How is the evaluation evaluated?

A first-principle approach to the evaluation of classifiers

K. Dyrland ⓘ   A. S. Lundervold   P.G.L. Porta Mana ⓘ

<kjetil.dyrland@gmail.com>   <***@***>   <pgl@portamana.org>

(listed alphabetically)

Draft. 4 March 2022; updated 18 May 2022

✎ Abstract to be written

0   Prologue: a short story

The manager of a factory which produces a kind of electronic component wishes to employ a machine-learning classifier to assess the durability of each produced component, which determines whether the it 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

$$\begin{array}{c} \textit{true class} \\ \begin{array}{cc} 0 & \quad 1 \end{array} \end{array}$$

$$\begin{smallmatrix}\textit{classifier} \\ \textit{output}\end{smallmatrix}\begin{smallmatrix}0\\1\end{smallmatrix}\begin{bmatrix} \text{True 0} & \text{False 0} \\ \text{False 1} & \text{True 1} \end{bmatrix}$$

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; blue bold indicates the classifier favoured by the metric, red the disfavoured:

---

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

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 favours 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 this kind of metrics and decides to employ both classifiers for a trial period, to see which factually leads to the best revenue. The two classifiers are integrated in 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

$$\text{classifier A:} \quad \textbf{3.5} \, \text{€/component} \, ,$$
$$\text{classifier B:} \quad -3.5 \, \text{€/component} \, . \tag{3}$$

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 developers' recommendations.

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:

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

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 obtained by multiplying matrices (4) and (1) element-wise, and then taking the grand sum. 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 current popular evaluation procedures for machine-learning classifiers:

(a) We are faced by an avalanche of possible evaluation metrics. Often it is not clear which is the most compelling. 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:

$$\begin{array}{c}\textit{true class}\\[2pt]\begin{array}{cc}0 & 1\end{array}\\[2pt]\begin{array}{c}\textit{classifier}\\\textit{output}\end{array}\begin{array}{c}0\\1\end{array}\left[\begin{array}{cc}45\,\text{€} & -335\,\text{€}\\-65\,\text{€} & 165\,\text{€}\end{array}\right].\end{array} \qquad (5)$$

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 which 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 and our intuitive reasoning failed us here.

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

3

(c) Most popular metrics were introduced by intuitive reasoning, ad hoc mathematical operations, special assumptions (such as gaussianity[2]), and analysis of special cases. ⚘ Possibly add note here about absence of literature Unfortunately such derivations do not guarantee generalization to all cases, nor that the proposed metric satisfies all basic requirements or that it is uniquely determined by them. 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 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). Correct or incorrect classification in this case leads to the following expected remaining life lengths for patients in a specific age range:

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

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

---

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

this amount. The value 0 means immediate death. It is easy to see that the metrics are exactly as is Table 1, the majority favouring classifier B. And yet the use of classifier A leads to a more than six-month longer expected remaining life than classifier B.

(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 situation as 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 higher frequency than another), even though they can become worse for imbalanced classes[7]. In our story the two classes were perfectly balanced.

But our story also points to a possible solution of 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 this '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 matrix by (U_{ij}) and the confusion matrix by (C_{ij}), this 'metric' would be

$$\sum_{i,j} U_{ij} \, C_{ij} \qquad (7)$$

where the sum extends to all matrix elements.

In the present work we argue that formula (7) is the only acceptable general 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 *expected utility* of a classifier having confusion matrix (C_{ij}).

Our argument is based on the fact that this formula is prescribed by *Decision Theory*. ✎ … An overview of decision theory is given in § 2.

---

[7] Jeni et al. 2013; Zhu 2020.

The expected utility (7) is a linear combination of the confusion-matrix elements, with coefficients independent of the elements themselves. We show that some common metrics such as precision, F_1-measure, Matthews correlation coefficient, balanced accuracy, 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). As a consequence each has a set of classification problems in which it does correctly rank the performance of all pairs of classifiers. We show this in §⚒

What happens if we are uncertain about the utilities appropriate to a classification problem? We show in §⚒ that this uncertainty still leads to a metric of the form (7), where the coefficients U_{ij} are the expected values of the possible utility matrices we are uncertain about.

What happens if the utilities U_{ij} in formula (7) are incorrectly specified? We show in §⚒ that an evaluation using incorrect utilities, even with relative errors as large as 25%, still leads to a higher amount of correctly ranked classifiers than the use of any other popular metric.

⚒ Add something on AUC.

⚒ Maybe move following to final discussion Our ultimate purpose in classification is often the choice of a specific course of action among several possible ones, rather than a simple guess of the correct class. This is especially true in medical applications. A clinician does not simply tell a patient "you will probably not contract the disease", but has to decide among dismissal or different kinds of preventive treatment[8]. In other words, our problem is often not *to guess the probable true class*, but *to make the optimal choice*. The two problems are not equivalent when classification takes place under uncertainty. For example, some test results may indicate a very low probability that a patient has a disease, or in other words that *the class 'healthy' is more probably true* than the class 'ill'. Yet the clinician may decide to give the patient some kind of treatment, that is, to behave *as if the patient belonged to the class 'ill'*, on the grounds that the treatment would cure the disease if present and only cause mild

---

[8] Sox et al. 2013; Hunink et al. 2014.

discomfort if the patient is healthy, and that the disease would have dangerous consequences if present and untreated. In this example the most probable class is 'healthy', but the optimal classification is 'ill'.

Most of the issues above are described in the context of binary classification, but they also affect multi-class problems. For simplicity our discussion in the present paper will focus on binary classification. In § 🔧 we shall discuss how it obviously generalizes beyond the binary case.

🔧 Add synopsis of rest of paper.

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.

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

---

[9] Kahneman et al. 2008; Gilovich et al. 2009; Kahneman 2011.

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 our discussion 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. The utility may for instance equal a gain or loss in money, energy, number of customers, life expectancy, or quality of life, measured in appropriate units; or a in combination of such quantities.

These utilities are collected into a *utility matrix* (U_{ij}) as 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 given 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 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:

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

A very important result in decision theory and game theory is that basic requirements of rational decision-making imply that there *must*

---

[10] Camerer & Kunreuther 1989; Kim & Markus 1999; Mittone & Savadori 2009.

be a set of utilities underlying the decisions of a rational agent, and the decisions must obey the rule 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 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. 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 metrics 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. We discuss and use it in our companion work ✎ ....

---

[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.   [13] Besides the general references already given:
Russell & Norvig 2022 § 15.5; Howard & Matheson 2005.

In the present work we are focusing on the stage where a large number of classifications has already been made by a classifier, for example on a test data-set 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 can also be written as

$$\sum_{ij} U_{ij} C_{ij} \qquad (9)$$

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

Formula (9) *is the natural metric to evaluate and compare the performance of classifiers on a test set.*

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 also 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,

$$U_{ij} \mapsto a \, U_{ij} + b \qquad a > 0 \,, \qquad (10)$$

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[14]. Such changes should not affect a decision problem. The fact that indeed they do not, is another example of the logical consistency of decision theory.
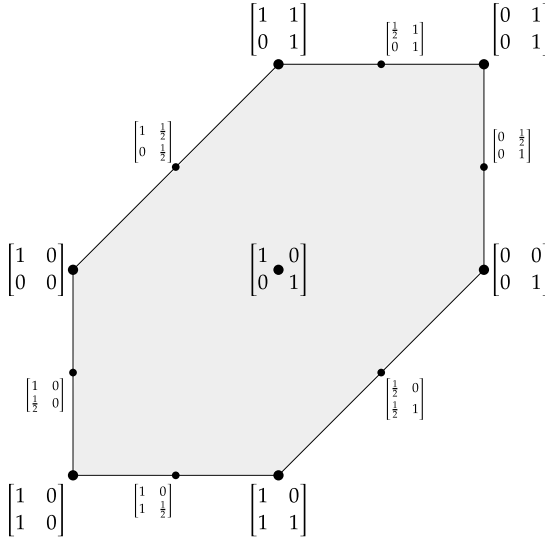
---

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

Figure 1   Space of utility matrices for binary classification. ⚲ add axes maybe

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 can still correspond for example to an actual monetary loss). That is, we are effecting the transformation

$$U_{ij} \mapsto \frac{U_{ij} - \min(U_{ij})}{\max(U_{ij}) - \min(U_{ij})} \ . \tag{11}$$

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 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 evaluation metric. Moving to the left from the centre, the utility for correct classification of class 1 decreases with respect of 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 the upper-left and lower-right corners. Fixing (x, y) axes through the centre of the set, a utility matrix has coordinates

$$\begin{bmatrix} 1 - x\ \delta(x > 0) & y\ \delta(y > 0) \\ -y\ \delta(y < 0) & 1 + x\ \delta(x < 0) \end{bmatrix}. \tag{12}$$

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 (11) and will be discussed in § ✎ ..., 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 and by 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

$$
\begin{array}{c}
\text{\textit{true class}} \\
\begin{array}{cc}
0 & \quad 1
\end{array} \\
\begin{array}{c}
\text{\rotatebox{90}{\textit{classifier output}}}
\end{array}
\begin{array}{c}
0 \\ 1
\end{array}
\begin{bmatrix}
C_{00} & C_{01} \\
C_{10} & C_{11}
\end{bmatrix} .
\end{array}
$$

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 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. (10), because such a modification corresponds to a change of unit and zero of the utility scale. With such a modification the evaluation metric (9) takes the equivalent form

$$
a \sum_{ij} U_{ij} C_{ij} + b \tag{13}
$$

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

$$
a\,(U_{00} - U_{10})\,C_{00} \; + \; a\,(U_{11} - U_{01})\,C_{11} \; + \; a\,f_0\,U_{10} + a\,f_1\,U_{01} + b \,. \tag{14}
$$

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:

$$
a = a(N, f_0, f_1) > 0 \,, \qquad b = b(N, f_0, f_1) \,. \tag{15}
$$

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*
>
> $$ a(N, f_0, f_1)\, X\, C_{00} + a(N, f_0, f_1)\, Y\, C_{11} + b(N, f_0, f_1) \tag{16} $$
>
> *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.*

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 (16):

- ✓ *Accuracy:* $C_{00} + C_{11}$. We have $a = 1$, $X = Y = 1$, $b = 0$. Indeed it corresponds to the evaluation metric based on the utility matrix $(U_{ij}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

- ✓ *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 $\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$.

- ✓ *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 $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$.

The following metrics instead *cannot* be written in the form (16):

- ✗ *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:* $\dfrac{f_1\, C_{00} + f_0\, C_{11}}{\sqrt{f_0\, f_1\, (f_1 + C_{00} - C_{11})\, (f_0 + C_{11} - C_{00})}}$. 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 (16), 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[15].

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 (16). Some confusion matrices will always be incorrectly ranked. Since any rational classification problem must be characterized by some underlying

---

[15] Hand & Christen 2018 discuss such biases in regard to the F_1-measure.

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 in some specific set of classification problems.

The second phenomenon is illustrated in figs 2–3, which we now explain.

Consider the uppermost left plot of fig. 2 as an example. It corresponds to a binary-classification problem with an underlying utility matrix equivalent to the identity $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, and a test set with a balanced number of classes, that is, each class appear with 50% frequency. We select a large number of confusion matrices having this class balance and true-positive and true-negative rates distributed between 1/2 and 1 with linearly increasing probabilities. These confusion matrices would thus arise mostly from good classifying algorithms. For each confusion matrix we calculate

1. its average utility yield, according to the utility matrix $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$; this number is the actual performance, appropriate to this specific problem, of the corresponding algorithm;

2. its score according to another metric, in this case the F_1-measure;

and we plot (blue disks) a corresponding point with these two numbers as x- and y-coordinates. This plot thus shows the general statistical relation between the appropriate utility score and the F_1-measure of this distribution of confusion matrices.

The other three plots in the upper panel of fig. 2 are analogous but for Matthews Correlation Coefficient, accuracy, and true-positive rate as y-coordinates. The lower panel is analogous to the upper panel, but for an imbalanced dataset with one class appearing with 80% frequency. Finally, fig. 3 is analogous to fig. 2 but with an underlying utility matrix equal to $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$.

In all these plots we see that it is always possible to find two or more confusion matrices (pairs of yellow non-triangular shapes) for which the F_1-measure or the Matthews Correlation Coefficient give completely reversed rankings, scoring the worst confusion matrix (and the algorithm yielding it) as the best, and the best as the worst. It is even possible to find at least two confusion matrices (pairs of red triangular shapes) for which both these metrics simultaneously give the wrong ranking. It turns out
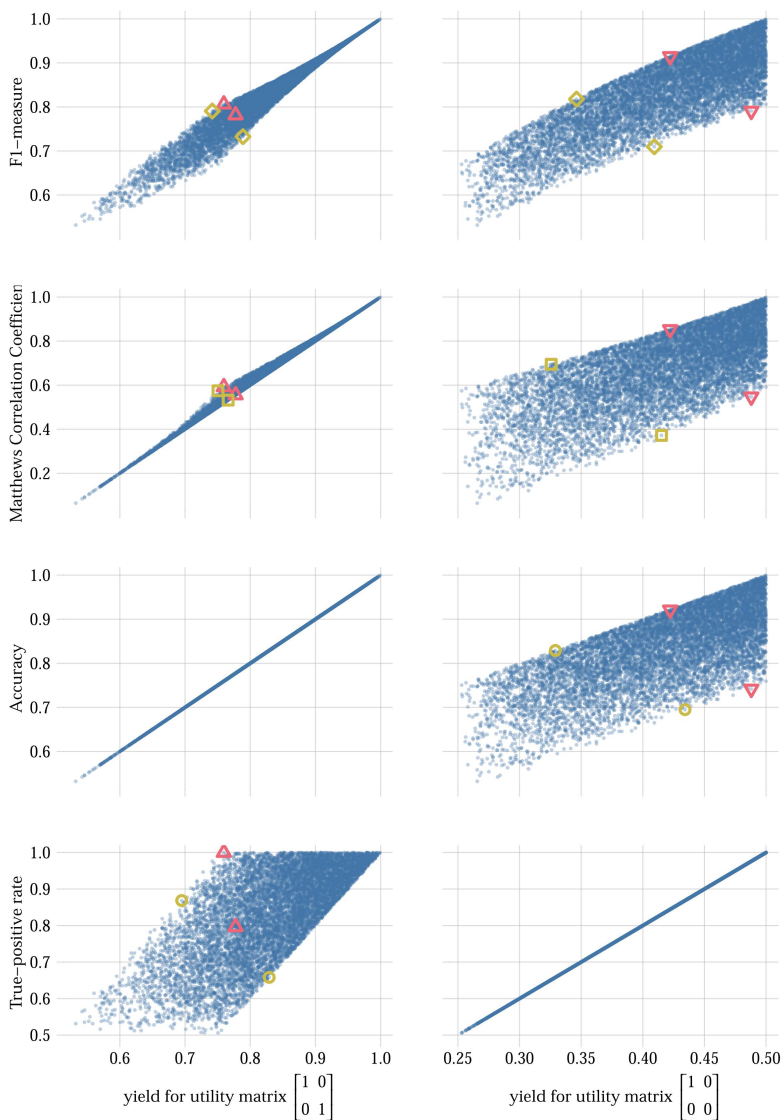
Figure 2 🔧 to be rewritten Relationship between various metrics and actual utility yield for a binary-classification problem with underlying utility matrix $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$. Upper panel: balanced dataset, either class with 50% frequency. Lower panel: imbalanced dataset, class 0 with 80% frequency. Pairs of yellow non-triangular shapes in a plot show two confusion matrices which are wrongly ranked by the metric of the plot's y-axis. Clearly there can be even three or more confusion matrices ranked in completely opposite order by the metric. The accuracy always give a correct rank, no matter the class balance. It is indeed the metric corresponding to the specific utility matrix of this classification problem. Red triangular pairs (almost overlapping in some plots) show two confusion matrices which are wrongly ranked by *all* metrics simultaneously except the accuracy.
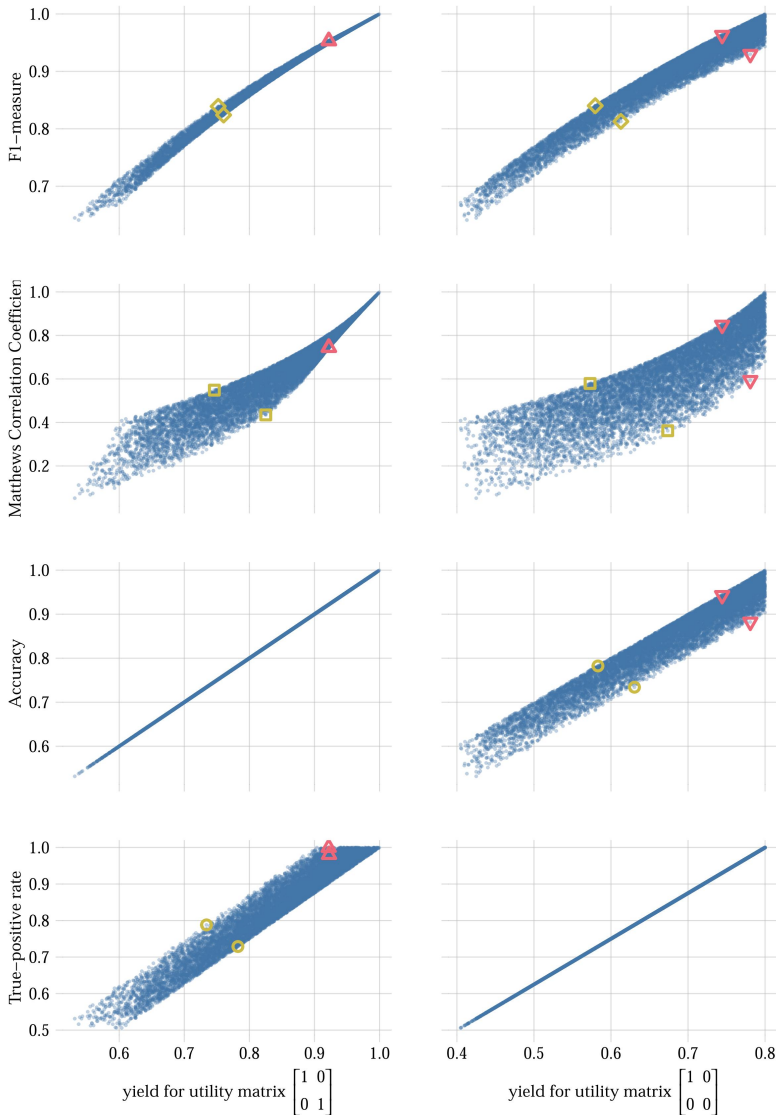
Figure 3   As for fig. 2 but for a binary-classification problem with underlying utility matrix $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$. The true-positive rate (recall) is the metric corresponding to this utility matrix, and therefore always leads to correct ranking, no matter the class balance.
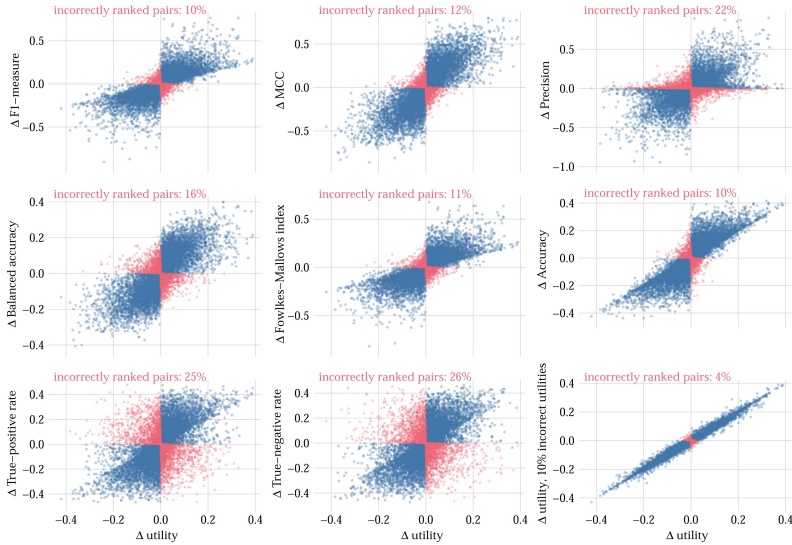
17

Figure 4   Relationship between various metrics and utility

that *there is no classification problem for which these wrong-classification cases do not occur*. This is the second phenomenon mentioned above.

🖉 add comment about illogical to say that MCC is almost good in the first case

4   Unknown or incorrect utilities

5   Remarks on the test set

It may not be amiss to emphasize that the *proportions N_c of classes in the test set should be representative of the proportions that will be encountered in the real application*. Otherwise the test-set results would misleading or even opposite to what the real performance will be. In the lottery example with utility matrix (**??**), suppose we have an algorithm that always makes the decision buy and another algorithm that always decides not-buy. In real instances of such lotteries the class win occurs 1% of the time, and lose, 99%. In a real application the first algorithm would thus yield −0.89, a loss, on average at each instance, and the second 0. The second algorithm is actually best. Suppose we test these algorithms on a test set

18

where the two classes appear 50%/50% instead. On this test set the first algorithm will yield 4.5 on average, the second 0. Thus according to the test the first algorithm is best – a wrong conclusion.

6   Discussion

✒ Utilities have a more immediate (problem-dependent) interpretation than other measures

✒ We recommend to avoid metrics not complying with decision theory – not of the form (16) – and to try to get an estimate of the utilities involved instead.

✒ Computationally more convenient (linear operation)

✒ maybe refer to[16]

---

[16] Varoquaux & Cheplygina 2022.

⚒
Pieces from old version below

6.1 Actual utility yield

The utility matrix is not only the basis for making optimal decisions by means of expected-utility maximization. It also provides the metric to rank a set of decisions already made – for example on a test set – by some algorithm, if we know the corresponding true classes. Suppose we have N test instances, in which each class c occurs N_c times, so that $\sum_c N_c = N$. A decision algorithm made decision d when the true class was c a number M_{dc} of times. These numbers form the confusion matrix (M_{dc}) of the algorithm's output. The numbers M_{dc} are must satisfy the constraints $\sum_d M_{dc} = N_c$ for each c.

For given decision d and class c, in each of the M_{dc} instances the algorithm yielded a utility U_{dc}. The actual average utility yield in the test set is then

$$\frac{1}{N} \sum_{dc} U_{dc} M_{dc} \ . \tag{17}$$

It is convenient to consider the average utility yield, rather than the total utility yield (without the $1/N$ factor), because if we shift the zero or change the measurement unity of our utilities then the yield changes in the same way.

The summary of decision theory just given suffices to address issues **i1**–**i4**.

In comparing, evaluating, and using machine-learning classifiers we face a number of questions and issues; some are well-known, others are rarely discussed:

i1 Choice of valuation metric. When we have to evaluate and compare different classifying algorithms or different hyperparameter values for one algorithm, we are avalanched by a choice of possible evaluation metrics: accuracy, area under curve, F_1-measure, mean square contingency[17] also known as Matthews correlation coefficient[18], precision, recall, sensitivity, specificity, and many others[19]. Only

---

[17] Yule 1912 denoted 'r' there.   [18] Matthews 1975; Fisher 1963 § 31 p. 183.   [19] Sammut & Webb 2017; see also the analysis in Goodman & Kruskal 1954; 1959; 1963; 1972.

vague guidelines are usually given to face this choice. Typically one computes several of such scores and hopes that they will lead to similar ranking.

i2 Rationale and consistency. Most or all of such metrics were proposed only on intuitive grounds, from the exploration of specific problems and relying on tacit assumptions, then heedlessly applied to new problems. The Matthews correlation coefficient, for example, relies on several assumptions of gaussianity[20], which for instance do not apply to skewed population distributions[21]. The area under the receiver-operating-characteristic curve is heavily affected by values of false-positive and false-negative frequencies, as well as by misclassification costs, that have nothing to do with those of the specific application of the classifier[22]. The F_1-measure implicitly gives correct classifications a weight that depends on their frequency or probability[23]; such dependence amounts to saying, for example, "this class is rare, *therefore* its correct classification leads to high gains", which is a form of scarcity cognitive bias[24].

We are therefore led to ask: are there valuation metrics that can be proven, from first principles, to be free from biases and unnecessary assumptions?

i3 Class imbalance. If our sample data are more numerous for one class than for another – a common predicament in medical applications – we must face the 'class-imbalance problem': the classifier ends up classifying all data as belonging to the more numerous class[25], which may be an undesirable action if the misclassification of cases from the less numerous class entails high losses. 🔧 discussion and refs about cost-sensitive learning

i4 Optimality vs truth.

All the issues above are manifestly connected: they involve considerations of importance, gain, loss, and of uncertainty.

In the present work we show how issues **i1–i4** are all solved at once by using the principles of *Decision Theory*. Decision theory gives a logically

---

[20] Fisher 1963 § 31 p. 183 first paragraph.   [21] Jeni et al. 2013; Zhu 2020.   [22] Baker & Pinsky 2001; Lobo et al. 2008.   [23] Hand & Christen 2018.   [24] Camerer & Kunreuther 1989; Kim & Markus 1999; Mittone & Savadori 2009.   [25] Sammut & Webb 2017; Provost 2000.

and mathematically self-consistent procedure to catalogue all possible valuation metrics, to make optimal choices under uncertainty, and to evaluate and compare the performance of several decision algorithms. Most important, we show that implementing decision-theoretic procedures in a machine-learning classifier does not require any changes in current training practices ♟ (possibly it may even make procedures like under- or over-sampling unnecessary!), is computationally inexpensive, and takes place downstream after the output of the classifier.

The use of decision theory requires sensible probabilities for the possible classes, which brings us to issue **??** above. In the present work we also present and use a computationally inexpensive way of calculating these probabilities from the ordinary output of a machine-learning classifier, both for classifiers such as ♟ example here that can only output a class label, and for classifiers that can output some sort of continuous score.

🔧 Write here a summary or outlook of the rest of the paper and a summary of results: • The admissible valuation metrics for a binary classifier form a two-dimensional family; that is, the choice of a specific metric corresponds to the choice of two numbers. Such choice is problem-dependent and cannot be given a priori. • Admissible metrics are only those that can be expressed as a linear function of the elements of the population-normalized confusion matrix. Metrics such as the F_1-measure or the Matthews correlation coefficient are therefore inadmissible

7   Classification from the point of view of decision theory

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 in number. 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'. In the following when we speak of 'classification' we mean a *decision* problem of this kind. The number of decisions thus equals that of classes: $n_\mathrm{d} = n_\mathrm{c}$.

♟ For simplicity we will focus on binary classification, $n_\mathrm{d} = n_\mathrm{c} = 2$, but the discussion generalizes to multi-class problems in an obvious way.

7.1   Choice of valuation metric, rationale and consistency (issues i1, i2)

7.2   Optimality vs truth (issue i4)

According to decision theory a classification algorithm should, at each application, calculate the probabilities ($p_{c|z}$) for the possible classes, given the feature z provided as input; calculate the expected utility of the available decisions according to eq. (8), using the probabilities and the utility matrix; and finally output the decision d^* having maximal expected utility:

$$d^* = \arg\max_d \sum_c U_{dc}\, p_{c|z} \; . \tag{18}$$

We assume here that the utilities are given and the same at each application – the latter assumption could be dropped, however; see the discussion in § 8.

   Current common practice with algorithms capable of outputting some sort of probability-like score is simply to output the class c^* having highest probability:

$$c^* = \arg\max_c p_{c|z} \; . \tag{19}$$

As discussed in § 0, issue **i4**, this means choosing the *most probable* class, not the *optimal* class, and the two are often different, the second being what we typically want. This choice is also the one that would be made with an identity utility matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

   How can we amend current practice for this kind of classifiers, so that they look for optimality rather than truth?

   A first idea could be to simply modify the standard output step (19) into (18). It is an easily implementable and computationally cheap modification: we just multiply the probability tuple by a matrix. Such simple modification, however, has a profound implication for the training procedure: we are modifying the algorithm to output the optimal class, and therefore it should also *learn what is optimal*, not what is true: *the targets in the training and validation phases should be the optimal classes, not the true classes*. But optimality depends on the value of sensible probabilities for the specific situation of uncertainty, in this case conditional on the input features. Determining the optimal classes would thus require a probabilistic analysis that is computationally unfeasible at present for problems that involve very high-dimensional spaces, such as image classification – if an exact probabilistic analysis were possible we would

23

not be developing machine-learning classifiers in the first place[26]. ⊹

Maybe useful to add a reminder that probability theory is the learning *theory par excellence (even if there's no 'learning' in its name)? Its rules are all about making logical updates given new data.*

8   Summary and discussion

✎ *maybe refer to*[27]

---

[26] Russell & Norvig 2022 chs 2, 12; Pearl 1988.　　[27] Varoquaux & Cheplygina 2022.

Appendix: broader overview of binary classification

Let us consider our binary-classification problem from a general perspective and summarize how it would be approached and solved from first principles[28] if our computational resources had no constraints.

In our long-term task we will receive 'units' of a specific kind; the units for example could be gadgets, individuals, or investment portfolios. Each new unit will belong to one of two classes, which we can denote $X=0$ and $X=1$; for example they could be 'defective' vs 'non-defective', 'ill' vs 'healthy'. The class will be unknown to us. For each new unit we shall need to decide among two possible actions, which we can denote $A=\hat{0}$ and $A=\hat{1}$; for example 'discard' vs 'keep', or 'treat' vs 'dismiss'. The utility of each action depends on the unknown class of the unit; we denote these utilities by $U(A \mid X)$. For each new unit we will be able to measure a 'feature' Z of a specific kind common to all units; for example Z could be a set of categorical and real quantities, or an image such as a brain scan. We have a set of units – our 'sample units' or 'sample data' – that are somehow "representative" of the units we will receive in our long-term task[29]. we know both the class and the feature of each of these sample units. Let us denote this sample information by D.

According to the principles of decision theory and probability theory, for each new unit we would proceed as follows:

1. Assign probabilities to the two possible values of the unit's class, given the value of the unit's feature $Z=z$, our sample data D, and any other available information:

   $$\mathrm{p}(X=0 \mid Z=z, D), \qquad \mathrm{p}(X=1 \mid Z=z, D) \equiv 1 - \mathrm{p}(X=0 \mid Z=z, D) , \tag{20}$$

   according to the rules of the probability calculus.

2. Calculate the expected utilities $\bar{U}$ of the two possible actions:

   $$\bar{U}(\hat{0}) := U(\hat{0} \mid X=0)\, \mathrm{p}(X=0 \mid Z=z, D) + U(\hat{0} \mid X=1)\, \mathrm{p}(X=1 \mid Z=z, D)$$
   $$\bar{U}(\hat{1}) := U(\hat{1} \mid X=0)\, \mathrm{p}(X=0 \mid Z=z, D) + U(\hat{1} \mid X=1)\, \mathrm{p}(X=1 \mid Z=z, D) \tag{21}$$

   and choose the action having maximal expected utility.

---

[28] Russell & Norvig 2022 part IV.   [29] for a critical analysis of the sometimes hollow term 'representative sample' see Kruskal & Mosteller 1979a,b,c; 1980.

How is the probability $p(X \mid Z=z, D)$ determined by the probability calculus? Here is a simplified, intuitive picture. First consider the case where the feature Z can only assume a small number of possible values, so that many units can in principle have the same value of Z.

Consider the collection of all units having $Z = z$ that we received in the past and will receive in the future. Among them, a proportion $F(X=0 \mid Z=z)$ belong to class 0, and a proportion $1 - F(X=0 \mid Z=z) \equiv F(X=1 \mid Z=z)$ to class 1. For example these two proportions could be 74% and 26%. Our present unit with $Z=z$ is a member of this collection. The probability $p(X=0 \mid Z=z)$ that our unit belongs to class 0, given that its feature has value z, is then intuitively equal to the proportion $F(X=0 \mid Z=z)$. Analogously for $X=1$.

The problem is that we do not know the proportion $F(X=0 \mid Z=z)$. However, we expect it to be roughly equal to the analogous proportion seen in our sample data; let us denote the latter by $F_s(X=0 \mid Z=z)$:

$$F(X=0 \mid Z=z) \sim F_s(X=0 \mid Z=z) . \tag{22}$$

this is indeed what we mean by saying that our sample data are 'representative' of the future units. Later we shall discuss the case in which such representativeness is of different kinds. We expect the discrepancy between $F(X=0 \mid Z=z)$ and $F_s(X=0 \mid Z=z)$ to be smaller, the larger the number of sample data. Vice versa we expect it to be larger, the smaller the number of sample data.

If Z can assume a continuum of values, as is the case for a brain scan for example, then the collection of units having $Z=z$ is more difficult to imagine. In this case each unit will be unique in its feature value – no two brains are exactly alike.

🔧

---

old text below

Given the unit's feature Z we will assign probabilities to the possible values of the unit's class: according to the rules of the probability calculus.

As mentioned in § 2, a decision problem under uncertainty is conceptually divided into two steps

The Suppose we have a population of units or individuals characterized by a possibly multidimensional variable Z and a binary variable $X \in \{0, 1\}$. Different joint combinations of (X, Z) values can appear in this population. Denote by $F(X=x, Z=z)$, or more simply $F(x, z)$ when there is no confusion, the number of individuals having specific joint

values $(X = x, Z = z)$. This is the absolute frequency of the values (x, z). We can also count the number of individuals having a specific value of $Z = z$, regardless of X; this is the marginal absolute frequency $F(z)$. It is easy to see that

$$F(z) = F(X=0, z) + F(X=1, z) \equiv \sum_x F(x, z) . \tag{23}$$

Analogously for $F(x)$.

Select only the subpopulation of individuals that have a specific value $Z = z$. In this subpopulation, the *proportion* of individuals having a specific value $X = x$ is $f(x \mid Z = z)$. This is the conditional relative frequency of x given that z. It is easy to see that

$$f(x \mid z) = \frac{F(x, z)}{F(z)} . \tag{24}$$

Now suppose that we know all these statistics about this population. An individual coming from this population is presented to us. We measure its Z and obtain the value z. What could be the value of X for this individual? We know that among all individuals having $Z = z$ (and the individual before us is one of them) a proportion $f(x \mid z)$ has $X = x$. Thus we can say that there is a probability $f(x \mid z)$ that our individual has $X = x$. And this is all we can say if we only know Z.

For this individual we must choose among two actions $\{a, b\}$. The utility of performing action a if the individual has $X = x$, and given any other known circumstances, is $U(a \mid x)$; similarly for b. If we knew the value of X, say $X = 0$, we would simply choose the action leading to maximal utility:

$$
\begin{aligned}
&\text{if} \quad U(a \mid X=0) > U(b \mid X=0) \quad \text{then choose action } a, \\
&\text{if} \quad U(a \mid X=0) < U(b \mid X=0) \quad \text{then choose action } b, \\
&\text{else} \quad \text{it does not matter which action is chosen.}
\end{aligned}
\tag{25}
$$

But we do not know the actual value of X. We have probabilities for the possible values of X given that $Z = z$ for our individual. Since X is uncertain, the final utilities of the two actions are also uncertain; but we can calculate their *expected* values $\bar{U}(a \mid Z = z)$ and $\bar{U}(b \mid Z = z)$:

$$
\begin{aligned}
\bar{U}(a \mid z) &:= U(a \mid X=0)\, f(X=0 \mid z) + U(a \mid X=1)\, f(X=1 \mid z) , \\
\bar{U}(b \mid z) &:= U(b \mid X=0)\, f(X=0 \mid z) + U(b \mid X=1)\, f(X=1 \mid z) .
\end{aligned}
\tag{26}
$$

Decision theory shows that the optimal action is the one having the maximal expected utility. Our choice therefore proceeds as follows:

$$\begin{aligned}
\text{if} \quad &\bar{U}(a \mid z) > \bar{U}(b \mid z) \quad \text{then choose action } a, \\
\text{if} \quad &\bar{U}(a \mid z) < \bar{U}(b \mid z) \quad \text{then choose action } b, \\
\text{else} \quad &\text{it does not matter which action is chosen.}
\end{aligned} \tag{27}$$

The decision procedure just discussed is very simple and does not need any machine-learning algorithms. It could be implemented in a simple algorithm that takes as input the full statistics $F(X, Z)$ of the population, the utilities, and yields an output according to (27).

Our main problem is that the full statistics $F(X, Z)$ is almost universally not known. Typically we only have the statistics $F_s(X, Z)$ of a sample of individuals that come from the population of interest or from populations that are somewhat related to the one of interest. This is where probability theory steps in. It allows us to assign probabilities to all the possible statistics $F(X, Z)$. From these probabilities we can calculate the *expected* value $\bar{f}(x \mid z)$ of the conditional frequencies $f(x \mid z)$. Decision theory says that the expected value $\bar{f}(x \mid z)$ should then be used, in this uncertain case, in eq. (26) in place of the unknown $f(x \mid z)$. The decision procedure (27) can then be used again.

Probability theory says that in this particular situation the probability of a particular possible statistics $F(X, Z)$ is the product of two factors having intuitive interpretations:

- the probability of observing the statistics $F_s(X, Z)$ of our data sample, assuming the full statistics to be $F(X, Z)$. With some combinatorics it can be shown that this probability is proportional to

$$\exp\left[\sum_{X,Z} F_s(X, Z) \ln F(X, Z)\right] \tag{28}$$

The argument of the exponential is the cross-entropy between $F_s(X, Z)$ and $F(X, Z)$; this is the reason of its appearance in the loss function used for classifiers[30].

This factor tells us how much the possible statistics *fit* the sample data; it gives more weight to statistics with a better fit.

---

[30] Bridle 1990; MacKay 1992.

- the probability of the full statistics $F(X, Z)$ for reasons not present in the data, for example because of physical laws, biological plausibility, or similar.

  This factor tells us whether the possible statistics should be favourably considered, or maybe even discarded instead, for reasons that go beyond the data we have seen; in other words, whether the hypothetical statistics would *generalize* well beyond the sample data.

The final probability comes from the balance between these 'fit' and 'generalization' factors. Note that the first factor becomes more important as the sample size and therefore $F_s(X, Z)$ increases; the sample data eventually determine what the most probable statistics is, if the sample is large enough.

A similar probabilistic reasoning applies if our sample data come not from the population of interest but from a population having at least the same *conditional* frequencies of as the one of interest, either $f(X \mid Z)$ or $f(Z \mid X)$. The latter case must be examined with care when our purpose is to guess X from Z. In this case we cannot use the conditional frequencies $f_s(X \mid Z)$ that appear in the data to obtain the expected value $\bar{f}(X \mid Z)$: they could be completely different from the ones of the population of interest. We must instead use the sample conditional frequencies $f_s(Z \mid X)$ to obtain the expected value $\bar{f}(Z \mid X)$, and then combine the latter with an appropriate probability $P(X)$ through Bayes's theorem:

$$\frac{\bar{f}(Z \mid X)\, P(X)}{\sum_X \bar{f}(Z \mid X)\, P(X)} \ . \tag{29}$$

The probability $P(X)$ cannot be obtained from the data, but requires a separate study or survey. In medical applications, where X represents for example the presence or absence of a disease, the probability $P(X)$ is the base rate of the disease. Direct use of $f_s(X \mid Z)$ from the data instead of (29) is the 'base-rate fallacy'[31].

In supervised learning the classifier is trained to learn the most probable $f(X \mid Z)$ from the data. The training finds the $f(X \mid Z)$ that most closely fits the conditional frequency $f_s(X \mid Z)$ of the sampled data; this roughly corresponds to maximizing the first factor (28) described above.

---

[31] Russell & Norvig 2022 § 12.5; Axelsson 2000; Jenny et al. 2018.

The architecture and the parameter regularizer of the classifier play the role of the second factor.

Bibliography

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

Axelsson, S. (2000): *The base-rate fallacy and the difficulty of intrusion detection*. ACM Trans. Inf. Syst. Secur. **3**[3], 186–205. DOI:10.1145/357830.357849, http://www.scs.carleton.ca/~soma/id-2007w/readings/axelsson-base-rate.pdf.

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.

Bridle, J. S. (1990): *Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition*. Neurocomputing **68**, 227–236. DOI:10.1007/978-3-642-76153-9_28.

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.

Goodman, L. A., Kruskal, W. H. (1954): *Measures of association for cross classifications*. J. Am. Stat. Assoc. **49**[268], 732–764. DOI:10.1080/01621459.1954.10501231. See corrections Goodman, Kruskal (1957; 1958) and also Goodman, Kruskal (1959; 1963; 1972).

— (1957): *Corrigenda: Measures of association for cross classifications*. J. Am. Stat. Assoc. **52**[280], 578. DOI:10.1080/01621459.1957.10501415. See Goodman, Kruskal (1954).

— (1958): *Corrigenda: Measures of association for cross classifications*. J. Am. Stat. Assoc. **53**[284], 1031. DOI:10.1080/01621459.1958.10501492. See Goodman, Kruskal (1954).

Goodman, L. A., Kruskal, W. H. (1959): *Measures of association for cross classifications. II: Further discussion and references*. J. Am. Stat. Assoc. **54**[285], 123–163. DOI:10.1080/01621459.1959.10501503. See also Goodman, Kruskal ([1954](); [1963](); [1972]()).

— (1963): *Measures of association for cross classifications. III: Approximate sampling theory*. J. Am. Stat. Assoc. **58**[302], 310–364. DOI:10.1080/01621459.1963.10500850. See correction Goodman, Kruskal ([1970]()) and also Goodman, Kruskal ([1954](); [1959](); [1972]()).

— (1970): *Corrigenda: Measures of association for cross classifications. III: Approximate sampling theory*. J. Am. Stat. Assoc. **65**[330], 1011. DOI:10.1080/01621459.1970.10481142. See Goodman, Kruskal ([1963]()).

— (1972): *Measures of association for cross classifications, IV: Simplification of asymptotic variances*. J. Am. Stat. Assoc. **67**[338], 415–421. DOI:10.1080/01621459.1972.10482401. See also Goodman, Kruskal ([1954](); [1959](); [1963]()).

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: Schwing, Albers ([1980]()): 89–113. With discussion. DOI:10.1007/978-1-4899-0445-4_5. Repr. in Howard, Matheson ([1984]()) pp. 481–506.

Howard, R. A., Matheson, J. E., eds. (1984): *Readings on the Principles and Applications of Decision Analysis. Vol. II: Professional Collection*. (Strategic Decisions Group, Menlo Park, USA).

— (2005): *Influence diagrams*. Decis. Anal. **2**[3], 127–143. DOI:10.1287/deca.1050.0020. First publ. 1984 in Howard, Matheson ([1984]()) 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.

Jenny, M. A., Keller, N., Gigerenzer, G. (2018): *Assessing minimal medical statistical literacy using the Quick Risk Test: a prospective observational study in Germany*. BMJ Open **8**, e020847, e020847corr2. DOI:10.1136/bmjopen-2017-020847, DOI:10.1136/bmjopen-2017-020847corr2.

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.

Kruskal, W., Mosteller, F. (1979a): *Representative sampling, I: Non-scientific literature*. Int. Stat. Rev. **47**[1], 13–24. See also Kruskal, Mosteller (1979b,c; 1980).

— (1979b): *Representative sampling, II: Scientific literature, excluding statistics*. Int. Stat. Rev. **47**[2], 111–127. See also Kruskal, Mosteller (1979a,c; 1980).

— (1979c): *Representative sampling, III: The current statistical literature*. Int. Stat. Rev. **47**[3], 245–265. See also Kruskal, Mosteller (1979a,b; 1980).

— (1980): *Representative sampling, IV: The history of the concept in statistics, 1895–1939*. Int. Stat. Rev. **48**[2], 169–195. See also Kruskal, Mosteller (1979a,b,c).

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.

MacKay, D. J. C. (1992): *The evidence framework applied to classification networks*. Neural Comput. **4**[5], 720–736. http://www.inference.phy.cam.ac.uk/mackay/PhD.html, DOI: 10.1162/neco.1992.4.5.720.

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.

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.

Pearl, J. (1988): *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, rev. 2nd pr. (Kaufmann, San Francisco). DOI:10.1016/C2009-0-27609-4.

Provost, F. (2000): *Machine learning from imbalanced data sets 101*. Tech. rep. WS-00-05-001. (AAAI, Menlo Park, USA). https://aaai.org/Library/Workshops/2000/ws00-05-001.php.

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.

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

Sammut, C., Webb, G. I., eds. (2017): *Encyclopedia of Machine Learning and Data Mining*, 2nd ed. (Springer, Boston). DOI:10.1007/978-1-4899-7687-1. First publ. 2011.

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

Schwing, R. C., Albers Jr., W. A., eds. (1980): *Societal Risk Assessment: How Safe is Safe Enough?* (Springer, New York). DOI:10.1007/978-1-4899-0445-4.

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.

Yule, G. U. (1912): *On the methods of measuring association between two attributes*. J. R. Stat. Soc. **75**6, 579–652. DOI:10.1111/j.2397-2335.1912.tb00463.x.

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.