen: Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters* 31(14):2248–2257 (2010)

4.2 Scaling up Spectral Clustering: The Case of Sparse Data Graphs

Claire Monteleoni (George Washington University – Washington, D.C., US)

License © Creative Commons BY 3.0 Unported license
© Claire Monteleoni

Joint work of Mahesh Mohan, Claire Monteleoni

Main reference M. Mohan, C. Monteleoni, “Effect of Uniform Sampling on Spectral Clustering”, manuscript, 2016.

Main reference M. Mohan, C. Monteleoni, “A Novel Sampling Algorithm for Speeding Up the Nystrom Approximation”, manuscript, 2016.

Main reference A. Choromanska, T. Jebara, H. Kim, M. Mohan, C. Monteleoni, “Fast spectral clustering via the Nystrom method”, in *Proc. of the 24th Int’l Conf. on Algorithmic Learning Theory (ALT 2013)*, LNCS, Vol. 8139, pp. 367–381, Springer, 2013.

URL http://dx.doi.org/10.1007/978-3-642-40935-6_26

While spectral methods for the unsupervised learning tasks of clustering and embedding have found wide success in a variety of practical applications, scaling them up to large data sets poses significant computational challenges. In particular, the storage and computation needed to handle the affinity matrix (a matrix of pairwise similarities between data points) can be prohibitive. An approach that has found promise is to instead approximate this matrix in some sense. In past work, we analyzed a variant of spectral clustering that uses the Nystrom approximation method, in which the columns are sampled uniformly. Exploiting a strong assumption of latent structure, namely that the (original) affinity matrix can be represented as block-diagonal with k blocks (or a perturbation of such), we provided bounds

on how well the clustering result approximates the result using the full-dimensional affinity matrix, with respect to the normalized-cut spectral clustering objective with k clusters.

We first pose an open question as to whether it is possible to design a sampling technique that performs better than uniform sampling, in terms of managing the tradeoff between its space and time complexity vs. the quality of the approximation. We recently provided a rejection sampling technique that addresses this goal by storing fewer and more “informative” columns. In experiments on a variety of real and synthetic data sets, our technique was able to speed up the computation and reduce the memory requirements of spectral methods, while simultaneously providing better approximations. Our observation that sparser data matrices led to decreased performance, not only for our rejection sampling technique but also for the standard uniform sampling, leads to a second open question: how to improve uniform sampling in the sparse case.

[Update: while interesting points were raised in the Dagstuhl Seminar discussion, for example, that if the matrix is sparse enough, one can avoid such sampling techniques altogether, there is still a continuum of sparsity levels which future work can address. On another note, it is worth further exploring which types of approximation guarantee on the affinity matrix imply good approximation of various spectral clustering objectives.]

Participants

- Sanjeev Arora
Princeton University, US
- Pranjal Awasthi
Rutgers University –
New Brunswick, US
- Shai Ben-David
University of Waterloo, CA
- Olivier Bousquet
Google Switzerland – Zürich, CH
- Kamalika Chaudhuri
University of California –
San Diego, US
- Sanjoy Dasgupta
University of California –
San Diego, US
- Debarghya Ghoshdastidar
Universität Tübingen, DE
- Barbara Hammer
Universität Bielefeld, DE
- Matthias Hein
Universität des Saarlandes, DE
- Christian Hennig
University College London, GB
- Adam Tauman Kalai
Microsoft New England R&D
Center – Cambridge, US
- Ravindran Kannan
Microsoft Research India –
Bangalore, IN
- Samory Kpotufe
Princeton University, US
- Marina Meila
University of Washington –
Seattle, US
- Claire Monteleoni
George Washington University –
Washington, D.C., US
- Lev Reyzin
University of Illinois –
Chicago, US
- Heiko Röglin
Universität Bonn, DE
- Sivan Sabato
Ben Gurion University –
Beer Sheva, IL
- Melanie Schmidt
Universität Bonn, DE
- Karin Schnass
Universität Innsbruck, AT
- Hans Ulrich Simon
Ruhr-Universität Bochum, DE
- Christian Sohler
TU Dortmund, DE
- Ingo Steinwart
Universität Stuttgart, DE
- Ilya Tolstikhin
MPI für Intelligente Systeme –
Tübingen, DE
- Ruth Urner
MPI für Intelligente Systeme –
Tübingen, DE
- Ulrike von Luxburg
Universität Tübingen, DE
- Robert C. Williamson
Australian National
University, AU

