An Enabling Technology

Evaluators like the independent trio evaluator can enable many applications that can increase the safety and fairness of AI systems. This notebooks explores some of them and gives suggestions for possible experiments to carry out.

Auto-ML

The vast majority of data that an AI system will process is unlabeled. Otherwise, why was the AI used? Can we leverage that unlabeled data somehow? Could we devise protocols that can train good models using unlabeled data? And can we do this robustly? The experiments in this section explore these questions.

Finding least correlated ensembles on large unlabeled test sets

The problem of correlated classifiers is the Achilles heel of algebraic evaluators. This presents the engineer of robust AI systems that uses them with a challenge - how do we build ensembles of classifiers that are, as best we can, independently correlated in their errors on a large unlabeled test set? We use the failures of the independent trio formula to reject possible candidates. Here are some suggestions on what sort of scans you could do
to see how well this works on the data and the classifiers of interest to you.

1. We expect classifiers trained differently to perform differently. Thus differences in training decorrelate their errors also. By training on disjoint sets of data, features, or algorithms you can get pair error correlations in the order of 5% or less. Carry out training protocols that scan for the most independent classifier ensembles by discarding the ones that make the evaluator fail.
    - **1.1.** Are there bounds on the correlations observed in the selected ones?
    - **1.2.** What about the rejected ones? Were good configurations rejected? What was the least error correlation that triggered the rejection? Is there a difference between failures that resulted in real algebraic numbers versus those that gave imaginary algebraic numbers?
    - **1.3.** Now measure the error in the evaluation estimates of the accepted configurations. Compare that to the error in the rejected ones.
    - **1.4.** Also compare it to various other random guessing that would be sensible - a real number between 0 and 1. These random guessers are good proxies for having no information except the one that you need to make in order to decode the multiple answers given by algebraic evaluators. For example, the comparison we made between the evaluator and ground truth in TheCoreTheorem.nb was based on picking the one based on our side knowledge about the value of the true prevalence. For binary classification this means that a fair comparison with random performers would mean that you guess a value of the prevalence with a width specified by your side knowledge. In our case, random guessing on the 0 to 1/2 interval.

Fairness Alarms

Different methods for ameliorating the harm of biased AI systems have been proposed and criticized. Common to some of them, is the use of purely observational statistics like false positives, etc. But how do we know that a system that was designed for X measure of fairness is actually working on unlabeled data? The experiment in this section suggests one possible way to use the independent trio evaluator to help with this problem.

Error-Correction Codes

Hamming famously quipped - “Damn it, if the computer knows I am wrong, why can’t it fix it?” He went on to invent error correcting codes. What error-correcting codes can you invent with the independent trio evaluator? Here we propose one way to build a code to help you get going with your own ideas.