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

Guessing what's true or choosing what's optimal?

A first-principle approach to some issues of classifiers

K. Dirland
<***@***>

A. S. Lundervold
<***@***>

P.G.L. Porta Mana
<pgl@portamana.org>

(or any permutation thereof)

Draft. 4 March 2022; updated 11 April 2022

abstract

[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

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. 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 scales: 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.

[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.

i2 Rationale and consistency. Most or all of such scales 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 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[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 scales 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 costs. 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*.

---

[4] Fisher 1963 § 31 p. 183 first paragraph. [5] Jeni et al. 2013; Zhu 2020. [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.

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 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.

i5 Probability. It is difficult to assess sensible probabilities for the possible classes. Some machine-learning algorithms only output a class label. Others, including deep networks, output real numbers that, even if normalized, cannot be reliably interpreted as degrees of belief or as frequencies in the population of interest[11]. Algorithms that internally do perform probabilistic calculations, such as naive-Bayes or logistic-regression classifiers[12], unfortunately rest on probabilistic assumptions – independence and particular shapes of distributions, for example – that are often unrealistic (and their consistency with the specific application is rarely checked).

All these issues are plainly connected: they involve considerations of importance or gain or cost, 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 scales, to make optimal choices under uncertainty, and to evaluate and compare the performance of several decision algorithms.

[11] MacKay 1992; Gal & Ghahramani 2016. [12] Murphy 2012 § 3.5, ch. 8; Bishop 2006 §§ 8.2, 4.3; Barber 2020 ch. 10, § 17.4.

Most important, the application of decision theory does not require any changes in current training practices, is computationally inexpensive, and takes place downstream after the output of the classifier.

The use of decision theory, however, requires the probabilities for the possible classes, which is issue **i5**. We present a computationally inexpensive way of calculating these probabilities from the ordinary output of a machine-learning classifier:

-

  old text:

  The three points above turn out to be tightly related and to have a common solution. We show that

  1. the admissible valuation scales for a binary classifier form a two-dimensional family; that is, the choice of a specific scale corresponds to the choice of two numbers. Such choice is problem-dependent and cannot be given a priori.
  2. admissible scales are only those that can be expressed as a linear function of the elements of the population-normalized confusion matrix. Scales such as the F_1-measure or the Matthews correlation coefficient are therefore inadmissible

2 Overview of decision theory

For a presentation 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 in 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 [13]

Decision theory makes a distinction between

a. the possible situations we are uncertain about
b. the possible choices we can make.

This distinction is important, in fact in some cases the numbers of possible uncertain situations

[13] Russell & Norvig 2022 ch. 15; Jeffrey 1965; North 1968.

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:

$$\mathrm{p}(X=0|Z=z,D), \qquad \mathrm{p}(X=1|Z=z,D) \equiv 1-\mathrm{p}(X=0|Z=z,D)\,, \quad (1)$$

   according to the rules of the probability calculus.

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

$$\begin{aligned}\bar{U}(\hat{0}) &\coloneqq U(\hat{0} \mid X=0)\,\mathrm{p}(X=0 \mid Z=z,D) + U(\hat{0} \mid X=1)\,\mathrm{p}(X=1 \mid Z=z,D)\\ \bar{U}(\hat{1}) &\coloneqq U(\hat{1} \mid X=0)\,\mathrm{p}(X=0 \mid Z=z,D) + U(\hat{1} \mid X=1)\,\mathrm{p}(X=1 \mid Z=z,D)\end{aligned} \quad (2)$$

   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 $\mathrm{p}(X \mid Z=z, D)$ 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 $\mathrm{p}(X=0 \mid Z=z)$ 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 $F_{\mathrm{s}}(X=0 \mid Z=z)$:

$$F(X=0 \mid Z=z) \sim F_{\mathrm{s}}(X=0 \mid Z=z)\,. \tag{3}$$

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 $F_{\mathrm{s}}(X=0 \mid Z=z)$ 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

$$F(z) = F(X=0, z) + F(X=1, z) \equiv \sum_x F(x, z) \,. \tag{4}$$

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

$$f(x \mid z) = \frac{F(x, z)}{F(z)} \,. \tag{5}$$

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:

$$\begin{aligned} &\text{if} \quad U(a \mid X=0) > U(b \mid X=0) \quad \text{then choose action } a, \\ &\text{if} \quad U(a \mid X=0) < U(b \mid X=0) \quad \text{then choose action } b, \\ &\text{else} \quad \text{it does not matter which action is chosen.} \end{aligned} \tag{6}$$

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

$$\begin{aligned} \bar{U}(a \mid z) &\coloneqq U(a \mid X=0)\, f(X=0 \mid z) + U(a \mid X=1)\, f(X=1 \mid z) \,, \\ \bar{U}(b \mid z) &\coloneqq U(b \mid X=0)\, f(X=0 \mid z) + U(b \mid X=1)\, f(X=1 \mid z) \,. \end{aligned} \tag{7}$$

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

$$\begin{aligned} &\text{if} \quad \bar{U}(a \mid z) > \bar{U}(b \mid z) \quad \text{then choose action } a, \\ &\text{if} \quad \bar{U}(a \mid z) < \bar{U}(b \mid z) \quad \text{then choose action } b, \\ &\text{else} \quad \text{it does not matter which action is chosen.} \end{aligned} \tag{8}$$

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 (8).

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. (7) in place of the unknown $f(x \mid z)$. The decision procedure (8) 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

$$\exp\left[\sum_{X,Z} F_s(X, Z) \ln F(X, Z)\right] \tag{9}$$

  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:

$$\frac{\bar{f}(Z \mid X)\, P(X)}{\sum_X \bar{f}(Z \mid X)\, P(X)} \,. \tag{10}$$

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 (10) 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 (9) 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.

Barber, D. (2020): *Bayesian Reasoning and Machine Learning*, online update. (Cambridge University Press, Cambridge). http://www.cs.ucl.ac.uk/staff/d.barber/brml. First publ. 2007.

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.

Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*. (Springer, New York). https://www.microsoft.com/en-us/research/people/cmbishop/prml-book.

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.

Gal, Y., Ghahramani, Z. (2016): *Dropout as a Bayesian approximation: representing model uncertainty in deep learning*. Proc. Mach. Learn. Res. **48**, 1050–1059. See also Appendix at arXiv doi:10.48550/arXiv.1506.02157.

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).

Goodman, L. A., Kruskal, W. H. (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/CBO9781139506779. 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.

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.

Murphy, K. P. (2012): *Machine Learning: A Probabilistic Perspective*. (MIT Press, Cambridge, USA). https://probml.github.io/pml-book/book0.html.

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.

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.