Is the evaluation evaluated?

A first-principle approach to use and evaluation of classifiers

K. Dirland

<***@***>

A. S. Lundervold

<***@***>

(or any permutation thereof)

P.G.L. Porta Mana ⊙

<pgl@portamana.org>

Draft. 4 March 2022; updated 6 May 2022

✏️

🔧[Luca] The two main points of the paper are about

- Deriving correct probabilities for classes by a simple, low-cost Bayesian analysis: from the confusion matrix, for machine-learning algorithms that only output class labels; and from the continuous output (e.g. last layer or softmax in deep nets), for algorithms that can provide some kind of continuous score.

- Implement decision-theory principles: (1) in the algorithm, so that it yields the *optimal* class label, (2) in the categorization of valuation scores, (3) in calculating valuation scores.

Maybe it'd be best to address these two points in two papers with subtitles 'I. . . . ' and 'II. . . . ', for a neater presentation. The two points depend on each other, though: to show the improvements by the Bayesian analysis we use the valuation scores of the second point; to implement decision theory in the algorithm we need the probabilities from the first point.

Let me know your thoughts about this.

1 Valuation metrics, amounts of data, inferences, and decisions

A factory which produces some kind of widgets wishes to employ a machine-learning classifier to assess whether each produced widget is good or faulty. The classifier takes some features of a widget as input and outputs one of the two labels 'positive' for putatively good widget, or 'negative' for a putatively faulty one.

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

$$
\begin{array}{c}
\textit{true widget condition}\\
+-\\
\underset{\textit{output}}{\overset{\textit{classifier}}{}}\;
\begin{array}{c}+\\-\end{array}
\left[\begin{array}{cc}
\text{True Positive} & \text{False Positive}\\
\text{False Negative} & \text{True Negative}
\end{array}\right]
\end{array}
$$

format, and normalized over the total number of evaluation data:

$$
\text{classifier A:}\quad
\begin{bmatrix}
0.37 & 0.22\\
0.13 & 0.28
\end{bmatrix}
\tag{1}
$$

$$
\text{classifier B:}\quad
\begin{bmatrix}
0.30 & 0.01\\
0.20 & 0.49
\end{bmatrix}
\tag{2}
$$

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[1] also known as Matthews correlation coefficient[2], precision, recall, sensitivity, specificity, and many others[3]. Only 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[4], which for instance do not apply to skewed population distributions[5]. The area under the receiver-operating-characteristic curve is heavily affected by

---

[1] Yule 1912 denoted 'r' there.   [2] Matthews 1975; Fisher 1963 § 31 p. 183.   [3] Sammut & Webb 2017; see also the analysis in Goodman & Kruskal 1954; 1959; 1963; 1972. [4] Fisher 1963 § 31 p. 183 first paragraph.   [5] Jeni et al. 2013; Zhu 2020.

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[6]. The F_1-measure implicitly gives correct classifications a weight that depends on their frequency or probability[7]; 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[8].

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[9], 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. 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[10].

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 very probably true*. 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 discomfort if the patient is healthy, and that the disease would have

---

[6] Baker & Pinsky 2001; Lobo et al. 2008.   [7] Hand & Christen 2018.   [8] Camerer & Kunreuther 1989; Kim & Markus 1999; Mittone & Savadori 2009.   [9] Sammut & Webb 2017; Provost 2000.   [10] Sox et al. 2013; Hunink et al. 2014.

dangerous consequences if present and untreated. In this example the most probable class is 'healthy', but the optimal classification is 'ill'.

This point of view has profound potential implications for the training of our algorithm: it means that its training targets ought to be the *optimal* class labels under that particular uncertain situation, not the *true* class labels. But how could such optimality be determined? – Luckily we shall see that no such change in the training process is necessary.

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 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

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 applications to be discussed later, 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, as argued under issue **i4**; in some cases even the number of classes and the number of decisions differ. This distinction prevents the appearance of various cognitive biases[11], for example the scarcity bias mentioned in **i2**, or plain wishful thinking: "this event is valuable, *therefore* it is more probable".

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, life expectancy, or number of customers, measured in appropriate units; or a in combination of such quantities.

As an example, imagine we are offered to buy a lottery ticket, which may be winning or not. The ticket costs 1 unit of some monetary currency,

---

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

and the lottery prize is 11 units. Our available decisions are whether to buy the ticket or not. We have four utilities, representing the total change in our money after the lottery, displayed in this self-explanatory table:

	win	lose
buy	+10	−1
not-buy	0	0

We denote $u(d \mid c)$ the utility of decision d if class c is true, or 'the utility of d given c', or 'the utility of d conditional on c'. One utility from the lottery example is $u(\text{buy} \mid \text{lose}) = -1$. If we have n_d available decisions and n_c possible classes, the utilities can be collected in a *utility matrix* $\boldsymbol{U} \equiv (U_{dc})$ having n_d rows and n_c columns. The utility matrix for the lottery is

$$\boldsymbol{U} = \begin{pmatrix} +10 & -1 \\ 0 & 0 \end{pmatrix} . \tag{3}$$

If we know which class is true, 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, decision theory states that the optimal decision is the one having maximal *expected* utility, denoted $\bar{u}(d)$ and defined as the expected value of the utility of decision d with respect to the probabilities of the various classes. For binary classification

$$\bar{u}(d) := u(d \mid c_1)\, p(c_1) + u(d \mid c_2)\, p(c_2) \tag{4}$$

where $p(c_1)$ and $p(c_2) \equiv 1 - p(c_1)$ are the probabilities of classes c_1 and c_2. The n_d expected utilities are therefore given by the matrix product of the utility matrix times the column matrix of probabilities.

For instance, if the ticket above has a 20% probability of winning and 80% of losing, that is $p(\text{win}) = 0.2$ and $p(\text{lose}) = 0.8$, then our two decisions have expected utilities

$$\bar{u}(\text{buy}) = +10 \cdot 0.2 - 1 \cdot 0.8 = +1.2 \, ,$$
$$\bar{u}(\text{not-buy}) = 0 \cdot 0.2 + 0 \cdot 0.8 = 0 \, . \tag{5}$$

The optimal choice is to buy the ticket, that is, to classify the ticket *as if* it belonged to class win, even if it most probably belongs to class lose. (Note that the utility of money is usually not equal to the amount of money, the relationship between the two being somewhat logarithmic[12].)

---

[12] e.g. North [1968](#) pp. 203–204; Raiffa [1970](#) ch. 4.

🔧 Add note about how (sequential) decision theory was used during World War I; see Good (1950) around § 6.2

2.4 Space of utility matrices

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 gain or losses of concrete quantities such as money or energy. In medical problems they can correspond to life expectancy and quality of life; see for example Sox et al. (2013) for a discussion of how such health factors are transformed into utilities. Decision theory, in the subfield of *utility theory*, gives rules that guarantee the mutual consistency of a set of utilities. In the present work we shall 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 chosen. 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.

It is nevertheless clear that every decision problem is completely determined by a set of $n_d \cdot n_c$ utilities. Actually if we change the elements of the utility matrix of a decision problem by a common additive constant or by a common positive multiplicative constant, then that decision problem is unchanged, in the sense that we reach the same decision by maximizing expected utility with the new utility matrix, given the same probabilities. This is evident from eq. (2): all expected utilities change by the same additive constant or the same positive factor, and therefore their ordering does not change. After all, an additive constant or a positive factor represents only changes in the zero or the measurement unit of our utility. 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.

Let us call *equivalent* two utility matrices that differ only by a constant additive term or by a positive factor or both. Inequivalent utility matrices represent inequivalent decision problems. Thus *all decision problems with n_d decisions and n_c classes are catalogued by $n_d n_c - 2$ parameters* (the set of such problems has the topology of an $(n_d n_c - 1)$-dimensional sphere). In the case of binary classification this means 2 parameters.

2.5   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

$$\frac{1}{N} \sum_{dc} U_{dc} \, M_{dc} \ . \tag{6}$$

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.

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 (3), 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 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.

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

3   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.

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

According to decision theory, a classification problem requires the specification of a utility matrix (U_{dc}). We saw in § 2.5 that the utility matrix should also be used in evaluating the decisions made by one or more classification algorithms in a test set with N datapoints. Each algorithm gives rise to a confusion matrix (M_{dc}), containing the number M_{dc} of times the algorithm made decision d when the true class was c.

The average utility obtained by an algorithm on the test set is

$$\frac{1}{N} \sum_{cd} U_{dc} \, M_{dc} \ . \tag{7}$$

This expressions is a linear combination of the confusion-matrix elements. Besides the factor $1/N$, the coefficients of the linear combination depend only on the utility matrix (U_{dc}).

If we are comparing several classifiers *on the same test set*, we can multiply the expression above by a generic positive function of the class frequencies, $a(N_c)$, and also add a generic function of the class frequencies, $b(N_c)$. These operations corresponds to changes in the zero and measurement unit of the utilities by amounts which depend on the class frequencies. Since the class frequencies are independent of the algorithms, these changes are the same for all algorithms and do not affect their final ranking. Note, however, that if we were to compare

utilities obtained on different test sets then such changes dependent on class frequencies would not be allowed.

We thus find the following important result according to decision theory: *A valuation metric should be a **linear combination** of the elements of the confusion matrix, possibly multiplied by a positive function of the class frequencies of the test set and with an additional term also depending only on the class frequencies. The coefficients of the linear combination are problem-specific and **cannot depend on the confusion-matrix elements**.* In formulae, the metric should have the form

$$a(N_c) \sum_{cd} x_{dc} M_{dc} + b(N_c) \tag{8}$$

for some coefficients x_{dc} and functions $a(\cdot) > 0, \ b(\cdot)$.

Let us see what this means in the case of binary classification. The decisions and classes are both typically denoted as positive '+' and negative '−', and we speak of the number of 'true positives', 'false positives', and so on, so that

$$M_{++} \equiv M_{\text{tp}}, \quad M_{+-} \equiv M_{\text{fp}}, \quad M_{-+} \equiv M_{\text{fn}}, \quad M_{--} \equiv M_{\text{tn}} .$$

Accordingly we denote the elements of the utility matrix as

$$U_{++} \equiv U_{\text{tp}}, \quad U_{+-} \equiv U_{\text{fp}}, \quad U_{-+} \equiv U_{\text{fn}}, \quad U_{--} \equiv U_{\text{tn}} .$$

With this notation, decision theory says that a valuation metric for this test set should have the following general form

$$a(N_+, N_-)(x_{\text{tp}} M_{\text{tp}} + x_{\text{fp}} M_{\text{fp}} + x_{\text{fn}} M_{\text{fn}} + x_{\text{tn}} M_{\text{tn}}) + b(N_+, N_-) . \tag{9}$$

for particular values of the coefficients $x_{\text{tp}}, x_{\text{fp}}, x_{\text{fn}}, x_{\text{tn}}$ and particular functions $a(\cdot) > 0$ and $b(\cdot)$. Let us examine some common valuation metrics from the point of view of this requirement.

- ✓ *Accuracy:* $(M_{\text{tp}} + M_{\text{tn}})/N$. It is a particular case of formula (7) with $x_{\text{tp}} = x_{\text{tn}} = 1, \ x_{\text{fp}} = x_{\text{fn}} = 0, \ a(N_+, N_-) = 1/(N_+ + N_-), \ b(\cdot) = 0$. It corresponds to using the utility matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.
- ✗ *Precision:* $M_{\text{tp}}/(M_{\text{tp}} + M_{\text{fp}})$. It cannot be written as a linear combination of the confusion-matrix elements.
- ✗ F_1*-measure:* $2M_{\text{tp}}/(2M_{\text{tp}} + M_{\text{fp}} + M_{\text{fn}})$. It cannot be written as a linear combination of the confusion-matrix elements.

- ✗ *Matthews correlation coefficient:* $\frac{M_{tp}\,M_{tn}-M_{fp}\,M_{fn}}{\sqrt{(M_{tp}+M_{fp})(M_{tn}+M_{fn})N_+N_-}}$. It cannot be written as a linear combination of the confusion-matrix elements.

- ✓ *True-positive rate:* M_{tp}/N_+. It is formula (7) with $x_{tp} = 1$, all other $x_{...} = 0$, $a(N_+, N_-) = 1/N_+$, and $b(\cdot) = 0$. It corresponds to using the utility matrix $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right)$.

- ✓ *True-negative rate:* M_{tn}/N_-. Analogous to true-positive rate.

- ✗ *Balanced accuracy:* $(N_-\,M_{tp} + N_+\,M_{tn})/(2N_+N_-)$. Despite being an average of two metrics (true-positive and true-negative rate) that agree with formula (7), it is not an instance of that formula. The reason is that the two averaged metrics involve different $a(\cdot)$ functions.

We see that many popular validation metrics, those marked by '✗' above, do not comply with decision theory. Their formulae involve an interdependence of utilities and probabilities which hides some forms of cognitive biases. 🔧 Very important: we should preferably add an example to illustrate this.

3.2 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. (2), using the probabilities and the utility matrix; and finally output the decision d^* having maximal expected utility:

$$d^* = \arg\max_d \sum_c U_{dc}\,p_{c|z} \ . \tag{10}$$

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 § 4.

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

$$c^* = \arg\max_c p_{c|z} \ . \tag{11}$$

As discussed in § 1, 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 (9) into (8). 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 not be developing machine-learning classifiers in the first place[13]. ⛏ 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.

4   Summary and discussion

---

[13] Russell & Norvig 2022 chs 2, 12; Pearl 1988.

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[14] 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[15]. 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:

$$p(X=0 \mid Z=z, D), \qquad p(X=1 \mid Z=z, D) \equiv 1 - p(X=0 \mid Z=z, D) , \tag{12}$$

   according to the rules of the probability calculus.

2. Calculate the expected utilities ū of the two possible actions:

$$\bar{u}(\hat{0}) := U(\hat{0} \mid X=0)\, p(X=0 \mid Z=z, D) + U(\hat{0} \mid X=1)\, p(X=1 \mid Z=z, D)$$
$$\bar{u}(\hat{1}) := U(\hat{1} \mid X=0)\, p(X=0 \mid Z=z, D) + U(\hat{1} \mid X=1)\, p(X=1 \mid Z=z, D) \tag{13}$$

   and choose the action having maximal expected utility.

---

[14] Russell & Norvig 2022 part IV.   [15] 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 $F_s(X=0 \mid Z=z)$:

$$F(X=0 \mid Z=z) \sim F_s(X=0 \mid Z=z) . \tag{14}$$

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

$$F(z) = F(X=0, z) + F(X=1, z) \equiv \sum_x F(x, z) \,. \qquad (15)$$

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

$$f(x \mid z) = \frac{F(x, z)}{F(z)} \,. \qquad (16)$$

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:

$$\begin{aligned}
&\text{if} \quad U(a \mid X=0) > U(b \mid X=0) \quad \text{then choose action } a, \\
&\text{if} \quad U(a \mid X=0) < U(b \mid X=0) \quad \text{then choose action } b, \qquad (17) \\
&\text{else} \quad \text{it does not matter which action is chosen.}
\end{aligned}$$

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 $\bar{U}(b \mid Z=z)$:

$$\begin{aligned}
\bar{U}(a \mid z) &\coloneqq U(a \mid X=0)\, f(X=0 \mid z) + U(a \mid X=1)\, f(X=1 \mid z), \\
\bar{U}(b \mid z) &\coloneqq U(b \mid X=0)\, f(X=0 \mid z) + U(b \mid X=1)\, f(X=1 \mid z).
\end{aligned} \qquad (18)$$

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

$$\begin{array}{llll} \text{if} & \bar{U}(a \mid z) > \bar{U}(b \mid z) & \text{then choose action } a, \\ \text{if} & \bar{U}(a \mid z) < \bar{U}(b \mid z) & \text{then choose action } b, & (19) \\ \text{else} & \text{it does not matter which action is chosen.} \end{array}$$

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 (17).

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. (16) in place of the unknown $f(x \mid z)$. The decision procedure (17) 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

$$\exp\left[\sum_{X,Z} F_s(X, Z) \ln F(X, Z)\right] \qquad (20)$$

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[16].

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

---

[16] 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:

$$\frac{\bar{f}(Z \mid X) \, P(X)}{\sum_X \bar{f}(Z \mid X) \, P(X)} \ . \tag{21}$$

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 (19) is the 'base-rate fallacy'[17].

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 (18) described above.

---

[17] 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.

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.

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

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. (1950): *Probability and the Weighing of Evidence*. (Griffin, London).

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).

— (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).

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.

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/CB09781139506779. First publ. 2001.

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/CB09780511809477. 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.

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.

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.