Open Science Framework DOI:10.31219/osf.io/***

Does the evaluation stand up to evaluation?
A first-principle approach to the evaluation of classifiers

K. Dyrland
<kjetil.dyrland@gmail.com>
A. S. Lundervold †
<alexander.selvikvag.lundervold@hvl.no>
P.G.L. Porta Mana
<pgl@portamana.org>
(listed alphabetically)
Dept of Computer science, Electrical Engineering and Mathematical Sciences
Western Norway University of Applied Sciences, Bergen, Norway
† & Mohn Medical Imaging and Visualization Centre, Bergen, Norway

Draft. 4 March 2022; updated 26 May 2022

Abstract to be written Machine Learning algorithms nowadays rely on a set of popular, well-known evaluation metrics, then trying to get the highest 'score' possible. Is it only the highest 'score' that matters? How can we evaluate the evaluation metrics?

We show that a machine learning classifier that is favored by the majority of evaluation metrics might not be the most optimal, thus leading to the classifier 'guessing' the most optimal class.

0 Prologue: a short story

The manager of a factory which produces a sort of electronic component wishes to employ a machine-learning classifier to assess the durability of each produced component. The durability determines whether the component will be used in one of two possible kinds of device. The classifier should take some complex features of the component as input, and output one of the two labels '0' for 'long durability', or '1' for 'short durability', depending on the component type.

Two candidate classifiers, let us call them A and B, are trained on available training data. When employed on a separate evaluation set, they yield the following confusion matrices, written in the format

$$\begin{array}{cc} & \textit{true class} \\ & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} \textit{classifier output} \\ \begin{array}{c} 0 \\ 1 \end{array} \end{array} & \begin{bmatrix} \text{True 0} & \text{False 0} \\ \text{False 1} & \text{True 1} \end{bmatrix} \end{array}$$

and normalized over the total number of evaluation data:

$$\text{classifier A:} \quad \begin{bmatrix} 0.27 & 0.15 \\ 0.23 & 0.35 \end{bmatrix}, \tag{1}$$

$$\text{classifier B:} \quad \begin{bmatrix} 0.43 & 0.18 \\ 0.07 & 0.32 \end{bmatrix}. \tag{2}$$

These matrices show that the factory produces, on average, 50% short- and 50% long-durability components.

The confusion matrices above lead to the following values of common evaluation metrics[1] for the two classifiers. Class 0 is 'positive', 1 'negative'. **Blue bold** indicates the classifier favoured by the metric, red the disfavoured:

Table 1

Metric	classifier A	classifier B
Accuracy (also balanced accuracy)	0.62	**0.75**
Precision	0.64	**0.70**
F_1 measure	0.59	**0.77**
Matthews Correlation Coefficient	0.24	**0.51**
Fowlkes-Mallows index	0.59	**0.78**
True-positive rate (recall)	0.54	**0.86**
True-negative rate (specificity)	**0.70**	0.64

The majority of these metrics favour classifier B, some of them by quite a wide relative difference. Only the true-negative rate favours classifier A, but only by a relative difference of 9%.

The developers of the classifiers, therefore, recommend the employment of classifier B.

The factory manager does not fully trust these metrics, asking, "how do I know they are appropriate?". The developers assure that these

[1] Balanced accuracy: Brodersen et al. 2010; F_1 measure: van Rijsbergen 1974; Matthews correlation coefficient: Matthews 1975; Fowlkes-Mallows index: Fowlkes & Mallows 1983.

metrics are widely used. The manager (of engineering background) comments, “I don’t remember ‘widely used’ being a criterion of scientific correctness – not after Galileo at least”, and decides to employ both classifiers for a trial period to see which factually leads to the best revenue. The two classifiers are integrated into two separate but otherwise identical parallel production lines.

During the trial period, the classifiers perform according to the classification statistics of the confusion matrices (1) and (2) above. At the end of this period, the factory manager finds that the average net gains per assessed component yielded by the two classifiers are

$$\begin{aligned} \text{classifier A:} &\quad \mathbf{3.5}\,\text{€/component}\,, \\ \text{classifier B:} &\quad -3.5\,\text{€/component}\,. \end{aligned} \tag{3}$$

That is, classifier B actually led to a *loss* of revenue. The manager therefore decides to employ classifier A, commenting with a smug smile that it is always unwise to trust the recommendations of developers, unacquainted with the nitty-gritty reality of a business. Nice! The average gains above are easy to calculate from some additional information. The final net gains caused by the correct or incorrect classification of one electronic component are as follows:

$$\begin{array}{cc} & \textit{true class} \\ & \begin{array}{cc} 0 & \quad 1 \end{array} \\ \begin{array}{c} \textit{classifier output} \\ 0 \\ 1 \end{array} & \begin{bmatrix} 15\,\text{€} & -335\,\text{€} \\ -35\,\text{€} & 165\,\text{€} \end{bmatrix} \end{array} \tag{4}$$

The reason behind these values is that short-durability components (class 1) provide more power and are used in high-end, costly devices; but they cause extreme damage and consequent repair costs and refunds if used in devices that require long-durability components (class 0). Long-durability components provide less power and are used in low-end, cheaper devices; they cause some damage if used in devices that require short-durability components, but with lower consequent costs.

Taking the sum of the products of the gains above by the respective percentages of occurrence – that is, the elements of the confusion matrix – yields the final average gain. The final average gain returned by the use of classifier A, for example, is

$$15\,\text{€} \times 0.27 - 335\,\text{€} \times 0.15 - 35\,\text{€} \times 0.23 + 165\,\text{€} \times 0.35 = 3.5\,\text{€}\,.$$

In the present case, the confusion matrices (1) and (2) lead to the amounts (3) found by the manager.

1 Issues in the evaluation of classifiers

The story above illustrates several well-known issues of currently popular evaluation procedures for machine-learning classifiers:

(a) We are swept by an avalanche of possible evaluation metrics. Often it is not clear which is the most compelling. For example, in the story above, one could argue that the true-negative rate was the appropriate metric, in view of the great difference in gains between correct and wrong classification for class 1, compared with that for class 0.

But at which point does this qualitative reasoning fail? Imagine that the net gains had been as follows instead:

$$\begin{array}{cc} & \textit{true class} \\ & \begin{array}{cc} 0 & \quad 1 \end{array} \\ \textit{classifier output} \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 45\,€ & -335\,€ \\ -65\,€ & 165\,€ \end{bmatrix} \end{array} . \tag{5}$$

One could argue that also this case there is a great economic difference between correct and wrong classification for class 1, as compared with class 0. The true-negative rate should, therefore, still be the appropriate metric. Yet a simple calculation shows that in this case, it is classifier B that actually leads to the best average revenue: 7.3 €/component, vs 4.7 €/component for classifier A. Hence the true-negative rate is *not* the appropriate metric here and our intuitive reasoning failed us.

(b) A classifier favoured by the majority of available metrics can still turn out *not* to be the best one in practice.

(c) Most popular metrics were introduced by intuitive reasoning, ad hoc mathematical operations, special assumptions (such as gaussianity[2]), and analysis of special cases. Unfortunately, such derivations do not guarantee generalization to all cases, nor that the proposed metric is uniquely determined by the assumptions, nor

---

[2] e.g. Fisher 1963 § 31 p. 183 for the Matthews correlation coefficient.

that it satisfies other basic but neglected requirements. By contrast, compare, for instance, the derivation of the Shannon entropy[3] as the *unique* metric universally satisfying a set of general, basic requirements for the amount of information; or the derivation of the probability calculus[4] as the *unique* set of rules satisfying general desiderata never heard of this word personally, do you mean desire or is it the same as desiring something? :) for inductive reasoning, learning, and prediction[5].

(d) Let us assume that some of the popular metrics identify the best algorithm 'in the majority of cases' – although it is difficult to statistically define such a majority, and no real surveys have ever been conducted to back up such an assumption. Yet, do we expect the end-user to simply *hope* not to belong to the unlucky minority? Is such uncertainty inevitable?

We cannot have a cavalier attitude towards this problem: life and death can depend on it in some machine-learning applications[6]. Imagine a story analogous to the factory one, but in a medical setting instead. The classifiers should distinguish between two tumour types, requiring two different types of medical intervention. The confusion matrices are the same (1) and (2). In this case, correct or incorrect classification leads to the following expected remaining life lengths[7] for patients in a specific age range:

$$\begin{array}{cc} & \textit{true class} \\ & \begin{array}{cc} \quad 0 \quad & \quad 1 \quad \end{array} \\ \textit{classifier output} \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} 350\,\text{months} & 0\,\text{months} \\ 300\,\text{months} & 500\,\text{months} \end{bmatrix} \end{array} . \tag{6}$$

This matrix is numerically equivalent to (4) up to a common additive constant of 335, so the final net gains are also simply shifted by this amount. The value 0 means immediate death. It is easy to see that the metrics are exactly as in Table 1, the majority favouring classifier B. And yet the use of classifier A leads to a more than

[3] Shannon 1948; Woodward 1964 § 3.2; also Good & Toulmin 1968. [4] Cox 1946; Fine 1973; Jaynes 2003 chs 1–2. Some literature cites Halpern 1999a as a critique of Cox's proof, but curiously does not cite Halpern's 1999b partial rebuttal of his own critique, as well as the rebuttals by Snow 1998; 2001. [5] Self & Cheeseman 1987; Cheeseman 1988; Russell & Norvig 2022 ch. 12. [6] cf. Howard 1980. [7] cf. the discussion in Sox et al. 2013 § 11.2.9.

six-month longer expected remaining life than classifier B. 🔧 Maybe a short reasoning for why the expected months left is as they are is good here, just something like: If the classifier predicts that the patient has a tomour, the patient goes straight into medical treatment, if the predictions is wrong, the patient immediately dies etc.

(e) Often it is not possible to temporarily deploy all candidate classifiers, as our fictitious manager did, in order to observe which factually leads to the best results. Or it may even be unethical: consider a situation like the medical one above, where a classifier may lead to more immediate deaths than another.

(f) Finally, all issues listed above are not caused by class imbalance (the occurrence of one class with a higher frequency than another), even though they can worsen for imbalanced classes[8]. For example, in our story, the two classes were perfectly balanced.

But our story also points to a possible solution to all these issues. The 'metric' that ultimately proved to be relevant to the manager was the average net monetary gain obtained by using a classifier. In the medical variation discussed in issue (d) above, it was the average life expectancy. In either case, such metric could have been easily calculated beforehand, upon gathering information about the average gains and losses of correct and incorrect classification, collected in the matrix (4) or (6), and combining these with statistics collected in the confusion matrix associated with the classifier. Denoting the former kind of matrix by (U_{ij}) and the confusion matrix by (C_{ij}), such a metric would have the formula

$$\sum_{i,j} U_{ij}\, C_{ij} \tag{7}$$

where the sum extends to all matrix elements. 🔧 Where i,j is?

In the present work, we argue that formula (7) is indeed the only acceptable metric for evaluating and comparing the performance of two or more classifiers, each with its own confusion matrix (C_{ij}) collected on relevant test data. The coefficients U_{ij}, called *utilities*, are problem-dependent. This formula is the *utility yield* of a classifier having confusion matrix (C_{ij}).

8 Jeni et al. 2013; Zhu 2020.

Our argument is based on *Decision Theory*, an overview of which is given in § 2.

The utility yield (7) is a linear combination of the confusion-matrix elements, with coefficients independent of the elements themselves. In § 3 we explore some properties of this formula and of the space of such metrics for binary classification problems. We also show that some common metrics such as precision, F_1-measure, Matthews correlation coefficient, balanced accuracy, and Fowlkes-Mallows index *cannot* be written as a linear combination of this kind. This impossibility has two consequences for such a metric. First, it means that the metric is always affected by some kind of cognitive bias. Second, there is *no* classification problem in which the metric correctly ranks the performance of all pairs of classifiers: using such a metric always leaves open the possibility that the evaluation is incorrect *a priori*. On the other hand, metrics such as accuracy, true-positive rate, true-negative rate can be written in the form (7). Consequently, each has a set of classification problems in which it correctly ranks the performance of all pairs of classifiers.

What happens if we are uncertain about the utilities appropriate to a classification problem? And what happens if the utilities are incorrectly assessed? We show in § 4 that uncertainty about utilities still leads to a metric of the form (7). We also show that an evaluation using incorrect utilities, even with relative errors as large as 20% of the maximal utility, still leads to a higher amount of correctly ranked classifiers than the use of any other popular metric.

Some remarks about the area under the curve of the receiver operating characteristic from the standpoint of our decision-theoretic approach is given in § 5.

In the final § 6, we summarize and discuss our results.

2 Brief overview of decision theory

2.1 References

Here we give a brief overview of decision theory. We only focus on the notions relevant to the problem of evaluating classifiers, and simply state the rules of the theory. These rules are quite intuitive, but it must be remarked that they are constructed in order to be logically and mathematically self-consistent: see the following references. For a presentation of decision theory from the point of view of artificial intelligence and machine learning, see Russell & Norvig 2022 ch. 15. Simple introductions are given by Jeffrey 1965; North 1968; Raiffa 1970, and a discussion of its foundations and history by Steele & Stefánsson 2020. For more thorough expositions see Raiffa & Schlaifer 2000; Berger 1985; Savage 1972; and Sox et al. 2013; Hunink et al. 2014 for a medical perspective. See also Ramsey's 1926 insightful and charming pioneering discussion.

2.2 Decisions and classes

Decision theory makes a distinction between

- the possible situations we are uncertain about, in our case, the possible classes;
- the possible decisions we can make.

This distinction is important because it prevents the appearance of various cognitive biases[9] in evaluating the probabilities and frequencies of the possible situations on the one hand, and the values of our decisions on the other. Examples are the scarcity bias[10] "this class is rare, *therefore* its correct classification must lead to high gains", and plain wishful thinking: "this event leads to high gains, *therefore* it is more probable".

Often even the number of classes and the number of decisions differ. But in using machine-learning classifiers, one typically considers situations where the set of available decisions and the set of possible classes have some kind of natural correspondence and equal cardinality. In a 'cat vs dog' image classification, for example, the classes are 'cat' and 'dog', and the decisions could be 'put into folder Cats' vs 'put into folder Dogs'. In a medical application the classes could be 'ill' and 'healthy' and the decisions 'treat' vs 'dismiss'. As already mentioned, for simplicity, most of our discussions and examples focus on binary classification.

2.3 Utilities and maximization of expected utility

To each decision, we associate several *utilities*, depending on which of the possible classes is actually true. A utility may, for instance, equal a gain or loss in money, energy, number of customers, life expectancy, or

[9] Kahneman et al. 2008; Gilovich et al. 2009; Kahneman 2011. [10] Camerer & Kunreuther 1989; Kim & Markus 1999; Mittone & Savadori 2009.

quality of life, measured in appropriate units; or a combination of such quantities.

These utilities are collected into a *utility matrix* (U_{ij}), like the ones shown in formulae (4), (5), (6). The component U_{ij} is the utility of the decision corresponding to class i if class j is true, or, briefly, the utility of class i conditional on class j.

In an individual classification instance, if we know which class is true, then the optimal decision is the one having maximal utility among those conditional on the true class. If, on the other hand, we are uncertain about which class is true, with probability p_j for class j such that $\sum_j p_j = 1$, then decision theory states that the optimal decision is the one having maximal *expected* utility $\bar{U}_i$, defined as the expected value of the utility of decision i with respect to the probabilities of the various classes:

$$\bar{U}_i \coloneqq \sum_j U_{ij}\, p_j \;. \tag{8}$$

In formulae, this principle of *maximization of expected utility* is

$$\text{choose class} \quad i^* = \arg\max_i \{\bar{U}_i\} \equiv \arg\max_i \left\{ \sum_j U_{ij}\, p_j \right\} . \tag{9}$$

A very important result in decision theory is that basic requirements of rational decision-making imply that there *must* be a set of utilities underlying the decisions of a rational agent, and the decisions must obey the principle of maximization of expected utility[11].

How are utilities determined? They are obviously problem-specific and cannot be given by the theory (which would otherwise be a model rather than a theory). Utilities can be obvious in decision problems involving gains or losses of measurable quantities such as money or energy (the utility of money is usually not equal to the amount of money, the relationship between the two being somewhat logarithmic[12]). In medical problems, they can correspond to life expectancy and quality of life; see for example Sox et al. 2013 esp. ch. 8 and § 11.2.9 and Hunink et al. 2014 esp. ch. 4 on how such health factors are transformed into utilities.

In some cases, the final utility of a single classification instance depends on a sequence of further uncertain events and further decisions.

[11] Russell & Norvig 2022 § 15.2; von Neumann & Morgenstern 1955 chs 2–3. [12] e.g. North 1968 pp. 203–204; Raiffa 1970 ch. 4.

In the story of § 0, for instance, the misclassification of a short-durability component as a long-durability one leads the final device to break only in a high fraction of cases, and in such cases, the end customer requires a refund in a high fraction of subcases; the refunded amount may even depend on further circumstances. The negative utility $U_{01} = -335\,€$ in table (4) comes from a statistical average of the losses in all these possible end results. This is the topic of so-called decision networks or influence diagrams[13]. The decision-theory subfield of *utility theory* gives rules that guarantee the mutual consistency of a set of utilities in single decisions or decision networks. For simple introductions to utility theory see Russell & Norvig 2022 § 15.2, North 1968 pp. 201–205, and the references given at the beginning of the present section.

In the present work, we do not worry about such rules in order not to complicate the discussion: they should be approximately satisfied if the utilities of a problem have been carefully assessed.

3 Evaluation of classifiers from a decision-theoretic perspective

3.1 Admissible evaluation metrics for classification problems

Maximization of expected utility is the ground rule for rational decision making[14]. In the present work, we focus on the stage where a large number of classifications have already been made by a classifier, for example, on a test dataset with N data. Denote by F_{ij} the number of instances in which the classifier chose class i and the true class was j. Then (F_{ij}) is the confusion matrix of the classifier on this particular test set. For all instances in which the classifier chose class i and the true class was j, a utility U_{ij} is eventually gained. The total utility yielded by the classifier on the test set is therefore $\sum_{ij} U_{ij}\, F_{ij}$. Dividing by N we obtain the average utility per datum, which we call the *utility yield*; it can be written as

$$\boxed{\sum_{ij} U_{ij}\, C_{ij}} \tag{10}$$

where $C_{ij} \coloneqq F_{ij}/N$ is the relative frequency of choice i and true class j, and (C_{ij}) is the normalized confusion matrix.

---

[13] Besides the general references already given: Russell & Norvig 2022 § 15.5; Howard & Matheson 2005. [14] We discuss and use it in our companion work, in preparation.

The utility yield, formula (10), is, therefore, the natural metric to evaluate and compare the performance of classifiers on a test set for a classification problem characterized by the utility matrix (U_{ij}).

Note how the utilities U_{ij} cannot depend on the frequencies F_{ij} or C_{ij}. If they did, it would mean that we had waited until *all* classification instances had been made in order to assess the value of each *single* instance. This would be a source of evaluation bias, such as the scarcity bias mentioned in § 2.2. It would, moreover, be an impossible procedure in contexts where the consequence of a single classification is manifest before the next classification is made.

If we modify the elements of a utility matrix by a common additive constant or by a common positive multiplicative constant,

$$U_{ij} \mapsto a\, U_{ij} + b \qquad a > 0\,, \tag{11}$$

then the final utilities yielded by a classifier with a particular confusion matrix are modified by the same constants. The ranking of any set of classifiers will therefore be the same. After all, an additive constant or a positive factor represent only changes in the zero or the measurement unit of our utility scale[15]. Such changes should not affect a decision problem. Indeed, the fact that they do not is another example of the logical consistency of decision theory.

3.2 Space of utility matrices for binary classification

Let us consider a binary classification problem. It is characterized by a matrix of 2×2 utilities. Let us suppose that they are not all equal; otherwise, the choice of class would be immaterial and the classification problem trivial. We can use the freedom of choosing a zero and measurement unit to bring the utility matrix to a standard form. Let us choose them such that the maximum utility is 1 and the minimum utility is 0 (note that this value may still correspond to an actual monetary loss, for example). That is, we are effecting the transformation

$$U_{ij} \mapsto \frac{U_{ij} - \min(U_{ij})}{\max(U_{ij}) - \min(U_{ij})}\,. \tag{12}$$

With this convention, it is clear that we only have two degrees of freedom in choosing the utility matrix of a binary classification problem. As a

[15] cf. Russell & Norvig 2022 § 15.2.2.

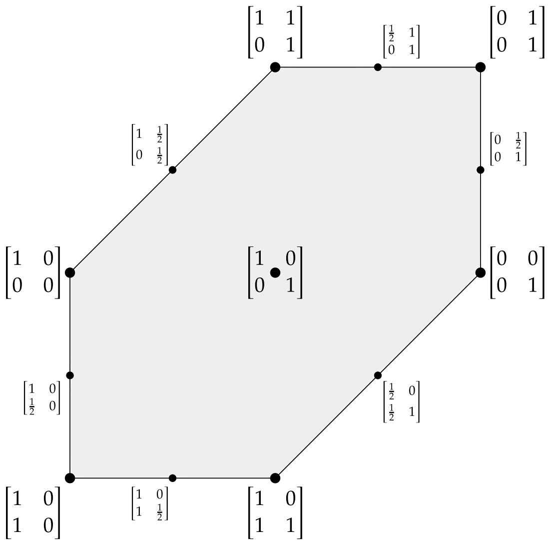

Figure 1 Space of utility matrices for binary classification.

consequence, *the space of possible evaluation metrics for binary classification is two-dimensional*. In order to evaluate candidate classifiers for a binary classification problem, we must choose a point from this space.

We can represent this space as in fig. 1. The centre is the utility matrix with equal maximum utilities for correct classification and equal minimum utilities for incorrect classification; we shall see later that it corresponds to the use of accuracy as the evaluation metric. Moving to the left from the centre, the utility for correct classification of class 1 decreases with respect to class 0; vice versa moving to the right. Moving upwards from the centre, the utility for misclassification of class 1 increases; moving downwards, the utility for misclassification of class 0 increases. We have excluded utility matrices in which misclassification has a higher utility than correct classification (although they may occur in some situations); they would appear in the missing upper-left and lower-right corners. Fixing (x, y) axes through the centre of the set, a utility matrix has coordinates

$$\begin{bmatrix} 1 - x\,\delta(x > 0) & y\,\delta(y > 0) \\ -y\,\delta(y < 0) & 1 + x\,\delta(x < 0) \end{bmatrix} . \tag{13}$$

Note that this representation is *not* meant to reflect any convex or metric properties, however. No metric or distance is defined in the space of utility matrices. Convex combination is defined if we drop the normalization (12) but it is not correctly reflected in the representation of fig. 1.

3.3 Relationship with common metrics

In § 3.1 we found that the most general evaluation metric according to decision theory must be a linear combination of the confusion-matrix elements. The coefficients of this linear combination cannot depend on the confusion-matrix elements themselves because such a dependence would reflect some sort of cognitive bias. Which common popular metrics adhere to this mathematical form? We want to answer this question in the binary classification case while giving as much allowance as possible in the typical context in which popular metrics are used.

Consider the case in which we are comparing several classifiers *on the same test set*. The number of data N and the relative frequencies f_0, f_1 with which the two classes '0', '1' occur in the test set are fixed and constant for all classifiers under evaluation.

A classifier yields a normalized confusion matrix (C_{ij}) which we write in the format

$$\begin{array}{cc} & \textit{true class} \\ & \begin{array}{cc} 0 & \quad 1 \end{array} \\ \textit{classifier output} \begin{array}{c} 0 \\ 1 \end{array} & \begin{bmatrix} C_{00} & C_{01} \\ C_{10} & C_{11} \end{bmatrix} \end{array} .$$

Owing to the constraints $C_{00} + C_{10} \equiv f_0$ and $C_{01} + C_{11} \equiv f_1$ we can always make two elements of the confusion matrix appear or disappear from any formula, replacing them with expressions involving the remaining two elements and the class frequencies. To avoid ambiguities in interpreting the functional form of mathematical formulae, let us agree to always express them in terms of C_{00} and C_{11} only, making the replacements $C_{10} = f_0 - C_{00}$, $C_{01} = f_1 - C_{11}$ wherever necessary.

Recall that given a utility matrix, we can always modify its elements by a common positive multiplicative constant a and by a common additive constant b, eq. (11), because such a modification corresponds to a change of unit and zero of the utility scale. With such a modification the evaluation metric (10) takes the equivalent form

$$a \sum_{ij} U_{ij}\, C_{ij} + b \tag{14}$$

because $\sum_{ij} C_{ij} \equiv 1$. Writing the sum explicitly and rewriting the elements C_{10}, C_{01} in terms of C_{00}, C_{11} as discussed above, this formula becomes

$$a\,(U_{00} - U_{10})\,C_{00} \;+\; a\,(U_{11} - U_{01})\,C_{11} \;+\; a\,f_0\,U_{10} + a\,f_1\,U_{01} + b\;. \quad (15)$$

Since in the present context N, f_0, f_1 are constants, we are free to construct the arbitrary constants $a > 0$ and b from them in any way we please:

$$a = a(N, f_0, f_1) > 0\;, \qquad b = b(N, f_0, f_1)\;. \quad (16)$$

We can also use this freedom to include the term $a\,f_0\,U_{10} + a\,f_1\,U_{01}$ into b in the formula above. We conclude that *an evaluation metric for binary classification complies with decision theory if and only if it can be written in the general form*

$$a(N, f_0, f_1)\;X\;C_{00} + a(N, f_0, f_1)\;Y\;C_{11} + b(N, f_0, f_1) \quad (17)$$

where X, Y are constants that do not depend on $C_{00}, C_{11}, N, f_0, f_1$; and $a(\,\cdot\,) >$ 0, $b(\,\cdot\,)$ are arbitrary functions of N, f_0, f_1 only.

A monotonic function (such as an exponential) of such form is also admissible if we only require a comparison score to rank several classifiers from best to worst.

Let us examine some common evaluation metrics for binary classification from this point of view. We write their formulae in terms of C_{00}, C_{11}.

The following metrics are particular instances of formula (17):

- ✓ *Accuracy:* $C_{00} + C_{11}$. We have $a = 1$, $X = Y = 1$, $b = 0$. Indeed it corresponds to the utility yield based on the identity utility matrix $(U_{ij}) = \left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$ (or equivalently a utility matrix that assigns the same utility to the correct classification of any class, and the same, lower utility to the misclassification of any class).
- ✓ *True-positive rate (recall):* C_{00}/f_0. Here $a = 1/f_0$, $X = 1$, $Y = 0$, $b = 0$. It corresponds to using the utility matrix $\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$.
- ✓ *True-negative rate (specificity):* C_{11}/f_1. Here $a = 1/f_1$, $X = 0$, $Y = 1$, $b = 0$. It corresponds to using the utility matrix $\left[\begin{smallmatrix}0&0\\0&1\end{smallmatrix}\right]$.

The following metrics instead *cannot* be written in the form (17), nor as monotonic functions of that form:

- ✗ *Precision:* $C_{00}/(C_{00} - C_{11} + f_1)$. Non-linear in C_{00}, C_{11}.
- ✗ *F_1-measure:* $2C_{00}/(C_{00} - C_{11} + 1)$. Non-linear in C_{00}, C_{11}. The same is true for the more general F_β-measures.
- ✗ *Matthews correlation coefficient:* $\frac{f_1\,C_{00}+f_0\,C_{11}}{\sqrt{f_0\;f_1\;(f_1+C_{00}-C_{11})\;(f_0+C_{11}-C_{00})}}$. Non-linear in C_{00}, C_{11}.
- ✗ *Fowlkes-Mallows index:* $C_{00}/\sqrt{f_0\,(f_1 + C_{00} - C_{11})}$. Non-linear in C_{00}, C_{11}.
- ✗ *Balanced accuracy:* $C_{00}/(2f_0) + C_{11}/(2f_1)$. Despite being linear in C_{00}, C_{11} and an average of two metrics (true-positive and true-negative rate) that are instances of formula (17), it is not an instance of that formula, because the two averaged metrics involve different $a(\,\cdot\,)$ functions.

We see that many popular evaluation metrics do not comply with the principles of decision theory. Any such metric suffers from two problems.

First, as discussed in § 2, the metric involves an interdependence of utilities and classification frequencies, which implies some form of cognitive bias[16].

Second, the ranking of confusion matrices yielded by the metric does not fully agree with that yielded by any utility matrix – a full agreement would otherwise imply that the metric could be written in the form (17). Some confusion matrices must therefore be incorrectly ranked. Since any rational classification problem is characterized by some underlying utility matrix, this means that the incompliant metric will always lead to some wrong evaluations. By contrast, compliant metrics such as the accuracy give completely correct rankings for all pairs of confusion matrices, at least in a specific set of classification problems.

The second phenomenon is illustrated in the plots of figs 2–3. Each blue dot in a plot represents a hypothetical confusion matrix obtained from a test dataset in a binary classification problem. The dot's coordinates are the utility yield of that confusion matrix according to a particular utility matrix underlying the classification problem, and the score of the confusion matrix according to another metric. The underlying utility matrix is $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$ for all plots in the left column, and $\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$ for all plots in the right column. The other metrics considered, one for each row

16 Hand & Christen 2018 discuss such biases regarding the F_1-measure.

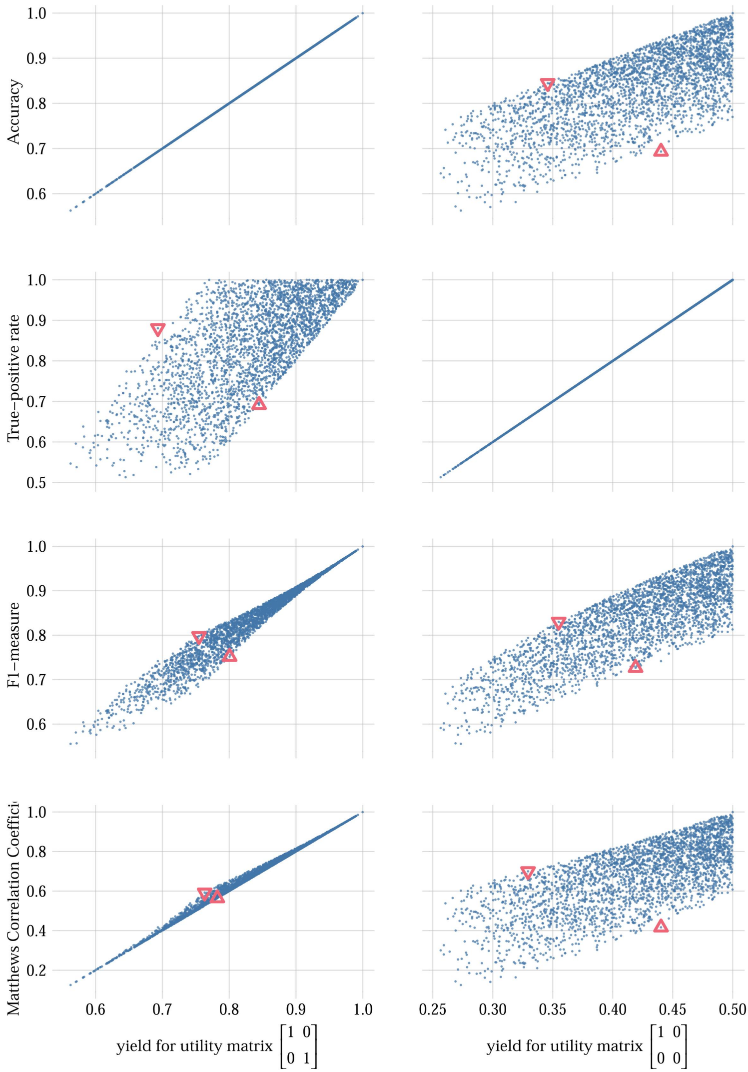

Figure 2 Relationship between various evaluation metrics and actual utility yields for a two binary classification problems with underlying utility matrices $\left[\begin{smallmatrix}1 & 0\\0 & 1\end{smallmatrix}\right]$ (left column) and $\left[\begin{smallmatrix}1 & 0\\0 & 0\end{smallmatrix}\right]$ (right column). All confusion matrices (blue dots) are obtained from a dataset with 50%/50% class balance. Pairs of red triangles in a plot show two confusion matrices that are wrongly ranked by the metric (y-axis) with respect to the actual utility yield (x-axis). Clearly, there can even be three or more confusion matrices ranked in completely reverse order by the metric. The accuracy yields correct evaluations the classification problem on the left column; and the true-positive rate, for the one on the right.

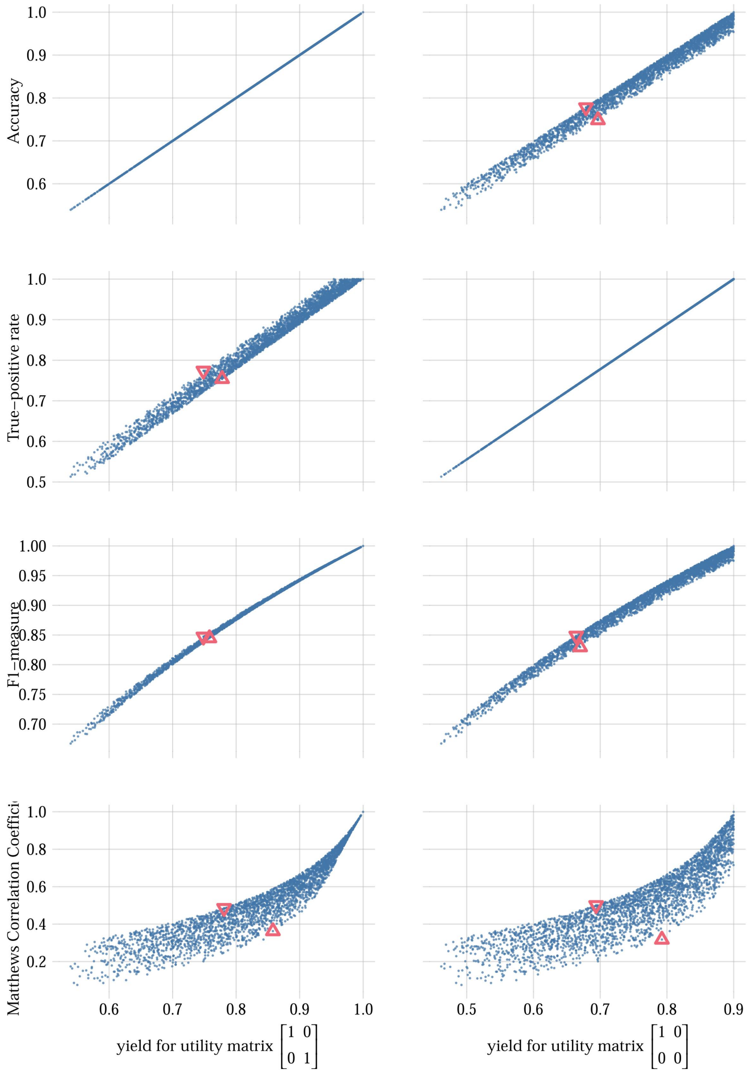

Figure 3 As for fig. 2 but for confusion matrices obtained from an imbalanced dataset with 90% occurrence of class 0 ('positive') and 10% of class 1 ('negative').

of plots, are accuracy, true-positive rate (recall, class 0 being 'positive'), F_1-measure, Matthews correlation coefficient.

The confusion matrices are selected by first fixing a proportion of classes in the dataset, which is 50%/50% (balanced dataset) for all plots in fig. 2 and 90%/10% (imbalanced dataset) for all plots in fig. 3; and then choosing true-positive and true-negative rates independently distributed between 1/2 and 1 This is a little unclear, what do you mean by 'choosing true-positive and true-negative rates independently distributed between 1/2 and 1' with linearly increasing probabilities. These confusion matrices therefore represent the classification statistics produced by classifiers that tend to have good performance – as is clear from the fact that the points tend to accumulate on the upper-right corners of the plots.

We see that the accuracy (first-row plots) always gives correct relative evaluations of all confusion matrices when the underlying utility matrix is equivalent to $\left[\begin{smallmatrix}1 & 0\\ 0 & 1\end{smallmatrix}\right]$ (left column): the y-coordinate is a monotonically increasing function – in fact a linear function – of the x-coordinate. Accuracy is indeed the utility yield corresponding to the identity utility matrix. The true-positive rate (second-row plots) always gives correct relative evaluations (provided the test set is the same) when the underlying utility matrix is equivalent to $\left[\begin{smallmatrix}1 & 0\\ 0 & 0\end{smallmatrix}\right]$ (right column). When each of these two metrics is used for a problem having a different underlying utility matrix, however, there is no deterministic relationship between the metric's score and the actual utility yield: the y-coordinate is not a function of the x-coordinate. Thus it is always possible to find two or more confusion matrices for which the metric gives completely reversed evaluations with respect to the actual utility yield, scoring the worst confusion matrix – and thus its associated algorithm – as the best, and the best as the worst. A couple of long sentences in a row, could rephrase it to something like: 'When each of these two metrics is used for a problem having a different underlying utility matrix, there is no deterministic relationship between the metric's score and the actual utility yield: the y-coordinate is not a function of the x-coordinate. In this case, it is always possible to find two or more confusion matrices for which the metric gives completely reversed evaluations based on the actual utility yield. Thus scoring the worst confusion matrix – and its associated algorithm – as the best, and the best as the worst. Pairs of red triangular shapes in a plot are examples of confusion matrices wrongly ranked by the y-axis metric.?' Pairs of red triangular shapes in a plot are examples of confusion matrices wrongly ranked by the y-axis metric.

Metrics such as accuracy and true-positive rate, complying with formula (17), thus require us to rely on evaluation *luck* only when they are used in the wrong classification problem.

The plots for the F_1-measure (third-row plots) and Matthews correlation coefficient (fourth-row plots) show that these two metrics do not have any functional relationship with the actual utility yield. It is again always possible to find two or more confusion matrices for which either metric gives completely reversed evaluations with respect to the actual utility yield. But for these two metrics, unlike accuracy and true-positive rate, cases of incorrect evaluation will *always* occur in every classification problem.

Metrics such as F_1-measure and Matthews correlation coefficient, not complying with formula (17), thus *always require us to rely on luck in our evaluations*. There are no classification problems for which these metrics lead to always correct evaluations.

A metric non-compliant with decision theory can lead to a large number of correct results for some classification problems and test sets. The bottom-left plot of fig. 2, for instance, shows that the Matthews correlation coefficient is almost a monotonically increasing deterministic function of the utility yield when the underlying utility matrix is the identity and the dataset is balanced (but it is not when the underlying utility matrix is $\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$ or the dataset is imbalanced; see corresponding plots). Such an occasional partial agreement is useless, however. Knowledge of the utility matrix is a prerequisite for relying on such partial agreement–but with this knowledge, we can directly use the actual utility yield instead, which has an exact agreement and is easier to compute.

4 Unknown or incorrect utilities

So far, we have argued that the natural evaluation metric for a classifier is the utility yield of its confusion matrix, according to the utilities underlying the classification problem of interest. We have also argued that many popular metrics, those not complying with formula (17), must always a priori lead to instances of incorrect evaluation. Our arguments are based on the principles of decision theory.

Several interrelated questions spring from our arguments, though:

- What to do when we are uncertain about the utilities underlying a classification problem?

- What happens if the utilities we use are actually wrong, that is, not the true ones underlying the problem?
- How often do uncompliant metrics such as F_1-measure or Matthews correlation coefficient lead to incorrect results, on average?

In fact, if a small error in the assessment of the utilities led to a large number of wrong evaluations, and at the same time, incompliant metrics led to a small number of wrong evaluations on average, then all the rigorousness of decision-theoretic metrics would be useless in practice, and incompliant metrics would be best for real applications.

This is not the case, however. We now discuss how to deal with uncertainty about the utilities and present an important result: Using wrong utilities, even with relative errors almost as large as 20% of the maximum utility, still leads to fewer incorrect relative evaluations on average than using some of the most common metrics.

4.1 Unknown utilities; average performance on several classification problems

Dealing with unknown utilities is straightforward. Suppose we are uncertain whether the utility matrix appropriate to a classification problem is $\boldsymbol{U}^{(1)} \equiv \bigl(U^{(1)}_{ij}\bigr)$, or $\boldsymbol{U}^{(2)}$, or $\boldsymbol{U}^{(3)}$, and so on, where the number of alternatives can even be infinite or continuous. Each alternative $\boldsymbol{U}^{(a)}$ has a probability q_a, or probability density $q(a)\,\mathrm{d}a$ in the continuous case. Then *for the classification problem, we should use the expected utility matrix*

$$\hat{\boldsymbol{U}} := q_1\, \boldsymbol{U}^{(1)} + q_2\, \boldsymbol{U}^{(2)} + q_3\, \boldsymbol{U}^{(3)} + \cdots \tag{18}$$

or $\hat{\boldsymbol{U}} := \int q(a)\, \boldsymbol{U}^{(a)}\, \mathrm{d}a$ *in the continuous case.*

We only give a sketch of the proof of this intuitive result. If we are uncertain about the utility matrix, then we have a double decision problem: choosing the optimal utility and choosing the optimal class. If the true utility matrix is, for instance, $\boldsymbol{U}^{(2)} \equiv \bigl(U^{(2)}_{ij}\bigr)$, and the true class is class 0, then choosing class 1 would yield a utility $U^{(2)}_{10}$; choosing class 0 would yield a utility $U^{(2)}_{00}$, and so on. Our double decision problem is thus characterized by a rectangular utility matrix that is the row-concatenation of the utility matrices $\boldsymbol{U}^{(a)}$. We make the realistic judgement that the probabilities q_a of the utility matrices and the probabilities p_j of the classes are independent, so that $q_a \cdot p_j$ is the probability that the true utility matrix is $\boldsymbol{U}^{(a)}$ and the true class is j. The principle of maximum expected utility, § 2.3 eq. (9), then leads to the maximization of the expected utilities

$$\bar{U}_i \coloneqq \sum_{j,a} U^{(a)}_{ij}\, q_a \cdot p_j \equiv \sum_j \underbrace{\left[\sum_a q_a\, U^{(a)}_{ij}\right]}_{\hat{\boldsymbol{U}}} p_j \tag{19}$$

in which the expected utility matrix (18) appears as the 'effective' utility matrix to be used for the class-decision problem alone.

If our uncertainty is symmetric with respect to the utilities conditional on the different classes – for instance, our uncertainty about the utilities conditional on class 0 is the same as on class 1 – then the expected utility matrix is equivalent to the identity matrix. The utility yield is in this case equal to the accuracy. The accuracy is, therefore, the natural evaluation metric to use if we are in a complete state of uncertainty regarding the underlying utilities. This fact is indeed reflected in some results discussed in § 4.2.

For a binary classification problem the set of possible utility matrices can be represented as in fig. 1, as discussed in § 3.2. Our uncertainty about the true underlying utility matrix corresponds to a discrete or continuous distribution of probability over this set. Note, however, that the expected utility matrix (18) does *not* correspond to the mass-centre of the distribution because of the peculiar coordinate system used in that figure. The actual mass-centre is obtained by representing the set of utility matrices as a two-dimensional surface (a tetrahedron) in three-dimensional space; for brevity, we do not discuss this representation in the present work.

The procedure of averaging utilities, formula (18), also applies if we want to evaluate how a classifier performs on average on several classification problems, which differ in their utility matrices. Again, what we need to use is the average of their utility matrices.

4.2 Consequences of wrong utility assessments and comparison with common metrics

It may happen that our assessment of the utility matrix of a classification problem is incorrect, especially if it has been made on semi-quantitative grounds owing to a lack of information. Then our comparative evaluations of classifiers may also end up being incorrect. What is the probability of an incorrect comparative evaluation, on average, in such cases? and how does it depend on the amount of error in the utilities? Is it higher than the probability of incorrect evaluation by other metrics?

A precise answer to these questions is extremely difficult, if not impossible, because to define 'on average' we would need to conduct a survey of classification problems of any kind, collecting statistics about their underlying utility matrices, about the confusion matrices of candidate classification algorithms for their solution, and about the errors committed in assessing utilities. We try to give a cursory answer to the questions above for the binary classification case, based on the following assumptions and judgements:

(i) Two possible distributions of true utility matrices on the set of fig. 1 (in that coordinate system): a uniform distribution; and a bivariate (truncated) gaussian distribution centred on the identity matrix $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$ and with standard deviation $1/3$ in the x and y coordinates of eq. (13), illustrated in fig. 4. Nice figures!

(ii) A distribution of confusion matrices for which the fraction of one class is uniformly distributed in $[0, 1]$, and the true-positive and true-negative rates are independently distributed in $[0.5, 1]$ with linearly increasing probabilities (median of 0.85, lower and upper quartiles at 0.75 and 0.93). This means that we consider problems with highly imbalanced data to be as common as problems with balanced data (a realistic assumption, according to our experience) and candidate classifiers to be generally good.

(iii) A truncated gaussian distribution of error around each true utility-matrix element, centred on the true utility value. We consider standard deviations ranging from 0 to 0.3. The gaussian must be truncated because each true utility has a value between 0 and 1, and moreover, we require correct classifications to have higher utilities than incorrect ones. Figure 5 illustrates the extent of such an error in the space of utility matrices, for standard deviations equal to 0.1 and 0.2.

Under these assumptions, we calculate how often a pair of classifiers, having two confusion matrices with the same class proportions, is evaluated in reverse order, with respect to their true utility yield, when

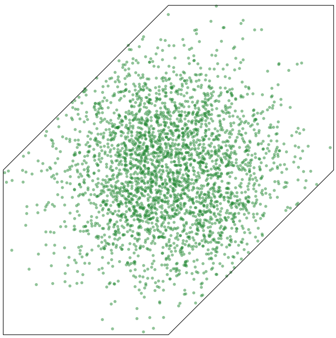

Figure 4 Truncated gaussian distribution in the space of utility matrices of fig. 1, described in item (i).

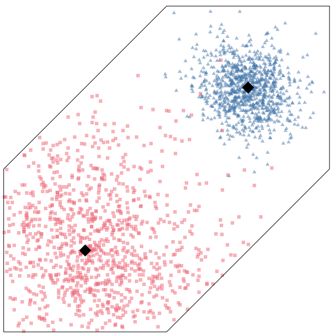

Figure 5 Extents of errors having standard deviations 0.1 (blue triangles) and 0.2 (red squares), around the utility matrices $\left[\begin{smallmatrix}0.5 & 0.5\\ 0 & 1\end{smallmatrix}\right]$ and $\left[\begin{smallmatrix}1 & 0\\ 0.5 & 0.1\end{smallmatrix}\right]$ (black diamonds).

an incorrect utility matrix or another metric is used for the evaluation. This calculation is an integration problem that we solve by Monte Carlo sampling. The procedure is intuitive:

1. Select a 'true' utility matrix according to the distribution (i).
2. Select errors around the elements of the true utility matrix, according to the distribution (iii), and add them to it.
3. Select a class proportion and then two confusion matrices having that class proportion (the class proportion must be the same since the matrices are obtained from the same data), according to the distributions (ii).
4. Calculate the difference in the true utility yield of the two confusion matrices, using the true utility from step 1.
5. Calculate the difference in

   5a. the scores according to several metrics

   5b. the yields according to the erroneous utility matrix from step 2

   for the two confusion matrices.
6. If the differences from steps 4 and 5 disagree in sign then this pair of confusion matrices was incorrectly ranked by the metric or the incorrect utility matrix. I dont fully understand this last process of

calculating the utility matrix, maybe you can explain in person? Or an exact example could be added.

The results of this sampling procedure for the case of uniform distribution of true utility matrices, several metrics, and utilities affected by errors with 0.1 standard deviation, are shown in fig. 6. Each point represents a pair of confusion matrices (step 3); its coordinates are the true utility yield and either the score given by a metric or (last plot) the yield according to the incorrect utility matrix. The red or yellow triangular points in the II and IV quadrants (discordant signs) are incorrectly ranked pairs. The percentages of incorrect rankings are calculated from 10^6 samples, giving slightly more than one decimal significant digit; fewer samples are shown in the plots.

The plots are displayed in order (left-right, top-bottom) of decreasing percentages of incorrect rankings. The accuracy metric proves to be the best among the ones considered, leading to 8.7% incorrect pairwise rankings. But we see that a utility matrix affected by gaussian errors with 0.1 standard deviation is even better, yielding 4% incorrect pairwise rankings.

The dependence of the fraction of incorrect rankings on the standard deviation of the error affecting the utilities is shown in the plots of fig. 7, for the case of uniform distribution (top plot) and gaussian distribution (bottom plot) of true utility matrices. It is approximately linear. The plots also report the fractions of incorrect rankings for the other metrics. We see that evaluations based on a utility matrix affected by errors with standard deviation up to 0.15 or even 0.25 are still more reliable than evaluations based on the other reported metrics. This is a remarkable fact, considering that errors with such standard deviations are quite large, as was shown in fig. 5.

A utility error with standard deviations around 0.25 covers the whole space of utility matrices almost uniformly (cf fig. 5). Such a large error means that we are almost completely uncertain about the utilities to start with. It therefore makes sense that the accuracy, equivalent to using the identity utility matrix, becomes a more reliable metric when this error level is reached: as we saw in § 4.1, the identity utility matrix is the natural one to use in a state of complete uncertainty about the utilities.

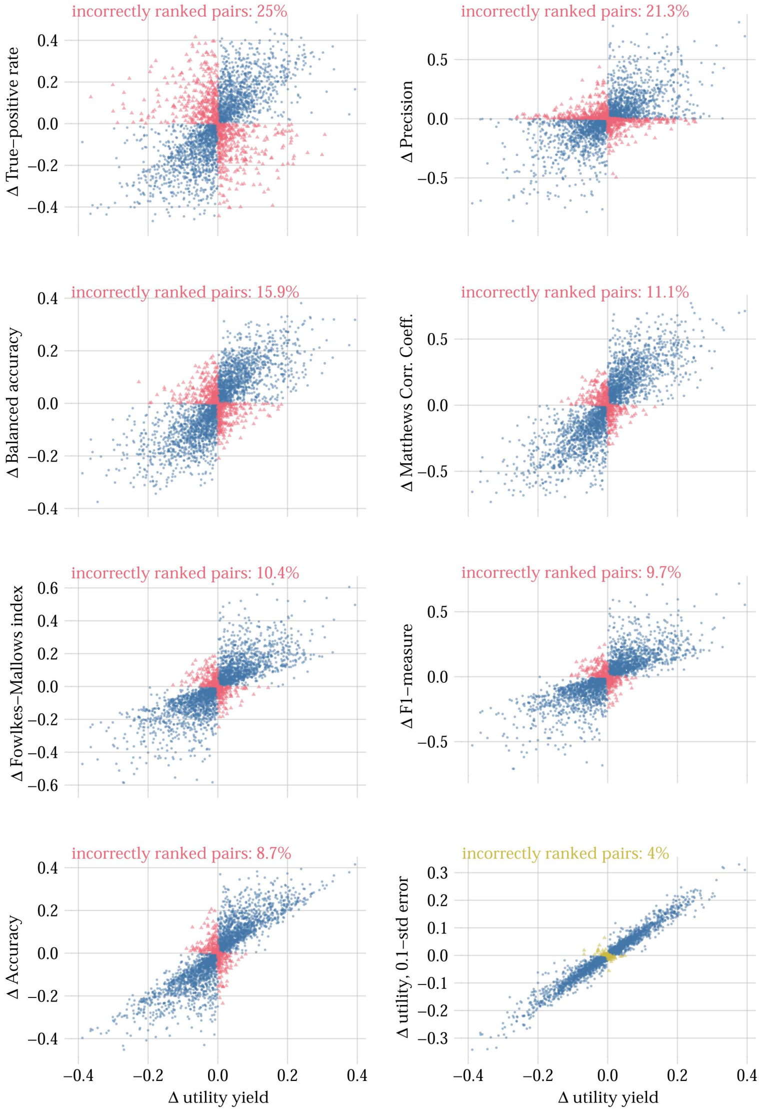

Figure 6 Relationship between difference in utility yields according to a 'true' utility matrix, and difference in scores according to other metrics including an incorrectly assessed utility matrix (error with 0.1 standard deviation). Points landing in the II or IV quadrants represent pairs of confusion matrices that were wrongly compared.

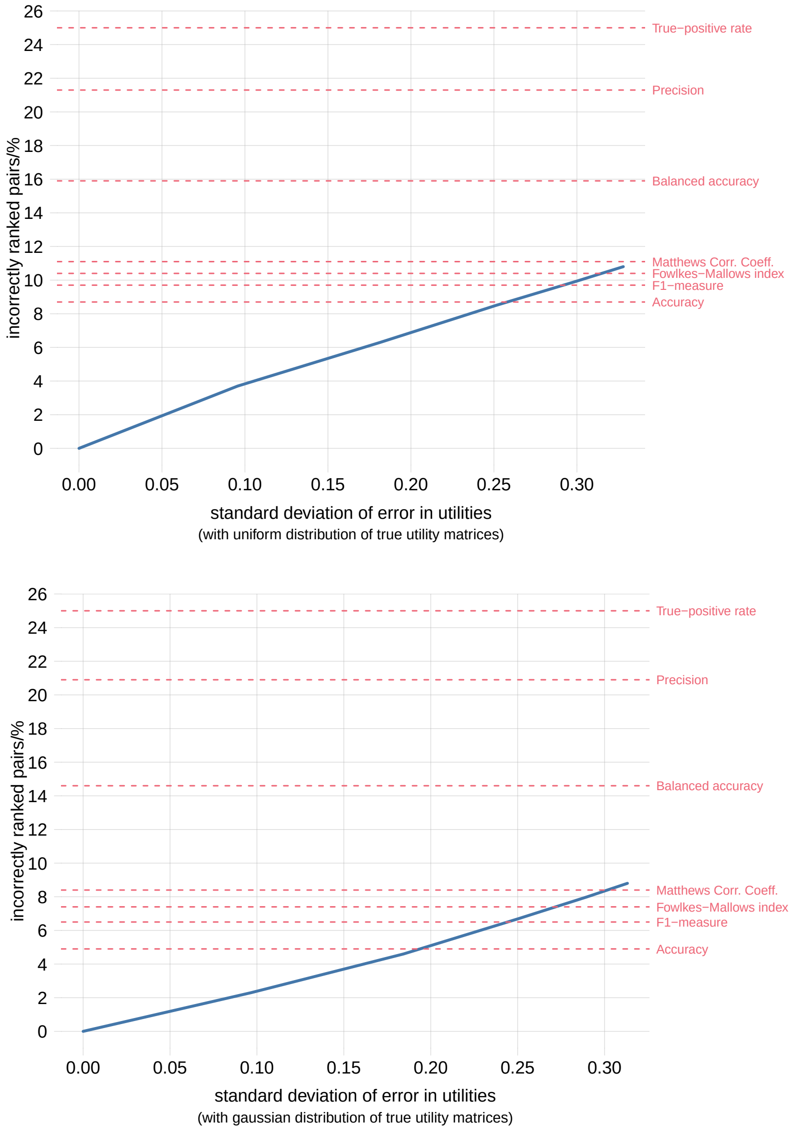

Figure 7 Dependence of the proportion of incorrectly ranked pairs of confusion matrices, on the standard deviation of the assessment error on the utilities. Top plot: case with uniform distribution of true utility matrices. Bottom plot: case with gaussian distribution of true utility matrices, as in fig. 4.

Very nice!

5 What about the area under the curve of the receiver operating characteristic?

Another very common metric for evaluating binary classifiers is the Area Under the Curve of the Receiver Operating Characteristic, or 'area under the curve' for short. This metric can only be used for particular classifying algorithms, and its meaning is different from that of the metrics reviewed so far. For these reasons, we leave a full discussion of it to future works and only offer a couple of remarks here.

The area under the curve can only be computed for classifiers that output a continuous variable rather than a class. A threshold for this variable determines whether its value predicts one class or the other. Different choices of threshold lead to different pairs of false-positive rate f (which is $1 -$ true-negative rate) and true-positive rate t in a given test set. These pairs can be plotted as a curve $f \mapsto t(f)$ on a graph with corresponding axes. Given the proportion of classes in the test set, every point on such curve corresponds to a possible confusion matrix $C_{ij}(f)$ that the classifier can produce depending on the threshold chosen. The area subtended by such a curve is a weighted average of true-positive rates with a peculiar choice of weights; the weights are uniform as a function of the false-positive rate, but generally not uniform as a function of the threshold, for example. The meaning and proper use of the receiver operating characteristic are discussed in a classic by Metz 1978, see especially p. 290.

From the standpoint of decision theory, two remarks can be made[17]. First, according to the principle of maximum expected utility, § 2.3, we should choose a threshold and corresponding false-positive rate f^* such as to maximize the utility yield, given by eq. (10):

$$\text{choose} \quad f^* = \arg\max_f \left\{ \sum_{i,j=0}^{1} U_{ij}\, C_{ij}(f) \right\} . \tag{20}$$

Any other values of f and of the threshold are irrelevant. Averages over f values are therefore irrelevant as well. Second, suppose our goal is to evaluate the average performance over several possible classification problems. In that case, the quantities to be averaged are the utility

[17] similar points are made by Baker & Pinsky 2001; Lobo et al. 2008.

matrices of those classification problems, as discussed in § 4.1, yielding a unique expected utility matrix. Once this is computed, we go back to a single choice of f according to our first remark.

Owing to these issues, the area under the curve suffers from the same problems as the non-compliant metrics discussed in § 3.3: in every classification problem, it always leads to cases of incorrect evaluation.

A correct use of the receiver-operating-characteristic curve $t(f)$ can be made, however. It is explained in Metz 1978 section *Cost/Benefit Analysis* p. 295, and in Sox et al. 2013 § 5.7.4 (curiously Sox et al. also mention the generally erroneous criterion of the area under the curve).

Denote the proportion of class 0 (positive) in the test set by B. The confusion matrix as a function of f is then

$$\begin{bmatrix} C_{00}(f) & C_{01}(f) \\ C_{10}(f) & C_{11}(f) \end{bmatrix} = \begin{bmatrix} B\,t(f) & (1-B)\,f \\ B\,[1-t(f)] & (1-B)\,(1-f) \end{bmatrix} . \tag{21}$$

The sum in formula (20) above can then be explicitly written, rearranging some terms,

$$\begin{aligned} \sum_{i,j=0}^{1} U_{ij}\, C_{ij}(f) \equiv (U_{00}-U_{10})\,B\,t(f) - (U_{11}-U_{01})\,(1-B)\,f \;+ \\ U_{10}\,B + U_{11}\,(1-B)\,. \end{aligned} \tag{22}$$

The principle of maximum expected utility (20) is then equivalent to the following condition, obtained using the explicit sum above but dropping the constant term on the second line for simplicity:

$$\text{choose} \quad f^* = \arg\max_f \bigl\{(U_{00}-U_{10})\,B\,t(f) - (U_{11}-U_{01})\,(1-B)\,f\bigr\}\,. \tag{23}$$

The function in braces is monotonically increasing because $t(f)$ is (we assume, as always, that the utility of correct classification of a class is higher than that of misclassification, so $U_{00}-U_{10} \geqslant 0$ and $U_{11}-U_{01} \geqslant 0$). Its maximum can thus be found by setting its derivative to zero:

$$\text{choose } f^* \text{ such that} \quad t'(f^*) = \frac{(U_{11}-U_{01})\,(1-B)}{(U_{00}-U_{10})\,B}\,. \tag{24}$$

If we have several classifiers, each with its own curve $t(f)$, then *the best is the one tangent to the line*

$$t = \frac{(U_{11}-U_{01})\,(1-B)}{(U_{00}-U_{10})\,B} f + \text{const.} \tag{25}$$

that has the highest intercept.

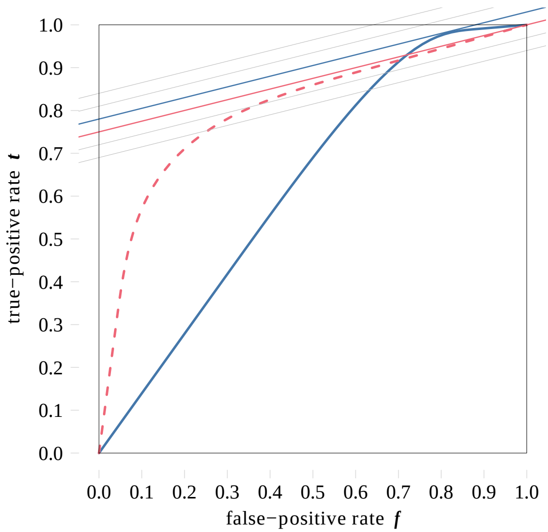

Figure 8 Receiver-operating-characteristic curves of two classifiers. The red dashed curve clearly subtends a larger area than the blue solid curve. Yet the classifier with the latter curve yields a higher utility, because it touches the family of parallel lines, eq. (25), at a higher point. This example arises for a utility matrix equal to $\left[\begin{smallmatrix}4&0\\0&1\end{smallmatrix}\right]$ and a test set with $B = 0.5$ (balanced), or for a utility matrix equal to $\left[\begin{smallmatrix}1&0\\0&1\end{smallmatrix}\right]$ and a test set with $B = 0.8$.

From this criterion it can be seen geometrically that if a classifier has its curve $t(f)$ *completely above* the curve of another classifier, then it must have a higher utility yield. But nothing, in general, can be said if the curves of the two classifiers cross. It is the tangent of a receiver-operating-characteristic curve that matters, not its subtended area. Figure 8 shows an example of this.

6 Summary and discussion

The evaluation and ranking of classification algorithms is a critical stage in their development and deployment. Without such evaluation, we cannot even say whether an algorithm is better than another, or whether a set of parameter values for a specific algorithm is better than another set.

And yet, at present, we have not an evaluation theory but only an evaluation folklore: different procedures, proposed only out of intuition and of analysis of special cases, with fuzzy criteria to decide which should be used, and without rigorous theoretical foundations that should guarantee uniqueness and universality properties and absence of biases. We believe that some of the surprising failures of machine learning *in actual applications*[18] come not only from biases in the choice of test datasets and other similar biases but also from the use of wrong evaluation metrics in the development stage.

In the present work, we have argued that theoretical foundations for the evaluation process are available in *Decision Theory*. Its main notions and principle – utilities and their maximization – are very intuitive, as shown (we hope) by the introductory story.

These are the main results of the application of decision theory to the evaluation of classifiers:

– The evaluation metric must depend on the specific classification problem.

– Such metric is completely defined by n^2 parameters, called utilities, collected in a utility matrix; n is the number of classes. Two parameters are arbitrary and represent a zero and measurement unit of the utility scale. In the binary classification case, this means that we have a two-dimensional set of possible metrics.

– The score of a classifier on a test set is simply given by its utility yield: the grand sum of the products of the elements of the utility matrix and the confusion matrix of the classifier. It is, therefore, a simple linear expression in the confusion-matrix elements.

– A utility matrix, obtained from an average, is also used when we are uncertain about the utilities underlying a classification problem or when we want to consider the average performance over several classification problems.

– Some popular metrics such as precision, balanced accuracy, Matthews correlation coefficient, Fowlkes-Mallows index, F_1-measure, and area under the receiver-operating-characteristic curve do not comply with decision theory. As a consequence, they are affected by cognitive biases and always lead to some erroneous comparative evaluations of classifiers in every classification problem, even when all utilities and frequencies are correctly assessed.

18 see e.g. Varoquaux & Cheplygina 2022.

– Using a utility matrix with incorrectly assessed utilities still leads, on average, to fewer wrong comparative evaluations than using other popular metrics.

We believe that the decision-theoretic evaluation of classifiers also has remarkable advantages:

First, it translates the fuzzy problem "which of the numerous scores should I rely on?" into a more structured, thus easier to confront, one: to assess, at least semi-quantitatively, how many times more valuable, desirable, or useful is the correct classification of a class than its incorrect classification, than the correct classification of another class, and so on. Such utilities usually have a more immediate, problem-dependent interpretation than other metrics.

Second, it leads to a mathematically simple, computationally convenient metric: a linear combination of confusion-matrix elements–no need for non-linear functions or integration of curves.

Third, the principles of the underlying theory guide us if we have to face new peculiar problems. Imagine, for instance, a classification problem where we cannot say, in general, whether true positives are more important than true negatives and so on, because such valuation can *vary from one tested item to another*. Decision theory, in this case, requires an item-wise assessment of utilities, and still provides an item-wise score, which can be accumulated across items to obtain a total evaluation score for the performance of candidate classifiers.

Author contributions

All authors have contributed equally to the present work, especially KD.

Thanks

PGLPM thanks Maja, Mari, Miri, Emma for continuous encouragement and affection; Buster Keaton and Saitama for filling life with awe and inspiration; and the developers and maintainers of LaTeX, Emacs, AUCTeX, Open Science Framework, R, Python, Inkscape, LibreOffice, Sci-Hub for making a free and impartial scientific exchange possible.

Bibliography

('de *X*' is listed under D, 'van *X*' under V, and so on, regardless of national conventions.)

Baker, S. G., Pinsky, P. F. (2001): *A proposed design and analysis for comparing digital and analog mammography special receiver operating characteristic methods for cancer screening*. J. Am. Stat. Assoc. **96**454, 421–428. DOI:10.1198/016214501753168136.

Berger, J. O. (1985): *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. (Springer, New York). DOI:10.1007/978-1-4757-4286-2. First publ. 1980.

Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M. (2010): *The balanced accuracy and its posterior distribution*. Proc. Int. Conf. Pattern Recognit. **20**, 3121–3124. DOI:10.1109/ICPR.2010.764.

Camerer, C. F., Kunreuther, H. (1989): *Decision processes for low probability events: policy implications*. J. Policy Anal. Manag. **8**4, 565–592. DOI:10.2307/3325045.

Cheeseman, P. (1988): *An inquiry into computer understanding*. Comput. Intell. **4**2, 58–66. DOI:10.1111/j.1467-8640.1988.tb00091.x.

Cox, R. T. (1946): *Probability, frequency, and reasonable expectation*. Am. J. Phys. **14**1, 1–13. DOI:10.1119/1.1990764.

Fine, T. L. (1973): *Theories of Probability: An Examination of Foundations*. (Academic Press, New York). DOI:10.1016/C2013-0-10655-1.

Fisher, R. A. (1963): *Statistical Methods for Research Workers*, rev. 13th ed. (Hafner, New York). First publ. 1925.

Fowlkes, E. B., Mallows, C. L. (1983): *A method for comparing two hierarchical clusterings*. J. Am. Stat. Assoc. **78**383, 553–569. DOI:10.1080/01621459.1983.10478008.

Gilovich, T., Griffin, D., Kahneman, D., eds. (2009): *Heuristics and Biases: The Psychology of Intuitive Judgment*, 8th pr. (Cambridge University Press, Cambridge, USA). DOI:10.1017/CBO9780511808098. First publ. 2002.

Good, I. J., Toulmin, G. H. (1968): *Coding theorems and weight of evidence*. IMA J. Appl. Math. **4**1, 94–105. DOI:10.1093/imamat/4.1.94.

Halpern, J. Y. (1999a): *A counterexample to theorems of Cox and Fine*. J. Artif. Intell. Res. **10**, 67–85. DOI:10.1613/jair.536. See also Snow (1998), Halpern (1999b).

— (1999b): *Cox's theorem revisited*. J. Artif. Intell. Res. **11**, 429–435. DOI:10.1613/jair.644. See also Snow (1998).

Hand, D., Christen, P. (2018): *A note on using the F-measure for evaluating record linkage algorithms*. Stat. Comput. **28**3, 539–547. DOI:10.1007/s11222-017-9746-6.

Howard, R. A. (1980): *On making life and death decisions*. In: **schwingetal1980**: 89–113. With discussion. DOI:10.1007/978-1-4899-0445-4_5. Repr. in **howardetal1984** pp. 481–506.

Howard, R. A., Matheson, J. E. (2005): *Influence diagrams*. Decis. Anal. **2**3, 127–143. DOI:10.1287/deca.1050.0020. First publ. 1984 in **howardetal1984** pp. 719–762.

Hunink, M. G. M., Weinstein, M. C., Wittenberg, E., Drummond, M. F., Pliskin, J. S., Wong, J. B., Glasziou, P. P. (2014): *Decision Making in Health and Medicine: Integrating Evidence and Values*, 2nd ed. (Cambridge University Press, Cambridge). DOI:10.1017/CBO9781139506779. First publ. 2001.

Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. DOI:10.1017/CBO9780511790423, https://archive.org/details/XQUHIUXHIQUHIQXUIHX2, http://www-biba.inrialpes.fr/Jaynes/prob.html.

Jeffrey, R. C. (1965): *The Logic of Decision*. (McGraw-Hill, New York).

Jeni, L. A., Cohn, J. F., De La Torre, F. (2013): *Facing imbalanced data: recommendations for the use of performance metrics*. Proc. Int. Conf. Affect. Comput. Intell. Interact. **2013**, 245–251. DOI:10.1109/ACII.2013.47.

Kahneman, D. (2011): *Thinking, Fast and Slow*. (Farrar, Straus and Giroux, New York).

Kahneman, D., Slovic, P., Tversky, A., eds. (2008): *Judgment under uncertainty: Heuristics and biases*, 24th pr. (Cambridge University Press, Cambridge). DOI:10.1017/CBO9780511809477. First publ. 1982.

Kim, H., Markus, H. R. (1999): *Deviance or uniqueness, harmony or conformity? A cultural analysis.* J. Pers. Soc. Psychol. **77**[4], 785–800. DOI:10.1037/0022-3514.77.4.785.

Lobo, J. M., Jiménez-Valverde, A., Real, R. (2008): *AUC: a misleading measure of the performance of predictive distribution models*. Glob. Ecol. Biogeogr. **17**[2], 145–151. DOI:10.1111/j.1466-8238.2007.00358.x, https://www2.unil.ch/biomapper/Download/Lobo-GloEcoBioGeo-2007.pdf.

Matthews, B. W. (1975): *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*. Biochim. Biophys. Acta **405**[2], 442–451. DOI:10.1016/0005-2795(75)90109-9.

Metz, C. E. (1978): *Basic principles of ROC analysis*. Semin. Nucl. Med. **VIII**[4], 283–298. DOI:10.1016/S0001-2998(78)80014-2.

Mittone, L., Savadori, L. (2009): *The scarcity bias*. Appl. Psychol. **58**[3], 453–468. DOI:10.1111/j.1464-0597.2009.00401.x.

North, D. W. (1968): *A tutorial introduction to decision theory*. IEEE Trans. Syst. Sci. Cybern. **4**[3], 200–210. DOI:10.1109/TSSC.1968.300114, https://stat.duke.edu/~scs/Courses/STAT102/DecisionTheoryTutorial.pdf.

Raiffa, H. (1970): *Decision Analysis: Introductory Lectures on Choices under Uncertainty*, 2nd pr. (Addison-Wesley, Reading, USA). First publ. 1968.

Raiffa, H., Schlaifer, R. (2000): *Applied Statistical Decision Theory*, repr. (Wiley, New York). First publ. 1961.

Ramsey, F. P. (1926): *Truth and probability*. In: **ramsey1931_r1950**: ch.VII:156–198. Repr. in **kyburgetal1964_r1980**, pp. 23–52. Written 1926.

Russell, S. J., Norvig, P. (2022): *Artificial Intelligence: A Modern Approach*, Fourth Global ed. (Pearson, Harlow, UK). First publ. 1995.

Savage, L. J. (1972): *The Foundations of Statistics*, 2nd rev. and enl. ed. (Dover, New York). First publ. 1954.

Self, M., Cheeseman, P. C. (1987): *Bayesian prediction for artificial intelligence*. In: *Proceedings of the third conference on uncertainty in artificial intelligence (uai'87)*, ed. by J. Lemmer, T. Levitt, L. Kanal (AUAI Press, Arlington, USA): 61–69. Repr. in arXiv DOI:10.48550/arXiv.1304.2717.

Shannon, C. E. (1948): *A mathematical theory of communication*. Bell Syst. Tech. J. **27**[3, 4], 379–423, 623–656. https://archive.org/details/bstj27-3-379, https://archive.org/details/bstj27-4-623, http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.

Snow, P. (1998): *On the correctness and reasonableness of Cox's theorem for finite domains*. Comput. Intell. **14**[3], 452–459. DOI:10.1111/0824-7935.00070.

— (2001): *The reasonableness of possibility from the perspective of Cox*. Comput. Intell. **17**[1], 178–192. DOI:10.1111/0824-7935.00138.

Sox, H. C., Higgins, M. C., Owens, D. K. (2013): *Medical Decision Making*, 2nd ed. (Wiley, New York). DOI:10.1002/9781118341544. First publ. 1988.

Steele, K., Stefánsson, H. O. (2020): *Decision theory*. In: *Stanford encyclopedia of philosophy*, ed. by E. N. Zalta (The Metaphysics Research Lab, Stanford). https://plato.stanford.edu/archives/win2020/entries/decision-theory. First publ. 2015.

van Rijsbergen, C. J. (1974): *Foundation of evaluation*. J. Doc. **30**^{4}, 365–373. DOI:10.1108/eb026584.

Varoquaux, G., Cheplygina, V. (2022): *Machine learning for medical imaging: methodological failures and recommendations for the future*. npj Digit. Med. **5**^{1}, 48. DOI:10.1038/s41746-022-00592-y.

von Neumann, J., Morgenstern, O. (1955): *Theory of Games and Economic Behavior*, 3rd ed., 6th pr. (Princeton University Press, Princeton). https://archive.org/details/in.ernet.dli.2015.215284. First publ. 1944.

Woodward, P. M. (1964): *Probability and Information Theory, with Applications to Radar*, 2nd ed. (Pergamon, Oxford). DOI:10.1016/C2013-0-05390-X. First publ. 1953.

Zhu, Q. (2020): *On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset*. Pattern Recognit. Lett. **136**, 71–80. DOI:10.1016/j.patrec.2020.03.030.