NIPS2017
Neural Information Processing Systems 2017
Dec 4 , 2017 - Dec 9 , 2017, Long Beach, United States

Reviews For Paper

Track	Blue
Paper ID	3587
Title	Limit of Learning and Efficient Strategies

Masked Reviewer ID: Assigned_Reviewer_1

Review:

Question	
Overall rating:	Clear rejection
Confidence score:	The reviewer is fairly confident that the evaluation is correct
	In this work, the authors suggest the moniker "Limit of Learning" (LoL) for the asymptotic (wrt data) loss of the optimal model within a model class. They propose a generalized version of Takeuchi's information criterion and recapitulate arguments that cross-validation is typically not asymptotically optimal. Finally, they propose an active set-esque algorithm which finds an optimal model under the assumption that the "best" model follows certain assumptions. This paper fails to clearly describe either of its main contributions (described in detail below). Additionally, it omits relevant recent developments in model selection. === Major comments As mentioned above, the authors do not explain their main contributions. In particular, while the authors claim that they generalize TIC, the main text does not make clear in what way this is the case. Likewise, the text offers no intuition on a proof for Theorem 1. While it is fine to put technical details in the supplement, a basic proof sketch would greatly improve the persuasiveness of this section. Similarly, the text never clarifies how the model class A_n is actually constructed. Since Algorithm 1 apparently relies on enumerating all models, this is an important omission. Presumably, this is different for different model spaces (e.g., GLMs compared to NNs). In Section 1, the authors suggest that an example is A_n = {{1}, {1,2}, … {1,2, …, 50}}. however, it is not obvious what these indices represent. Is model i a GLM which considers interactions up to degree i? some other arbitrary model? This is not explained. A second major problem is that the authors appear unaware of relevant recent work. For example, the notion of "bigger model for bigger data" is rigorously analyzed in the context of Bayesian nonparametrics; indeed, this has even been extended to the sequential setting. For example, all of the following references consider this setting. Hu, T. Online Regression with Varying Gaussians and Non-Identical Distributions.

Analysis and Applications, 2011, 9, 395-408.
Nguyen-tuong, D.; Peters, J. R. & Seeger, M. Local Gaussian Process Regression for Real Time Online Model Learning. Advances in Neural Information Processing Systems 22, 2009.
Xiao, H. & Eckert, C. Lazy Gaussian Process Committee for Real-Time Online Regression. Proceedings of the 27^th AAAI Conference on Artificial Intelligence, 2013.
Hughes, M. C. & Sudderth, E. B. Memoized Online Variational Inference for Dirichlet Process Mixture Models. Advances in Neural Information Processing Systems 26, 2013.
Lin, D. Online Learning of Nonparametric Mixture Models via Sequential Variational Approximation. Advances in Neural Information Processing Systems, 2013.

Detailed Comments:

Additionally, work on sequential information theoretic measures, such as sequential normalized maximum likelihood models are also highly relevant recent works which go unmentioned in this manuscript.

While it is possible the current work has some novel contribution, the complete lack of context wrt that existing work makes it very difficult to judge.

=== Minor comments

In general, I believe this paper dwells too long on basic machine learning and model selection concepts (e.g., Figures 1 and 2) at the expense of space for its actual contributions (e.g., what the A_n for the example actually is, the second term of Equation 4, Algorithm 1). Further, in several cases (e.g., Figure 3, the "tax" discussion in Section 3.1), the authors propose analogies which do not clearly connect to their methods. While such colorful language can prove helpful as an introduction in some settings, it does not completely replace the need for at least sketches and concrete intuitions.

Throughout the paper, the notation is sloppy to the point of being confusing. For example, in Definition 2, LoL is described as an "optimal prediction loss", while the definition uses "arg min" over the model class. Is LoL a model or a number? What does "approaches p" mean in Equation 3? What is meant by the "tuples" and "switching" discussed just above Section 3.1?

The paper does not make reference to neural networks other than to describe them in Example 2. Even for non-deep MLPs, naively training a modest-sized set of models is not feasible. As described above, the main contribution of this work is a bit muddled, but it seems to be an information theoretic in nature. If the experimental results on NNs is not important enough to even include in the main manuscript, I am not sure of the benefit of mentioning them.

What sort of assumptions are placed on the data? to what extent is the method affected by non-stationarity? For example, one could imagine a data stream which is complex in the beginning (say, a patient in an emergency shortly after undergoing trauma) but which becomes less complex over time (say, after the patient stabilizes). The assumptions built-in to the proposed approach ("selected models [should] evolve… from the small to the large") seem to fail in such a setting.

	"High dimensionality and limited data size are equivalent." I am not convinced of this claim. The rest of the paper posits that the size of the data affects what might be considered "high dimensionality"; I do not believe these two ideas are equivalent. Why not allow self loops in the graph since the best expert can stay the same anyway? === Typos, etc. "We shall this" -> "We shall call this" "that determine the" -> "that determines the" Table 2 should be an algorithm. The case of the references ("bayesian" in Casella et al., etc.) are not correct.

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Overall rating:	Ok but not good enough - rejection
Confidence score:	The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper
Detailed Comments:	The authors propose a new model selection algorithm. Additionally the authors present a selection algorithm aimed at a scenario where the data set size grows over time. The authors claim that their selection criterion is optimal in terms of the criterion defined in equation 1 of the paper. While a lot of space is spent on describing overfitting a clear connection to existing algorithms is not made. It would be interesting to state how existing algorithms would perform in terms of the proposed evaluation metric. Early stopping seems highly relevant to avoiding overfitting and thereby avoiding poor out-of-sample performance, and should be discussed in the related work section as well. The main point of criticism is that in various places the authors claim their method to be "practically useful" and applicable to NNs. Yet, the paper is lacking any kind of experimental evaluation to back up these claims. The connection to how the brain works in Section 3 and Figure 3 seems entirely unnecessary and does not add any value to the work. Minor comments/typos: * line 87: "we shall this" -> "we shall call this" * line 96: determine -> determines

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question	

Overall rating:	Good paper, accept
Confidence score:	The reviewer is confident but not absolutely certain that the evaluation is correct
Detailed Comments:	Machine learning techniques have to make bias-variance trade-off. Model selection is a good tool. This paper develops a penalized model selection technique of selecting the model from a set of candidates (e.g. model class) to achieve the optimal possible prediction loss (e.g. the limit of learning). The authours also proposed a computationally efficient algorithm that sequentially expands the model complexity. This paper is well-written and well-organized. The theoretical insight is very impressive.

Masked Reviewer ID: Assigned_Reviewer_4

Review:

Question	
Overall rating:	Marginally below acceptance threshold
Confidence score:	The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper
Detailed Comments:	This paper first proposes a modified version of the Takeuchi Information Criterion (TIC), which they call the Generalized TIC (GTIC), then uses this criteria in an algorithm for sequential model selection in a scenario where training data arrives in batches over time. This paper contains interesting mathematical results and experiments. However, I found it difficult to understand it's significance for practical applications. Furthermore, the organization and grammar could be improved. Regarding the GTIC, the authors' provide a proof that GTIC model-selection converges in probability to the optimal model for general, non-linear models under certain assumptions and large training data set size. It is difficult for me to evaluate the significance of this result. In practical model-selection scenarios, tools such as the TIC, AIC, and BIC, are used with simpler statistical models when training data is severely limited while hold-out validation sets are used when training data is more abundant. I am more familiar with the later scenario, for which it would take more experiments to convince me that this criterion would be useful. The second part of the paper proposes a sequential model selection algorithm that makes use of the assumption that larger models can be expected to perform better with more training data. While this algorithm was quite interesting, it's usefulness is again not entirely clear to me. In particular, I don't see why it is useful to reuse information from time t-1 when performing model selection (or rather, candidate model-weighting) at time t. Is the point that we can save computation by incrementally updating our model?