Ulisses M. Braga-Neto and Edward R. Dougherty

# Machine Learning Requires Probability and Statistics

The contemporary practice of machine learning often involves the application of deterministic, computationally intensive algorithms to iteratively minimize a criterion of fit between a discriminant and sample data. There is often little interest in using probability to model the uncertainty in the problem and statistics to characterize the behavior of predictors derived from data, with the emphasis being on computation and coding. It follows that little can be stated about performance on future data, beyond perhaps a simple error count on a given test set. In this article, we argue that the knowledge imparted by deterministic computational methods is not rigorously related to the real world and, in particular, future events. This connection requires rigorous probabilistic modeling and statistical inference as well as an understanding of the proper role of computation and an appreciation of epistemological issues.

## Gauss and the least-squares method

We illustrate the issue with a brief historical excursion into the development of the least-squares method, which is usually credited to an 1809 paper by Gauss [1]—even though Legendre published it in 1805 [2]. Around 1795 (thus, before Legendre's publication), Gauss became preoccupied with the inaccuracy of the classical model of planetary motion (due to Kepler and refined by Newton) in predicting the orbit of the asteroid Ceres. The classical planetary model does not take into account the uncertainty introduced by noise in the observations and the presence of unmodeled variables. Namely, astronomical observations contain inaccuracies, such as human error, atmospheric interference, and optical imperfections in telescopes, and the orbits of planets are determined not only by the sun but also by a superposition of the effects of all the other planets, which creates an intractable analytical problem (known as the *n-body problem*). To address these issues, Gauss introduced the least-squares method, which can be summarized in the famous passage [1]:

> The most probable value of the unknown quantities will be that in which the sum of the squares of the differences between the actually observed and the computed values multiplied by numbers that measure the degree of precision is a minimum.

Thus, the "most probable value" is the one that minimizes the sum of squared deviations between the observations and a candidate in a given family (for example, the family of all ellipses). The least-squares approach has proven to be extremely influential in science and engineering. However, its basic formulation has a significant limitation—it cannot say anything about the performance of the method on future data. This is because no knowledge about the properties of the noise in the model is assumed or sought. This makes the least-squares approach essentially deterministic.

Gauss was not unaware of this issue. In a later work [3], he gave the conditions on the observation noise under which the approach is optimal: if the noise random variables are uncorrelated and have zero mean with constant variance across all observations, then the least-squares solution is unbiased and has minimum variance among all linear estimators; i.e., it is the best linear unbiased estimator (BLUE)—a result known today as the Gauss–Markov theorem. Even though Laplace also worked on the theory of the least-squares method around the same time as Gauss, and Markov later clarified many of its issues, R. Plackett credits Gauss' 1821 paper fully for the Gauss–Markov theorem (which should perhaps then be called the Gauss theorem) [4]. Gauss's 1821 approach is fully stochastic, where the unmeasurable and uncontrollable disturbances are modeled as random variables.

The contrast between the 1809 deterministic least-squares method by Gauss and his 1821 fully stochastic approach represents a significant epistemological transition and illuminates the entire issue we discuss in this article. The 1809 result appeared to be a useful computational method that produced a "good-looking" result given the data. However,

*Digital Object Identifier 10.1109/MSP.2020.2985385*
*Date of current version: 26 June 2020*

its properties could not be established until the result in 1821, which required conditions on the probability distribution of the noise to lead to the statistical optimality of the procedure on future data (and not optimality based merely on the minimization of a sum of squared errors on the current data). In a similar fashion, probability theory and statistics are indispensable components of statistical signal processing, stochastic control, and information theory.

## Computers: Thinking the unthinkable or not thinking at all?

Computers are fascinating because of their superhuman speed and accuracy in executing rote tasks. This feeling about automation is old and precedes computers. For example, Denis Diderot, who made the mechanical arts one of the central parts of the *Encyclopedie*, includes the following quote by a certain M. Perault in the famous article on the stocking machine [5], [6]:

> When one sees stockings being knit, one marvels at the suppleness and dexterity of the worker's hand, though he only makes one stitch at a time. What then when one sees a machine that makes hundreds of stitches at once, that is, makes in one moment all the diverse movements that the human hands would take many hours to make? [...] and all that without the worker who operates the machine understanding anything, knowing anything, or even dreaming of it.

The impression that automation always produces results that are not just faster but superior to manual labor is very strong. The philosopher W. Barrett put it this way in his book *The Illusion of Technique* [7]:

> In the popular imagination the faith in hardware expresses itself in the images of technological gigantism: just make the computer mammoth enough and it will solve all problems. But the intrinsic logic of a problem remains what it is even if we had at our disposal a computer gigantic enough to cover a modern city. The absence of an intelligent idea in the grasp of a problem cannot be redeemed by the elaborateness of the machinery one subsequently employs.

A purely data-driven approach is naturally computationally expensive, and this was a key reason why its use was not frequent before the advent of fast and cheap computers. Now computation and storage are relatively cheap and widely available, which makes it very attractive to apply computation indiscriminately. B. Efron put the matter thus in his paper, "Computers and the Theory of Statistics: Thinking the Unthinkable" [8]: "The 'unthinkable' mentioned in the title is simply the thought that one might be willing to perform 500,000 numerical operations in the analysis of 16 data points. Or one might be willing to perform a billion operations to analyze 500 numbers. Such statements would have seemed insane thirty years ago."

The propensity of using computation indiscriminately was very much in the minds of the pioneers of the information age. The very first manual of the BASIC programming language, invented by J. Kemeny and T. Kurz at Dartmouth in 1964, had a piece of advice from digital signal processing pioneer Richard Hamming of Bell Labs: "Typing is no substitute for thinking" [9].

Our point is not that computation should be avoided but that an exaggerated reliance on it can create an illusion of excellence and independence of human supervision. In the case of machine learning, it has created the expectation that vast amounts of computation can produce accurate predictions from data, without a specification of conditions that provide the possibility of this knowledge.

## On prediction, validation, and experimental design

We arrive at the fundamental question: How does one know that one has a predictive model that is strongly connected to the real world and future events? In classification, a model is predictive if the classification error rate is small. However, how do we know that we have a predictive classifier? This question can only be answered in practice by estimating the classification error using an error-estimation rule applied to the training data, a distinct set of test data, or a mixture of training and test data [10]. The accuracy of the error estimation rule must be measured by a validity criterion, the most common one being the root-mean-square error between estimated and true error, which in turn depends on the feature–label distribution, i.e., the joint probability distribution between the feature vector and the label. Without an underlying probability model, classification validity cannot be established. One could apply any existing error estimation rule, such as cross-validation or test-set error estimation, but this would simply provide a number that relates to the sample training and testing data used and has no quantifiable relation with future performance of the classifier. One might claim to expect a proportion of errors on future applications that agree with the estimate, but this statement is not quantifiable in terms of prediction versus observation and therefore lacks scientific content.

The next logical question is: How can one know that the feature–label distribution reflects the true relationship between feature vector and label? If one has the wrong feature–label distribution, then the trained model will perform poorly on future data. In practice, it is not possible to know completely the feature–label distribution at work in a specific problem. However, assumptions about the feature–label distribution can be enforced by sound experimental design, i.e., the way the data gathering process is planned and executed. In a recent paper, F. Mazzocchi characterizes the issue as follows [11]:

> Science does not collect data randomly. Experiments are designed and carried out within theoretical, methodological and instrumental limitations. Instruments are designed based on prior theories and knowledge, which determine what these instruments indicate

> with respect to the object under investigation. Research does not examine each possible manipulation that could occur, but selects what is relevant in light of a given perspective, sometimes in order to match theoretical predictions with experience.

For instance, for the Gauss–Markov theorem to hold and the BLUE to be valid, the data must be acquired and conditioned/transformed so that the disturbances are, at least approximately, uncorrelated and zero mean with constant variance. Another often overlooked example of an important experimental design issue is the requirement by many machine learning algorithms that the training data be independent and identically distributed. For example, cross-validation is approximately unbiased only under this assumption: if it is violated, cross-validation can be grossly biased [12].

A last objection could be raised by a skeptic: How can we be sure that we can learn from the present data about events that will happen in the future? In other words, are we not always measuring performance on existing test data, as it becomes available? This is the radical empiricist challenge to science, also known as the problem of induction, which was first raised by David Hume in the 18th century [13]. Many attempts have been made to answer this question in the affirmative, but this principle cannot be proved logically. Instead, we must adopt it as a postulate. In the preface to "Scientific Inference," Sir Harold Jeffreys puts it this way [14]: "Discussions from the philosophical and logical point of view have tended to the conclusion that this principle cannot be proved by logic alone, which is true, and have left it at that. […] In the present work the principle is frankly adopted as a primitive postulate and its consequences are developed." As is the case in science at large, Jeffreys' "primitive postulate" must also be adopted in machine learning to avoid the radical empiricist perspective.

## Deep neural networks

Like other machine learning methods based on optimization, neural networks learn from data by iteratively adjusting the parameters of a discriminant to fit a set of labeled data points. The justification often cited for such approaches is that the discriminant of a neural network with a sufficiently large number of parameters can produce results that are arbitrarily close to the optimal predictor in a distribution-free manner; that is, neural networks are universal function approximators. If the large number of parameters is organized over a large number of layers, one has a deep neural network.

A classical theorem by G. Cybenko [15] shows that the classifiers produced by a neural network discriminant with continuous sigmoids are dense in the space of all classifiers, and therefore can be arbitrarily close to the optimal classifier. Cybenko's theorem applies to depth-bound ("shallow") neural networks, with a depth of 2 (one hidden layer and one output layer), but the number of neurons $k$ in the hidden layer, i.e., the width of the neural network, must be allowed to increase without bound for arbitrary approximation. A recent result by Z. Lu and collaborators [16] provides a comparable denseness result for width-bound ("deep") networks, where the maximum number of neurons per layer is fixed, but the number of layers must be allowed to increase freely.

However, these deterministic results do not address the performance of neural networks trained from data. In particular, they do not weigh directly on the issue of consistency, i.e., on the stochastic convergence of the error of the trained classifier to the optimal error as sample size increases to infinity. To bear out the promise of distribution-free classification, consistency has to be universal; i.e., it must hold under any feature–label distribution. The few universal consistency results of which we are aware apply to shallow neural networks and make unrealistic demands on training, such as the requirement of training error or $L_1$-error minimization (for which gradient descent cannot be applied); e.g., see Theorems 30.7 and 30.9 in [17], respectively.

It turns out that even universal consistency is not enough. It was shown by L. Devroye and collaborators [17, Th. 7.2] hat for any universally consistent classification rule, a feature–label distribution exists such that convergence of the classification error to the optimal error is as slow as desired. In other words, universal consistency can say nothing about the problem of selecting a good classifier using finite training data: under no assumptions about the feature–label distribution, any classification algorithm can be arbitrarily bad. The situation changes if assumptions are made about the feature–label distribution. For example, it was shown by N. Glick [18, Th. A] that the difference between the expected and optimal classification error rates converges exponentially fast to zero in discrete histogram classification, with a rate that depends on the feature–label distribution: the more separable the classes are under the feature–label distribution (in a precise sense), the faster the rate of convergence is guaranteed to be.

Bayesian deep learning provides an alternative to deterministic neural networks and has attracted significant attention in recent years [19], [20]. The classical approach to Bayesian neural networks has been based on placing prior distributions on the weights of the network [21], [22]. In [23], an alternative approach to Bayesian deep learning was developed where dropout training of deep neural networks was formulated as approximate Bayesian inference in deep Gaussian processes. These are welcome developments. However, we point out that these approaches to Bayesian deep learning are not probabilistically related to the feature–label distribution, nor do they even require a feature–label distribution. The mechanism has a statistical dimension relative to the prior distribution, but the output is not necessarily statistically related to nature, only to the actual data. If there is an underlying model distribution representing scientific knowledge, then the issue arises as to the connection between it and the prior distribution. In the classical Bayesian approach, there is none, the priors being chosen ad hoc, perhaps according to some general information–theoretical criteria. This disconnect has been referred to as a "scientific gap" in [24].

In that reference, an alternative approach was proposed, where the model distribution is treated as uncertain, with the uncertainty occurring in its parameters. This model uncertainty is propagated to uncertainty in the classifier or regressor, and so the latter uncertainty is directly related to the underlying model.

## Conclusions

In this article, we have attempted to make the case that the success of machine learning in science and engineering depends essentially on the use of rigorous probabilistic modeling and statistics. We discussed the problem in the context of general machine learning and then looked more closely at the currently important topic of deep neural networks.

In our view, the issues discussed here concern not only the theoretician or scientist but also affect the practitioner, because practical application requires consistency with the demands of the theory. This view is expressed this way by Y. Gal and Z. Ghahramani in [23]: "Model uncertainty is indispensable for the deep learning practitioner as well," whereas N. Wiener put it this way in the context of biology [25]: "The physiologist need not be able to prove a certain mathematical theorem, but he must be able to grasp its physiological significance and tell the mathematician for what he should look for."

In the last few decades, however, there has been open opposition to the use of probability and statistics in predictive modeling. For example, this can be observed in the well-known polemic by L. Breiman [26], in which he writes: "The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems."

There is some evidence that this might be due to the lack of widespread mathematical literacy in the research community at large. For instance, J. Simon writes [27], "In the mid-1960's, I noticed that most graduate students—among them many who had had several advanced courses in statistics—were unable to apply statistical methods correctly in their social science research. I sympathized with them. Even many experts are unable to understand intuitively the formal mathematical approach to the subject. Clearly, we need a method free of the formulas that bewilder almost everyone."

The authors have pondered on such issues elsewhere [28]–[31]; others have done so as well [11], [32]–[35]. The crucial question is: Do we want knowledge about the real world in the sense of modern engineering and science, or do we merely want knowledge of specific events, the latter being more understandable and requiring simpler mathematics? Before answering the question, one should consider the enormous benefits that we, as modern engineers, have accrued from the probabilistic-statistical approach, beginning with the Wiener–Kolmogorov theory of linear systems in the 1930s and flowing forward in the development of optimal filtering, stochastic control, statistical signal processing, and information theory. Deep thought should be given as to whether abandoning this epistemology is desirable.

## Authors

***Ulisses M. Braga-Neto*** (ulisses@ece.tamu.edu) received his Ph.D. degree in electrical and computer engineering from The Johns Hopkins University, Baltimore, Maryland. He is currently a professor in the Department of Electrical and Computer Engineering at Texas A&M University, College Station. His research interests include pattern recognition, machine learning, and statistical signal processing. He is the author of the recent textbook, *Fundamentals of Pattern Recognition and Machine Learning* (Springer, 2020) and coauthor of *Error Estimation for Pattern Recognition* (IEEE-Wiley, 2015) with Edward Dougherty. He received the National Science Foundation CAREER Award for his work in this area. He is a Senior Member of the IEEE.

***Edward R. Dougherty*** (edward@ece.tamu.edu) received his M.S. degree in computer science from Stevens Institute of Technology, Hoboken, New Jersey and his Ph.D. degree in mathematics from the Rutgers University, Camden, New Jersey. He is a distinguished professor with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, where he holds the R.M. Kennedy '26 Chair in electrical engineering and is the scientific director of the Center for Bioinformatics and Genomic Systems Engineering. He was awarded the Doctor Honoris Causa by the Tampere University of Technology, Finland, the SPIE President's Award, and served as the editor of the Society for Optical Engineering/Society for Imaging Science and Technology *Journal of Electronic Imaging*. At Texas A&M University, he received the Association of Former Students Distinguished Achievement Award in Research. He is a Fellow of the IEEE and SPIE.

## References

[1] C. F. Gauss, *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium*, vol. 7. Hamburg, Germany: Perthes et Besser, 1809.

[2] S. M. Stigler, "Gauss and the invention of least squares," *Ann. Statist.*, vol. 9, no. 3, pp. 465–474, 1981. doi: 10.1214/aos/1176345451.

[3] C. F. Gauss, *Theoria Combinationis Observationum Erroribus Minimis Obnoxiae*, vol. 1. Göttingen, Germany: H. Dieterich, 1823.

[4] R. L. Plackett, "A historical note on the method of least squares," *Biometrika*, vol. 36, nos. 3–4, pp. 458–460, 1949. doi: 10.2307/2332682.

[5] D. Diderot, "Bas (bonneterie—)," in *Encyclopédie, ou Dictionnaire raisonné des sciences, des arts et des métiers, par une Société de Gens de lettres*, Autumn 2017 ed., R. Morrissey and G. Roe, Eds. Chicago: Univ. Chicago, ARTFL Encyclopédie Project, 2017, p. 2:98. [Online]. Available: https://encyclopedie.uchicago.edu/

[6] J. Stalnaker, *The Unfinished Enlightenment: Description in the Age of the Encyclopedia*. Ithaca, NY: Cornell Univ. Press, 2010.

[7] W. Barrett, *The Illusion of Technique: A Search for Meaning in a Technological Civilization*. Garden City, NY: Anchor Press, 1978.

[8] B. Efron, "Computers and the theory of statistics: Thinking the unthinkable," *SIAM Rev.*, vol. 21, no. 4, pp. 460–480, 1979. doi: 10.1137/1021092.

[9] M. J. Lorenzo, *Endless Loop: The History of the BASIC Programming Language*. Scotts Valley, CA: CreateSpace Independent Publishing Platform, 2017.

[10] U. M. Braga-Neto and E. R. Dougherty, *Error Estimation for Pattern Recognition*. Hoboken, NJ: Wiley, 2015.

[11] F. Mazzocchi, "Could big data be the end of theory in science?" *EMBO Rep.*, vol. 16, no. 10, pp. 1250–1255, 2015. doi: 10.15252/embr.201541001.

[12] U. Braga-Neto, A. Zollanvari, and E. R. Dougherty, "Cross-validation under separate sampling: Strong bias and how to correct it," *Bioinformatics*, vol. 30, no. 23, pp. 3349–3355, 2014. doi: 10.1093/bioinformatics/btu527.

[13] D. Hume, *An Enquiry Concerning Human Understanding: A Critical Edition*, vol. 3. Oxford, U.K.: Oxford Univ. Press, 2000.

[14] H. Jeffreys, *Scientific Inference*, 3rd ed. Cambridge, U.K.: Cambridge Univ. Press, 1973.

[15] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals Syst.*, vol. 2, no. 4, pp. 303–314, 1989. doi: 10.1007/BF02551274.

[16] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proc. 31st Int. Conf. Advances Neural Information Processing Systems*, 2017, pp. 6231–6239. doi: 10.5555/3295222.3295371.

[17] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[18] N. Glick, "Sample-based multinomial classification," *Biometrics*, vol. 29, no. 2, pp. 241–256, 1973. doi: 10.2307/2529389.

[19] H. Wang and D.-Y. Yeung, "Towards Bayesian deep learning: A framework and some existing methods," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3395–3408, 2016. doi: 10.1109/TKDE.2016.2606428.

[20] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. 31st Int. Conf. Advances Neural Information Processing Systems*, 2017, pp. 5574–5584. doi: 10.5555/3295222.3295309.

[21] D. J. MacKay, "A practical Bayesian framework for backpropagation networks," *Neural Comput.*, vol. 4, no. 3, pp. 448–472, 1992. doi: 10.1162/neco.1992.4.3.448.

[22] A. Graves, "Practical variational inference for neural networks," in *Proc. 24th Int. Conf. Advances Neural Information Processing Systems*, 2011, pp. 2348–2356. doi: 10.5555/2986459.2986721.

[23] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. 33rd Int. Conf. Machine Learning*, 2016, pp. 1050–1059. doi: 10.5555/3045390.3045502.

[24] X. Qian and E. R. Dougherty, "Bayesian regression with network prior: Optimal Bayesian filtering perspective," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6243–6253, 2016. doi: 10.1109/TSP.2016.2605072.

[25] N. Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, MA: MIT Press, 1948.

[26] L. Breiman, "Statistical modeling: The two cultures (with comments and a rejoinder by the author)," *Stat. Sci.*, vol. 16, no. 3, pp. 199–231, 2001. doi: 10.1214/ss/1009213726.

[27] J. Simon, *Resampling: The New Statistics*. Arlington, VA: Resampling Stats, 1997.

[28] U. Braga-Neto, "Small-sample error estimation: Mythology versus mathematics," in *Proc. Mathematical Methods Pattern and Image Analysis*, 2005, vol. 5916, p. 59160V. doi: 10.1117/12.619331.

[29] E. R. Dougherty and U. Braga-Neto, "Epistemology of computational biology: Mathematical models and experimental prediction as the basis of their validity," *J. Biol. Syst.*, vol. 18, no. 14, pp. 65–90, 2006. doi: 10.1142/S0218339006001726.

[30] E. R. Dougherty and M. L. Bittner, *Epistemology of the Cell: A Systems Perspective on Biological Knowledge*, vol. 35. Hoboken, NJ: Wiley, 2011.

[31] P. V. Coveney, E. R. Dougherty, and R. R. Highfield, "Big data need big theory too," *Phil. Trans. Roy. Soc. A, Math. Phys. Eng. Sci.*, vol. 374, p. 20160153, Nov. 2016. doi: 10.1098/rsta.2016.0153.

[32] D. J. Glass and N. Hall, "A brief history of the hypothesis," *Cell*, vol. 134, no. 3, pp. 378–381, 2008. doi: 10.1016/j.cell.2008.07.033.

[33] M. Frické, "Big data and its epistemology," *J. Assoc. Inform. Sci. Technol.*, vol. 66, no. 4, pp. 651–661, 2015. doi: 10.1002/asi.23212.

[34] J. P. Ioannidis, "Why most published research findings are false," *PLoS Med.*, vol. 2, no. 8, p. e124, 2005. doi: 10.1371/journal.pmed.0020124.

[35] W. Lee Kraus, "Editorial: Would you like a hypothesis with those data? Omics and the age of discovery science," *Mol. Endocrinol.*, vol. 29, no. 11, pp. 1531–1534, 2015. doi: 10.1210/me.2015-1253.