Evaluación de probabilidades a posteriori: teoría de decisión, proper scoring rules y calibración

Luciana Ferrer[1] and Daniel Ramos[2]

[1] Instituto de Ciencias de la Computación, CONICET-UBA, Argentina, ORCID:0000-0002-0426-8683, lferrer@dc.uba.ar
[2] Escuela Politécnica Superior, Universidad Autónoma de Madrid, Spain, ORCID: 0000-0001-5998-1489, daniel.ramos@uam.es

Resumen. La mayoría de los clasificadores de aprendizaje automático están diseñados para generar probabilidades a posteriori para las clases, dadas las muestras de entrada. Estas probabilidades pueden utilizarse para tomar una decisión categórica sobre la clase de la muestra; proporcionarse como entrada a un sistema posterior; o entregarse a un humano para su interpretación. Evaluar la calidad de las probabilidades a posteriori generadas por estos sistemas es un problema esencial que fue abordado hace décadas con la invención de las proper scoring rules (PSRs). Desafortunadamente, gran parte de la literatura reciente en aprendizaje automático utiliza métricas de calibración —más comúnmente, el error de calibración esperado (ECE)— como un sustituto para evaluar el rendimiento de las probabilidades a posteriori. El problema con este enfoque es que las métricas de calibración reflejan solo un aspecto de la calidad de las probabilidades, ignorando el rendimiento en discriminación. Por esta razón, argumentamos que las métricas de calibración no deberían tener ningún papel en la evaluación de la calidad de las probabilidades a posteriori, y que en su lugar deberían utilizarse las PSRs esperadas para este propósito. Aunque no son útiles para evaluar el rendimiento, las métricas de calibración pueden usarse como herramientas de diagnóstico durante el desarrollo del sistema. Con este objetivo en mente, discutimos una métrica de calibración simple y práctica, llamada pérdida de calibración. Comparamos esta métrica con el ECE y con la divergencia de puntuación esperada, y argumentamos que la pérdida de calibración es superior a estas dos métricas.

Palabras clave: Teoría de decisión, proper scoring rules, calibración, sistemas de clasificación

Evaluating posterior probabilities: decision theory, proper scoring rules, and calibration

Abstract. Most machine learning classifiers are designed to output posterior probabilities for the classes given the input sample. These probabilities may be used to make the categorical decision on the class of

the sample; provided as input to a downstream system; or provided to a human for interpretation. Evaluating the quality of the posteriors generated by these systems is an essential problem which was addressed decades ago with the invention of proper scoring rules (PSRs). Unfortunately, much of the recent machine learning literature uses calibration metrics—most commonly, the expected calibration error (ECE)—as a proxy to assess posterior performance. The problem with this approach is that calibration metrics reflect only one aspect of the quality of the posteriors, ignoring the discrimination performance. For this reason, we argue that calibration metrics should play no role in the assessment of posterior quality and expected PSRs should instead be used for this job. While not useful for performance assessment, calibration metrics may be used as diagnostic tools during system development. With this purpose in mind, we discuss a simple and practical calibration metric, called calibration loss. We compare this metric with the ECE and with the expected score divergence metric and argue that calibration loss is superior to these two metrics.

Keywords: Decision theory, proper scoring rules, calibration, classification systems

DOI: El journal no genera números DOI.
Publicado en: Transactions on Machine Learning Research, 2025
Factor de impacto: No disponible aún.