Mach Learn (2008) 72: 1–4
DOI 10.1007/s10994-008-5062-x

Guest Editors' Introduction: Special issue on Learning Theory (COLT-2007)

Nader H. Bshouty · Claudio Gentile

Received: 12 May 2008 / Accepted: 14 May 2008 / Published online: 28 May 2008

This special issue of *Machine Learning* is dedicated to the 20th Annual Conference on Learning Theory (previously known as the Conference on Computational Learning Theory), held in San Diego, CA, USA, June 13–15, 2007, as part of the 2007 Federated Computing Research Conference (FCRC). The authors of seven papers were invited to submit expanded versions of their conference papers. These papers then went through the standard reviewing process of *Machine Learning*. The papers have mostly been selected according to theoretical significance, with an eye towards their potential ability to influence future research directions within the context of Learning Theory.

Learning Theory is an influential field of Machine Learning aimed at providing a theoretical underpinning to the basic intuition of what *learning* really means, but also a more "readily accessible" mathematical support to learning systems applied to a wide range of domains. Stated differently, whereas the intuitive notion of "learning" can be formalized in many different ways, the application domain one has in mind does play a fundamental role in suggesting specific learning models, mathematical assumptions, and performance measures. Hence, Learning Theory is more like a collection of mathematical theories for Machine Learning, whose major part concerns the formal quantification of the performance of algorithms operating within the assumed models.

Altogether the papers in this special issue represent a snapshot of a variety of current lines of research in theoretical aspects of Machine Learning, ranging from on-line learning of individual sequences to active learning models, from Statistical Learning to Inductive Inference models.

Below we briefly introduce each of the papers, and provide some background information.

N.H. Bshouty (✉)
Department of Computer Science, Technion, Haifa 32000, Israel
e-mail: bshouty@cs.technion.ac.il

C. Gentile
Dipartimento di Informatica e Comunicazione, Università dell'Insubria, Varese, Italy
e-mail: claudio.gentile@uninsubria.it

Consider an input stream consisting of numerical "readings" or "observations" from an integer range $\{1, \ldots, n\}$, with each reading labeled as either "blue" or "red". This implicitly defines two frequency distributions (and hence, empirical probability distributions), a blue distribution and a red distribution. Consider the problem of estimating the "distance" between the red and the blue distributions, using only a small amount of space, much less than the length of the input stream.

Estimation of distances allows us to construct approximate representations, e.g., histograms, wavelets, Fourier summaries, or equivalently, find models of the input stream. It also has applications in pattern matching, image analysis and statistical learning.

There are various natural definitions for "distance". Two broad classes of such distances are f-divergences and Bregman divergences. These include, as special cases, such well-studied notions as ℓ_1, ℓ_2 and ${\ell_2}^2$ distances, Hellinger distance and relative entropy (Kullback-Leibler divergence). It has been known that ℓ_1 and ℓ_2 distances admit efficient (i.e., small space) algorithms (i.e., sketchable) whereas certain other natural distances do not (not sketchable).

In "Sketching Information Divergences", by Guha, Indyk and McGregor, the authors prove two general results showing that f-divergences are not sketchable except for ℓ_1, and that Bregman divergences are not sketchable except for ${\ell_2}^2$. They then characterize a large family of distances that cannot be sketched. They also present data-stream algorithms for the additive approximation of a wide range of information divergences.

In "Regret to the best vs. regret to the average", by Even-Dar et al., the authors investigate the well-known setup of online regret minimization for individual sequences (the "expert setting"). In this sequential setting, the learning algorithm is asked to compute a distribution over a set of experts, and receives a gain equal to the average instantaneous gain over the experts. The paper considers different notions of regret: the standard regret to the best expert, and regrets to averages of all experts. This investigation gives new insights into the properties and limitations of known no-regret algorithms, but also provides new regret-minimization procedures, whose performances, according to combined regret minimization criteria, are more satisfactory. For instance, the paper shows that any algorithm (like the standard exponential weighting scheme) suffering cumulative regret $O(\sqrt{T})$ to the best expert is forced to suffer $\Omega(\sqrt{T})$ regret to the average. Alternative algorithms are then analyzed whose regret to the best is $O(\sqrt{T}\log T)$ and have only constant (i.e., independent of T) cumulative regret to any fixed distribution over the experts.

It is known that the exponential weighting scheme we mentioned above can also be applied in statistical learning settings. In the paper "Aggregation by exponential weighting, sharp oracle inequalities and sparsity", by Dalalyan and Tsybakov, the authors consider the problem of aggregating predictors operating under the square loss, in the case when instance points are nonrandom, whereas the associated labels are statistically related via i.i.d. random noise. Sharp PAC-Bayesian risk bounds are proven when the aggregation scheme is based on exponential weights under general assumptions on the functions to aggregate. Several new sharp oracle inequalities (i.e., regret inequalities with leading constant 1 and optimal rate in the remainder term) are derived for the aggregated predictor. These results are then used to prove sparsity inequalities. The latter bounds are very interesting, and meaningfully apply when the optimal linear aggregation is made up of a small number of predictors. To this end, special care is taken into the choice of the prior distribution governing the PAC-Bayesian averages.

The paper "U-Shaped, Iterative, and Iterative with Counter Learning", by Case and Moelius, studies U-Shaped Learning and iterative learning. U-Shape learning has been observed in cognitive science as a typical behavior of human learning within the setting of inductive inference of formal languages from positive data. In U-Shaped Learning the learner

first *learns* (e.g., a child that learns that the past tense of "speak" is "spoke"), then *unlearn* (the child over-regularize and incorrectly uses "speaked") and finally *relearns* (the child returns to correctly using "spoke"). In Iterative learning each of a learner's conjectures can depend only upon the learner's most recent conjecture and input element.

While it has been known that U-shaped learning is not necessary in Gold's model of learning in the limit and necessary in the setting of behaviorally correct learning, it remained open whether or not iterative learners must perform U-shapes when learning languages from positive data. Since humans cannot memorize all examples received during the learning process, this is an important problem.

The authors resolve this problem and show that U-shapes are unnecessary in iterative learning. Furthermore, a new model for iterative learning is introduced, i.e., iterative-with-counter learning. Here, the learner is iterative but has additionally access to a counter telling it how many examples have been presented so far.

In "A theory of learning with similarity functions", by Balcan, Blum, and Srebro, the authors develop a whole theory of learning classification problems through functions that generalize the notion of kernel functions in a quite interesting and practically relevant way.

Kernel functions are one of the most popular tools in Machine Learning. The related theory has by now reached full maturity (as evinced by the publication of many books, and the organization of many events in leading Machine Learning conferences). A kernel function operates by implicitly mapping data points into a very high dimensional space where the algorithms essentially view data only through inner products. These inner products might be seen as a measure of pairwise similarity between data points. For a given classification problem at hand, a "good" kernel function is one which leads to a large margin separation of the data in the mapped space. However, in many cases it might be difficult for a domain expert to use this theory to help design an appropriate kernel. In addition, the requirement of positive semidefiniteness can rule out the most natural similarity functions for the given domain. In their paper, Balcan, Blum, and Srebro develop an alternative and more general theory of learning with similarity functions, delivering sufficient conditions for such functions to allow one to learn satisfactorily in a statistical learning setting. This theory need not refer to implicit high-dimensional spaces, nor does it require the similarity function be positive semidefinite (or even symmetric), and is formulated in terms of more tangible quantities than the standard theory of kernel functions. These results also generalize the standard theory in the sense that any good kernel function under the standard definition can be viewed as a good similarity function under the authors' theory.

In "Learning Large-Alphabet and Analog Circuits with Value Injection Queries" by Angluin, Aspnes, Chen and Reyzin, the authors describe several results on efficient learning algorithms for acyclic discrete circuits with bounded fan-in and alphabet size s using Value Injection Queries (VIQ). In this model the target circuit is specified by a set of n wires and the learner is allowed to specify the values of a subset of the input wires and receives as feedback only the value of the circuit (output wire). VIQ generalizes the notion of Membership Queries in the standard Exact Learning model.

This paper is an improvement of an earlier result by Angluin, Aspnes, Chen and Wu (STOC'06), where an efficient learning algorithm for AC^0 and NC^1 Boolean circuits using VIQs is given. The present paper addresses the question of learning similar circuits with larger alphabet size. The main results of the present paper include: (i) An algorithm for learning a transitively reduced circuit of n wires, fan-in bounded by k, and alphabet size s using $(ns)^{O(k)}$ VIQs. (ii) An algorithm for learning an acyclic circuit with bounded shortcut width b, n input wires, fan-in bounded by k, and alphabet size s using $(ns)^{O(k+b)}$ VIQs. The bounded shortcut width condition on the circuit is a generalization of a transitively reduced circuit (where $b = 0$). (iii) A hardness result for learning circuits with large alphabets

when there are no restrictions on their topologies. (iv) An algorithm for learning an analog $O(\log n)$-depth circuit whose gates satisfy a Lipschitz condition. The learning algorithm outputs a circuit whose "behavior" is close to the target circuit, where the accuracy parameter is a parameter of the learning problem. (v) A learning algorithm for similar circuits in an exact model with Equivalence Queries and VIQs.

In "Robust reductions from ranking to classification", by Balcan et al., the authors consider efficient ways of reducing a ranking problem, whose performance is measured by the so-called AUC (Area Under the operating characteristic Curve) criterion, to a binary classification problem. In the most basic version, in a ranking problem we are given a set of unlabeled data points belonging to two classes 0 and 1, and the goal is to rank all points from class 0 before any point from class 1. In this context, the quantity $1 - AUC$ measures how many pairs of neighboring points would have to be swapped to repair the ranking, normalized by the number of 0s times the number of 1s. The authors consider a statistical setting, and show through nice combinatorial arguments that a given binary classification regret on the induced binary problem can be turned into an AUC regret at most twice as large, thereby proving that a good algorithm for binary classification can be efficiently transformed into one that is good for solving the ranking problem.

We are grateful to the authors for their contributions, to the members of the COLT 2007 Program Committee for their help in selecting a good sample of high-level papers that witness the richness and diversity within the field of Learning Theory, and to the anonymous reviewers for their valuable help in bringing the papers to their current form. We would like to thank the editorial staff of *Machine Learning* for their support, and the Editor-in-Chief, Foster Provost, for the opportunity he gave us to compile this special issue.

Haifa and Varese, May, 2008