Author response to reviews of “Decision trees as partitioning machines to characterize their generalization properties”

We first address criticisms relevant to all reviewers and then follow with individual answers.

A concern shared by all reviewers relates to the scope of the framework, particularly to the limitation that only continuous features are considered. As we mentioned in the conclusion, the case of categorical features will be the object of future work. As a matter of fact, since the submission, we have figured out how to apply the partitioning framework to decision trees with such features. The extension can handle categorical and real-valued features at the same time, so we are hoping that the proposed framework will apply to more general settings. The new results are non-trivial and require the introduction of substantially more content which would not fit in the current paper. Therefore, we plan to publish these results in a subsequent paper.

Reviewer 1:

1. About the scope of the framework: (1) The results of the paper are multiclass since we consider the growth function and not only the VC dimension. In fact, 8 out of the 19 datasets used in the experiments were multiclass tasks. (2) Unfortunately, we were not able to think of other hypothesis classes for which the framework could be applied in a useful manner. However, our intention was specifically to treat decision trees, since they are extensively used in practice.
2. About “VC bounds based on parametrized function classes”: We do not think that these bounds are relevant for decision trees since they are essentially non-parametric. Indeed, decision trees are identified by the tree architecture, the decision rule used at each internal node, and the class value at each leaf. Hence, this representation does not map naturally to a set of adjustable parameters. Moreover, the VC dimension of a function class is not always related to the number of adjustable parameters. Indeed, Vapnik, in his "Statistical Learning Theory" book (section 4.11), provides examples of parametrized function classes where the VC dimension is less than, equal to, or exceeds the number of adjustable parameters.

Reviewer 2:

1. About the comparison to CART: We agree that it could have been interesting to compare against other pruning methods such as pessimistic error pruning (Breiman 1987) and error-based pruning (Quinlan 1992) or to some newer variants of them. We chose the CART algorithm (AKA cost-complexity pruning) for two main reasons: (1) Even though the CART algorithm is old, it is still widely used in practice. As a matter of fact, it is the algorithm that is implemented in the popular `scikit-learn` Python package (as explained in the online documentation of Decision Trees in section 1.10.8. Minimal Cost-Complexity Pruning). (2) Furthermore, the way the cost-complexity pruning algorithm works is by adding a penalty to the accuracy of a subtree which depends on some ad hoc notion of complexity (and finding the optimal penalty with cross-validation). Since the VC dimension and the growth function are theoretically valid quantifiers of the said complexity, it is natural to compare to this type of pruning.

Reviewer 3:

1. About the comparison to “Calculating the VC-dimension of decision trees”: The paper of Aslan et al. focuses on decision trees with binary features, while our paper focuses on real-valued features, which limits the possible links between the two papers. Their exhaustive search algorithm finds the exact value of the VC dimension of trees built from binary features, but it is a brute force algorithm, exponential in the tree size. Thus, they can only apply it to trees having a maximum depth of 4. Moreover, they try to estimate the VC dimension via a regression approach, hoping it would also apply to larger trees. On the other hand, Corollary 10 gives the asymptotic behavior of the VC dimension of trees on real-valued features (which Aslan et al. do not provide).
2. About the partially addressed broader impact: We do not see in what way we have not addressed the broader impact of our work. Could you tell us what you think is missing from our discussion?