How Good Are Machine Learning Clouds for Binary Classification with Good Features?

Hantian Zhang[†]   Luyuan Zeng[†]   Wentao Wu[‡]   Ce Zhang[†]

[†]ETH Zurich   [‡]Microsoft Research, Redmond

{hantian.zhang, ce.zhang}@inf.ethz.ch   wentao.wu@microsoft.com

In spite of the recent advancement of machine learning research, modern machine learning systems are still far from easy to use, at least from the perspective of business users or even scientists without a computer science background. Recently, there is a trend toward pushing machine learning onto the cloud as a "service," a.k.a. *machine learning clouds*. By putting a set of machine learning primitives on the cloud, these services significantly raise the level of abstraction for machine learning. For example, with Amazon Machine Learning, users only need to upload the dataset and specify the type of task (classification or regression). The cloud will then train machine learning models without any user intervention.

Raising the level of abstractions and building a system to automatically manage all physical decisions, however, rarely comes free. We investigate the usability of state-of-the-art machine learning clouds. Specifically, we ask:

To what extent can existing declarative machine learning clouds support real-world machine learning tasks?

More concretely, *what would users lose by resorting to declarative machine learning clouds instead of using a best-effort, non-declarative machine learning system?* To answer this question, we conduct an empirical study with `mlbench`, a novel benchmark consisting of real-world datasets *and* best-effort solutions harvested from Kaggle.[1]

Highlight of Methodology   In our study, we use a methodology that allows us to separate measuring the performance of the machine learning clouds themselves from other external factors that may have significant impact on quality, such as feature selection and hyper-parameter tuning. We use a performance metric that fairly measures the strength and weakness of current machine learning clouds by comparing their relative performance with top-ranked solutions in Kaggle competitions.

MLBENCH   We constructed `mlbench`, a benchmark for evaluating machine learning models on the clouds. We collected all winning codes for Kaggle competitions that are available online. We then ran the winning code
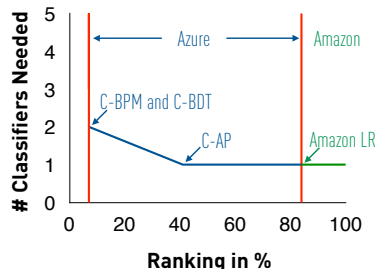
Figure 1: The worst ranking one would get in *all* five Kaggle competitions using machine learning models provided by **Azure** and **Amazon**. C-* represent classifiers on **Azure**; LR represents logistic regression on **Amazon**.

and used their feature engineering code to extract features. Those features are arguably the best features one can get, as the winners used them to win the Kaggle competitions. `mlbench` consists of 7 sets of features from 6 Kaggle competions. We restrict the scope of our study to binary classification. We plan to extend it to multi-class classification and regression and add more datasets to `mlbench`.

Highlight of Results   We evaluate the two most popular machine learning clouds, Azure Machine Learning Studio (**Azure**) and Amazon Machine Learning (**Amazon**), using `mlbench`. We find that: (1) Diversity of models is beneficial. **Azure** provides more alternative models than **Amazon** — **Amazon** only uses logistic regression for binary classification. Figure 1 suggests that the additional models provided by **Azure** help to achieve significant higher ranking; (2) Model selection is challenging. Using a different model can often lead to very different rankings. Although **Azure** provides more models to users, model selection is a difficult job for **Azure** users; and (3) Hyper-parameter tuning is time-consuming but beneficial. We use grid search to systematically explore hyper-parameters and observe huge variance of quality across different hyper-parameter settings.

A full version of this paper is available at `https://ease.ml/mlbench`.