

The foundations of machine learning on probability theory and decision theory

Luca 
<pgl@portamana.org>

(or any permutation thereof)

Alexander
<***@***>

6 November 2021; updated 16 December 2021

1 Inferential vs decision algorithms

There are redundancies and repetitions to be fixed

A predictive, non-probabilistic machine-learning algorithm is typically designed to be used multiple times to output a quantity y , which we call ‘predictand’, given some input quantity x , called the ‘predictor’. At bottom, such an algorithm is the implementation of a function $x \mapsto y$. (We do not make any assumptions regarding x and y : either could be discrete, continuous, or belong to some complicated multidimensional manifold; thus the machine-learning algorithm could be doing classification or regression.)

The internal parameters of the algorithm are usually explored at the training phase and finally kept fixed at values to be used thereafter.

At each use, the algorithm is making both a forecast and a decision under the hood. A forecast because it must consider probable predictand values related to the predictor. A decision because it chooses as its output only one among such values.

In training and selecting such an algorithm we are also making a forecast and a decision. A forecast because we consider several possible algorithms, architectures, and internal-parameter values, and we must evaluate which of them will most probably yield optimal outputs in their forthcoming use. A decision because we must finally choose one such algorithm (which can be a combination of different kinds of algorithms, but such combination is itself an algorithm), with definite internal-parameter values, to be shipped and employed by the final user.

The processes of training and selecting a machine-learning algorithm are thus a forecast & decision problem. These processes should therefore

follow the rules of probability theory and decision theory. If they did not, they would be marred by intrinsic logical inconsistencies or have a mathematically guaranteed suboptimal long-run performance¹. We shall interchangeably speak of forecast or ‘inference’, meaning the assessment of a probability distribution for some quantity or scenario of interest, *without commitment* to any specific value of such quantity.

The purpose of the present note is to make a step-by-step derivation of a general machine-learning training and selection process, according to probability theory and decision theory, and to interpret some common machine-learning algorithms and practices from this point of view.

The foundations of machine-learning algorithms on probability theory have been explored at the very least since the 1980s²: these algorithms can generally be seen as probability calculations based on exchangeability assumptions³. The output of a trained machine-learning algorithm, however, is often a specific value taken at face value; that is, treated as “the truth”⁴. From this point of view the algorithm is making a *decision*: choosing a specific value to be used in the problem at hand. Moreover, once the algorithm has been trained it is often used for future decisions without any further training; that is, its internal parameters (the weights of a neural net, for example) remain fixed at some specific value. This also represents a choice from our part.

When we not only assess the probabilities for the values of a quantity, but also choose a specific value or other course of action based on such assessment, we enter the domain of decision theory⁵. We shall consider this theory as normative and use its principles to approach the problem of training and choosing the internal parameters of a machine-learning algorithm that performs regression or classification, that is, that yields an output which should ideally be equal to a true value y ,

¹ Pratt et al. 1996; Berger 1985; Raiffa & Schlaifer 2000; Bernardo & Smith 2000; Jaynes 2003 esp. chs 13–14; DeGroot 2004; for early works see Wald 1964; Savage 1972; von Neumann & Morgenstern 1955; Bernoulli 1738; for introductions and summaries: Good 1952; Pratt et al. 1964; North 1968; and the brilliant Raiffa 1970; it is known that human beings often do not follow rational decision theory (nor logic for that matter); this is why logic, probability theory, decision theory are *normative*, not descriptive, theories; see e.g. Tversky 1975; Tversky & Kahneman 1981. ² Tishby et al. 1989; Levin et al. 1990; MacKay 1992a,b,c,d; 2005 esp. Part V; Neal 1996. ³ Bernardo & Smith 2000 ch. 4. ⁴ cf. MacKay 1992b § 3. ⁵ Savage 1972; Raiffa & Schlaifer 2000; Berger 1985; Bernardo & Smith 2000 ch. 2; Pratt et al. 1996; Jaynes 2003 chs 13–14; for a charming introduction see Raiffa 1970.

called the predictand, given an input x , called the predictor. Predictor and predictand quantities can be real-valued, categorical, or belong to some general manifold. A brief summary of decision theory is given in the next section.

2 A simplified overview of decision theory

 to be written

3 Training, parameter choice, and algorithm choice as a combined decision problem

Let us now examine our problem from a decision-theoretic perspective.

3.1 Decisions

A machine-learning algorithm with internal parameters θ typically yields an output that is a function $t(x \mid \theta)$ of the input x and of the (fixed) parameter θ . For example, for a neural net t is a composition of nonlinear functions of x , and θ is a set of connection weights; in the case of a linear-regression algorithm, t is a linear function of x and θ its coefficients.

Our goal is to choose one among several machine-learning algorithms, and specific internal-parameter values for that algorithm.

This nested choice can actually be combined into a single choice. Label the candidates algorithms with 1, 2, etc., and denote their parameter spaces by Θ_1 , Θ_2 , etc.. The choice of algorithm and internal parameter can then be seen as the choice of a parameter value θ in the union space $\Theta_1 \cup \Theta_2 \cup \dots$, denoted Θ . We shall denote the output produced by the machine-learning algorithm & parameter θ operating on the input x simply as $\theta(x)$, getting rid of the symbol “ t ”. Later we shall explore how the separation into two different kinds of choices, algorithm and parameters separately, is made.

Our possible choices or decisions therefore consist of the possible values $\theta \in \Theta$.

We can alternatively see our set of choices as the set of possible sequences of future outputs in response to future predictor values. Since a specific output sequence is determined by a specific value of θ , the two

points of view are equivalent. In § 3.3 this equivalence will be apparent in the mathematical expression for the utility.

3.2 Scenarios

Besides the space of choices we must specify the space of possible scenarios, of which only one will turn out to be true. Our scenarios consist of all possible sequences of predictor-predictand pairs $((x_1, y_1), (x_2, y_2), \dots)$ that our algorithm will encounter in its lifetime: (x_n) will be the known inputs fed to the algorithm, and (y_n) the unknown values that the algorithm will try to predict. Let us assume that this sequence is finite although very large.

For typographical convenience any pair (x_n, y_n) is briefly denoted $x_n y_n$; this juxtaposition does not represent any mathematical operation. Denote $\mathbf{x} := (x_1, x_2, \dots)$, analogously for \mathbf{y} , and $\mathbf{x}\mathbf{y} := (x_1 y_1, x_2 y_2, \dots)$. If the quantity x takes values in the manifold X and y in Y , our possible scenarios live in the space $\prod_n (X \times Y)$.

Besides the sequence $(x_n y_n)$ we also have a sequence $(\tilde{x}_m \tilde{y}_m)$ of predictors (\tilde{x}_m) and corresponding *known* predictands (\tilde{y}_m) : our training data. One may adopt the point of view that our set of scenarios should consist of all possible sequences $(x_n y_n, \tilde{x}_m \tilde{y}_m)$ (the subsequence $(\tilde{x}_m \tilde{y}_m)$ being common to all of them), on the grounds that one wants the choice of algorithm and parameters to be optimal not only for future predictions, but also for past ones. In the following steps we shall also consider this alternative point of view. Let us denote $\tilde{\mathbf{x}} := (\tilde{x}_1, \tilde{x}_2, \dots)$, analogously for $\tilde{\mathbf{y}}$, and $\tilde{\mathbf{x}}\tilde{\mathbf{y}} := (\tilde{x}_1 \tilde{y}_1, \tilde{x}_2 \tilde{y}_2, \dots)$.

✚ It's maybe less confusing to proceed here and in the next sections without considering the possibility of including training data in the scenarios. Then in a later section the results are adapted to this assumptions. Will try this change later.

3.3 Utilities

For each combination of decision (algorithm & internal parameter) θ and scenario $\mathbf{x}\mathbf{y}$ we must now specify the utility $U(\theta \mid \mathbf{x}\mathbf{y})$.

We make the realistic assumptions that this utility is the sum of utilities $u(\theta \mid x_n y_n)$ for each single application n of the algorithm to the

sequence of data \mathbf{xy} , and that such individual utilities have identical functional forms:

$$U(\theta \mid \mathbf{xy}) = \sum_n u(\theta \mid x_n y_n) . \quad (1)$$

A more general approach, where the functional form of the utility changes with each instance (even if x_n and y_n assume the same pair of values), is also possible and may be more appropriate in particular situations.

In each single instance what we are actually choosing is an output value $\theta(x)$, determined by the input x and by the algorithm and its parameter θ . The single-instance unknown is the true value y . So the single-instance utility $u(\theta \mid x_n y_n)$ can actually be rewritten as

$$u[y_n \mid \theta(x_n)] , \quad (2)$$

so that

$$U(\theta \mid \mathbf{xy}) = \sum_n u[y_n \mid \theta(x_n)] . \quad (3)$$

This equation expresses the fact, remarked in § 3.1, that the choice of parameter θ is equivalent to a choice *en masse* of future outputs (y_n), since the latter are determined by the former.

If we also want to include the training data in the set of possible scenarios, as discussed in § 3.2, then the total utility is

$$U(\theta \mid \mathbf{xy}\tilde{\mathbf{x}}\tilde{\mathbf{y}}) = \sum_n u[y_n \mid \theta(x_n)] + \sum_m u[\tilde{y}_m \mid \theta(\tilde{x}_m)] . \quad (4)$$

The functional form of the single-instance utility $u(\cdot \mid \cdot)$ depends on the specific problem – it is in fact no less problem-specific than the choice of machine-learning algorithm – so we do not make any more specific assumptions about it.

3.4 Probabilities for the scenarios and exchangeability

Lastly we need to assess the distribution of probability over the possible scenarios \mathbf{xy} . This probability distribution is conditional on some hypotheses, assumptions, or background knowledge H , and on the sequence

$(\tilde{x}_m \tilde{y}_m)$ of known predictors and predictands discussed in § 3.2. It can be written in two equivalent ways:

$$p(xy \mid \tilde{x}\tilde{y}, H) dx y \equiv p(y \mid x, \tilde{x}\tilde{y}, H) p(x \mid \tilde{x}\tilde{y}, H) dx y . \quad (5)$$

Two alternative assumptions can be made about this distribution.

(I) *Joint exchangeability*. The prior distribution is jointly exchangeable in predictors and predictands. That is, the probability is the same for any sequence obtained from $xy\tilde{x}\tilde{y}$ by simultaneously exchanging predictor-predictand pairs between different instances in the sequence, even across future and training subsequences.

(II) *Conditional exchangeability*.⁶ The prior distribution is conditionally exchangeable in the predictands given the predictors, but not in the predictors across future and training subsequences. That is, the probability is the same for any sequence obtained from $xy\tilde{x}\tilde{y}$ by exchanging predictand values between different instances in the sequence *that have the same predictor values*; however, the probability is generally not the same if we exchange predictor values, especially if across future data and training data.

These two assumptions can also be understood in terms of “populations”⁷ and their frequency distributions. (I) says that future and training predictor-predictand pairs all come from the same population. In other words, the *joint* frequency distribution observed in the training data should be similar to the joint frequency distribution of future data our machine-learning algorithm will be applied to. (II) says that future and training predictands come from the same subpopulations conditional on the same predictor values. The predictor values in future and training data, however, come from potentially different populations. In other words, every frequency distribution, observed in the training data, of the predictand *conditional* on each predictor value should be similar to every frequency distribution of future predictands, *conditional* on the same predictor value; the (marginal) frequency distribution of predictor values in the training data, however, may differ from that of predictor values in future data. This difference is related to the distinction between “generative” and “discriminative” approaches.

⁶ Bernardo & Smith 2000 § 4.6.2; Diaconis 1988 § 3; Lindley & Novick 1981 § 3, Appendix 2.

⁷ see the extremely insightful discussion in Lindley & Novick 1981.

Assumption (I) applies to cases where our training predictor-predictand pairs come from the same source (or, again, “population” if you like) as the future pairs. Assumption (II) applies to cases where our training predictors and future predictors come from different sources, possibly because of constraints in their choice – for example, the algorithm will be applied to predictor values expensive to realize artificially, and the training is based on predictor values cheaper to realize artificially.

It is important to note that joint exchangeability (I) implies conditional exchangeability (II), but not vice versa. So the conditional exchangeability of “predictand | predictor” is assumed in any case. It is indeed a necessary assumptions for our regression problem to make sense at all.

We shall use a notation that allows us to consider assumptions (I) and (III) at the same time.

$$p(y | x, \tilde{x}\tilde{y}, H) = \int \left[\prod_n F(y_n | x_n) \right] p(F | \tilde{x}\tilde{y}, H) dF \quad (6a)$$

with

$$p(F | \tilde{x}\tilde{y}, H) = \frac{\left[\prod_m F(\tilde{y}_m | \tilde{x}_m) \right] p(F | H)}{\int \left[\prod_m F(\tilde{y}_m | \tilde{x}_m) \right] p(F | H) dF} . \quad (6b)$$

This distribution is typically assumed to be exchangeable in the whole sequence of data (known and unknown) and therefore by de Finetti’s theorem⁸ and Bayes’s theorem its density must have the form

$$p(xy | \tilde{x}\tilde{y}, H) = \int \left[\prod_n F(x_n y_n) \right] p(F | \tilde{x}\tilde{y}, H) dF \quad (7a)$$

with

$$p(F | \tilde{x}\tilde{y}, H) = \frac{\left[\prod_m F(\tilde{x}_m \tilde{y}_m) \right] p(F | H)}{\int \left[\prod_m F(\tilde{x}_m \tilde{y}_m) \right] p(F | H) dF} . \quad (7b)$$

These expressions can be intuitively interpreted as follows⁹. The known and unknown sequences of data together constitute a “population” where the different values in X and Y appear with joint frequency density $F(xy) dx dy$. If we knew such density, then our probability assessment for any new pair of values would simply be $F(xy)$ owing to symmetry

⁸ De Finetti 1930; 1937; Hewitt & Savage 1955; Bernardo & Smith 2000 ch. 4; Dawid 2013 for an insightful summary see. ⁹ cf. Lindley & Novick 1981.

reasons. But since we do not know the density F , we must marginalize over all possible such densities, each given a probability, as in eq. (7a). The prior probability density at frequency F is $p(F | H) dF$, which is updated to $p(F | \tilde{x}\tilde{y}, H)$, eq. (7b), when the training data $\tilde{x}\tilde{y}$ are known.

If the training data are considered part of the scenarios, then our probability distribution is

$$p(\mathbf{x}\mathbf{y} | \tilde{x}\tilde{y}, H) \delta(\tilde{x}'\tilde{y}' - \tilde{x}\tilde{y}) d\mathbf{x}\mathbf{y} d\tilde{x}'\tilde{y}' \quad (8)$$

since the training data are known and their probability is one; the term $p(\mathbf{x}\mathbf{y} | \tilde{x}\tilde{y}, H)$ is still given by eqs (7).

3.5 Expected utilities and final choice

According to decision theory every action θ has an associated expected utility

$$E(\theta | \tilde{x}\tilde{y}, H) := \iint U(\theta | \mathbf{x}\mathbf{y}) p(\mathbf{x}\mathbf{y} | \tilde{x}\tilde{y}, H) d\mathbf{y} d\mathbf{x} \quad (9)$$

which, using eqs (3) and (7), becomes

$$\begin{aligned} E(\theta | \tilde{x}\tilde{y}, H) = & \frac{1}{Z(\tilde{x}\tilde{y})} \iiint \sum_n u[y_n | \theta(x_n)] \times \\ & \left[\prod_m F(y_m | x_m) G(x_m) \right] \left[\prod_m F(\tilde{y}_m | \tilde{x}_m) \right] \times \\ & p(F | H) p(G | \tilde{x}\tilde{y}, H) d\mathbf{y} d\mathbf{x} dF dG \end{aligned} \quad (10a)$$

with

$$Z(\tilde{x}\tilde{y}) := \int \left[\prod_m F(\tilde{y}_m | \tilde{x}_m) \right] p(F | H) dF. \quad (10b)$$

This expression can be simplified exchanging integrals and the sum in n , and then integrating over the pairs $d\mathbf{y}_m d\mathbf{x}_m$ for which $m \neq n$; such

integrals give unity since G and each F are normalized. We obtain

$$\begin{aligned}
 E(\theta \mid \tilde{x}\tilde{y}, H) &= \frac{1}{Z(\tilde{x}\tilde{y})} \sum_n \iiint u[y_n \mid \theta(x_n)] F(x_n \mid y_n) G(x_n) \times \\
 &\quad \left[\prod_m F(\tilde{y}_m \mid \tilde{x}_m) \right] p(F \mid H) p(G \mid \tilde{x}\tilde{y}, H) dy_n dx_n dF dG \\
 &\propto \iiint u[y \mid \theta(x)] F(y \mid x) \left[\prod_m F(\tilde{y}_m \mid \tilde{x}_m) \right] \times \\
 &\quad p(F \mid H) p(x \mid \tilde{x}\tilde{y}, H) dy dx dF .
 \end{aligned} \tag{11}$$

In the last expression we have renamed the dummy integration variables $x_n y_n$ to $x y$, and performed a formal integration over dG . The terms of the sum in n are therefore all equal, and the utility is a multiple of any such term. We also omit the θ -independent factor $Z(\tilde{x}\tilde{y})$. Thus the final expected utility of θ , besides a constant factor, is

$$\begin{aligned}
 E(\theta \mid \tilde{x}\tilde{y}, H) &= \iiint u[y \mid \theta(x)] F(y \mid x) \left[\prod_m F(\tilde{y}_m \mid \tilde{x}_m) \right] p(F \mid H) \times \\
 &\quad p(x \mid \tilde{x}\tilde{y}, H) dy dx dF .
 \end{aligned} \tag{12}$$

The formal solution to our decision problem is finally this: choose the algorithm and internal parameters θ^* given by

$$\begin{aligned}
 \theta^* &:= \arg \sup_{\theta} \iiint u[y \mid \theta(x)] \times \\
 &\quad F(y \mid x) \left[\prod_m F(\tilde{y}_m \mid \tilde{x}_m) \right] p(F \mid H) \times \\
 &\quad p(x \mid \tilde{x}\tilde{y}, H) dy dx dF .
 \end{aligned}$$

(13)

In the next section we analyse and discuss this formula, and study possible approximations to it.

4 Observations on the decision-theoretic solution


Formula (13) presents three noteworthy points:

(A) The optimization can be separated into an optimization of internal parameters for each algorithm i , and then an optimization among algorithms, because we can first take $\theta_i^* := \arg \sup_{\theta \in \Theta_i}$ for each i and then $\theta^* := \arg \sup_{\theta \in \{\theta_i^*\}}$.

(B) There is a utility part $u[y \mid \theta(x)]$, and a clearly distinct inferential part $p(\mathbf{x}y \mid \tilde{\mathbf{x}}\tilde{\mathbf{y}}, H)$ rewritten in terms involving the conditional distributions F and the probability distribution for x . The variable θ , which runs over the candidate machine-learning algorithms and their internal-parameter values, appears only in the utility part, not in the inferential part. The latter is the same for all candidate algorithms. This fact has several important implications.

First, it is inconsistent to use different algorithms, depending on the value of θ , to compute the inferential part $p(\mathbf{x}y \mid \tilde{\mathbf{x}}\tilde{\mathbf{y}}, H)$. For example, it would be inconsistent to use a deep network H_{dn} to compute $p(\mathbf{x}y \mid \tilde{\mathbf{x}}\tilde{\mathbf{y}}, H_{\text{dn}})$, multiplying this probability by the deep network's utility $u[y \mid \theta_{\text{dn}}(x)]$; and then use a probabilistic random forest H_{rf} to compute $p(\mathbf{x}y \mid \tilde{\mathbf{x}}\tilde{\mathbf{y}}, H_{\text{rf}})$, multiplying this probability by the random forest's utility $u[y \mid \theta_{\text{rf}}(x)]$; and so on. Only one algorithm should be used to compute $p(\mathbf{x}y \mid \tilde{\mathbf{x}}\tilde{\mathbf{y}}, H)$; this could be either an average over all candidate algorithms (with appropriate weights, as given by the probability calculus; see MacKay 2005 § 28.1 p. 347), or the most probable among them (which would be equivalent to a maximum a-posteriori approximation).

This fact may appear to conflict with analyses¹⁰ focused on an inferential-only interpretation. That kind of analysis involves a comparison among different models, based on the probabilities that they give to the data. However, such comparison is not meant as *choice*¹¹, so there is no real conflict.

Second, a probabilistic variant of a candidate algorithm might be most appropriate for the inferential part (for example because it is a non-parametric algorithm), and yet the final optimization might select a different candidate algorithm.  This must be explained much better.

¹⁰ for example MacKay 1992a § 5; 1992b § 2.2; Bishop 2006 § 1.3; Barber 2020 ch. 12; Murphy 2012 § 1.4.8. ¹¹ “This paper concerns inference alone and no loss functions or utilities are involved” MacKay 1992a footnote 1; “When we discuss model comparison, this should not be construed as implying model *choice*” MacKay 2005 § 28.1 p. 347.

(C) The optimization crucially depends on the distribution of probability over the future predictors: $p(x \mid \tilde{x}\tilde{y}, H)$. This fact makes sense. One algorithm might give correct predictions for a very wide range \mathcal{W} of predictor values x , and poor predictions in the complementary, very narrow range \mathcal{N} ; another algorithm might give poor predictions in \mathcal{N} and correct predictions in \mathcal{W} . If future predictor values are much more probable to fall in \mathcal{N} than \mathcal{W} , $p(x \in \mathcal{N}) \gg p(x \in \mathcal{R})$, then the second algorithm will yield a higher utility in the long run.

The optimal choice of algorithm and internal parameters therefore requires an inference of probable future predictor values. Current practice in machine learning often neglects such inference. Indeed, the importance of such inference should be contrasted with the common routine of balancing the amount of training data among predictor domains. Balancing could actually be counter-productive.

5 Approximations and interpretations

5.1 Many-data approximation

If we have many training data, with frequencies ϕ , the predictive conditional probabilities will be roughly equal to these frequencies:

$$\begin{aligned} \arg \sup_{\theta} \iiint u[y \mid \theta(x)] F(y \mid x) \left[\prod_m F(\tilde{y}_m \mid \tilde{x}_m) \right] p(F \mid H) \times \\ p(x \mid \tilde{x}\tilde{y}, H) \, dy \, dx \, dF \\ \approx \arg \sup_{\theta} \iint u[y \mid \theta(x)] \phi(y \mid x) p(x \mid \tilde{x}\tilde{y}, H) \, dy \, dx \\ \approx \arg \sup_{\theta} \sum_m u[\tilde{y}_m \mid \theta(\tilde{x}_m)] \quad (14) \end{aligned}$$

According to this interpretation *the loss function used in machine learning does not represent the logarithm of our uncertainty or of the distribution of noise: it represents the utility.*

5.2 Single algorithm and approximate inference

 Work in progress

$$\begin{aligned}
 & \arg \sup_{\theta} \iiint u[y \mid \theta(x)] F(y \mid x) \left[\prod_m F(\tilde{y}_m \mid \tilde{x}_m) \right] p(F \mid H) \times \\
 & \quad p(x \mid \tilde{x}\tilde{y}, H) \, dy \, dx \, dF \\
 & = \arg \sup_{\theta} \iiint u[y \mid \theta(x)] f(y \mid x, w) \left[\prod_m f(\tilde{y}_m \mid \tilde{x}_m, w) \right] p(w \mid H) \times \\
 & \quad p(x \mid \tilde{x}\tilde{y}, H) \, dy \, dx \, dw \\
 & \approx \arg \sup_{\theta} \iiint u[y \mid \theta(x)] f(y \mid x, w^*) \times \\
 & \quad p(x \mid \tilde{x}\tilde{y}, H) \, dy \, dx \, dw .
 \end{aligned} \tag{15}$$

6 Discussion

🔑 To be written.

Utility becomes unimportant when the inference is almost deterministic. In modern machine-learning applications in particular fields such as medicine this assumption may not be valid, owing either to the intrinsic, noisy complexity of the phenomena involved, or to the scarcity of training data.

Bibliography

- (“de X” is listed under D, “van X” under V, and so on, regardless of national conventions.)
- 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](https://doi.org/10.1007/978-1-4757-4286-2). First publ. 1980.
- Bernardo, J.-M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M., eds. (1988): *Bayesian Statistics 3*. (Oxford University Press, Oxford).
- Bernardo, J.-M., Smith, A. F. (2000): *Bayesian Theory*, repr. (Wiley, New York). [doi:10.1002/9780470316870](https://doi.org/10.1002/9780470316870). First publ. 1994.
- Bernoulli, D. (1738): *Specimen theoriae novae de mensura sortis*. Commentarii academiae scientiarum imperialis petropolitanae V, 175–192. <https://archive.org/details/SpecimenTheoriaeNovaeDeMensuraSortis>. Transl. in Bernoulli (1954).
- (1954): *Exposition of a new theory on the measurement of risk*. *Econometrica* **22**¹, 23–36. [doi:10.2307/1909829](https://doi.org/10.2307/1909829). Transl. by Louise Sommer of Bernoulli (1738).
- Bishop, C. M. (2006): *Pattern Recognition and Machine Learning*. (Springer, New York). <https://www.microsoft.com/en-us/research/people/cmbishop/prml-book>.

- 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.
- de Finetti, B. (1930): *Funzione caratteristica di un fenomeno aleatorio*. Atti Accad. Lincei: Sc. Fis. Mat. Nat. IV⁵, 86–133. <http://www.brunodefinetti.it/Opere.htm>.
- (1937): *La prévision: ses lois logiques, ses sources subjectives*. Ann. Inst. Henri Poincaré 7¹, 1–68. http://www.numdam.org/item/AIHP_1937__7_1_1_0. Transl. in Kyburg, Smokler (1980), pp. 53–118, by Henry E. Kyburg, Jr.
- DeGroot, M. H. (2004): *optimal statistical decisions*, repr. (Wiley, New York).
- Diaconis, P. (1988): *Recent progress on de Finetti's notions of exchangeability*. In: Bernardo, DeGroot, Lindley, Smith (1988): 111–125. With discussion by D. Blackwell, Simon French, and author's reply. <http://statweb.stanford.edu/~cgates/PERSI/year.html>, <https://statistics.stanford.edu/research/recent-progress-de-finettis-notions-exchangeability>.
- Good, I. J. (1952): *Rational decisions*. J. Roy. Stat. Soc. B XIV¹, 107–114. doi:10.1111/j.2517-6161.1952.tb00104.x. Repr. in Good (1983) ch. 1.
- (1983): *Good Thinking: The Foundations of Probability and Its Applications*. (University of Minnesota Press, Minneapolis, USA).
- Hewitt, E., Savage, L. J. (1955): *Symmetric measures on Cartesian products*. Trans. Am. Math. Soc. 80², 470–501. doi:10.1090/S0002-9947-1955-0076206-8.
- Jaynes, E. T. (2003): *Probability Theory: The Logic of Science*. (Cambridge University Press, Cambridge). Ed. by G. Larry Bretthorst. First publ. 1994. doi:10.1017/CBO9780511790423, <https://archive.org/details/XQUHIUXHIQUHIQXUIHX2>, <http://www.biba.inrialpes.fr/Jaynes/prob.html>.
- Kyburg Jr., H. E., Smokler, H. E., eds. (1980): *Studies in Subjective Probability*, 2nd ed. (Robert E. Krieger, Huntington, USA). First publ. 1964.
- Levin, E., Tishby, N., Solla, S. A. (1990): *A statistical approach to learning and generalization in layered neural networks*. Proc. IEEE 78¹⁰, 1568–1574. doi:10.1109/5.58339.
- Lindley, D. V., Novick, M. R. (1981): *The role of exchangeability in inference*. Ann. Stat. 9¹, 45–58. doi:10.1214/aos/1176345331.
- MacKay, D. J. C. (1992a): *Bayesian interpolation*. Neural Comput. 4³, 415–447. <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>, doi:10.1162/neco.1992.4.3.415.
- (1992b): *A practical Bayesian framework for backpropagation networks*. Neural Comput. 4³, 448–472. <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>, doi:10.1162/neco.1992.4.3.448.
- (1992c): *Information-based objective functions for active data selection*. Neural Comput. 4⁴, 590–604. <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>, doi:10.1162/neco.1992.4.4.590.
- (1992d): *The evidence framework applied to classification networks*. Neural Comput. 4⁵, 720–736. <http://www.inference.phy.cam.ac.uk/mackay/PhD.html>, doi:10.1162/neco.1992.4.5.720.
- (2005): *Information Theory, Inference, and Learning Algorithms*, Version 7.2 (4th pr.) (Cambridge University Press, Cambridge). <https://www.inference.org.uk/itila/book.html>. First publ. 1995.
- Murphy, K. P. (2012): *Machine Learning: A Probabilistic Perspective*. (MIT Press, Cambridge, USA). <https://problml.github.io/pml-book/book0.html>.

- Neal, R. M. (1996): *Bayesian Learning for Neural Networks*. (Springer, New York). DOI: [10.1007/978-1-4612-0745-0](https://doi.org/10.1007/978-1-4612-0745-0), <https://www.cs.toronto.edu/~radford/bnn.book.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://doi.org/10.1109/TSSC.1968.300114), <https://stat.duke.edu/~scs/Courses/STAT102/DecisionTheoryTutorial.pdf>.
- Pratt, J. W., Raiffa, H., Schlaifer, R. (1964): *The foundations of decision under uncertainty: an elementary exposition*. J. Am. Stat. Assoc. **59**³⁰⁶, 353–375. DOI:[10.1080/01621459.1964.10482164](https://doi.org/10.1080/01621459.1964.10482164).
- (1996): *Introduction to Statistical Decision Theory*, 2nd pr. (MIT Press, Cambridge, USA). First publ. 1995.
- 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.
- Savage, L. J. (1972): *The Foundations of Statistics*, 2nd rev. and enl. ed. (Dover, New York). First publ. 1954.
- Tishby, N., Levin, E., Solla, S. A. (1989): *Consistent inference of probabilities in layered networks: predictions and generalizations*. Int. Joint Conf. Neural Netw. **1989**, II-403–II-409. DOI: [10.1109/IJCNN.1989.118274](https://doi.org/10.1109/IJCNN.1989.118274).
- Tversky, A. (1975): *A critique of expected utility theory: descriptive and normative considerations*. Erkenntnis **9**², 163–173. DOI:[10.1007/BF00226380](https://doi.org/10.1007/BF00226380).
- Tversky, A., Kahneman, D. (1981): *The framing of decisions and the psychology of choice*. Science **211**⁴⁴⁸¹, 453–458. DOI:[10.1126/science.7455683](https://doi.org/10.1126/science.7455683).
- von Neumann, J., Morgenstern, O. (1955): *Theory of Games and Economic Behavior*, 3rd ed., 6th pr. (Princeton University Press, Princeton). <https://archive.org/details/in.ernet.dli.2015.215284>. First publ. 1944.
- Wald, A. (1964): *Statistical Decision Functions*, 5th pr. (Wiley, New York). <https://hdl.handle.net/2027/uc1.b4979640>. First publ. 1950.