Abstract

There are strong incentives to build classification systems that show outstanding performance on various datasets and benchmarks. This can encourage a narrow focus on models and the performance metrics used to evaluate and compare them—resulting in a growing body of literature to evaluate and compare metrics. This paper strives for a more balanced perspective on binary classifier performance metrics by showing how uncertainty in these metrics can easily eclipse differences in empirical performance. We emphasise the discrete nature of confusion matrices and show how they can be well represented in a 3D lattice whose cross-sections form the space of receiver operating characteristic (ROC) curves. We develop novel interactive visualisations of performance metric contours within (and beyond) ROC space, showing the discrete probability mass functions of true and false positive rates and how these relate to performance metric distributions. We aim to raise awareness of the substantial uncertainty in performance metric estimates that can arise when classifiers are evaluated on empirical datasets and benchmarks, and that performance claims should be tempered by this understanding.

1. Introduction

Today’s algorithmic modeling culture [Breiman 2001] rewards those whose models outperform all others. Performance optimisation is central to statistical, ML and AI models [Thomas & Uminsky 2020], and numerical metrics are often regarded as objective, valid indicators of performance; a model’s performance on a test dataset or benchmark task is seen as indicative of its ability to perform “in the wild” on new, real-world data. This paper strives for a more balanced perspective on classifier performance metrics by visualising their distributions under different models of uncertainty. We aim to draw attention towards the role of more data—and data that is more representative of a classifier’s intended application—to characterise and improve different classifiers, rather than judging them through performance shoot-outs alone.

Confusion matrices summarise the empirical performance of classifiers. Simplest are those that tally the four outcomes of binary decisions (Figure 1). A desire to further summarise each matrix with one number—which, necessarily loses information because the matrices have three degrees of freedom—has led to many metrics with different meanings, interpretations and ranges (Appendix A).

These different performance metrics produce different classifier performance rankings, and there are often strong incentives to top these ranks [Maier-Hein et al. 2018]. Those seeking to develop or promote “the best classifier” may wonder which is “the best performance metric”; several studies argue the merits of different metrics, even though what is “best” in practice depends on a classifier’s specific real-world application [Hand 2006; Rudin & Radin 2019]. Others have sought to understand and characterise the behaviour and interpretation of these different metrics; we follow this philosophy to reveal further insight.

In essence, we show how uncertainty in (discrete) confusion matrices manifests in various (continuous) performance. Our aim is to encourage more attention towards reducing uncertainty in performance estimates before attempting to argue the merits of a particular classifier or metric.

2. Prior motivating work

2.1. Studies that argue the merits of specific metrics

Our research was stimulated by a series of papers whose titles suggest that Matthews Correlation Coefficient (MCC) is better than other metrics [Chicco & Jurman 2020; Chicco 2019].
Figure 1: (Left) A binary confusion matrix shows the counts of a classifier’s predictions in response to a set of examples whose actual classes are known. With fixed total ($N$), these four-element matrices have 3 degrees of freedom. (Right) 3D projections of confusion matrices with $N = 100$. Each point corresponds to a unique confusion matrix and is coloured by its Matthews Correlation Coefficient. We label the four extreme points ($TP = 100$, $FN = 100$, etc.) of the regular simplex. Rather than show all 176851 possible points, we show three slices corresponding to matrices where $p = 20, 50, 90$, from back to front. This graphic is a composite from our interactive visualisation (see Appendix D.1).

et al., 2021a,b). While the story in these papers is more nuanced than their titles suggest, their high citation rates suggest that they have caught people’s attention, especially in bioinformatics (Chicco, 2017). A recent opinion piece may see MCC’s popularity rise in robotic vision (Chicco & Jurman, 2022). However, Zhu (2020) challenges the idea that MCC should be “generally regarded as a balanced measure which can be used even if the classes are of very different sizes”. Note that Chicco et al. (2021b) clearly state that MCC is perhaps not the best measure in all situations.

Instead of arguing for specific metrics, we seek to enable practitioners to explore and understand the behaviour of performance metrics and the uncertainty in their empirical values. “Several rates that summarize...the confusion matrix exist nowadays; none of them, however, has reached consensus in the computer science” (Chicco et al., 2021a).

The idea of “the best” metric oversimplifies the fact that “each of the indicators serves a different purpose” (Glas et al., 2003). Practitioners have a responsibility to understand the strengths and limitations of different indicators for the task at hand, and that metrics only summarise empirical performance; they do not provide more precise estimates of performance—that demands more data, representative of the context into which a classifier would be deployed.

2.2. Studies that relate various performance metrics

Powers (2011) describes algebraic relationships between several metrics before using simulation to explore their behaviour; Ferri et al. (2009) use cluster analysis to group metrics that behave similarly. It is difficult to draw strong conclusions from the values metrics take on simulated or experimental data, especially when comparisons are averaged or clustered across different datasets.

Other authors seek to characterise performance metrics more directly by exploring how they satisfy various properties. Sokolova & Lapalme (2009) consider how the values of different metrics change as the counts in a confusion matrix change. Brzezinski et al. (2018) consider ten desirable properties of metrics on confusion matrices with fixed totals, but different prevalence. Gösgens et al. (2021) consider further desirable properties and show that some of them are incompatible. Luque et al. (2019) characterise different performance metrics as class balance changes.

2.3. Studies that visualise performance metric geometry

Broadly speaking, there have been two main approaches to visualising the values that performance metrics take over the space of possible binary confusion matrices. The first uses the rates that characterise different confusion matrices (e.g., true positive rate, false positive rate, prevalence (Flach, 2003; Luque et al., 2019)). This yields visualisations that are essentially “a collection of stacked-up ROC spaces, with the z-coordinate corresponding to the proportion of the positive class” (Brzezinski et al., 2018).

The second approach uses projections that preserve distances between confusion matrices. To explain the notion of distance here, consider the confusion matrix in Figure 1. By adding 1 to one element and subtracting 1 from another, we can construct 12 different adjacent confusion matrices, equidistant from the original matrix. Barycentric projection
With these issues in mind, we propose an isometric projection that preserves information about the underlying counts:

\[
\begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 \\
-\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & 1
\end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}.
\]

We refer to the projected points as a confusion simplex, to distinguish it from the confusion tetrahedron of (Chicco et al., 2021a). The vertices of a confusion simplex for matrices of size \(N\) are \(TP = (N, 0, 0, 0)\), \(FP = (0, N, 0, 0)\), \(FN = (0, 0, N, 0)\) and \(TN = (-\frac{N}{3}, -\frac{N}{3}, -\frac{N}{3}, N)\) as shown in Figure 1(b). The Euclidean distance between each pair of vertices is \(N\sqrt{2}\). This projection preserves information about the counts of confusion matrices and, hence, how certain we can be about classifier performance.

While drafting this paper, we learned that a similar isometric projection had been previously proposed by (Brzezinski et al., 2018) which maps the extreme confusion matrices with a given total (i.e., the matrices where \(TP = N\), \(FP = N\), etc.) to four of the corners of the cube \([-1, +1]^3\). Like the confusion tetrahedron, Brzezinski et al.’s projection carries only relative information about confusion matrices but could be rescaled to provide a rotated, translated version of Equation (1).

2.4. Studies that consider uncertainty

Classifier performance is estimated using finite amounts of test data; the more data we use, and the more representative that data is, the more certain we are about the classifier’s performance on new data. Also, in multinomial classification (Lovell et al., 2021), it is important to consider the amounts we have observed a classifier make predictions, and thus, how certain we can be about its predictive performance.

2.3. Slicing the 3D confusion simplex into ROCs

Classifiers are often evaluated and compared by presenting them with a fixed set of \(N\) examples, \(p\) positive, and \(n\) negative. Interest centres on how many actual positives are correctly identified (TP) and how many actual negatives are correctly identified (TN). Thus, the performance of classifiers can be represented in two dimensions using a rectangular lattice of \((p + 1) \times (n + 1)\) points.

Figure 1(b) shows perspective views of three different 2D slices (rectangular lattices) of points in a confusion simplex: a tall, skinny lattice \((p = 90, n = 10)\); a square lattice where classes are balanced \((p = 50, n = 50)\); and a short, wide lattice \((p = 20, n = 80)\). These slices can be shown in rectangular 2D orthographic projections (Figure 2(a)). When the axes of these projections are scaled so that they plot the true positive rate (TPR) against the false positive rate (FPR) of each point, we see these points in the space of the receiver operating characteristic (ROC) curve (Figure 2).

ROC curves provide a compact summary of empirical true and false positive rates but, by using rates, they omit information about the underlying numbers of positive and negative examples classified. These numbers indicate how certain we are about a classifiers’ performance: the more positive and negative cases the classifier has seen, the more certain we are about its performance: true and false positive rates alone do not carry this information. To inject
that information back into visualisations, we could use orthographic projection (Figure 2(a)) so that the aspect ratio clearly shows the balance of positive and negative examples. However, when classes are highly imbalanced, this approach becomes challenging to print and inspect. To deal with this, we suggest faintly plotting all possible points in the ROC space (as in Figure 2(b)) to reveal when there are few positive or negative examples. This approach can also be used with precision-recall plots (Davis & Goadrich, 2006) and we provide an interactive visualisation at to show how reference points in ROC and Precision-Recall plots relate (Appendix D.2). When $p$ or $n$ are so large that the lattice of ROC points appears continuous, we suggest labeling the axis ticks with the numbers (not rates) of false and true positives to show the amount of underlying data (as in Figure 5).

Having visualised the discrete lattice of points achievable in ROC space, we now show how to visualise continuous performance metrics in that space, and beyond.

5. The geometry of performance metrics

There are many confusion matrix performance metrics to make sense of (Appendix A). It is not obvious how these metrics behave when the four entries of the confusion matrix are not zero, which is typically of interest when people want to rank or compare the performance of different classifiers.

One way to better understand these metrics is to plot their contours, the isolines along which they take a particular value $k$. We have used this to illustrate the curved contours of the MCC above (Figure 2) and we go further by deriving algebraic expressions for the exact contours of many popular metrics—both prevalence-dependent (Appendix B) and independent (Appendix C). We provide interactive visualisations of these contours (Appendix D.3) and animations that show how contours change with prevalence for [Accuracy, Balanced Accuracy, $F_1$ score] and Matthews Correlation Coefficient (Appendix D.4). The advantage of these algebraic expressions is that they can show the exact performance metric contours for any size of confusion matrix.

The algebraic expressions of these contours also allow us to appreciate the geometry of performance metrics beyond ROC space. For example, MCC can be understood as a set of elliptical contours, whose eccentricity depends on class imbalance (Figure B.2). These contours show how performance metrics vary (and also their symmetries (Brzezinski et al., 2018; Luque et al., 2019)) in a way that our visual system can readily apprehend. To complete the picture, we need to model how confusion matrices are likely to vary.

6. Modeling uncertainty in confusion matrices

Using the notation in Figure 1, suppose a classifier is presented with $p_1$ possible examples to classify and that it gets $a_1$ of these correct (true positives) and the remainder $c_1 = p_1 - a_1$ incorrect (false negatives). On the basis of these observations, what do we believe is the probability ($\theta_a$) that this classifier correctly identifies positive examples? A frequentist approach would estimate that $\theta_a = a_1/p_1 = \text{TPR}$ is the empirical true positive rate of the classifier and that the distribution of a future correct classifications with $p$ new examples is

$$a \mid p, \theta_a \sim \text{Binomial}(p, \theta_a)$$

so that

$$P(a \mid p, \theta_a) = \binom{p}{a} \theta_a^a (1 - \theta_a)^{p-a}. \quad (3)$$
A Bayesian approach (Navarro & Perfors, 2010; Murphy, 2012; Agresti, 2013) allows us to express our prior uncertainty about a classifier’s true positive rate—this is particularly important when we have small amounts of data. We can assign a prior distribution to \( \theta_a \), and it is mathematically convenient to do that with a beta distribution:

\[
\theta_a | u, v \sim \text{Beta}(u, v).
\]

As this prior is conjugate to the binomial distribution, after observing \( a_1 \) true positive (and \( c_1 \) false negative) classifications of \( p_1 \) examples, the posterior distribution of \( \theta_a \) remains a beta distribution:

\[
\theta_a | a_1, p_1, u, v \sim \text{Beta}(u + a_1, v + c_1)
\]

and the posterior predictive distribution of seeing \( a \) correct classifications of \( p \) further examples is

\[
P(a | p, a_1, c_1, u, v) = \binom{p}{a} \frac{\text{Beta}(u + a_1 + a, v + c_1 + c)}{\text{Beta}(u + a_1 + a, v + c_1)}
\]

\[
(4)
\]

\( a \) is distributed according to a beta-binomial distribution:

\[
a | p, a_1, c_1, u, v \sim \text{BB}(a, p, u + a_1, v + c_1).
\]

Equations (4) and (5) give us the basis of a binomial model for uncertainty in confusion matrices and Equations (4) and (5) give us the basis of a beta-binomial model. Using the same logic, we can express models for the number of true negatives (\( d \)) returned by classifier presented with \( n \) negative examples when the probability of the classifier correctly identifying these negative examples is \( \theta_d \).

Using beta-binomial models for true positives and true negatives demands that we declare our prior uncertainty about \( \theta_a \) and \( \theta_d \). For demonstration, and in the absence of other relevant information, we choose the uninformative uniform priors: \( \theta_a \sim \text{Beta}(1, 1) \) and \( \theta_d \sim \text{Beta}(1, 1) \).

A common scenario in classifier development and evaluation is to present a classifier with a test set of \( p_1 \) positive and \( n_1 \) negative examples. In this situation, we can assume that the probabilities of correctly classifying a further \( a \) positive and \( d \) negative examples are statistically independent, so that

\[
P(a, d | a_1, p_1, d_1, n_1) = P(a | a_1, p_1) \cdot P(d | d_1, n_1)
\]

(6)

for the binomial model, where \( d_1 \) is the number of true negative classifications observed. The joint distribution of true positives and negatives under a beta-binomial model can be factored similarly, enabling confusion matrix uncertainty to be visualised. We note that there are potential situations where this independence assumption may not hold, for example in cytopathology analysis, if apparently normal cells are extracted from a sample obtained from a positive patient in addition to abnormal cells (Burger et al., 1981).

7. Visualising confusion matrix variation

The ideas set out in the previous section are the same as those set out by Tötsch & Hoffmann (2021), who proceed to use simulation to estimate the posterior distribution of various performance metrics under beta-binomial models of true positives and negatives. However, as Tötsch & Hoffmann point out, “the posterior distribution can be derived analytically. There is no need for Markov chain Monte Carlo sampling”. So, rather than using time-consuming simulation, we have used the pmfs of Equations (3) and (4) to develop an interactive visualisation of the impact of uncertainty on confusion matrices and their performance metrics in ROC space and in Precision-Recall space (Appendix D.5).

Figure 3 shows two screenshots from this visualisation. Unlike the histograms of samples from the posterior predictive distributions of true positive and true negative rates in Tötsch & Hoffmann (2021), Figure 3 exposes the underlying discreteness of ROC space. Note that the pmfs of the beta-binomial model are broader than those of the binomial, reflecting the additional uncertainty in true and false positive rates embodied in this Bayesian model. Appendix E uses a real-world example to illustrate the differences between these models when data are scarce (Rodrigues et al., 2013).

8. Visualising performance metric variation

We can now visualise the distribution of a given performance metric by summing the probability masses that lie along each contour of the performance metric in ROC space. The geometry of performance metric contours in conjunction with the layout of the \((n + 1) \times (p + 1)\) possible points in ROC space determines which probability masses are summed together (as illustrated in Appendix F).

Figure 4 shows the posterior predictive pmfs of MCC, BA and \( F_1 \) values, given observations of \([16 \ 8 \ 4 \ 32] \). Note the spread of these pmfs about the maximum a posteriori (MAP) value of each performance metric. This corresponds to the distributions of true and false positive rates shown in Figure 3 and puts performance comparisons into perspective: MCC, BA and \( F_1 \) each show substantial uncertainty about their observed values.

Note also the shape of the pmfs in Figure 4 and that both BA and \( F_1 \) have a few points lying above that bell-curve. These occur where several ROC points lie on the same performance metric contour (e.g., \( F_1 = \frac{4}{16} \) and \( F_1 = \frac{2}{3} \), Figure F.1). The numbers of points on each contour are shown in the bottom row of Figure 3. Adding one extra negative example markedly changes the confluence of ROC points and performance metric contours for MCC and BA, less so for \( F_1 \) (Figure F.2). The discrete nature of confusion matrices can lead to jumps in performance metrics pmfs,
even though those metrics are smooth continuous functions.

The more labelled data evaluated by a classifier, the more certain we can be about its true and false positive rates. Figure 5 shows this with the posterior predictive pmfs of confusion matrices of increasing totals but constant true and false positive rates: more data yields more precise estimates. Roughly speaking, under the binomial and beta-binomial models of uncertainty, the standard deviation of the true and false positive rate pmfs will be proportional to \( \frac{1}{\sqrt{N}} \).

Our interactive visualisations aim to foster understanding of how uncertainty in the discrete domain of confusion matrices gives rise to distributions of continuous performance metrics. Armed with that understanding, practitioners can then use more compact summaries (e.g., box plots, violin plots) to report estimated performance metric distributions. The main point to note is that there can be significant uncertainty in performance metrics, uncertainty which depends on the amount of data used in evaluating empirical performance, not the performance metrics themselves. To reduce that uncertainty requires more data that is representative of the cases the classifier will see in production.

9. Discussion

Our goal is to ensure that practitioners can understand, visualise and put into perspective the magnitude and nature of uncertainty in classifier performance estimates to inform more meaningful discussion of the strengths and limitations of specific classification systems. This work has its limitations. We appreciate that performance evaluation and model selection can involve a host of competing considerations beyond predictive performance, such as model transparency and fairness to different groups affected by model predictions; these are not within the scope of our work here. Nor do we address the common assumption that evaluation data is representative, i.e., that future data are expected to come from the same process as past data and have roughly the same range of values (McElreath 2016).

Related to that last point is the question of how data diversity and representativeness may manifest in confusion matrix uncertainty. Obviously, the more representative data we can use for evaluation the more accurate and precise our classifier performance estimates (Figure 5). But whether more diverse data results in higher or lower performance depends on how separable the two classes are by the classification system at hand. Here both data quality and quantity are important: creating synthetic data points through perturbing or interpolating real data (Chawla et al. 2002) will certainly increase the quantity of of data, it won’t necessarily increase its quality, i.e., how well it represents future examples.

The Bayesian approach we have described combines prior beliefs with observed data to form posterior distributions of plausible classifier performance metrics. A frequentist approach would hold that uncertainty is a consequence of sampling variation, leading to a focus on the sampling distribution of classifier performance metrics. Murphy (2012) contrasts and critiques these two paradigms, mentioning the bootstrap as a means to approximate the sampling distribution and confidence intervals as a way to characterise its spread. The Bayesian approach however is ideal for the situation we want to highlight: where a lack of data creates substantial uncertainty about the future performance of a classifier. Frequentist approaches are challenged by small data sets, as illustrated in Appendix E. Furthermore the Bayesian approach can yield a the functional form of a posterior (e.g., beta-binomial) enabling us to visualise its pmf precisely, whereas a frequentist approach would involve empirical approximations to the sampling distribution.

We began by mentioning the strong incentives to build classification systems that show outstanding performance and we acknowledge the frequent use of evaluation metrics selecting the “best” model. Unfortunately, uncertainty in performance metrics does not make this task any easier. Rather, it forces us to confront the limitations arising from restricted data, and the need to balance multiple competing considerations, including issues beyond predictive performance such as system fairness and transparency.

While we clearly don’t believe that there is one “best” performance metric, one could argue that the best classifier minimises the expected costs (or maximizes the benefits) of its future decisions (Section B.3). This is a conceptually appealing way to describe classifier performance with a single number but, in practice, adds further uncertainty into the evaluation process, i.e., uncertainty about the costs or benefits of different decisions. Eliciting and quantifying these costs is difficult, subjective and generally ignored.

Much attention has been given to the “problem” of class imbalance (Luque et al. 2019; Mullick et al. 2020; Lovell et al. 2021). We use quotes to emphasise that class imbalance is mainly problematic to those trying to build automated decision making systems—the real problem of rare but highly adverse cases (e.g., life threatening disease) is that they occur at all; we would not want them to happen more frequently. Also, class imbalance is a necessary consequence of multinomial classification: with \( C \) possible classes and \( N \) examples, at least one class will have equal or fewer than \( N/C \) examples.

There are two types of confusion matrix performance metric: prevalence-independent (or balanced) metrics, whose values depend only on rates of true positives and negatives (Appendix C); and prevalence-dependent metrics whose values depend on these rates and prevalence (Appendix B). (Note that balanced versions of prevalence dependent metrics can be derived—see Luque et al. 2019 and Section C.12.)
Posterior predictive pmfs: beta-binomial model

Figure 3: Joint and marginal posterior predictive pmfs under different models of uncertainty for the confusion matrix \[ \begin{bmatrix} 16 & 8 \\ 4 & 32 \end{bmatrix} \]. These are the distributions we would expect to see if the classifier that produced that confusion matrix was given a further 20 positive and 40 negative examples to classify. Top left panels show the pmfs of the false positive rate; bottom right panels show the pmfs of the true positive rate; bottom left panels show the joint pmfs of true and false positive rates in ROC space. Point areas in the joint distribution plots are proportional to the probability masses they are centred on. These are screenshots from our interactive visualisation (see Appendix D.5).

Figure 4: There is potential for substantial variation about the performance metric values of the observed confusion matrix as shown by these posterior predictive pmfs of MCC, BA, F_1 (left to right) under beta-binomial (red) and binomial models (blue) of uncertainty, given the observed confusion matrix used in Figure 3. Vertical lines show the MAP performance metric values. The green points show the number of times each value of a particular performance metric is observed: these counts are proportional to the prior pmfs of performance metric values. F_1 values of \( \frac{1}{2} \) and \( \frac{2}{3} \) are observed 11 times (Figure F.1). These are screenshots from our interactive visualisation (Appendix D.5).
This has led to discussion about which metrics are best for scenarios where classes are balanced or imbalanced. However, neither balanced nor prevalence-dependent performance metrics reduce the uncertainty arising from finite amounts of evaluation data—only additional data can do that. We note that augmenting examples of rare classes to assist with learning and performance evaluation will distort class prior probabilities, and the trained classifier’s outputs will have to be adjusted to provide accurate predictions for real-world class abundance (Saerens et al., 2002).

The ideas we have presented for binary confusion matrices could be extended to multinomial classification by treating a $C \times C$ multinomial confusion matrix as a set of $C$ binary classification problems. This one-versus-all strategy is common in multinomial classification and has been used to provide compact, informative visualisations of confusion matrices in terms of their prior and posterior odds (Lovell et al., 2021). The models of uncertainty we have presented could be applied to each class versus all others, yielding posterior predictive pmfs of various performance metrics. One advantage of this approach would be to reveal the degree of uncertainty associated with each class so that those developing classification systems could consider where best to direct their attention in making improvements. An alternate approach would be to change the probabilistic model of confusion matrix distributions from binomial (or beta-binomial) to multinomial (or Dirichlet-multinomial) (Murphy, 2012). But with $C > 2$ classes, we can no longer visualise the space of $C \times C$ confusion matrices in 3 dimensions, nor their performance metrics. Handling uncertainty in multinomial or multi-label (Gortler et al., 2022) classification scenarios is certainly an open challenge.

10. Conclusion

Publications advocating specific performance metrics have motivated us to study the (continuous) geometry of performance metrics and the (discrete) geometry of the space of confusion matrices they are applied to. Through this, we have gained a clearer understanding of the cause and effect of uncertainty in performance metrics: the primary cause is uncertainty about the confusion matrices that will be produced by a trained classifier; this depends on the numbers of actual positive and negative examples we have observed the classifier determine. Using binomial and (more conservative) beta-binomial models of uncertainty, we can calculate the exact pmfs of the predictive distribution of confusion matrices, given the classifications we have observed.

Using the contours of various performance metrics, we have demonstrated how the posterior predictive pmfs of confusion matrices can be transformed into posterior predictive pmfs of different performance metrics. We have provided a range of static and interactive visualisations for researchers and practitioners to explore uncertainty in confusion matrices and various performance metrics derived from them. These contributions aim to put performance metrics and their uncertainty into perspective, specifically, when observations of positive or negative classes are few, uncertainty in performance metrics can easily eclipse differences in performance between classifiers. Arguments about which classifier performs best, or which performance metric is best, need to take this uncertainty into account, and the visualisations we provide enable researchers to do that.

Our work also provides a useful perspective on performance evaluation where classes are imbalanced—a common scenario in binary decision making and a necessary situation in
multinomial classification. Some metrics depend on class imbalance; others do not; and “balanced” metrics can be created from prevalence-dependent ones. However, the fundamental issue in performance evaluation is not so much the relative abundance of different classes as their absolute abundance: our estimates of a classifier’s ability to correctly detect a class will be highly uncertain when there are few instances that class, regardless of balance. Performance metrics can’t address this: more data is needed and, until it arrives, we must acknowledge the uncertainty in our findings.

Finally, and with a view to the ethical dimensions of this work, we acknowledge that classifier performance evaluation goes far beyond purely quantitative metrics. It is heartening to see the breadth of issues in performance evaluation, benchmarking and datasets gaining more attention. Still, quantitative metrics will always play a prominent role in considering the strengths and limitations of classification systems. We hope that this visualisation of confusion matrix performance metrics and their uncertainties will inform their use in practice.

References


Never mind the metrics—what about the uncertainty?


Rudin, C. and Radin, J. Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition. *Harvard Data Science Review*, 1(2), 2019. doi: 10.1162/9960892.5a8a3a3d.


A. Performance metrics and other definitions

Many sums, products, ratios and functions of the four elements of a binary confusion matrix have been defined. Unfortunately, the notation varies between authors as does the row and column layout of the confusion matrix itself. We define performance metrics using the notation and layout of the confusion matrix in Figure 1 in the main paper:

<table>
<thead>
<tr>
<th>Total</th>
<th>actual class</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>predicted class</td>
<td>positive</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p = a + c</td>
<td>n = b + d</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>True Positives</td>
<td>False Positives</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(a)</td>
<td>(b)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FN</td>
<td>TN</td>
</tr>
<tr>
<td></td>
<td></td>
<td>False Negatives</td>
<td>True Negatives</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(c)</td>
<td>(d)</td>
</tr>
</tbody>
</table>

The following equations define performance metrics and other quantities used in this paper. We use an asterisk (*) to indicate quantities that depend on prevalence (Eq. 12).

Here are the definitions of the row, column and overall totals of the confusion matrix:

- Total: \( N = TP + FN + FP + TN \) (7)
- Condition Positive: \( p = TP + FN \) (8)
- Condition Negative: \( n = FP + TN \) (9)
- Predicted Positive*: \( \hat{p} = TP + FP \) (10)
- Predicted Negative*: \( \hat{n} = FN + TN \). (11)

**Prevalence** refers to the proportion of positive cases in a dataset:

- Prevalence*: \( \text{Prev} = \frac{p}{N} \). (12)

True and False Positive rates form the Positive Likelihood Ratio:

- True Positive Rate: \( TPR = \frac{TP}{p} = 1 - FNR \) (Sensitivity, Recall) (13)
- False Positive Rate: \( FPR = \frac{FP}{n} = 1 - TNR \) (14)
- Positive Likelihood Ratio: \( LR_+ = \frac{TPR}{FPR} \) (15)

while True and False Negative rates form the Negative Likelihood Ratio:

- True Negative Rate: \( TNR = \frac{TN}{n} \) (Specificity) (16)
- False Negative Rate: \( FNR = \frac{FN}{p} \) (17)
- Negative Likelihood Ratio: \( LR_- = \frac{FNR}{TNR} \) (18)

and these Likelihood Ratios form the Diagnostic Odds Ratio:

- Diagnostic Odds Ratio: \( DOR = \frac{LR_+}{LR_-} = \frac{TP \cdot TN}{FP \cdot FN} \) (19)
Never mind the metrics—what about the uncertainty?

To help visualise $LR_+$, $LR_-$ and $DOR$ in comparison to other performance metrics, we introduce the following scaled versions of their logarithms:

\[
\begin{align*}
\text{scaled log } LR_+ & = \frac{1}{\log(n(p-1)/p)} \log \left( \frac{TP}{p} \cdot \frac{n}{FP} \right) \quad \text{Section C.4} \\
\text{scaled log } LR_- & = \frac{1}{\log(pn/(n-1))} \log \left( \frac{FN}{p} \cdot \frac{n}{TN} \right) \quad \text{Section C.5} \\
\text{scaled log } DOR & = \frac{1}{\log(p-1)(n-1)} \log \left( \frac{TP \cdot TN}{FP \cdot FN} \right) \quad \text{Section C.8}
\end{align*}
\]

True Positive and True Negative rates are the basis of the following prevalence-independent performance metrics:

\[
\begin{align*}
\text{Balanced Accuracy} & = \frac{TPR + TNR}{2} = \frac{BM + 1}{2} \quad (23) \\
\text{Bookmaker Informedness} & = \frac{TPR + TNR}{2} - 1 \\
\text{Youden’s } J & = \text{Bookmaker Informedness} \\
\text{Geometric Mean} & = \sqrt{TPR \cdot TNR} \quad (25) \\
\text{Prevalence Threshold} & = \frac{\sqrt{FPR}}{\sqrt{TPR} + \sqrt{FPR}} \quad \text{See (Balayla, 2020)} \quad (26)
\end{align*}
\]

Next come ratios that relate to a classifier’s predictions. These depend on prevalence, but (Luque et al., 2019) have proposed prevalence-independent (“balanced”) versions:

\[
\begin{align*}
\text{Positive Predictive Value}^* & = \frac{TP}{p} \quad \text{Precision} \quad (27) \\
\text{Balanced PPV} & = \frac{TPR}{1 + TPR - TNR} \quad \text{See (Luque et al., 2019)} \quad (28) \\
\text{Negative Predictive Value}^* & = \frac{TN}{n} \quad \text{See (Powers, 2011)} \quad (29) \\
\text{Balanced NPV} & = \frac{TNR}{1 + TNR - TPR} \quad \text{See (Luque et al., 2019)} \quad (30)
\end{align*}
\]

and these predictive values form Markedness and its balanced version:

\[
\begin{align*}
\text{Markedness}^* & = \text{PPV} + \text{NPV} - 1 \quad (31) \\
\text{Balanced Markedness} & = \text{PPV}_{bal} + \text{NPV}_{bal} - 1 \quad \text{See (Luque et al., 2019)} \quad (32)
\end{align*}
\]

Accuracy refers to the proportion of correctly classified examples and is prevalence-dependent, unlike $BA$ and $BM$:

\[
\begin{align*}
\text{Accuracy}^* & = \frac{TP + TN}{N} \quad (33)
\end{align*}
\]

Like Accuracy, both $F_1$ and Threat Score involve ratios of sums of confusion matrix elements:

\[
\begin{align*}
\text{F}_1^* & = \frac{2 \cdot \text{PPV} \cdot TPR}{\text{PPV} + TPR} \quad \text{Sørensen–Dice coefficient} \\
& = \frac{2TP}{2TP + FP + FN} \quad (34) \\
\text{Balanced } F_1 & = \frac{2 \cdot \text{TPR}}{2 + TPR - TNR} \quad \text{See (Luque et al., 2019)} \quad (35) \\
\text{Threat Score}^* & = \frac{TP}{TP + FN + FP} \quad \text{Jaccard index, Critical Success index} \quad (36) \\
\text{Balanced Threat Score} & = \frac{TPR}{2 - TNR} \quad \text{Section C.13} \quad (37)
\end{align*}
\]
We define Decision Benefits as a weighted sum of confusion matrix elements:

\[
\text{Decision Benefits}^* \quad \text{DB} = \beta_a \cdot \text{TP} + \beta_b \cdot \text{FP} + \\
\beta_c \cdot \text{FN} + \beta_d \cdot \text{TN}.
\]

The remaining metrics combine the four elements of the confusion matrix in more complex ways:

- **Matthews Correlation Coefficient**
  \[
  \text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{\hat{p} \cdot p \cdot \hat{n} \cdot n}}
  \]
  \[
  = \text{sgn}(\text{BM}) \sqrt{\text{BM} \cdot \text{MK}}
  \]
  See (Powers, 2011) (39)

- **Balanced MCC**
  \[
  \text{MCC}_{\text{bal}} = \frac{\text{BM}}{\sqrt{1 - (\text{TPR} - \text{TNR})^2}}
  \]
  See (Luque et al., 2019) (40)

- **Fowlkes-Mallows Index**
  \[
  \text{FM} = \sqrt{\text{PPV} \cdot \text{TPR}}
  \]
  See (Fowlkes & Mallows, 1983) (41)

- **Balanced FM**
  \[
  \text{FM}_{\text{bal}} = \frac{\text{TPR}}{\sqrt{1 + \text{TPR} - \text{TNR}}}
  \]
  Section C.12 (42)

- **Cohen’s kappa**
  \[
  \kappa = \frac{2(\text{TP} \cdot \text{TN} - \text{FN} \cdot \text{FP})}{\hat{p} \cdot n + p \cdot \hat{n}}
  \]
  Section C.16 (43)
B. Performance metric contours that depend on prevalence

B.1. Matthews Correlation Coefficient

Using the notation of Figure 1, the Matthews Correlation Coefficient (Eq. (39)) can be rewritten in terms of $a, d, p, n$ as

$$\text{MCC}(a, b, c, d) = \frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$$

$$= \frac{ad - (n - d)(p - a)}{\sqrt{(a + n - d)pn(p - a + d)}}$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $-1 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \begin{cases} 
\frac{1}{2(k^2p + n)} \left( + \frac{k^2p(n + p)^2(4d(n - d) + k^2np)}{n} + 2dp(k^2 - 1) + k^2p(p - n) + 2np \right) & k \geq 0 \\
\frac{1}{2(k^2p + n)} \left( - \frac{k^2p(n + p)^2(4d(n - d) + k^2np)}{n} + 2dp(k^2 - 1) + k^2p(p - n) + 2np \right) & k < 0 
\end{cases}$$

(45)

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta) = \begin{cases} 
\frac{1}{2p(k^2p + n)} \left( + \frac{k^2p(n + p)^2(4n^2\delta(n - \delta) + k^2np)}{n} + 2n\delta p(k^2 - 1) + k^2p(p - n) + 2np \right) & k \geq 0 \\
\frac{1}{2p(k^2p + n)} \left( - \frac{k^2p(n + p)^2(4n^2\delta(n - \delta) + k^2np)}{n} + 2n\delta p(k^2 - 1) + k^2p(p - n) + 2np \right) & k < 0 
\end{cases}$$

(46)

in which case, all contours where $k < 0$ intersect at $(\alpha, \delta) = (0, 1)$ and all contours where $k > 0$ intersect at $(\alpha, \delta) = (1, 0)$, as is the case for Markedness. These intersections become apparent when we visualise the contours of the Matthews Correlation Coefficient within and beyond the ROC space (Figure B.2) which reveals that the contours describe a series of concentric ellipses whose eccentricity depends on prevalence.
Never mind the metrics—what about the uncertainty?

Figure B.2: Contours of the Matthews Correlation Coefficient within and beyond the ROC space describe a series of concentric ellipses whose eccentricity depends on prevalence and which intersect where $(FPR, TPR) = (0, 0)$ and $(FPR, TPR) = (1, 1)$.
Never mind the metrics—what about the uncertainty?

Figure B.3: Contours of the $F_1$ score in the ROC space.

B.2. $F_1$ Score

Using the notation of Figure 1, the $F_1$ score (Eq. (34)) can be rewritten in terms of $a, d, p, n$ as

$$F_1(a, b, c, d) = \frac{2a}{2a + b + c} = \frac{2a}{2a + n - d + p - a}.$$ 

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \frac{k(d - n - p)}{k - 2} \quad (47)$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta) = \frac{k(n\delta - n - p)}{p(k - 2)} \quad (48)$$

in which case, all contours intersect at $(\alpha, \delta) = (0, (p + n)/n)$. These intersections become apparent when we visualise the contours of the $F_1$ score within and beyond the ROC space (Figure B.4) which reveals how the slopes of the linear contours depend on prevalence, similar to those of the Threat Score.
Never mind the metrics—what about the uncertainty?

Figure B.4: Contours of the $F_1$ score in and beyond the ROC space are straight lines that intersect at $(\alpha, \delta) = (0, (p + n)/n)$ or, equivalently $(\text{FPR}, \text{TPR}) = (1 - (p + n)/n, 0)$.
Never mind the metrics—what about the uncertainty?

Figure B.5: Contours of the Threat Score (Jaccard Index, Critical Success Index) in the ROC space.

B.3. Threat Score (Jaccard Index, Critical Success Index)

Using the notation of Figure 1, the Threat Score (Eq. (36)) can be rewritten in terms of $a, d, p, n$ as

$$
TS(a, b, c, d) = \frac{a}{a + b + c} = \frac{a}{p + n - d}.
$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$
a(k, p, n, d) = k(p + n - d) \quad (49)
$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$
\alpha(k, p, n, \delta) = k(p + n(1 - \delta))/p \quad (50)
$$

in which case, all contours intersect at $(\alpha, \delta) = (0, (p + n)/n)$. These intersections become apparent when we visualise the contours of the $F_1$ score within and beyond the ROC space (Figure B.6) which reveals how the slopes of the linear contours depend on prevalence, similar to those of the $F_1$ Score.
Never mind the metrics—what about the uncertainty?

Figure B.6: Contours of the Threat Score score in and beyond the ROC space are straight lines that intersect at \((\alpha, \delta) = (0, (p + n)/n)\) or, equivalently \((\text{FPR}, \text{TPR}) = (1 - (p + n)/n, 0)\).
Never mind the metrics—what about the uncertainty?

Figure B.7: Contours of accuracy in the ROC space.

B.4. Accuracy

Using the notation of Figure 1, the Accuracy (Eq. (33)) can be rewritten in terms of $a, d, p, n$ as

$$TS(a, b, c, d) = \frac{a + d}{a + b + c + d} = \frac{a + d}{p + n}.$$  

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k < 1$ along the contour lines with

$$a(k, p, n, d) = k(p + n) - d$$  

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta) = \frac{k(p + n) - n\delta}{p}.$$  

These contour lines are parallel and planar.
Never mind the metrics—what about the uncertainty?

![Figure B.8: Contours of Decision Benefits in the ROC space using $\beta = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$.](image)

![Figure B.9: Contours of Decision Benefits in the ROC space using $\beta = \begin{bmatrix} 4 \\ 1 \\ 7 \end{bmatrix}$.](image)

B.5. Decision Costs (Benefits)

While it is common to refer to the costs of correct and incorrect classifications, we work here equivalently in terms of benefits in keeping with other performance metrics in this paper where “bigger is better” and to ensure the colour scales used in our graphics can be interpreted consistently (“blue is better”).

We define the matrix of benefits associated with the confusion matrix of Figure 1 as

$$\beta = \begin{bmatrix} \beta_a \\ \beta_b \\ \beta_c \\ \beta_d \end{bmatrix}.$$

Literature that concentrates on decision costs for a fixed prevalence treats this matrix as having only two degrees of freedom, e.g., the ratio (or difference) of costs for true positive and false negative classifications, and the ratio (or difference) of costs for true negative and false positive classifications. This paper considers confusion matrices with fixed totals, but varying prevalence, so our parameterisation of the benefits matrix affords three degrees of freedom.
Never mind the metrics—what about the uncertainty?

Using the notation of Figure 1, the Decision Benefits (Eq. (38)) can be rewritten in terms of $a, d, p, n$ and benefits $\beta$ as

$$DB(a, b, c, d, \beta) = a \cdot \beta_a + b \cdot \beta_b + c \cdot \beta_c + d \cdot \beta_d$$

$$= a \cdot \beta_a + (n - d)\beta_b + (p - a)\beta_c + d \cdot \beta_d$$

$$= a(\beta_a - \beta_c) + d(\beta_d - \beta_b) + p \cdot \beta_c + n \cdot \beta_d$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $k$ along the contour lines with

$$a(k, p, n, d, \beta) = -d\frac{\beta_d - \beta_b}{\beta_a - \beta_c} + \frac{k - p \cdot \beta_c - n \cdot \beta_d}{\beta_a - \beta_c}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta, \beta) = -\delta \frac{n(\beta_d - \beta_b)}{p(\beta_a - \beta_c)} + \frac{k - p \cdot \beta_c - n \cdot \beta_d}{p(\beta_a - \beta_c)}.$$

These contour lines are parallel and planar.

To help with plotting the contours for all confusion matrices of size $N$, we use a shifted and scaled version of the benefits matrix, $\beta^*$, whose minimum element is 0 and whose maximum element is $\frac{1}{N}$

$$\beta^* = \frac{1}{N} \cdot \frac{\beta - \min(\beta)}{\max(\beta)}.$$ 

This ensures feasible contours range between $[0, 1]$. In interactive plotting, we enforce the constraints

$$\beta_a > \beta_c \geq 0$$

$$\beta_d > \beta_b \geq 0$$

to ensure that the contours have finite, positive slope.

When

$$\beta = \begin{bmatrix} \beta & 0 \\ 0 & \beta \end{bmatrix}$$

for some positive $\beta$, the Decision Cost contours are the same as those of Accuracy.
B.6. Positive Predictive Value (Precision)

Using the notation of Figure 1, the Positive Predictive Value (Eq. (27)) can be rewritten in terms of \( a, d, p, n \) as

\[
PPV(a, b, c, d) = \frac{a}{a + b} = \frac{a}{a + n - d}.
\]

For given numbers of positives \( (p) \) and negatives \( (n) \), this performance metric achieves a value of \( 0 \leq k \leq 1 \) along the contour lines with

\[
a(k, p, n, d) = \frac{k(d - n)}{k - 1} \tag{55}
\]

or, in terms of true positive rate \( \alpha = a/p \) and true negative rate \( \delta = d/n \)

\[
\alpha(k, p, n, \delta) = \frac{kn(\delta - 1)}{p(k - 1)} \tag{56}
\]

in which case, all contours intersect at \( (\alpha, \delta) = (0, 1) \).
Never mind the metrics—what about the uncertainty?

Figure B.11: Contours of Negative Predictive Value in the ROC space.

B.7. Negative Predictive Value

Using the notation of Figure 1, the Negative Predictive Value (Eq. (27)) can be rewritten in terms of $a, d, p, n$ as

$$\text{NPV}(a, b, c, d) = \frac{d}{c + d} = \frac{d}{p - a + d}.$$  

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = p + \frac{d(k - 1)}{k} \quad (57)$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta) = 1 + \frac{n\delta(k - 1)}{pk} \quad (58)$$

in which case, all contours intersect at $(\alpha, \delta) = (1, 0)$. 
B.8. Markedness

Using the notation of Figure 1, Markedness (Eq. (31)) can be rewritten in terms of $a, d, p, n$ as

$$MK(a, b, c, d) = \frac{a}{a + b} + \frac{d}{c + d} - 1$$

$$= \frac{a}{a + n - d} + \frac{d}{p - a + d} - 1.$$ 

For given numbers of positives $(p)$ and negatives $(n)$, this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \begin{cases} 
\frac{\sqrt{(kn + kp + n)^2 - 4dk(n + p) + 2dk - kn + kp - n}}{2k} & k \geq 0 \\
-\frac{\sqrt{(kn + kp + n)^2 - 4dk(n + p) - 2dk + kn - kp + n}}{2k} & k < 0
\end{cases} \quad (59)$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta) = \begin{cases} 
\frac{\sqrt{(kn + kp + n)^2 - 4n\delta k(n + p) + 2n\delta k - kn + kp - n}}{2pk} & k \geq 0 \\
-\frac{\sqrt{(kn + kp + n)^2 - 4n\delta k(n + p) - 2n\delta k + kn - kp + n}}{2pk} & k < 0
\end{cases} \quad (60)$$

in which case, all contours where $k < 0$ intersect at $(\alpha, \delta) = (0, 1)$ and all contours where $k > 0$ intersect at $(\alpha, \delta) = (1, 0)$, as is the case for the Matthews Correlation Coefficient.
Never mind the metrics—what about the uncertainty?

Figure B.13: Contours of Markedness within and beyond the ROC space describe a series of polynomial curves whose shape depends on prevalence and which intersect where $(FPR, TPR) = (0, 0)$ and $(FPR, TPR) = (1, 1)$.
Never mind the metrics—what about the uncertainty?

(a) Cohen’s kappa with \( \text{Prev} = 0.1 \)
(b) Cohen’s kappa with \( \text{Prev} = 0.5 \)
(c) Cohen’s kappa with \( \text{Prev} = 0.9 \)

Figure B.14: Contours of Cohen’s kappa in the ROC space.

B.9. Cohen’s kappa

Using the notation of Figure 1, Cohen’s kappa (Eq. (44)) can be rewritten in terms of \( a, d, p, n \) as

\[
\kappa(a, b, c, d) = \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}
\]

\[
= \frac{2(ad - (n - d)(p - a))}{(a + n - d)n + p(p - a + d)}. \tag{61}
\]

For given numbers of positives \( p \) and negatives \( n \), this performance metric achieves a value of \( 0 \leq k \leq 1 \) along the contour lines with

\[
a(k, p, n, d) = \frac{dk(n - p) + 2dp - k(n^2 + p^2) - 2np}{(k - 2)n - kp} \tag{62}
\]

or, in terms of true positive rate \( \alpha = a/p \) and true negative rate \( \delta = d/n \)

\[
\alpha(k, p, n, \delta) = \frac{n\delta k(n - p) + 2n\delta p - k(n^2 + p^2) - 2np}{p(k - 2)n - kp} \tag{63}
\]

in which case, all contours intersect at

\[
(\alpha, \delta) = \left( \frac{n}{n - p}, 1 - \frac{n}{n - p} \right)
\]

when \( p \neq n \).
Never mind the metrics—what about the uncertainty?

Figure B.15: Contours of Fowlkes-Mallows index in the ROC space.

B.10. Fowlkes-Mallows index

Using the notation of Figure 1 the Fowlkes-Mallows Index (Eq. (42)) can be rewritten in terms of $a, d, p, n$ as

$$FM(a, b, c, d) = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}}$$

$$= \sqrt{\frac{a}{a + n - d} \cdot \frac{a}{p}}$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \frac{1}{2} \left( \sqrt{k^2 p(-4d + k^2 p + 4n)} + k^2 p \right) \quad (64)$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, p, n, \delta) = \frac{1}{2p} \left( \sqrt{k^2 \delta(-4n\delta + k^2 \delta p + 4\delta)} + k^2 \delta \right). \quad (65)$$

29
C. Performance metric contours that are independent of prevalence

C.1. True Positive Rate (Sensitivity)

Using the notation of Figure 1, the True Positive Rate (Eq. (13)) can be rewritten in terms of \( a, d, p, n \) as

\[
TPR(a, b, c, d) = \frac{a}{a + c} = \frac{a}{p}. 
\]

For given numbers of positives \((p)\), this performance metric achieves a value of \( 0 \leq k \leq 1 \) along the contour lines with

\[
a(k, p) = pk \quad (66)
\]

or, in terms of true positive rate \( \alpha = a/p \) and true negative rate \( \delta = d/n \)

\[
\alpha(k) = k \quad (67)
\]
C.2. True Negative Rate (Specificity)

Using the notation of Figure 1, the True Negative Rate (Eq. (16)) can be rewritten in terms of $a, d, p, n$ as

$$TNR(a, b, c, d) = \frac{d}{b + d} = \frac{d}{n}.$$

For given numbers of negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$d(k, n) = nk \quad (68)$$

or, in terms of true negative rate $\delta = d/n$

$$\delta(k) = k \quad (69)$$
Never mind the metrics—at what about the uncertainty?

Figure C.3: Contours of the geometric mean of true positive rate (sensitivity) and true negative rate (specificity) in the ROC space.

C.3. Geometric Mean

Using the notation of Figure 1, the Geometric Mean (Eq. 25) can be rewritten in terms of $a, d, p, n$ as

$$GM(a, b, c, d) = \sqrt[2]{\frac{a}{a+c} \cdot \frac{d}{b+d}} = \sqrt[2]{\frac{a}{p} \cdot \frac{d}{n}}.$$  

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \frac{k^2 np}{d}.$$  \hspace{1cm} (70)

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{k^2}{\delta}.$$  \hspace{1cm} (71)
C.4. Positive Likelihood Ratio ($LR_+$) and scaled log Positive Likelihood Ratio

Using the notation of Figure 1, the logarithm of the Likelihood Ratio of a positive outcomes (Eq. (15)) can be rewritten in terms of $a, d, p, n$ as

$$\log LR_+(a, b, c, d) = \log \left( \frac{a \ n}{p \ b} \right) = \log \left( \frac{a \ n}{p \ n - d} \right).$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $k$ along the contour lines with

$$a(k, p, n, d) = e^{k \ p(n - d)}/n \quad (72)$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = e^{k \ p(1 - \delta)} \quad (73)$$

$\log LR_+$ has range $(-\infty, \infty)$. To visualise the finite values of this function, it is convenient to work with a scaled version of this function whose contours lie between $[-1, 1]$. The largest finite value of $LR_+$ is $n(p - 1)/p$ so we can produce a scaled version of Eq. (72) by dividing $\log LR_+(a, b, c, d)$ by

$$M = \log(n(p - 1)/p) \quad (74)$$

to give

$$\text{scaled } \log LR_+(a, b, c, d) = \frac{1}{\log(n(p - 1)/p)} \log \left( \frac{a \ n}{p \ b} \right) \quad (75)$$

which yields the contour equations

$$a(k, p, n, d) = e^{MKp(n - d)}/n \quad (76)$$

and

$$\alpha(k, \delta) = e^{MKp(1 - \delta)} \quad (77)$$
Never mind the metrics—what about the uncertainty?

Figure C.5: Contours of the scaled logarithm of the negative likelihood ratio in the ROC space.

C.5. Negative Likelihood Ratio ($LR_-$) and scaled log Negative Likelihood Ratio

Using the notation of Figure 1, the logarithm of the Likelihood Ratio of a negative outcomes (Eq. (18)) can be rewritten in terms of $a, d, p, n$ as

$$\log LR_-(a, b, c, d) = \log \left( \frac{c}{p} \frac{n}{d} \right) = \log \left( \frac{p - a n}{p - d} \right).$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $k$ along the contour lines with

$$a(k, p, n, d) = p - \frac{d e^k p}{n}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = 1 - \delta e^k$$

$\log LR_-$ has range $(-\infty, \infty)$. To visualise the finite values of this function, it is convenient to work with a scaled version of this function whose contours lie between $[-1, 1]$. The largest finite value of $LR_-$ is $pn/(n - 1)$ so we can produce a scaled version of Eq. (78) by dividing $\log LR_-(a, b, c, d)$ by

$$M = \log(pn/(n - 1))$$

to give

$$\text{scaled } \log LR_-(a, b, c, d) = \frac{1}{\log(pn/(n - 1))} \log \left( \frac{c}{p} \frac{n}{d} \right)$$

which yields the contour equations

$$a(k, p, n, d) = p - \frac{d e^{Mk} p}{n}$$

and

$$\alpha(k, \delta) = 1 - \delta e^{Mk}$$
Never mind the metrics—what about the uncertainty?

Figure C.6: Contours of Bookmaker Informedness in the ROC space.

C.6. Bookmaker Informedness, BA

Using the notation of Figure 1, the Bookmaker Informedness (Eq. (24)) can be rewritten in terms of $a, d, p, n$ as

$$BM(a, b, c, d) = \frac{a}{a + c} + \frac{d}{b + d} - 1$$

$$= \frac{a}{p} + \frac{d}{n} - 1. \quad (84)$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$\alpha(k, p, n, d) = p \left( k + 1 - \frac{d}{n} \right) \quad (84)$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = k + 1 - \delta \quad (85)$$
C.7. Prevalence Threshold

Using the notation of Figure 1, the Prevalence Threshold (Eq. (26)) can be rewritten in terms of \( a, d, p, n \) as

\[
\text{PT}(a, b, c, d) = \frac{\sqrt{b/(b+d)}}{\sqrt{a/(a+c)} + \sqrt{b/(b+d)}}
= \frac{\sqrt{(n-d)/n}}{\sqrt{a/p} + \sqrt{(n-d)/n}}.
\]

For given numbers of positives \((p)\) and negatives \((n)\), this performance metric achieves a value of \( 0 \leq k \leq 1 \) along the contour lines with

\[
a(k, p, n, d) = \frac{(k-1)^2 p(n-d)}{k^2 n}
\]  

(86)

or, in terms of true positive rate \( \alpha = a/p \) and true negative rate \( \delta = d/n \)

\[
\alpha(k, \delta) = \frac{(k-1)^2 (1-\delta)}{k^2}
\]  

(87)
Never mind the metrics—what about the uncertainty?

Figure C.8: Contours of scaled log Diagnostic Odds Ratio in the ROC space.

C.8. log Diagnostic Odds Ratio and scaled log Diagnostic Odds Ratio

Using the notation of Figure [1] the logarithm of the Diagnostic Odds Ratio (Eq. 19) can be rewritten in terms of \( a, d, p, n \) as

\[
\log \text{DOR}(a, b, c, d) = \log \left( \frac{ad}{bc} \right) \quad (88)
\]

\[
= \log \left( \frac{ad}{(p-a)(n-d)} \right).
\]

For given numbers of positives \((p)\) and negatives \((n)\), this performance metric achieves a value of \( k \) along the contour lines with

\[
a(k, p, n, d) = \frac{e^k(d-n)p}{d(e^k - 1) - e^k n} \quad (89)
\]

or, in terms of true positive rate \( \alpha = a/p \) and true negative rate \( \delta = d/n \)

\[
\alpha(k, \delta) = \frac{e^k(\delta - 1)}{\delta(e^k - 1) - e^k} \quad (90)
\]

\[\log \text{DOR} \text{ has range } (-\infty, \infty). \]

To visualise the finite values of this function, it is convenient to work with a scaled version of this function whose contours lie between \([-1, 1]\). The largest finite value of DOR is \((p - 1)(n - 1)\) so we can produce a scaled version of Eq. [90] by dividing \( \log \text{DOR}(a, b, c, d) \) by

\[
M = \log(p - 1)(n - 1) \quad (91)
\]

to give

\[
\text{scaled } \log \text{DOR}(a, b, c, d) = \frac{1}{\log(p - 1)(n - 1)} \log \left( \frac{ad}{bc} \right) \quad (92)
\]

which yields the contour equations

\[
a(k, p, n, d) = \frac{e^{Mk}(d-n)p}{d(e^{Mk} - 1) - e^{Mk} n} \quad (93)
\]

and

\[
\alpha(k, \delta) = \frac{e^{Mk}(\delta - 1)}{\delta(e^{Mk} - 1) - e^{Mk}}. \quad (94)
\]
Never mind the metrics—what about the uncertainty?

Figure C.9: Contours of the balanced Matthews Correlation Coefficient in the ROC space.

C.9. Balanced Matthews Correlation Coefficient

Using the notation of Figure 1, the balanced Matthews Correlation Coefficient (Eq. (41)) can be rewritten in terms of \( a, d, p, n \) as

\[
\text{MCC}_{\text{bal}} = \frac{a/p + d/n - 1}{\sqrt{(1 - \left(\frac{a}{p} \cdot \frac{d}{n}\right)^2}}}
\]  

(95)

For given numbers of positives \((p)\) and negatives \((n)\), this performance metric achieves a value of \( k \) along the contour lines with

\[
a(k, p, n, d) = \frac{p}{n} \cdot \frac{k \sqrt{-4d^2 + 4dn + k^2n^2 + n + d(k^2 - 1)}}{k^2 + 1}
\]  

(96)

or, in terms of true positive rate \( \alpha = a/p \) and true negative rate \( \delta = d/n \)

\[
\alpha(k, \delta) = \frac{k \sqrt{-4\delta^2 + 4d\delta + k^2} + 1 + \delta(k^2 - 1)}}{k^2 + 1}
\]  

(97)
Figure C.10: Contours of the balanced Markedness in the ROC space.

C.10. Balanced Markedness

Using the notation of Figure 1, balanced Markedness (Eq. (32)) can be rewritten in terms of $a, d, p, n$ as

$$MK_{bal} = \frac{a}{a + p} + \frac{d}{d + n} - 1$$

or

$$MK_{bal} = \frac{a}{a + p - pd/n} + \frac{d}{d + n - na/p} - 1$$

(98)

(99)

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $k$ along the contour lines with

$$a(k, p, n, d) = p \left( \frac{d}{n} + \frac{\sqrt{(2k + 1)^2 - 8dk/n - 1}}{2k} \right)$$

(100)

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \delta + \frac{\sqrt{(2k + 1)^2 - 8\delta k - 1}}{2k}$$

(101)
Figure C.11: Contours of $F_{1\text{bal}}$ in the ROC space.

C.11. Balanced $F_1$ ($F_{1\text{bal}}$)

Using the notation of Figure 1, balanced $F_1$ (Eq. (35)) can be rewritten in terms of $a, d, p, n$ as

$$F_{1\text{bal}} = \frac{2a/p}{2 + a/p - d/n}$$  \hspace{1cm} (102)

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $k$ along the contour lines with

$$a(k, p, n, d) = \frac{k(p(d - 2n))}{(k - 2)n}$$  \hspace{1cm} (103)

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{k(\delta - 2)}{k - 2}$$  \hspace{1cm} (104)

in which case, all contours intersect at $(\alpha, \delta) = (0, 2)$. 


Figure C.12: Contours of $FM_{bal}$ in the ROC space.

C.12. Balanced Fowlkes-Mallows index

Using the notation of Figure 1 and the same approach as [Luque et al., 2019], a balanced version of the Fowlkes-Mallows Index (Eq. 43) can be rewritten in terms of $a, d, p, n$ as

$$FM_{bal}(a, b, c, d) = \frac{a}{p} \left(1 + \frac{a}{p} - \frac{d}{n}\right)$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \frac{kp}{2} \left(\sqrt{k^2 + 4d/n} + k\right)$$

(105)

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{k}{2} \left(\sqrt{k^2 + 4 - 4\delta} + k\right).$$

(106)
C.13. Balanced Threat Score

Using the notation of Figure 1 and the same approach as (Luque et al., 2019), a balanced version of the Threat Score (Eq. (37)) can be rewritten in terms of $a, d, p, n$ as

$$TS_{bal}(a, b, c, d) = \frac{ap}{2 - \frac{d}{n}}$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = kp \left(2 - \frac{d}{n}\right)$$

(107)

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = k (2 - \delta)$$

(108)
Never mind the metrics—what about the uncertainty?

C.14. Balanced Positive Predictive Value

Using the notation of Figure 1, the balanced Positive Predictive Value (Eq. (28)) can be rewritten in terms of $a, d, p, n$ as

$$PPV_{bal}(a, b, c, d) = \frac{a p}{a p + (1 - d/n)}$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$a(k, p, n, d) = \frac{kp(d - n)}{(k - 1)n}$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{k(\delta - 1)}{k - 1}. \quad (110)$$
C.15. Balanced Negative Predictive Value

Using the notation of Figure 1, the balanced Negative Predictive Value (Eq. (30)) can be rewritten in terms of $a, d, p, n$ as

$$\text{NPV}_{\text{bal}}(a, b, c, d) = \frac{d}{n} + \frac{1}{1 - \frac{a}{p}}$$

For given numbers of positives ($p$) and negatives ($n$), this performance metric achieves a value of $0 \leq k \leq 1$ along the contour lines with

$$\alpha(k, p, n, d) = \frac{dp(k - 1)}{kn} + p \quad (111)$$

or, in terms of true positive rate $\alpha = a/p$ and true negative rate $\delta = d/n$

$$\alpha(k, \delta) = \frac{\delta(k - 1)}{k} + 1. \quad (112)$$
C.16. Balanced Cohen’s Kappa is Bookmaker Informedness

Using the relationships:

\[
\begin{align*}
TP &= p \cdot TPR \\
FP &= n(1 - TNR) \\
FN &= p(1 - TPR) \\
TN &= n \cdot TNR
\end{align*}
\]

we can rewrite Cohen’s Kappa (Eq. (44)) as

\[
\kappa = \frac{2(p \cdot TPR \cdot n \cdot TNR - p(1 - TPR)n(1 - TNR))}{(p \cdot TPR + n(1 - TNR))n + p(p(1 - TPR) + n \cdot TNR)}.
\]

Using the same approach as [Luque et al., 2019], we can create a class-balanced version of this metric by setting \( p = n \) to give an expression in terms of the true positive and true negative rate alone:

\[
\kappa_{bal} = \frac{2(TPR \cdot TNR - (1 - TPR)(1 - TNR))}{TPR + 1 - TNR + 1 - TPR + TNR} = \frac{2(TPR \cdot TNR - 1 + TPR + TNR - TPR \cdot TNR)}{2} = TPR + TNR - 1
\]

which is the same as Bookmaker Informedness (Eq. (24)).
D. Links to interactive visualisations, animations and source code

We have used R and its `plotly` and `tidyverse` libraries (R Core Team, 2020; Sievert, 2020; Wickham et al., 2019), as well as Desmos’ Graphing Calculator (Desmos, Inc., n.d.) to provide interactive visualisations for several key concepts in this paper. These visualisations are described in this section along with links to specific figures in the main paper. Desmos automatically ensures that the underlying code is available to copy and develop further and we provide RMarkdown for all other figures.

Source code (Rmarkdown) is available from Github at [https://github.com/DavidRLovell/never-mind-the-metrics](https://github.com/DavidRLovell/never-mind-the-metrics) under the GNU General Public License v3.0.
Never mind the metrics—what about the uncertainty?

Figure D.1: Screenshot of the interactive 3D confusion simplex [http://bit.ly/see-confusion-simplex].

D.1. Interactive 3D confusion simplex

[http://bit.ly/see-confusion-simplex] shows an interactive visualisation of the 3D projection of binary confusion matrices of size 100. Each point corresponds to a unique confusion matrix and is coloured by the value of that matrix’s Matthews Correlation Coefficient (MCC). In total, there are \( \binom{100+4-1}{4-1} = 176,851 \) different binary confusion matrices of size 100. Rather than show all of these, we have taken three slices through the lattice: from back to front, the rectangular lattices of points correspond to confusion matrices where \( p = 20, 50, 90 \), respectively.

Users can mouse over the tetrahedron, then click and drag to change its orientation. Clicking on the text ‘Pos==20’ will toggle that slice of the confusion matrix.
Never mind the metrics—what about the uncertainty?

D.2. All possible ROC and Precision-Recall reference points

http://bit.ly/see-ROC-reference-points shows all possible \((p + 1) \times (n + 1)\) points in ROC and Precision-Recall spaces corresponding to confusion matrices of size \(N = p + n\), coloured from red (low) to blue (high) Balanced Accuracy. Users can change \(N\) and \(p\) by adjusting the sliders in the left hand side of the Desmos window.

This pointillist approach can be used with ROC curves and Precision-Recall plots (Davis & Goadrich, 2006) which map \((FPR, TPR)\) points in ROC space to \((TPR, PPV)\) according to

\[
(x, y) \mapsto \left( y, \frac{1 + \frac{n}{p} \cdot \frac{x}{y}}{y} \right)^{-1},
\]

a mapping which clearly depends on class balance through the factor \(\frac{n}{p}\). While (Saito & Rehmsmeier, 2015) regard these precision-recall plots as more informative than ROC curves because their achievable shape depends on prevalence, we find them hard to interpret without reference points.
Never mind the metrics—what about the uncertainty?

Figure D.3: Using reference points in the background of ROC and Precision-Recall plots can indicate the number of positive and negative examples used in a confusion matrix while maintaining a 1:1 plot aspect ratio. The points on these plots correspond to confusion matrices of size 50. We have used colours and reference contours corresponding to Balanced Accuracy (see Section C.6) to show the mapping between ROC and Precision-Recall plots, but when plotting actual data, these reference points could be made faint and unobtrusive. These screenshots are taken from our interactive visualisation of ROC and Precision-Recall reference points.

(a) All possible ROC and Precision-Recall points with 10 positive and 40 negative cases (Prev = 0.2).

(b) All possible ROC and Precision-Recall points with 25 positive and 25 negative cases (Prev = 0.5).

(c) All possible ROC and Precision-Recall points with 40 positive and 10 negative cases (Prev = 0.8).
Never mind the metrics—what about the uncertainty?

Figure D.4: Screenshot of the Desmos visualisation of confusion matrix performance metric contours (http://bit.ly/see-confusion-metrics). There are many things that users can switch on and off in this visualisation by clicking on the small round circles at the left edge of the screen.

D.3. Confusion matrix performance metric contours

http://bit.ly/see-confusion-metrics enables us to interactively visualise a range of confusion matrix performance metrics by plotting their contours, coloured from red (low) to white (middle) to blue (high). This visualisation was used to produce all of the figures in Appendices B and C.

Users can change $N$ and $p$ by adjusting the sliders in the left hand side of the Desmos window, and can set the position of a test point by adjusting the $a_1$ and $d_1$ sliders. There are many things that users can turn on and off by clicking on the small round circles at the left edge of the screen:

**Contours of prevalence-dependent and prevalence independent metrics.** These switches are titled Show Accuracy, Show MCC, through to Show Geometric Mean and, when activated, display the contours of the chosen performance metrics

**Additional information and decoration** switches allow users to show all possible ROC points; a movable test point whose corresponding confusion matrix and performance metric values can be displayed; and various titles. Importantly, users can toggle the limits of what is displayed, so that performance metric contours beyond ROC space can be visualised (as in Figure B.2).
Never mind the metrics—what about the uncertainty?

D.4. Animated performance metric contours

These animated plots show how the contours of various performance metrics change with class balance, i.e., as the number of negative examples \( n \) and positive examples \( p \) vary in confusion matrices of fixed size \( N \).

We have created animations of

- Accuracy: https://bit.ly/see-animated-accuracy
- Balanced Accuracy: https://bit.ly/see-animated-BA
- F1 Score: https://bit.ly/see-animated-F1

Each animation frame shows a two dimensional slice through the tetrahedral confusion simplex, a projection of the four dimensional confusion matrices of size 100 into three dimensions. The animation shows slices sweeping from the edge of the simplex where \( TP = p, TN = n \) through to the edge where \( FN = p, FP = n \).

Each coloured point corresponds to a specific confusion matrix in which

\[
\begin{bmatrix}
TP & FP \\
FN & TN
\end{bmatrix} = \begin{bmatrix}
TP & FP \\
p - TP & n - FP
\end{bmatrix}
\]

and \( N = p + n = 100 \). Hence, for a given \( p \) and \( n \), we can plot the \( (p + 1) \times (n + 1) \) points whose TP values range from 0 to \( p \) and whose FP values range from 0 to \( n \) while overlaying the contours of the ‘r metric.name’ performance metric ranging from \(-0.9, -0.8, \ldots, 0.9\).

Note that

- The contours of the performance metrics are defined continuously, but empirical confusion matrices can only take on values at the discrete points in these plots.
- The left hand plot shows these points and performance metric contours in ROC space in which a classifier’s true positive rate is plotted against its false positive rate in the space of rational numbers from \([0, 1] \times [0, 1]\).
Never mind the metrics—what about the uncertainty?

- The right hand plot shows these points and `r metric.name` contours as an orthographic projection of the slice of points from the confusion simplex.

- The left hand ROC plot is equivalent to re-scaling the $x$-axis of the right hand plot by a factor of $\frac{1}{n}$ and the $y$-axis by $\frac{1}{p}$.
Never mind the metrics—what about the uncertainty?

Figure D.6: Screenshot of the Desmos visualisation of confusion matrix uncertainty models (http://bit.ly/see-confusion-uncertainty). There are many things that users can switch on and off in this visualisation by clicking on the small round circles at the left edge of the screen.

D.5. Uncertainty in confusion matrices and their performance metrics

http://bit.ly/see-confusion-uncertainty enables interactive exploration of the posterior predictive pmfs of confusion matrices and three performance metrics (MCC, BA, $F_1$) under binomial and beta-binomial models of uncertainty. This visualisation was used to produce Figures 3, E.1, 4 and F.2.

Users can change $N$ and $p$ by adjusting the sliders in the left hand side of the Desmos window, and can set the position of a test point by adjusting the $a$ and $d$ sliders. There are many things that users can turn on and off by clicking on the small round circles at the left edge of the screen:

- Marginal and joint pmfs of True and False Positive rates. Users can show these posterior predictive probability mass functions for confusion matrices of size $N = p + n$ under binomial and beta-binomial models of uncertainty, given that $a$ True Positives and $d$ True Negatives have been observed.

- Posterior predictive pmfs of MCC, BA and $F_1$ can be shown using the Show PMF... switches for each performance metric. There are also switches to show the unique performance metric values (Show rug...), the number of times these unique values are observed (Show count...) and a histogram summary of the probability mass functions (Show histogram...).

- Additional information and decoration switches allow users to show all possible ROC points; a movable test point whose corresponding confusion matrix and performance metric values can be displayed; and various labels.

- Axis and point size scales are sliders that allow users to adjust the size of the points used in the joint pmf display, the maximum of the performance metric pmfs y-axis ($P_{max}$), and the maximum of the performance metric counts y-axis ($C_{max}$).

As noted on the visualisation, Desmos supports lists of up to 10 000 elements, so $np$ must be < 10 001 for the joint probability mass distributions to plot. This visualisation runs in your web browser and interaction becomes slow with larger values of $np$: we recommend starting with $N = 60, p = 20$. 

53
Never mind the metrics—what about the uncertainty?

Figure E.1: Joint and marginal posterior predictive probability mass functions (pmfs) of the cocaine purity classifier confusion matrix reported by Rodrigues et al. (2013) and used by Tötsch & Hoffmann (2021) to illustrate confusion matrix uncertainty. These are screenshots from our interactive visualisation (see Appendix D.5).

### E. Differences between binomial and beta-binomial models of uncertainty when data are scarce

To further illustrate the differences between the uncertainties conveyed by binomial and beta-binomial models, we consider the drug purity data from Rodrigues et al. (2013) used by Tötsch & Hoffmann (2021). Figure E.1 visualises the confusion matrix of 26 positive examples and 8 negative examples in which no false negatives were observed. The binomial model of the true positive rate places the entire probability mass at $TPR = 1$ (Figure E.1a, bottom right), while the beta-binomial is more moderate, suggesting that a range of true positive values from 0.8–1.0 are plausible (Figure E.1b, bottom right). Tötsch & Hoffmann (2021) explored uncertainty in this data by simulating 20,000 draws from the posterior predictive distribution of the beta-binomial model and using histograms to summarise the true positive and negative rates that were sampled. This work and the foundation that Caelen (2017) provided highlight the importance of modeling uncertainty in interpreting confusion matrices. We think that calculating and visualising the exact discrete probability mass functions conveys an even more meaningful and accurate appreciation of that uncertainty.

Do situations of data scarcity arise in the present era of “big data”? We believe so, and see two possible causes. The first is when the phenomenon of interest is genuinely (and often, mercifully) rare, as might occur in medical or health settings. The second arises in multinomial classification where the number of classes is relatively large in comparison to the total number of observations, leading to some classes with relatively few examples. Certainly, the number of publications devoted to learning from rare events and class-imbalanced data suggests that there are many researchers who are interested in working with scarce data (Haixiang et al., 2017). And, in the absence of additional information, the performance estimates of the classification models they build will be uncertain as a result of this scarce data, regardless of the performance metrics used.
Figure F.1: Two ways to show confusion matrix pmfs and performance metric contours in ROC space. Both plots show the posterior predictive pmf of confusion matrices under a beta-binomial model of uncertainty for a classifier observed to produce the confusion matrix \( \begin{bmatrix} 16 & 8 \\ 4 & 32 \end{bmatrix} \) (the same as in Figure 4(b)). Like Figure 4, the left plot uses circle areas to represent probability mass; the right plot uses ridge lines. In the background are the contours of the \( F_1 \) performance metric and in black are the contours \( F_1 = \frac{4}{11} \) and \( F_1 = \frac{2}{3} \), along each of which lie 11 points in ROC space.

F. Visualising the conjunction of ROC points and metric contours

Having calculated and visualised the posterior predictive pmf of confusion matrices of size \( N = p + n \) after observing a classifier’s empirical performance, we can visualise the distribution of a given performance metric. Let \( C \) represent the confusion matrix random variable whose pmf is \( P_C(c) \), and let \( M = \mu(C) \) represent the random variable we get by applying a performance metric function \( \mu(\cdot) \) to a confusion matrix. The distribution of that performance metric is

\[
P_M(m) = P(\mu(C) = m) = \sum_{C : \mu(C) = m} P_C(c)
\]

i.e., the probability mass at performance metric value \( m \) is sum of the probability masses where \( \mu(C) = m \). In other words, we find the pmf of the performance metric by summing the probability masses that lie along each contour of the performance metric in ROC space.

The geometry of performance metric contours in conjunction with the layout of the \( (n + 1) \times (p + 1) \) possible points in ROC space determines which probability masses are summed together. Using \( F_1 \) for demonstration, Figure F.1 visualises the posterior predictive pmfs of confusion matrices in relation to performance metric contours. Plotting these probability masses against performance metric values gives posterior predictive distributions such as shown in the top and middle rows of Figures 4 and F.2.

Figure F.3 counts the number of ROC points that occur on the same performance metric contour. With confusion matrices of 20 positive and 40 negative examples (top row) certain MCC, BA and \( F_1 \) contours intersect multiple points in ROC space, most noticeably the 0 contours, which intersect 21 points with MCC and BA (along the \((0, 0), (1, 1)\) diagonal), and 41 points with \( F_1 \) (along the \((0, 0), (1, 0)\) horizontal).

One additional negative example (Figure F.3, bottom row) removes this confluence of points and contours in ROC space for MCC and BA, but not \( F_1 \). The MCC and BA pmfs change smoothly, following bell-shaped curves as we sweep across each row (i.e., true positive rate) of the joint pmfs in the bottom left panels of Figure F.2(a) and (b)—we emphasise this by using lines to connect these probabilities in Figure F.2 just like the ridge line plot in Figure F.1. The posterior pmf of \( F_1 \) remains much the same, due to the particular linear relationship between its contours in ROC space and the values of \( p \) and \( n \) (see Section B.2).
Never mind the metrics—what about the uncertainty?

Figure F.2: These plots visualise the same quantities as Figure 4 but with a slightly different observed confusion matrix of 20 positive and 41 negative examples: [16 4; 9 32], i.e., one more false negative. Note the changes in the counts of discrete MCC and BA values in comparison to Figure 4. To emphasise the now smoothly-changing pmf functions for MCC and BA, we use lines to connect points corresponding to the same true positive rate, (i.e., horizontal slices of the joint pmfs in Figure 3).

Figure F.3: Each panel shows all the possible points in the ROC space of confusion matrices of 20 positive and 40 negative examples (top row) and 20 positive and 41 negative examples (bottom row). Points are coloured by the number of times the performance metric value at that point is observed in the confusion matrices of those totals. Three different performance metrics are presented: MCC (left), BA (middle), F\textsubscript{1} (right). Performance metric contours are shown in the background, coloured by their value. Note that one additional negative example changes the configuration of possible points in ROC space so that each possible MCC and BA value is unique (bottom left and middle); the multiplicity of different F\textsubscript{1} values remains much the same. (bottom right).