An open mathematical problem in multiclass classification

Shashank Singh

September 2022

This note describes an interesting open problem in the statistical theory of multiclass classification. The problem is easy to state, but seems challenging to solve, so I thought it would be good to get some more eyes on it. ☺

Please don't hesitate to contact me at shashankssingh@gmail.com with ideas or questions.

Background: The binary case A famous result in the theory of binary classification states that accuracy (i.e., the proportion of samples labeled correctly) is maximized by the "Bayes" classifier

$$\widehat{Y}_{\text{Bayes}}(x) = \begin{cases} 0 & \text{if } \eta(x) \leq 0.5 \\ 1 & \text{if } \eta(x) > 0.5 \end{cases}, \tag{1}$$

where $\eta(x) := \mathbb{E}[Y|X = x]$ denotes the true probability that a sample with covariate x lies in class 1. Although η is unknown in practice, this result motivates a simple recipe for binary classification: estimate the conditional class probability η (e.g., using logistic regression, random forests, nearest neighbors, or something else), and then threshold this estimate at 0.5.

In many real-world classification problems, accuracy is a poor measure of performance; classifiers with high accuracy may fail to distinguish the classes well. For example, if Class 0 is generally more common than Class 1, such that $\sup_x \eta(x) \leq 0.5$, then the Bayes classifier will classify *all* inputs as Class 0. A host of alternative performance measures, such as precision/recall, F_β scores, AUROC, AUPR, etc., have been proposed. However, theoretical results for classification in terms of these more general performance measures are quite limited. Notably, it is not clear when thresholding an estimate of η performs well in terms of general performance measures. Theorem 3 of Singh and Khim [2021] showed that optimizing general measures of binary classification performance is not always possible with *deterministic classifiers* (which always predict the same label for a given covariate value), but may require *stochastic classifiers* (which may guess a class randomly for some covariate values). In particular, we showed that there always exists an optimal stochastic classifier of the form

$$\widehat{Y}_{p,t}(x) = \begin{cases} 0 & \text{if } \eta(x) < t \\ \text{Bernoulli}(p) & \text{if } \eta(x) = t \\ 1 & \text{if } \eta(x) > t \end{cases}, \quad \text{for some} \quad p, t \in [0, 1]. \tag{2}$$

For most values of η, $\widehat{Y}_{p,t}$ returns a deterministic class, but when $\eta(x) = t$, $\widehat{Y}_{p,t}$ guesses Class 0 with probability $1 - p$ and Class 1 with probability p. Similar to (1), this motivates a simple recipe for binary classification under more general performance measures: estimate the conditional class probability η, and then threshold this estimate at a threshold $(p, t) \in [0, 1]^2$ that optimizes training performance. Proving this result (see Appendix A of Singh and Khim [2021]) was surprisingly challenging, involving an elementary but non-trivial degree of measure theory.

Open problem: The multivariate case Understanding performance in terms of general performance measures is especially important in multi-class classification, where class imbalance is the rule and accuracy is rarely used. However, it is not clear to me how to generalize the above problem to the multiclass case; in particular, it is not clear to me what form the optimal stochastic multi-class classifier should take. Does the number of random parameters p needed scale linearly with the number k of classes? Or with the number of pairs of classes? Or with the number of possible subsets of classes?

References

Shashank Singh and Justin Khim. Statistical theory for imbalanced binary classification. *arXiv preprint arXiv:2107.01777*, 2021.