Beyond Mitchell: Multi-Objective Machine Learning – minimal entropy, energy and error

Johan Loeckx

Artificial Intelligence Lab
Vrije Universiteit Brussel, Belgium
jloeckx@ai.vub.ac.be

Abstract

In this paper the need for a truly multi-objective view on machine learning processes is put forward. Though methodologies exist that take into account multiple objectives like model complexity or generalization, they eventually serve to achieve higher prediction accuracies only — albeit with good generalization properties. The fundamental conjecture of this paper, yet to be experimentally validated, is that the machine learning strategies that aim to optimize the structural and energetic properties of models, are the processes that lead to hierarchy building and abstraction respectively, reflected ultimately in the internal representations. In this sense, this paper attempts to give impetus to create a holistic vision in which the impact of different algorithms on abstraction and hierarchy can be investigated, especially in the context of Deep Learning.

1 Introduction

Machine Learning is undeniably one of the most successful domains of Artificial Intelligence and has provided a solution to many interesting real-world problems like optical character recognition (OCR), fraud detection, real-time control, search etc. However, some interesting challenges remain, especially what concerns the formation of higher-order concepts and abstract representations. The topic has come under new attention since the coming of so-called “Deep Learning”, a set of techniques and algorithms to model high-level abstractions by constraining model architectures and using adapted training mechanisms [1]. Although finding the right abstractions and learning different levels of representations are central in the discussions about Deep Learning, it is still common practice to follow Tom M. Mitchell’s operational definition of Machine Learning. Unfortunately, this definition does not take computational architectures or inner representations explicitly into account and even previous efforts in multi-objective machine learning considers representational strategies as a means to improve (externally measured) performance rather than a goal on its own [3]. Just like behaviourist approaches ignore the inner processes, a similar phenomenon is too often observed with machine learning. A crucial difference, however, is that computers are measurable.

> The central theorem of this paper is that machine learning is an intrinsically multi-objective process and that the strategies that focus on improving structural, energetic and information-theoretic aspects of internal representations, lead to abstraction and hierarchy building.

2 Link with deep learning

Until recently, machine learning techniques have exploited shallow-structured (and often fixed) architectures, primarily because their behaviour could be understood and training of deeper architectures failed due to a complexity explosion. Recent research in Deep Learning on the contrary, focuses mainly on training algorithms that are adapted to new kinds of ”deep” architectures [2]. As such, they aim to minimize a performance target P in deeply structured models, based on a dataset D:

$$\underset{x}{\text{minimize}} \quad P(M_x, D)$$

M_x being the computer model, x being the model parameter space and D the dataset of experiences. Often a combination of unsupervised (learning from X, not taking into account P) and supervised learning (strict optimization towards minimal P) is employed. Though these new developments meant an important leap forward in Machine Learning and yielded impressive results, there is a fundamental problem with the approach: the algorithms do not explicitly take the structural model properties into account when trying to learn from data.

In fact, the implicit presumptions in Deep Learning algorithms are heuristic strategies to attain specific structural properties, but there is no measured feedback from the output (model structure) to input (decision variables). It is similar to implementing a machine learning algorithm based on heuristics, without measuring the training or test error, without feedback to improve the algorithm's performance. Structural properties are mainly considered a by-product, a (desirable) side effect of the applied training mechanisms as the methodology ignores the *learning process* and internal representation, ignores what is happening "inside" the computer — its computation and memory. One exception is when the "interpretability" of models is discussed [3]. Though existing research on multi-objective machine learning acknowledges its positive effect on performance (e.g. the generalization properties) [3], the other objectives are typically "meta"-properties of the accuracy, like generalization [4].

Our stance is that computation and representation are two sides of the same coin (computation transforms one representation in another) and that learning efficient representations and model structures are thus as important as successfully predicting the external (unseen) data.

3 Extending Mitchell's definition

In other words, finding a more efficient or simple way to represent information or to perform a machine learning task, is a part as central to learning as is predicting correct data. Imagine three classifiers for modelling the XOR operator: one that implements the exact (hierarchical) formula, one based on nearest-neighbour 4 prototypes and a nearest-neighbour algorithm based on all existing samples in the dataset as pictured in Fig. 1. Though any of these methods may deliver the same performance, it is clear that the exact formula or 4-prototype example do a better job in capturing the underlying relationships in the data. There are many advantages to moving away from a purely "external" definition of Machine Learning towards a cognitively inspired operational definition:

> "A computer program is said to learn from experience X if its performance at task in T, as measured by P, improves with experience X or when the internal entropy S or its executing energy E decreases with respect to the training data D associated with experience X".

The author, therefore proposes to formulate machine learning problems as optimization or search process not only to optimize the model performance P, but also the Energy (and information-theoretic) efficiency E and model entropy/hierarchy S:

$$\underset{x}{\text{minimize}} \quad P(M_x, D), E(M_x, D), S(M_x, D)$$

A lot of algorithms already implicitly employ these kind of optimization strategies: using prototypes is a strategy to compact information; regularization a strategy to reducing energy consumption; weight sharing to decrease the complexity / entropy of a model; penalty functions to constrain model complexity; Occam's razor, pruning in decision trees, ...

4 Conclusions and Future Work

The advantages of employing a multi-objective approach to machine learning are manifold. First, the artificial distinction between supervised and unsupervised learning disappears as a particular model is

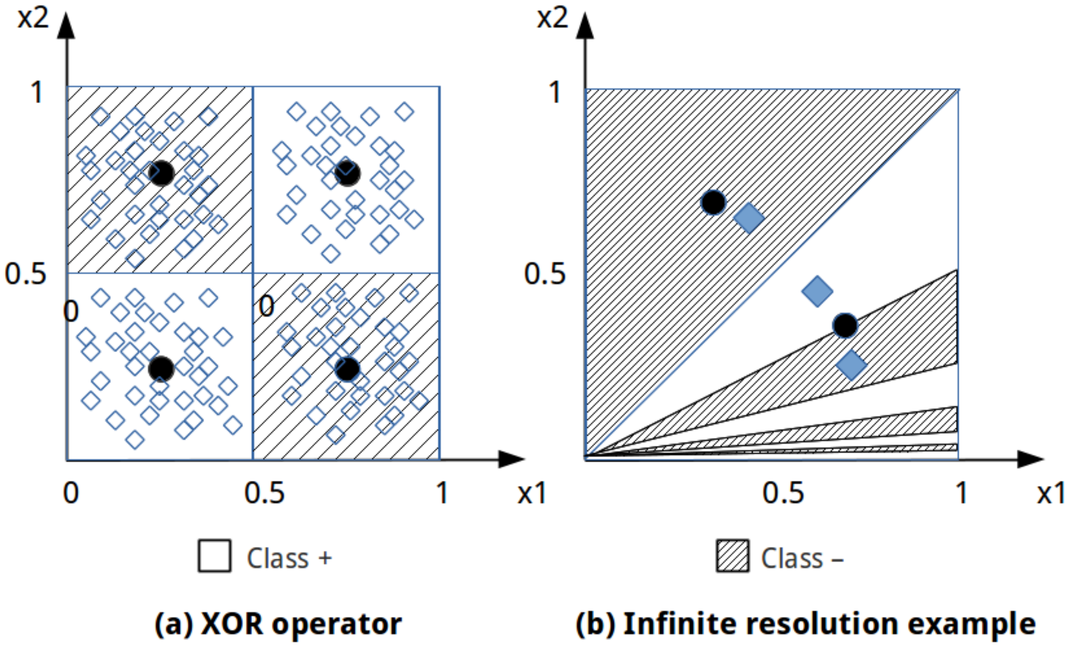

Figure 1: (a) The XOR operator is a prototype example of how a hierarchical composition of two simple classifiers have extensive expressive power. Though the performance is similar for two nearest-neighbour algorithms with a dense grid of prototypes (diamonds) or only 4 (circles) , it is clear that the classifier with four prototypes if preferred because it exhibits a higher abstraction. A traditional comparison of accuracy only does not reveal such information, a situation which becomes dramatic when considering more complex and higher-dimensional problems. Figure (b) illustrates the energy-error trade-off: an “infinite resolution” problem requires ever more prototypes (and thus, energy) for decreasing error if the structural complexity remains equal.

still being optimized, but towards minimal entropy (condensing data, not unlike auto-encoding in Restricted Boltzmann Machines) rather than minimal error. Next, the multi-objective approach allows a more deep comparison between different machine learning models, particularly between different kinds of algorithms. Thirdly, it allows a more focused approach towards deep learning techniques. Lastly, it opens up opportunities to quantize biases in machine learning algorithms.

Clearly, operational evidence to quantify the claims that have been made in this paper is needed. Exploring computational measures to quantify abstract quantities like “entropy” and “energy” will lead to new horizons.

Acknowledgments

This research has been supported by the EU FP7 PRAISE project #318770.

References

[1] Yoshua Bengio. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127, 2009.

[2] Li Deng and Dong Yu. *Deep Learning*. Now Publishers Incorporated, 2014.

[3] Yaochu Jin and Bernhard Sendhoff. Pareto-based multiobjective machine learning: An overview and case studies. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(3):397–415, 2008.

[4] Thorsten Suttorp and Christian Igel. Multi-objective optimization of support vector machines. In *Multi-objective machine learning*, pages 199–220. Springer, 2006.