LSST DESC Notes

The Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC): Metrics

Tarek Alam, Anita Bahmanyar, Rahul Biswas, Mi Dai, Lluis Galbany, Renée Hložek, Emille Ishida, Saurabh Jha, David Jones, Rick Kessler, Michelle Lochner, Ashish Mahabal, Kaisey Mandel, Juan Rafael Martinez Galarza, Alex Malz, Daniel Mutukrishna, Gautham Narayan, Tina Peters, Hiranya Peiris, and Kara Ponder

We describe and illustrate the process by which a global performance metric was chosen for Photometric LSST Astronomical Time-series Classification Challenge (PLAsTiCC), a Kaggle competition aiming to identify promising transient and variable classifiers for LSST by involving the broader community outside astronomy.
This note is the brief introduction to the metrics used for the PLAsTiCC data challenge.

1. Introduction

The main goal of the PLAsTiCC competition is to answer the following: *can one classify a large test set of transients and variables using their photometric data, given a small and unbalanced training set?*

In this challenge, classification over the full range of classes is preferred, hence the main metric that will be used to evaluate the challenge is the *Brier metric*, as described below.

The metric of this note is for the first version of the Kaggle competition, though there are future plans for an early classification challenge and identification of class-specific metrics for different science goals. This note serves only to summarize the results and code online in the ProClam repository. Interactive notebooks and calculations are provided there.

The criteria for the metric included:

- The metric must return a single scalar value.
- The metric must be well-defined for non-binary classes.
- The metric must balance diverse science use cases in the presence of heavily nonuniform class prevalence.
- The metric must respect the information content of probabilistic classifications.
- The metric must be able to evaluate deterministic classifications.
- The metric must be interpretable, meaning it gives a more optimal value for “good” mock classifiers and a less optimal value for mock classifiers plagued by anticipated systematic errors; in other words, it must pass basic tests of intuition.
- The metric must be reliable, giving consistent results for different instantiations of the same test case.

2. Methods

We considered two metrics of classification probabilities, each of which is interpretable and avoids reducing probabilities to point estimates

The Brier score is defined as

$$B = \sum_{m=1}^{M} \frac{w_m}{N_m} \sum_{n=1}^{N_m} \left((1 - p_n(m|m))^2 + \sum_{m' \neq m}^{M} (p_n(m'|m))^2 \right),$$

where the sum over M is a sum over the M classes defined, while the sum over N_m is the sum over the individual objects in a given class m in M. The w_m are the weights defined per class. These weighted averages of the per-class metrics maybe be considered in terms of the systematics we discussed, by upweighting or downweighting the chosen ‘class’ most affected by the systematics.

The log-loss is defined as

$$L = - \sum_{m=1}^{M} \frac{w_m}{N_m} \sum_{n=1}^{N_m} \ln[p_n(m|m)],$$

where again the sum over M is computed across classes, while the sum over N_m is within a class.

We define a weight vector across classes N_m, which will be provided to the Kaggle team separately, since it contains model-specific information. For both the Brier metric and the log-loss metric, the goal is to minimise the metric score.

In addition to providing the output of some classifiers evaluated on the metric, as shown in Figure 1, we also include various 'systematics' across which we test the metric performance.

These systematics include:

- idealized: highly accurate on all classes
- guessing: random classifications across all classes
- tunnel vision: classifies one class well and others randomly
- cruise control: classifies all objects as a single class
- subsumed: consistently misclassifies one class as one other class

3. Results

We show the performance of the metric on the various classifiers included in Figure 1.

4. Contributions

The contributions of those directly involved in the metrics are included in the online version of this note in the ProClam repository.

5. Conclusion

Acknowledgments

The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules in France; the Science & Technology Facilities Council in the United Kingdom; and the Department of Energy, the National Science Foundation, and the LSST Corporation in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3–Lyon/Villeurbanne - France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract

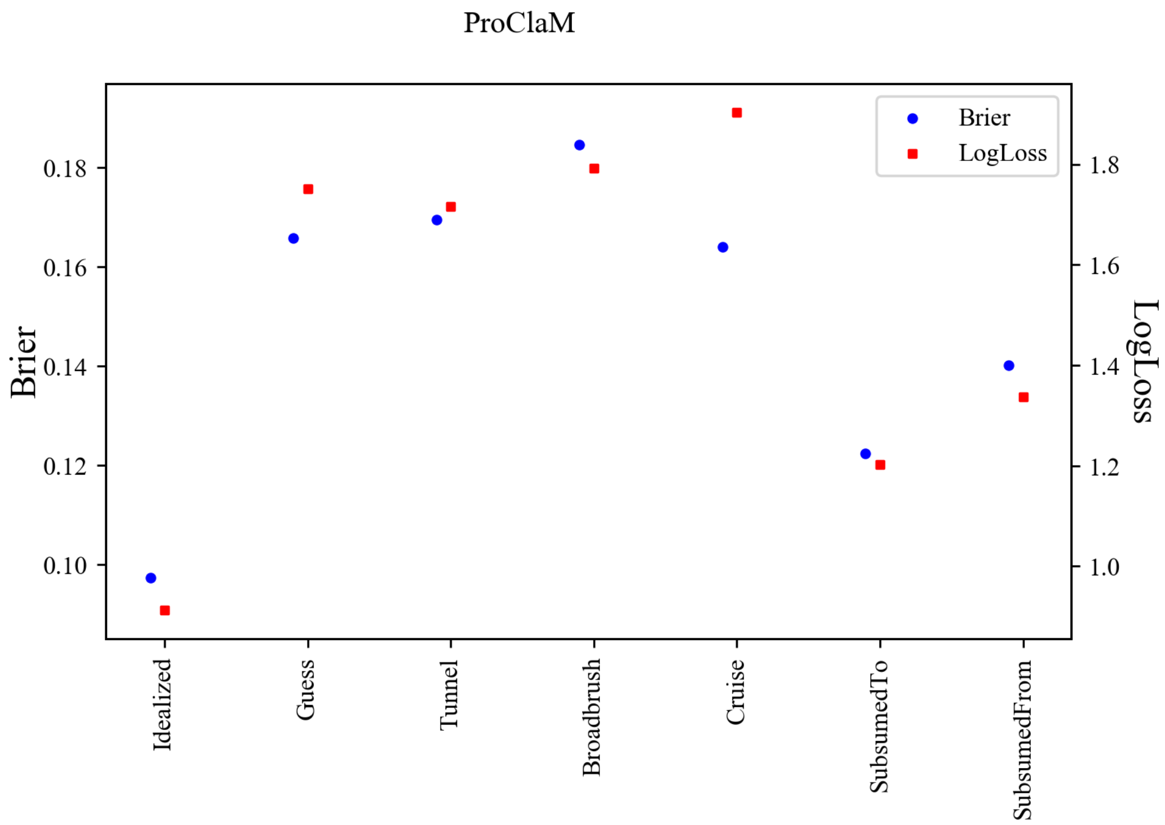

Figure 1. The Brier and Log-loss metrics evaluated against example classifications, and 'systematic' classification types. The code used to produe these plots is online as ProClam, and example ipython notebooks are provided for use by PLAsTiCC participants.

No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515.