View Reviews

Paper ID

8

Paper Title

Sorted Shift: A large-scale benchmark for out-of-distribution model predictive performance evaluation under gradual distribution shift

Reviewer #1

Questions

1. Please provide a few sentences summarizing the work, describing its strengths and/or weaknesses, and explaining your recommendation.

This work proposes a method for using existing ML benchmark datasets to evaluate a particular form of distribution shift. This involves picking a feature, sorting the dataset along that features, and splitting into thirds, using only the middle third for training and IID test evaluation, and the remainder for out-of-distribution (OOD) evaluation. They present some empirical results comparing the average performance of various sklearn algorithms on this type of shift, across 88 datasets in the Penn ML Benchmark, where they repeat the process once for each feature in each dataset.

The simplicity of the approach could be viewed as a strength. As for weaknesses, two jump out:
* First, it is unclear what kind of real-world distribution shift this is meant to model (it is certainly not what I would associate with "gradual" distribution shift). Here, the support of X is (by design) completely disjoint between the training and OOD distribution for a single feature. Moreover, the shift is not limited to X, as conditioning on e.g., certain subsets of age should presumably result in a change in the distribution of other features / labels, making interpretation somewhat difficult.
* Second, on a related point, there is not much in the way of providing interpretation / deeper analysis of the results, which are simply presented as averages over all experiments for each of 6 sklearn algorithms. On questions like "is logistic regression better at this kind of extrapolation", the answers are unclear: It is noted that no hyperparameter tuning was performed, and that the default settings for logistic regression imply substantial regularization.

Overall, this work fits the theme of the workshop, but feels quite thin in terms of content.

2. Recommendation

Reject

Reviewer #2

Questions

1. Please provide a few sentences summarizing the work, describing its strengths and/or weaknesses, and explaining your recommendation.

The authors propose a method to split tabular data into train/validation/test splits that induces a covariate shift between the data splits based on values of a specific feature. The paper purports to tackle the problem of “gradual” distribution shift, which certainly has practical relevance and would be on topic for the workshop. I am concerned, however, that the method discussed by the authors is too synthetic to give us a sense for how methods would fare against distribution shifts in the wild.

The notion of a “gradual distribution shift” is not clearly defined. In absence of a formal definition, I would guess it means a distribution shift that occurs over a large time interval and intensifies as time passes. To me, sorting data based on non-overlapping intervals of feature values doesn’t seem like a “gradual” shift. My original interpretation brings to mind simple image corruptions that can be realized at varying intensities [1] so you have a sort of control knob over how severe the distribution shift is. Perhaps this kind of controllability could be combined with the feature sorting method, but as is I don’t consider a dramatic covariate shift to an unseen area of feature space as “gradual”. I also don't consider this type of shift as realistic.

Minor comments:
* I would prefer that the abstract make clear that the focus is on tabular data. Given this scope, I was left wondering what can be done with categorical features that do not have a natural ordering.
* Lines 071-072 (left) appear to be an incomplete sentence.
* For tabular data with timestamps, the sorting idea seems related to training on some number of years, then testing on held-out years (with the difference being that the model typically wouldn’t have the year as an input feature). Some discussion of this would be interesting.
* It’s good that the authors acknowledge the lack of hyperparameter search as a limitation of their study. But this brings up the thorny issue of model selection for OOD generalization [2]. Have the authors considered how one *should* tune hyperparameters given this splitting setup? Without a methodology (or a few candidate methodologies) for model selection, this splitting technique would have very limited utility in my opinion.

[1] https://arxiv.org/abs/1903.12261
[2] https://arxiv.org/abs/2007.01434

2. Recommendation

Reject