On the Value of ML Models

Workshop on Human and Machine Decisions @ NeurIPS 2021

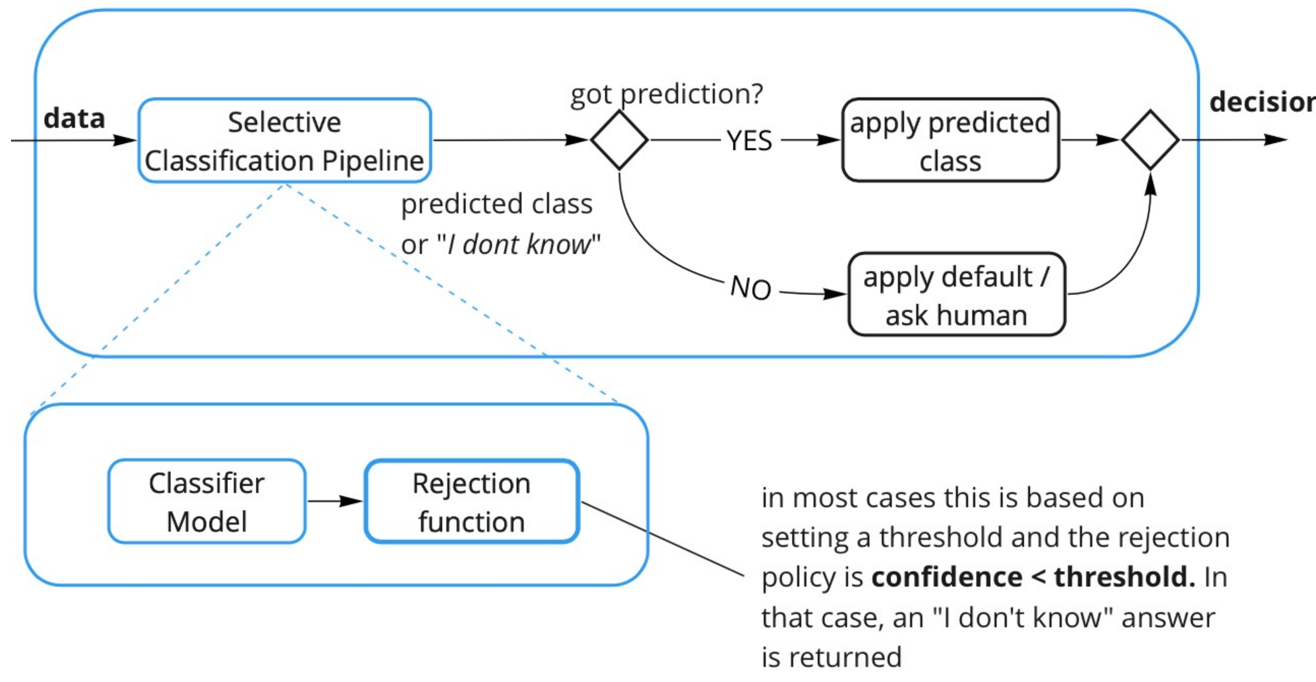
Fabio Casati, Pierre-André Noël
Element AI, a ServiceNow company

Jie Yang
TU Delft

NEURAL INFORMATION PROCESSING SYSTEMS

ML is applied in a selective fashion

- In practice, **ML models are almost always used as selective models**: a default, safe outcome is selected when the ML prediction is rejected.
- Selectivity is often implemented by filtering based on a confidence threshold.
- So, predictions can be **correct** or **wrong**, or the prediction workflow **abstains,** that is, the prediction is rejected.
- This is the rule, not the exception. And it comes with massive theoretical and practical consequences.



Value, not Accuracy

- Commonly used measures such as accuracy, F1, true positives, and AUC can be misleading: What matters is a notion of *value*, which depends on the "utility" of correct predictions, wrong predictions, and rejections,
- On the one hand this is obvious. On the other, this is constantly overlooked both in research and in practice.
- While the full diversity/complexity of real use cases cannot be accounted for by research benchmarks, value-estimating metrics may be designed to capture high-level commonalities among classes of real use cases.

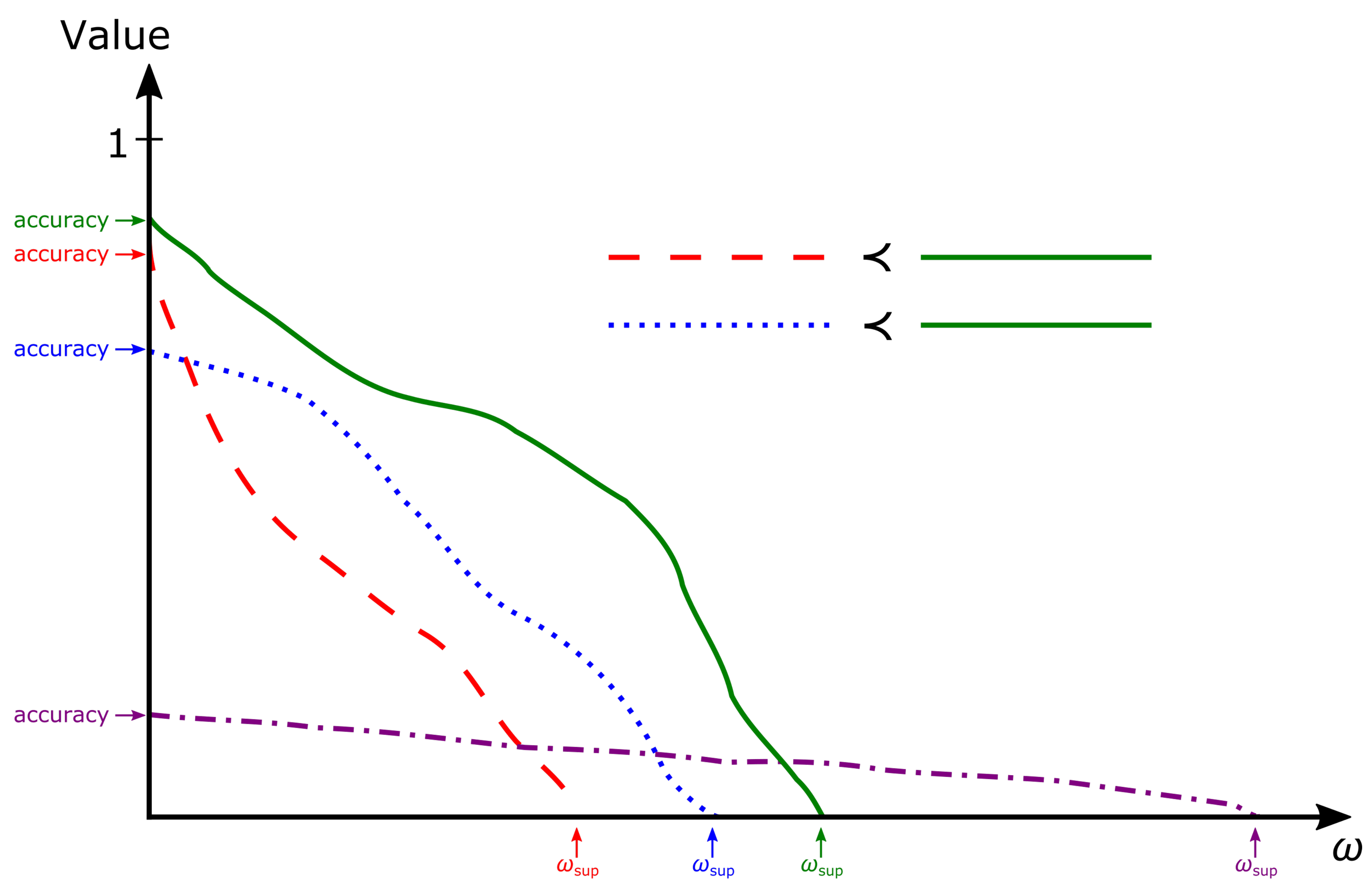
On Learning and Confidence Distributions

- The ability of a model to provide meaningful confidence measures—and of a deployment to use the right threshold—is central to value.
- Using a validation dataset representative of the use case, we can maximize the model value by tuning the rejection function so that **even models with arbitrarily low accuracy bring better or equal value than no model**.
- Calibration does not affect value if the rejection threshold is tuned with a validation dataset. Also, commonly used measures of calibration (such as ECE) may be misleading.
- **Learning does not mean better accuracy or better calibration**. Given a calibrated model, its value can be increased without altering its prediction (and thus its accuracy) by ***learning*** a confidence distribution that is more **discriminating.**
- Such concerns are not part of the main ML research narrative. We argue that they should, and that benchmarks should aim to account for them.

	Use the Top Prediction	**Abstain** (don't Use It)
Top Prediction Is **Correct**	We used the top prediction and it was correct (*i.e.*, the most desirable outcome). Value: 1 (or V_{correct} in arbitrary units)	We abstained from using the top prediction, instead defaulting to some "safe" behaviour (*i.e.*, the outcome of intermediate desirability). Value: 0 (or V_{abstain} in arbitrary units) Whether the top prediction was correct or wrong is completely irrelevant for the present use case.
Top Prediction Is **Wrong**	We used the top prediction but it was wrong (*i.e.*, the least desirable outcome). Value: $-\omega$ (or V_{wrong} in arbitrary units)	

Above: We assume that the value derived from a model's prediction may only depend on which of these three cases occurs. Three arbitrary values may be ascribed to each of these cases, but a change of variable takes this down to a single parameter, ω, determining the cost of making a ωrong prediction in terms of the value of a correct one.

Below: Each curve traces the value of a *fictious* model as a function of ω. Like ROC curves, a model whose value is everywhere above another one's is strictly "better" for all use cases. We argue for this kind of plot (and/or derived quantities) to be used in benchmarks.



Mathematical Details

- Traditional classifier $\mathbf{f} : X \to [0,1]^C$.
  - Predicted class $\text{argmax}_i\, f_i(\mathbf{x}) \in \{1, \cdots, C\}$.
  - Confidence $\max_i f_i(\mathbf{x}) \in [0,1]$.
- Abstaining classifier $g : X \mapsto \{0, 1, \cdots, C\}$.
  - Predicted class $g(\mathbf{x}) \in \{0, 1, \cdots, C\}$, 0 means "abstain".
  - No access to any confidence score.
- Common approach in applications: apply threshold t.

$$g_{\mathbf{f},t}(\mathbf{x}) = \begin{cases} \underset{i\in\{1,\cdots,C\}}{\text{argmax}}\ f_i(\mathbf{x}) & \text{if } \underset{i\in\{1,\cdots,C\}}{\max} f_i(\mathbf{x}) \geq t, \\ 0 & \text{otherwise.} \end{cases}$$

- Given a dataset $\mathfrak{D}$ and an abstaining classifier g, we may count that model's total number N_{correct} of correct predictions, N_{abstain} of abstentions, and N_{wrong} of incorrect predictions.
- Suppose that we know $\mathfrak{U} = (V_{\text{correct}}, V_{\text{abstain}}, V_{\text{wrong}})$ for a given use case such that the total value (in dollars or any other utility unit) provided by an abstaining classifier g for a dataset $\mathfrak{D}$ is

$$V(g, \mathfrak{U}, \mathfrak{D}) = V_{\text{correct}} \cdot N_{\text{correct}} + V_{\text{abstain}} \cdot N_{\text{abstain}} + V_{\text{wrong}} \cdot N_{\text{wrong}} \quad .$$

- We define the change of variables

$$\mathcal{V}(g, \omega, \mathfrak{D}) = \frac{V(g, \mathfrak{U}, \mathfrak{D}) - V_{\text{abstain}}}{|\mathfrak{D}|(V_{\text{correct}} - V_{\text{abstain}})} \quad \text{where} \quad \omega = \frac{V_{\text{abstain}} - V_{\text{wrong}}}{V_{\text{correct}} - V_{\text{abstain}}}$$

- If $V_{\text{wrong}} < V_{\text{abstain}} < V_{\text{correct}}$, then $\omega > 0$ and the dimensionless value is

$$\mathcal{V}(g, \omega, \mathfrak{D}) = \frac{N_{\text{correct}} - \omega N_{\text{wrong}}}{N_{\text{correct}} + N_{\text{abstain}} + N_{\text{wrong}}} \quad .$$

- ω-aware: given traditional classifier $\mathbf{f}$ and validation dataset $\mathfrak{D}'$, the optimal threshold for a fixed ω is

$$t_\omega = \underset{t\in\mathbb{R}}{\text{argmax}}\ \mathcal{V}(g_{\mathbf{f},t}, \omega, \mathfrak{D}') \quad .$$

- Calibrated: if $\mathbf{f}^{\text{cal}}$ returns confidence c, then probability c to be correct.
  - *Calibration doesn't affect value*: if $\mathbf{f}^{\text{cal}}$ obtained from $\mathbf{f}$ using monotonously increasing function $\tilde{c} : [0,1] \to [0,1]$ on its confidence, then same optimal value at threshold $t_\omega^{\text{cal}} = \tilde{c}(t_\omega)$.
  - *Calibration grounds threshold*: if $\rho : [0,1] \to \mathbb{R}_{\geq 0}$ is PDF for confidence of $\mathbf{f}^{\text{cal}}$ in $\mathfrak{D}'$, then $t_\omega^{\text{cal}} = \omega/(\omega+1)$ maximizes

$$\mathcal{V}(g_{\mathbf{f}^{\text{cal}},t}, \omega, \mathfrak{D}') = \int_t^1 \left[c - \omega(1-c)\right]\rho(c)\, \mathrm{d}c \quad .$$

- Trivial: "better" model by moving mass in $\rho(c)$ to higher confidence.
  - This operation increases the accuracy $\int_0^1 c\, \rho(c)\, \mathrm{d}c$.
- Less trivial: "better" model by moving mass both up and down.
  - Increase discrimination $\int_0^1 (\frac{1}{2} - c)^2 \rho(c)\, \mathrm{d}c$ with same accuracy.
  - Intuition: make value curve fall more slowly.