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

Probabilities for machine-learning classifiers

Classifiers as diagnostic tests

K. Dirland <***@***>
A. S. Lundervold <***@***>
P.G.L. Porta Mana <pgl@portamana.org>
(or any permutation thereof)

Draft. 14 April 2022; updated 16 April 2022

1 Sensible probabilities for classifiers

Some machine-learning algorithms for classification, such as support-vector machines, typically output a class label. Others, such as deep networks, output a set of real numbers. These real numbers can be positive, normalized to unity, and can bear some qualitative relation to the plausibilities of the classes. But they cannot be reliably interpreted as sensible probabilities, that is, as the degrees of belief assigned to each possible class by a rational agent[1]; or, in terms of 'populations'[2], as the expected frequencies of the classes in the hypothetical population of units (degrees of belief and frequencies being related by de Finetti's theorem[3]).

Algorithms that internally do perform probabilistic calculations, such as naive-Bayes or logistic-regression classifiers[4], unfortunately rest on probabilistic assumptions, such as independence and particular shapes of distributions, that are often unrealistic (and their consistency with the specific application is rarely checked). Only particular classifiers such as Bayesian neural networks[5] output sensible probabilities.

Why are probability values important? As we argue in a companion work , our ultimate purpose in classification is seldom only to guess a class; most often it is to choose a specific course of action, or to make a decision, among several available ones. A clinician, for example, does not

[1] MacKay 1992; Gal & Ghahramani 2016; Russell & Norvig 2022 chs 2, 12, 13. [2] Lindley & Novick 1981; Fisher 1967 § II.4. [3] Bernardo & Smith 2000 ch. 4; Dawid 2013. [4] Murphy 2012 § 3.5, ch. 8; Bishop 2006 §§ 8.2, 4.3; Barber 2020 ch. 10, § 17.4. [5] Neal & Zhang 2006; Bishop 2006 § 5.7.

simply tell a patient "you will probably not contract the disease", but has to decide among dismissal or different kinds of preventive treatment[6]. Said otherwise, in classification we must choose the *optimal* class, not the probably true one. Making optimal choices in situations of uncertainty is the domain of Decision Theory[7]. In order to make an optimal choice, decision theory requires the use of probability values that properly reflect our state of uncertainty.

Determining class probabilities conditional on the input features is unfortunately computationally unfeasible at present for problems that involve very high-dimensional spaces, such as image classification; in fact if an exact probabilistic analysis were possible we would not be developing machine-learning classifiers in the first place[8]. 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.

In the present work we propose an alternative solution that has a low computational cost and that can be applied to all commonly used classifiers, even those that only output class labels. The idea is *to treat the classifier as a diagnostic test*, such as any common diagnostic or prognostic test used in medicine for example. We consider its output, discrete or continuous, as a quantity that has some correlation (in the general sense of the word, not the linear Pearson one) with the true class. This correlation can be analysed and quantified with probability theory in a test set and used to calculate a sensible probability for the class given the output of the test – of the classifier. This analysis needs to be made only once and is computationally cheap, because the classifier output takes values in a discrete or low-dimensional space.

This approach differs from the computationally infeasible one discussed above in that we are calculating the computationally easier probability

$$\mathrm{p}(\text{class} \mid \text{classifier output}) \tag{1}$$

rather than

$$\mathrm{p}(\text{class} \mid \text{feature})\,. \tag{2}$$

[6] Sox et al. 2013; Hunink et al. 2014. [7] Russell & Norvig 2022 ch. 15; Jeffrey 1965; North 1968; Raiffa 1970. [8] Russell & Norvig 2022 chs 2, 12; Pearl 1988.

Note that the former probability is the marginal

$$\mathrm{p}(\text{class} \mid \text{classifier output}) = \sum_{\text{feature}} \mathrm{p}(\text{class} \mid \text{feature})\, \mathrm{p}(\text{feature} \mid \text{classifier output})\,. \quad (3)$$

We can thus think of this approach as a marginalization over the possible features, which is necessary because we have no effective access to them. From this point of view we are not calculating 'approximate probabilities', but exact probabilities conditional on reduced information.

There are indeed many parallels in the way machine-learning classifiers and diagnostic tests, a flu test for example, are devised and work. For both, in principle we would like to assess some situational variable – class, pathological condition – by means of its correlation (which includes deterministic dependence as a limiting case) with a set of 'difficult' variables that are either very complex or not easily observable – image's pixels, presence of replicating viral agents –:

$$\text{situational variable} \leftrightarrow \text{feature variables}$$

We devise an auxiliary variable – output, test result – to be correlated with the difficult set of variables:

$$\text{situational variable} \leftrightarrow \text{difficult variables} \leftrightarrow \text{aux variable}$$

so that we can now assess the situational variable by observing the more easily accessible auxiliary variable. The correlation of the auxiliary variable is achieved by the training process in the case of the machine-learning algorithm, and by exploiting biochemical processes or reactions in the case of the flu test. Note that the situational variable is *informationally screened* from the auxiliary variable by the difficult variables. That is, the auxiliary variable does not – in fact, cannot – contain any more information about the situational variable than that contained in the difficult variables. This means that the probability relationship between the three variables is as follows:

$$\mathrm{p}\left(\begin{smallmatrix}\text{situational}\\ \text{variable}\end{smallmatrix} \,\middle|\, \begin{smallmatrix}\text{aux}\\ \text{variable}\end{smallmatrix}\right) = \sum_{\substack{\text{difficult}\\ \text{variables}}} \mathrm{p}\left(\begin{smallmatrix}\text{situational}\\ \text{variable}\end{smallmatrix} \,\middle|\, \begin{smallmatrix}\text{difficult}\\ \text{variables}\end{smallmatrix}\right) \mathrm{p}\left(\begin{smallmatrix}\text{difficult}\\ \text{variables}\end{smallmatrix} \,\middle|\, \begin{smallmatrix}\text{aux}\\ \text{variable}\end{smallmatrix}\right)\,. \quad (4)$$

In the case of the diagnostic test we do not take its output at face value. Its application on a test set of cases gives us a table of joint statistics – a so-called contingency table[9], akin to a confusion matrix – between the test's output and the true situation, say, regarding the presence of the flu. It is from this contingency table that we derive the *probability* of the virus's presence, given the test result. Analogously we can use the statistics of the application of machine-learning classifier on a test set to derive the probability of a class, given the classifier output. This can be seamlessly done for both a discrete and a continuous output.

The probability $\mathrm{p}(\text{class} \mid \text{classifier output})$ can be calculated nonparametrically, that is, without making any assumptions such as linearity or normality, besides very mild and reasonable assumptions of continuity.

In this work we actually calculate

$$\mathrm{p}(\text{class}, \text{feature} \mid \text{test data}) \tag{5}$$

using the versatile computational approach by Dunson & Bhattacharya (2011), and obtain the probability (1) by conditionalization. The calculation requires Monte Carlo sampling ✎ refs here but needs to be made only once.

This point of view has the following advantages:

- It does not require any changes of the standard training procedures.
- It is easily implemented as an additional low-cost computation of a function at the end of the classifier's output.
- For one classifier, the assessment of this function needs to be done only once.
- It does not make any assumptions such as linearity or gaussianity, except for the (unavoidable) assumption of continuity ✎ need to explain this better
- It yields not only the probability distribution for the classes, but also a measure of how much this distribution could change if we collected more test data (the 'probability of the probability', so to speak).
- It allows us to use the classifier both in a discriminative and generative way. That is, we can use either $\mathrm{p}(\text{class} \mid \text{classifier output})$,

[9] Fienberg 2007; Mosteller et al. 2013.

or $\mathrm{p}(\text{classifier output} \mid \text{class})$ in conjunction with Bayes's theorem. The latter approach enables us to avoid possible base-rate fallacies[10].

- It can be seamlessly integrated with a utility matrix to compute the optimal class, as shown in the companion work ✎.

✎ Add a summary of the main results here. The approach should lead to an improvement in performance.

We could show that even if we used a biased test set, the method corrects the bias (provided we know what the bias is).

✎ Add summary of rest of paper.

2 Implementation of the idea

3 Results

4 Mathematical details of the nonparametric density regression

5 Summary and discussion

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

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[11] 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[12]. 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:

$$\mathrm{p}(X=0|Z=z,D), \qquad \mathrm{p}(X=1|Z=z,D) \equiv 1-\mathrm{p}(X=0|Z=z,D)\,, \tag{6}$$

   according to the rules of the probability calculus.

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

$$\begin{aligned}\bar{\mathrm{u}}(\hat{0}) &\coloneqq U(\hat{0} \mid X=0)\,\mathrm{p}(X=0 \mid Z=z,D) + U(\hat{0} \mid X=1)\,\mathrm{p}(X=1 \mid Z=z,D)\\ \bar{\mathrm{u}}(\hat{1}) &\coloneqq U(\hat{1} \mid X=0)\,\mathrm{p}(X=0 \mid Z=z,D) + U(\hat{1} \mid X=1)\,\mathrm{p}(X=1 \mid Z=z,D)\end{aligned} \tag{7}$$

   and choose the action having maximal expected utility.

[11] Russell & Norvig 2022 part IV. [12] 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 $F_{\mathrm{s}}(X=0 \mid Z=z)$:

$$F(X=0 \mid Z=z) \sim F_{\mathrm{s}}(X=0 \mid Z=z)\,. \tag{8}$$

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.

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

$$F(z) = F(X=0, z) + F(X=1, z) \equiv \sum_x F(x, z) \,. \tag{9}$$

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

$$f(x \mid z) = \frac{F(x, z)}{F(z)} \,. \tag{10}$$

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:

$$\begin{aligned}
&\text{if} \quad U(a \mid X=0) > U(b \mid X=0) \quad \text{then choose action } a,\\
&\text{if} \quad U(a \mid X=0) < U(b \mid X=0) \quad \text{then choose action } b,\\
&\text{else} \quad \text{it does not matter which action is chosen.}
\end{aligned} \tag{11}$$

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

$$\begin{aligned}
\bar{U}(a \mid z) &\coloneqq U(a \mid X=0)\, f(X=0 \mid z) + U(a \mid X=1)\, f(X=1 \mid z) \,,\\
\bar{U}(b \mid z) &\coloneqq U(b \mid X=0)\, f(X=0 \mid z) + U(b \mid X=1)\, f(X=1 \mid z) \,.
\end{aligned} \tag{12}$$

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

$$\begin{aligned} &\text{if} \quad \bar{U}(a \mid z) > \bar{U}(b \mid z) \quad \text{then choose action } a, \\ &\text{if} \quad \bar{U}(a \mid z) < \bar{U}(b \mid z) \quad \text{then choose action } b, \\ &\text{else} \quad \text{it does not matter which action is chosen.} \end{aligned} \tag{13}$$

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 (13).

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. (12) in place of the unknown $f(x \mid z)$. The decision procedure (13) 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

$$\exp\left[\sum_{X,Z} F_s(X, Z) \ln F(X, Z)\right] \tag{14}$$

  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[13].

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

[13] 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:

$$\frac{\bar{f}(Z \mid X)\, P(X)}{\sum_X \bar{f}(Z \mid X)\, P(X)} \,. \tag{15}$$

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 (15) is the 'base-rate fallacy'[14].

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 (14) described above.

[14] 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.

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.

Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M., eds. (2011): *Bayesian Statistics 9*. (Oxford University Press, Oxford). DOI:10.1093/acprof:oso/9780199694587.001.0001.

Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, repr. (Wiley, New York). DOI:10.1002/9780470316870. First publ. 1994.

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.

Damien, P., Dellaportas, P., Polson, N. G., Stephens, D. A., eds. (2013): *Bayesian Theory and Applications*. (Oxford University Press, Oxford). DOI:10.1093/acprof:oso/9780199695607.001.0001.

Dawid, A. P. (2013): *Exchangeability and its ramifications*. In: Damien, Dellaportas, Polson, Stephens (2013): ch. 2:19–29. DOI:10.1093/acprof:oso/9780199695607.003.0002.

Dunson, D. B., Bhattacharya, A. (2011): *Nonparametric Bayes regression and classification through mixtures of product kernels*. In: Bernardo, Bayarri, Berger, Dawid, Heckerman, Smith, West (2011): 145–158. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.178.1521, DOI:10.1093/acprof:oso/9780199694587.003.0005, older version at https://www.researchgate.net/publication/228447342_Nonparametric_Bayes_Regression_and_Classification_Through_Mixtures_of_Product_Kernels.

Fienberg, S. E. (2007): *The Analysis of Cross-Classified Categorical Data*, 2nd ed. (Springer, New York). DOI:10.1007/978-0-387-72825-4. First publ. 1980.

Fisher, R. A. (1967): *Statistical Methods and Scientific Inference*, repr. of 2nd rev. ed. (Oliver and Boyd, Edinburgh). First publ. 1956.

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).

Goodman, L. A., Kruskal, W. H. (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).

Guyon, I., Gunn, S., Nikravesh, M., Zadeh, L. A., eds. (2006): *Feature Extraction: Foundations and Applications*. (Springer, Berlin). doi:10.1007/978-3-540-35488-8.

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).

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.

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).

Lindley, D. V., Novick, M. R. (1981): *The role of exchangeability in inference*. Ann. Stat. **9**[1], 45–58. doi:10.1214/aos/1176345331.

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.

Mosteller, F., Fienberg, S. E., Rourke, R. E. K. (2013): *Beginning statistics with data analysis*, repr. (Dover, Mineola, USA). First publ. 1983.

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

Neal, R. M., Zhang, J. (2006): *High dimensional classification with Bayesian neural networks and Dirichlet diffusion trees*. In: Guyon, Gunn, Nikravesh, Zadeh (2006): p. 10:265–296. doi:10.1007/978-3-540-35488-8_11.

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.

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

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

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.