Smooth ECE: Principled Reliability Diagrams via Kernel Smoothing

Jarosław Błasiok, Preetum Nakkiran
ICLR 2024 · Apple, Columbia University

Abstract

We give improved methods for measuring calibration error of binary classifiers:

- New calibration metric: SmoothECE (smECE)
- New reliability diagrams: Smooth Reliability (reflects smECE)

Simple method with **strong theoretical guarantees**, and **open-sourced**.

Background: What is Calibration?

Calibration: How "reliable" are predicted probabilities?
"90% chance of rain" $\Longrightarrow$ Actually rains ~90% of time

Setting: Binary classification. Given distribution over
$f(x) \in [0,1]$: Prediction
$y \in \{0,1\}$: True outcome

Calibration measure: How mis-calibrated is this distribution over $(f(x), y)$?

Problem

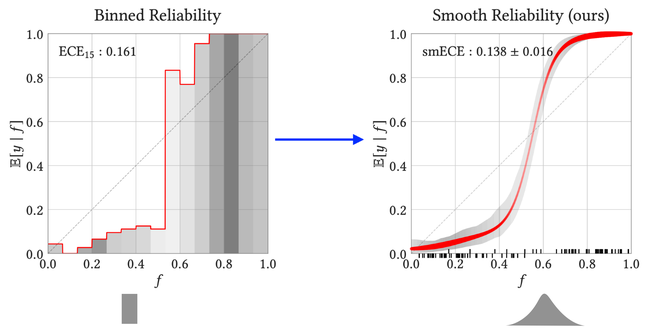
Common methods for measuring & plotting calibration (ECE, binning):
discontinuous, unclear theoretical guarantees, unspecified hyperparameters.

Better calibration measures exist, but **not with associated diagrams.**
[Błasiok, Gopalan, Hu, Nakkiran 2023]

Our Method

`> pip install relplot`

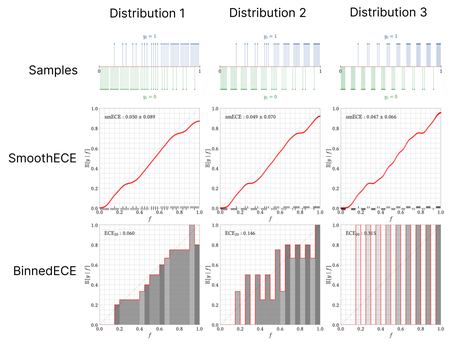
Summary: Kernel-smoothing with *very particular* choice of bandwidth



smECE := "average deviation from diagonal"

instead of rounding predictions... ***...smoothly round them***

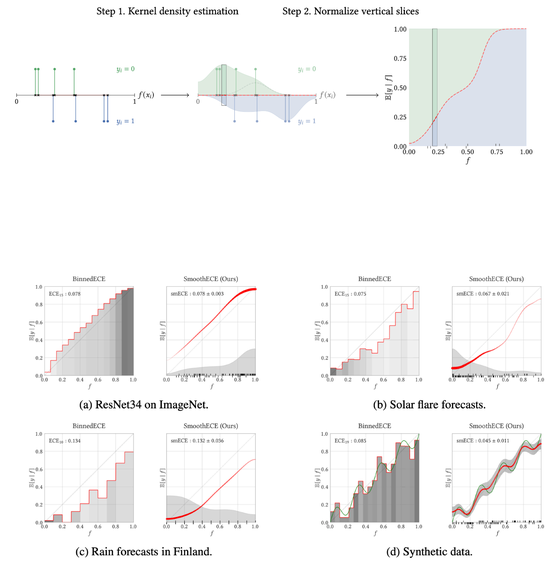
Example: Discontinuity of Binned ECE



Theoretical guarantees

If we pick bandwidth σ such that: smECE$_\sigma \approx \sigma$, then
smECE is a *consistent calibration metric*
i.e. smECE ~ poly(distance-to-calibration).

Examples



Conclusion

Better ways of **measuring** and **visualizing** calibration of classifiers.

Python package: simple, hyperparameter-free, numpy & matplotlib-compatible

```
import relplot as rp
…
calib_error = rp.smECE(f, y)
fig, ax = rp.rel_diagram(f, y)
```