Mach Learn (2007) 66:115–118
DOI 10.1007/s10994-007-0753-2

Guest editorial: Learning theory

Olivier Bousquet · André Elisseeff

Published online: 1 February 2007

The theoretical study of the learning ability of machines was initiated in the sixties with the works of Gold, Solomonoff, Vapnik and Chervonenkis among a few others. In almost half a century, this discipline has developed into various directions and several sub-areas have branched out. The notion of a unique theory of learning has disappeared in favor of a set of machine learning theories.[1] Two broad directions have been pursued: the study of tasks or concepts (and their possible representations) that can be learned by a machine, and the study of the properties of particular learning algorithms. Both are tied together since learnability usually is defined as the existence of an algorithm that satisfies certain properties.

Different paradigms have been explored. They differ in the type of assumptions they make about how the data is processed and generated: the data is either processed sequentially, i.e., one piece at a time (online learning, reinforcement learning) or as a set (off-line or batch learning); it can be generated by a deterministic (possibly adversarial) or stochastic process (probability distribution).

The most common paradigm is the one studied in Mathematical Statistics: the independently and identically distributed (i.i.d.) off-line setting where a set of examples is assumed to be sampled independently from a fixed (but unknown to the algorithm) distribution. This is also the model at the heart of *Statistical Learning Theory* (Vapnik, 1982) and Valiant's PAC learning model (Valiant, 1984). This model has been largely explored with various slightly different focuses, e.g., consistency (Devroye, Györfi, & Lugosi, 1996), rates of convergence (Anthony & Bartlett, 1999; Boucheron, Bousquet, & Lugosi, 2005) or computational aspects (Kearns, 1990; Kearns & Vazirani, 1994).

O. Bousquet (✉)
Pertinence, 32, rue des Jeneurs, F-75002 Paris, France
e-mail: obousquet@gmail.com

A. Elisseeff
IBM Zürich Research Lab, Säumerstrasse 4, CH-8803 Rüschlikon, Switzerland
e-mail: ael@zurich.ibm.com

[1] We use the term machines as opposed to humans—indeed, psychologists, cognitive scientists or education researchers also have their own Learning Theories. The term *machine* should be thought of as an *algorithm* since the physical device is of no relevance.

To deal with the case where the examples are presented sequentially, another paradigm called *Online Learning Theory* has been developed. It does not make any assumption about how data is generated. It is often called the adversarial setting as one can imagine that the data is generated by a malicious adversary whose goal is to maximize the errors made by the learning algorithm. Useful references are (Cesa-Bianchi & Lugosi, 2006) and (Vovk, Gammerman, & Shafer, 2005).

When examples are supposed to be presented sequentially and are discrete objects, another paradigm called *Inductive Inference* or *Computational Learning Theory*[2] should be considered (see e.g., Jain et al., 1999). It is particularly well suited to the study of languages and more generally to settings where the data generating process is strongly constrained (e.g., the number of examples is finite, only certain sequence of letters are valid, etc.).

The contributions to this special issue consider the statistical learning theory framework where examples are not provided sequentially.

The paper "Suboptimal Behavior of Bayes and MDL in Classification under Misspecification" by Peter Grünwald and John Langford focuses on the asymptotic consistency of several (known) algorithms. They prove that these algorithms may not be consistent for classification even in situations that appear naturally. The inconsistency comes from a misspecification problem (i.e., the model used by the algorithm does not contain the true distribution generating the data). The key point here is that in the context of classification, the application of Bayesian inference necessarily leads to misspecified probability models because there is no direct (and exact) correspondence between probability distributions and classifiers. This work brings new insights about the limitations of such algorithms and discusses possible workarounds.

In the paper "A New PAC Bound for Intersection-Closed Concept Classes" by Peter Auer and Ronald Ortner, the authors consider the PAC framework where one is interested in the number of examples that are needed for an algorithm to achieve a given accuracy in its predictions. Their result refines and extends several previous bounds: they prove an improved bound for intersection closed concept classes (i.e., a class of binary-valued functions closed by product) of finite Vapnik-Chervonenkis dimension (a combinatorial parameter describing the richness of the class). As it turns out, investigating these refinements yields a deeper understanding of the relationship between the structure of the classes of concepts and their learning complexity. Besides the well-known VC property and the intersection-closed property under which they obtain a better bound than previously known, they are led to introduce a new combinatorial property which gives an even tighter bound that matches the optimal one known for classes of hyper-rectangles. Although not quite the end of the story, this paper is a significant additional step towards the characterization of distribution-free sample complexity of concept classes.

The paper "Model Selection by Bootstrap Penalization for Classification" by Magalie Fromont is deriving finite sample error bounds (equivalent to sample complexity bounds) for model selection, that is the problem of automatically choosing the best class of concepts in a collection of such classes. This problem originates in Vapnik's Structural Risk Minimization idea where it was proposed to minimize the sum of the minimal empirical error in the class and a penalty term accounting for the *complexity* of the class. Vapnik originally proposed a penalty based on the VC dimension. This was later refined and other sharper penalties involving Rademacher averages (see e.g., (Boucheron, Bousquet, & Lugosi, 2005) for definitions and

[2] Note the ambiguity in the naming of these paradigms. PAC-learning also focuses on computational aspects and is sometimes referred as computational learning theory.

examples) were proposed. Fromont presents another kind of penalty based on bootstrap samples, thus making a step towards bridging the gap between theoretical (penalty-based) and practical (bootstrap or cross-validation based) model selection.

In "Optimal Dyadic Decision Trees" by Gilles Blanchard, Christin Schäfer, Yves Rozenholc and Klaus-Robert Müller, the authors propose a comprehensive study of a new algorithm that builds a dyadic partition of the space (a decision tree with the restriction of cutting intervals exactly in the middle) for classification. This study involves both computational and statistical aspects. In particular they show that the algorithm obtains (with very high probability) a dyadic decision tree which is almost as good as the best such tree (given the distribution of the data). They also show the optimality of this algorithm with respect to the learning of certain Hölder classes of functions. Besides these statistical guarantees, the proposed algorithm has the additional advantage of providing a simple and understandable model, which is one of the foremost requirements in many applications areas.

In a slightly different setting the paper "A Framework for Statistical Clustering with Constant Time Approximation Algorithms for K-Median and K-Means Clustering" by Shai Ben-David attacks a problem that has been much less explored. Indeed, the bulk of the work in Statistical Learning Theory has concerned the supervised setting (where examples are pairs and the goal is to find a function that predicts the second element of the pair from the first one) rather than the unsupervised one (where examples are simple objects and the goal is to provide a description of their distribution). One reason is that it is much harder to give formal definitions of the goal to be achieved in this setting. This paper provides an elegant answer to this problem, and explores various aspects (statistical and computational) of the defined framework.

Another paper dealing with the unsupervised setting is "Statistical Properties of Kernel Principal Component Analysis" by Gilles Blanchard, Olivier Bousquet and Laurent Zwald (edited by Nicolò Cesa-Bianchi). The main focus of this paper is the analysis of the convergence in the statistical framework of the reconstruction error of the kernel PCA algorithm. Unlike the clustering algorithms studied in the previous paper which aim at describing the data via a set of clusters, kernel PCA aims at building a low dimensional description of the data (i.e., it produces a small set of features that are non-linear functions of the data). The main novelty is to use techniques that were first applied to the precise analysis of classification algorithms in order to understand, in the non-linear (kernel-based) case, what is the influence of the geometry of the feature space on the rates of convergence.

Finally, in "Feature Space Perspectives for Learning the Kernel" by Charles Micchelli and Massimiliano Pontil, the authors consider the classical setting of regularization-based algorithms which minimize, over a set of functions, a functional involving the empirical error and a norm of the function. They go one step further by considering this minimization over a collection of sets of functions with different associated norms. This problem relates to the automatic choice of the kernel. The main result of the paper shows that a kernel can be automatically tuned by solving a variational problem over the dual of a space of certain continuous functions. Some examples (finitely many kernels, single feature kernels) are then treated in detail showing the applicability and the generality of the main result.

These papers exemplify typical research activities in statistical learning theory from the proof of theoretical bounds to the design of optimal algorithms. They show how diverse theories can be when it comes to the analysis of machine learning. Although apparently disparate, they all aim at using a theoretical framework to either design new algorithms or analyze existing methods. In both cases, they share the same spirit: to understand the nature of learning. From that respect, they should not be considered in isolation from other fields that address a similar problem. Research fields like Information Theory (in particular

compression and universal coding), Game Theory, Logic (Proof Theory and Model Theory in connection with Inductive Logic Programming where concepts are represented using first order logic), Language Theory, Statistical Physics, among others, provide many useful tools to investigate the properties of machine learning algorithms. The variety of approaches also shows that the nature of learning is still far from being understood.

Besides the simple and extensively explored setting of binary classification with i.i.d. data for which a wealth of results has been obtained, many other problems starting from the multi-class extension and going towards other more complex and realistic settings remain largely under-investigated. Machine learning theories have therefore much to do. This is the characteristic of a young field that we expect to be evolving at a fast pace in the coming years. With the advent of new technologies such as sensor networks, pervasive computing or massive storage systems, new settings and new data requiring novel computational approaches will appear. We expect this practical changes to open further directions of research.

References

Anthony, M. & Bartlett, P. L. (1999). *Neural network learning: theoretical foundations*. Cambridge University Press.

Boucheron, S., Bousquet, O., & Lugosi, G. (2005). Theory of classification: a survey of recent advances. *ESAIM: Probability and Statistics*, *9*, 323–375.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge University Press: New York.

Devroye, L., Györfi, L., & Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. Springer: New York.

Jain, S., Osherson, D., Royer, J. S., & Sharma, A. (1999). *Systems that learn*, 2nd edition, MIT Press.

Kearns, M. (1990). *Computational complexity of machine learning*. MIT Press.

Kearns, M., & Vazirani, U. (1994). *An introduction to computational learning theory*. MIT Press.

Valiant, L. G. (1984). A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on theory of computing*. ACM Press: New York.

Vapnik, V. N. (1982). Estimation of dependencies based on empirical data Springer.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. New York: Springer.