The Science of Rejection: Learn When to Reject ML Inferences

Burcu Sayin Günel[1], Jie Yang[2], Andrea Passerini[1] and Fabio Casati[1]

[1]University of Trento, [2]Delft University of Technology

Why the science of learning to reject model predictions is central to ML?

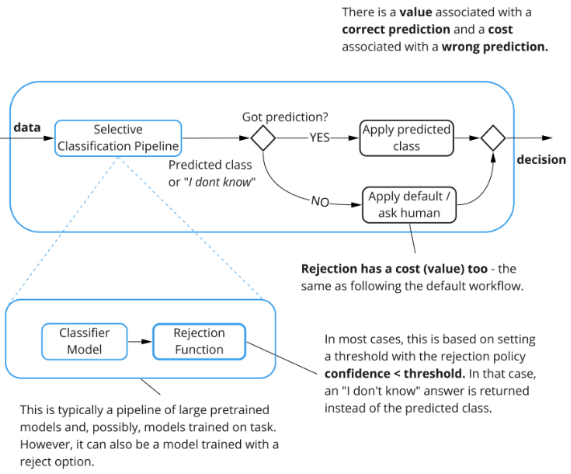


Figure 1. ML models in Enterprise Workflows, taken from [1]

AI Workflows and the Metrics

When we deploy an AI solution in an end-to-end enterprise workflow, we have some ML classifier m that, given an input $i \in I$ (where I is a possibly infinite set of items to classify), produces a predicted class and a confidence (or a distribution of predicted class with confidences). There is then a filtering based on whether the confidence is greater than some threshold, and if so the prediction is applied, else a default path is followed. From this simple description we can draw a few observations:

Value matters

The threshold and the system behavior depend on the "cost" of machine errors and its relation to the cost of a rejection and the value of a correct machine prediction. **We propose a value function to re-evaluate the value of ML models**.

Measuring the "Value"

Let's refer to the value of a correct prediction as V, to the value (cost) of following the default flow as C_d and to the cost of a wrong prediction as $C_w = K \cdot C_d$ (that is, we express C_w in terms of how "bad" is an erroneous prediction compared to the default flow). Also, for simplicity, let's assume that $V = -C_d$, and let's normalize by taking $V = 1$, again for simplicity. If the enterprise has a sense of K, then the optimal threshold T is $T = \frac{K-1}{K+1}$, *assuming the model is well calibrated*. Similarly, we can show that the *expected value* for each prediction with confidence c is

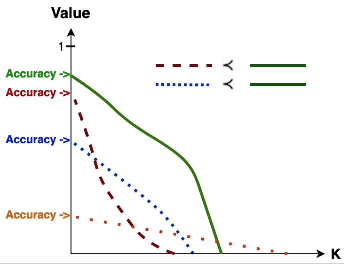
$$E[value] = -1 \cdot \rho_t + (1 - \rho_t) \cdot (c(K+1)K) \quad (1)$$

where ρ_t is the probability of a prediction confidence being below the selected threshold t (see [1, 3]).

Calibration matters

If we have a well-calibrated model with arbitrarily bad accuracy α, we can still get value from it.

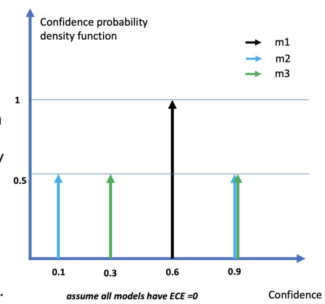
- K=0 corresponds to the accuracy of the model.
- Value=0 is the state before applying the model (we reject any prediction).
- Any model is valuable if we have a validation set to tune our rejection policy.
- Because customers can set K, they can also control the *risk*, and be arbitrarily conservative.



Metrics matter

Commonly adopted calibration metrics, such as the Expected Calibration Error (ECE) [2] and its variation (eg, based on how we bin the samples) do not correspond to the metric we want to improve. They help us to get a sense of the model calibration as a whole and they are independent of any confidence threshold T or cost structure. However, when we apply a model as per the workflow in Figure 1, we only care about calibration around T. Current calibration techniques, such as temperature scaling, show spectacular ECE results but if our threshold is 0.8, we really don't care about error in the 0.1-0.2 range, nor we care if a confidence is 0.999 or 0.85.

On the right, we show the probability density function for three models over a dataset. $m3$ has the same accuracy of $m1$, and $m2$ has *worse* accuracy, but both $m2$ and $m3$ deliver better value for a wide range of K. Even in the case $ECE \neq 0$, models $m1$ and $m2$ may have same accuracy but the one with worse ECE would have better value. If we tweak a little $m3$ to make it slightly underconfident or overconfident for the impulse at 0.3: we have higher ECE but still higher value for most values of K.



The better we are able to identify subset of items for which our model m is calibrated, the lower is the cost for our deployment of m in an AI workflow. **Our work in progress [3] builds a novel calibration metric that considers the joint distribution of confidence and accuracy.**

References

[1] Fabio Casati, Pierre-André Noël, and Jie Yang. On the value of ml models. In *NeurIPS Workshop on Human and Machine Decisions (WHMD)*, 2021.

[2] Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. *ArXiv*, abs/1904.01685, 2019.

[3] Burcu Sayin, Jie Yang, Andrea Passerini, and Fabio Casati. The science of rejection: A research area for human computation. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2021.