Classifier Uncertainty Beyond Calibration

Sébastien Melo, Gaël Varoquaux, and Marine Le Morvan.

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores themselves.

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores themselves.

Decomposing Uncertainty

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores themselves.

Decomposing Uncertainty

- Proper scoring rules: measure the total prediction error. Decomposed into:

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores themselves.

Decomposing Uncertainty

- Proper scoring rules: measure the total prediction error. Decomposed into:
    - **Aleatoric Loss**: Irreducible error from task randomness.

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores themselves.

Decomposing Uncertainty

- Proper scoring rules: measure the total prediction error. Decomposed into:
  - **Aleatoric Loss**: Irreducible error from task randomness.
  - **Epistemic Loss**: Reducible error from the model's lack of knowledge. This is what we need to look at!

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores themselves.

Decomposing Uncertainty

- Proper scoring rules: measure the total prediction error. Decomposed into:
  - **Aleatoric Loss**: Irreducible error from task randomness.
  - **Epistemic Loss**: Reducible error from the model's lack of knowledge. This is what we need to look at!
- Epistemic loss: two key components:

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores themselves.

Decomposing Uncertainty

- Proper scoring rules: measure the total prediction error. Decomposed into:
  - **Aleatoric Loss**: Irreducible error from task randomness.
  - **Epistemic Loss**: Reducible error from the model's lack of knowledge. This is what we need to look at!
- Epistemic loss: two key components:
  - **Calibration Loss**: Predicted probabilities and event frequencies don't match: large literature.

Why Reliable Confidence Scores Matter

The Need for Trustworthy Confidence

- For black-box models in high-stakes settings: need to know the uncertainty behind a decision = **confidence scores**.
- Simple metrics (e.g. accuracy) aren't enough: need to evaluate the probabilistic quality of the confidence scores themselves.

Decomposing Uncertainty

- Proper scoring rules: measure the total prediction error. Decomposed into:
  - **Aleatoric Loss**: Irreducible error from task randomness.
  - **Epistemic Loss**: Reducible error from the model's lack of knowledge. This is what we need to look at!
- Epistemic loss: two key components:
  - **Calibration Loss**: Predicted probabilities and event frequencies don't match: large literature.
  - **Grouping Loss**: Variance in true probability among samples that were given the same confidence score : one known estimator.

Our Results: Better Estimators for Better Decisions

1. More Sample-Efficient Estimators

- We introduce binning-free estimators for the grouping loss and its associated decision risk.

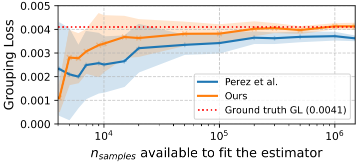

Figure 1: Our estimator converges faster and more tightly than prior work.

Our Results: Better Estimators for Better Decisions

1. More Sample-Efficient Estimators

- We introduce binning-free estimators for the grouping loss and its associated decision risk.

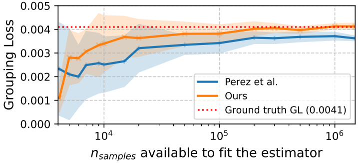

Figure 1: Our estimator converges faster and more tightly than prior work.

2. Improved Individual Decisions with LLM Cascades

- We use our risk estimates as a per-query quality score to build intelligent LLM cascades.

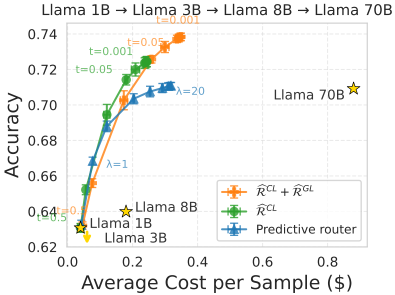

Figure 2: Our cascade improves accuracy while reducing cost compared to baselines.