CrossValidated

Is accuracy an improper scoring rule in a binary classification setting?

Asked 7 years, 1 month ago Modified 1 year, 2 months ago Viewed 10k times

31

I have recently been learning about proper scoring rules for probabilistic classifiers. Several threads on this website have made a point of emphasizing that accuracy is an improper scoring rule and should not be used to evaluate the quality of predictions generated by a probabilistic model such as logistic regression.

However, quite a few academic papers I have read have given misclassification loss as an example of a (non-strict) proper scoring rule in a binary classification setting. The clearest explanation I could find was in this paper, at bottom of page 7. To the best of my understanding, minimizing misclassification loss is equivalent to maximizing accuracy, and the equations in the paper make sense intuitively.

For example: using the notation of the paper, if the true conditional probability (given some feature vector **x**) of the class of interest is $\eta = 0.7$, any forecast $q > 0.5$ would have an an expected loss $\mathbf{R}(\eta|q) = 0.7(0) + 0.3(1) = 0.3$, and any $q \leq 0.5$ would have an expected loss of 0.7. The loss function would therefore be minimized at $q = \eta = 0.7$ and consequently proper; the generalization to the entire range of true conditional probabilities and forecasts seems straightforward enough from there.

Assuming the above calculations and statements are correct, the drawbacks of a non-unique minimum and all predictions above 0.5 sharing the same minimum expected loss are obvious. I still see no reason to use accuracy over the traditional alternatives such as log score, Brier score, etc. However, is it correct to say that accuracy is a proper scoring rule when evaluating probabilistic models in a binary setting, or am I making a mistake - either in my understanding of misclassification loss, or in equating it with accuracy?

probability accuracy scoring-rules

Share Cite Improve this question Follow

asked Jul 31, 2018 at 2:54

Zyzzva

311 4 3

2 Answers

Sorted by:

33

TL;DR

Accuracy is an improper scoring rule. Don't use it.

The slightly longer version

Actually, accuracy is not even a scoring rule. So asking whether it is (strictly) proper is a category error. The most we can say is that *under additional assumptions*, accuracy is consistent with a scoring rule that is improper, discontinuous and misleading. (Don't use it.)

Your confusion

Your confusion stems from the fact that misclassification loss as per the paper you cite is not a scoring rule, either.

The details: scoring rules vs. classification evaluations

Let us fix terminology. We are interested in a binary outcome $y \in \{0, 1\}$, and we have a probabilistic prediction $\hat{q} = \widehat{P}(Y = 1) \in (0, 1)$. We know that $P(Y = 1) = \eta > 0.5$, but our model $\hat{q}$ may or may not know that.

A *scoring rule* is a mapping that takes a probabilistic prediction $\hat{q}$ and an outcome y to a loss,

$$s\colon (\hat{q}, y) \mapsto s(\hat{q}, y).$$

s is *proper* if it is optimized in expectation by $\hat{q} = \eta$. ("Optimized" usually means "minimized", but some authors flip signs and try to maximize a scoring rule.) s is *strictly proper* if it is optimized in expectation *only* by $\hat{q} = \eta$.

We will typically evaluate s on many predictions $\hat{q}_i$ and corresponding outcomes y_i and average to estimate this expectation.

Now, what is *accuracy*? Accuracy does not take a probabilistic prediction as an argument. *It takes a classification* $\hat{y} \in \{0, 1\}$ and an outcome:

$$a\colon (\hat{y}, y) \mapsto a(\hat{y}, y) = \begin{cases} 1, & \hat{y} = y \\ 0, & \hat{y} \neq y. \end{cases}$$

Therefore, *accuracy is not a scoring rule*. It is a classification evaluation. (This is a term I just invented; don't go looking for it in the literature.)

Now, of course we can take a probabilistic prediction like our $\hat{q}$ and turn it into a classification $\hat{y}$. But to do so, we will need the additional assumptions alluded to above. For instance, it is very common to use a threshold θ and classify:

$$\hat{y}(\hat{q},\theta) := \begin{cases} 1, & \hat{q} \geq \theta \\ 0, & \hat{q} < \theta. \end{cases}$$

A very common threshold value is $\theta = 0.5$. Note that if we use this threshold and then evaluate the accuracy over many predictions $\hat{q}_i$ (as above) and corresponding outcomes y_i, then we arrive exactly at the misclassification loss as per Buja et al. Thus, misclassification loss is also not a scoring rule, but a classification evaluation.

If we take a classification algorithm like the one above, we can turn a classification evaluation into a scoring rule. The point is that we need the additional assumptions of the classifier. And that accuracy or misclassification loss or whatever other classification evaluation we choose may then depend less on the probabilistic prediction $\hat{q}$ and more on the way we turn $\hat{q}$ into a classification $\hat{y} = \hat{y}(\hat{q},\theta)$. So optimizing the classification evaluation may be chasing after a red herring if we are really interested in evaluating $\hat{q}$.

Now, what is improper about these scoring-rules-under-additional-assumptions? Nothing, in the present case. $\hat{q} = \eta$, under the implicit $\theta = 0.5$, will maximize accuracy and minimize misclassification loss over all possible $\hat{q} \in (0,1)$. So in this case, our scoring-rule-under-additional-assumptions is proper.

Note that what is important for accuracy or misclassification loss is only one question: *do we classify ($\hat{y}$) everything as the majority class or not?* If we do so, accuracy or misclassification loss are happy. If not, they aren't. What is important about this question is that it has only a very tenuous connection to the quality of $\hat{q}$.

Consequently, our scoring-rules-under-additional-assumptions are not *strictly* proper, as *any* $\hat{q} \geq \theta$ will lead to the same classification evaluation. We might use the standard $\theta = 0.5$, believe that the majority class occurs with $\hat{q} = 0.99$ and classify everything as the majority class, because $\hat{q} \geq \theta$. Accuracy is high, but we have no incentive to improve our $\hat{q}$ to the correct value of η.

Or we might have done an extensive analysis of the asymmetric costs of misclassification and decided that the best classification probability threshold should actually be $\theta = 0.2$. For instance, this could happen if $y = 1$ means that you suffer from some disease. It might be better to treat you even if you don't suffer from the disease ($y = 0$), rather than the other way around, so it might make sense to treat people even if there is a low predicted probability (small $\hat{q}$) they suffer from it. We might then have a horrendously wrong model that believes that the true majority class only occurs with $\hat{q} = 0.25$ - but because of the costs of misclassification, we still classify everything as this (assumed) minority class, because again $\hat{q} \geq \theta$. If we did this, accuracy or misclassification loss would make us believe we are doing everything right, even if our predictive model does not even get which one of our two classes is the majority one.

Therefore, accuracy or misclassification loss can be misleading.

In addition, accuracy and misclassification loss *are* improper under the additional assumptions in more complex situations where the outcomes are not iid. Frank Harrell, in his blog post Damage Caused by Classification Accuracy and Other Discontinuous Improper Accuracy Scoring Rules cites an example from one of his books where using accuracy or misclassification loss will lead to a misspecified model, since they are *not* optimized by the correct conditional predictive probability.

Another problem with accuracy and misclassification loss is that they are discontinuous as a function of the threshold θ. Frank Harrell goes into this, too.

More information can be found at Why is accuracy not the best measure for assessing classification models?.

The bottom line

Don't use accuracy. Nor misclassification loss.

The nitpick: "strict" vs. "strictly"

Should we be talking about "strict" proper scoring rules, or about "strictly" proper scoring rules? "Strict" modifies "proper", not "scoring rule". (There are "proper scoring rules" and "strictly proper scoring rules", but no "strict scoring rules".) As such, "strictly" should be an adverb, not an adjective, and "strictly" should be used. As is more common in the literature, e.g., the papers by Tilmann Gneiting.

Share Cite Improve this answer Follow

answered Jul 31, 2018 at 8:28

Stephan Kolassa

137k 22 273 532

2 The threshold of 0.5 *is* the assumption. The probabilistic prediction q is mapped to a *classification* by using the threshold, and the misclassification loss is then only a function of this classification. You could calculate the misclassification loss equally for any other classification, e.g., one that rolls a die and assigns an instance to class A if we roll a 1 or 2. I did my best to explain what is a complicated and often misunderstood topic (and I do feel that everything I write about *is* relevant); I am sorry if I did not succeed. I would be happy to discuss any remaining points. – Stephan Kolassa Aug 1, 2018 at 6:13

1 As for the relevancy comment, I apologize if it came off the wrong way. I tried to focus the scope of the question to be specifically about proper vs. improper, not discontinuous/misleading/etc. I am well acquainted with the links you provided and have no issues with your comments on misclassification costs or bottom line. I am just seeking a more rigorous explanation of the statement "accuracy is improper", especially given that this paper suggests otherwise for the common use case of binary outcomes. I appreciate you taking the time to discuss this with me and share your detailed thoughts. – Zyzzva Aug 1, 2018 at 9:47

1 After further reflection, I think I have a clearer grasp of the point you are making. If we consider the

2

As already stated by Stephan Kolassa, accuracy is not even a scoring rule, so it makes no sense to ask for its propriety.

There are several scoring rules which are connected to accuracy, in that they punish misclassification, thus they are often called misclassification losses (and one is also mentioned in the Buja et al. paper). **Some of them are proper, and some are improper.** I know of none which is strictly proper.

To see this, let's assume we have observations in $y \in \{0, 1\}$ and a probabilistic prediction $p = P(Y = 1)$. A common way to transform the probability p into a classification is to set a threshold $\theta \in (0, 1)$ and classifiy as 1 if $p \geq \theta$. For $\theta = 0.5$ the misclassification loss is given by

$$s(p, y) = |\mathbf{1}(p \geq 0.5) - y|$$

and it is a proper scoring rule, but it is not strictly proper.

However, propriety completely depends on the choice $\theta = 0.5$ here. For other values this is not the case, for example the scoring rule

$$\tilde{s}(p, y) = |\mathbf{1}(p \geq 0.6) - y|$$

is improper.

To get a proper misclassification loss for a given threshold θ we need to introduce asymmetry $(1 - \theta)/\theta$ to ensure propriety, i.e. the loss is no longer symmetric. For example for $\theta = 0.6$ as above, we have that

$$s(p, y) = \mathbf{1}(p \geq 0.6)(1 - y) + \frac{2}{3}\mathbf{1}(p < 0.6)y$$

is a proper scoring rule. Again, it is not strictly proper.

In summary, not only the question whether accuracy is improper is misleading, but also the question whether misclassification loss is improper is too imprecise. It depends on the definitions you choose and unfortunately the term misclassification loss can be interpreted in several ways.

Proofs and more details can be found e.g. in these peer-reviewed papers:

1. Gneiting & Raftery (2007) Strictly Proper Scoring Rules, Prediction, and Estimation (The scoring rule is called zero-one loss here)
2. Parry (2016) Linear scoring rules for probabilistic binary classification (Section 4.2)

Share Cite Improve this answer Follow

edited Jun 16, 2024 at 13:02
Richard Hardy
71.3k 13 129 288

answered Jul 30, 2023 at 16:02
picky_porpoise
1,764 6 18

Start asking to get answers

Find the answer to your question by asking.

Ask question

Explore related questions

probability accuracy scoring-rules

See similar questions with these tags.