CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: http://amstat.tandfonline.com/loi/ucha20

Machine Learning, a Probabilistic Perspective

Christian Robert

To cite this article: Christian Robert (2014) *Machine Learning, a Probabilistic Perspective*, CHANCE, 27:2, 62-63, DOI: 10.1080/09332480.2014.914768

To link to this article: https://doi.org/10.1080/09332480.2014.914768

Published online: 23 Apr 2014.

The book also introduces the notion of a Bayesian likelihood function (p.228), which "differs slightly from that in classical statistics."The only difference I can spot is in the interpretation: Both functions of (θ, x) are numerically the same. Overall, the chapter on Bayesian inference does not spend much time on prior specification. There is a section on conjugate priors that does not mention picking the hyperparameters. While improper priors are introduced as limits of proper priors and as conveying "the least amount of information about [the parameters]" (p.236), the difficulty in using improper priors for hypothesis testing is not mentioned. Both Chib's method and the Savage-Dickey density ratio are suggested for the approximation of marginal likelihoods.

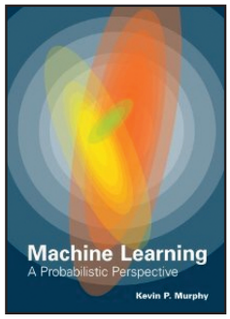
"It may be too time-consuming, or simply not feasible, to obtain such replications. An alternative is to resample the original data." (p.205)

Somewhat in agreement with my own R course, the Monte Carlo chapter starts with the empirical cdf and bootstrap. I found the explanation of the empirical cdf (p.198) confusing in that the authors distinguish between a deterministic version and a stochastic version of the empirical cdf. In a statistical setting, there is no deterministic version! The chapter naturally includes a mention of the Kolmogorov Smirnov test, as well as a short section on density estimation, which recommends using the "theta KDE" method of Botev, Grotowski, and Kroese (2010) without defining it. The bootstrap is introduced through the empirical cdf, except for the above quote that I find puzzling: Observing a sample from an unknown cdf prohibits replicating iid simulations from this cdf. As in my course, bootstrapping linear regression is given as an example. However, the authors suggest resampling the pairs (xi, yi), while I tell my students to resample from the estimated residuals, since those are the iid variables. The MCMC section introduces Metropolis-Hastings and Gibbs sampling algorithms (that are used in the later chapters), but fails to warn about the calibration for the former.

In conclusion, and despite the rather gruff tone of my review, this book is a fairly decent and quick introduction to the practice of statistical analysis that manages to reach the complexity of models like stochastic volatility in a few hundred pages. It could thus be helpful when teaching an undergraduate statistics class for nonspecialists.

Machine Learning, a Probabilistic Perspective

Kevin P. Murphy



Hardcover: 1104 pages
Year: 2012
Publisher: The MIT Press
ISBN-13: 978-0262018029

I have to admit the rather embarrassing fact that *Machine Learning, a Probabilistic Perspective* is the first machine learning book I have read in full detail. This is a massive book with close to 1,100 pages, so I hesitated taking it with me to Warwick for a week. It is also massive in its contents, as it covers most (all?) of what I call statistics (but visibly corresponds to machine learning, as well), with a Bayesian bent most of the time (which is the secret meaning of probabilistic in the title).

"... [W]e define machine learning as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!)." (p.1)

Apart from the introduction—which I deem rather confusing for not dwelling on the nature of errors and randomness and on the reason for using probabilistic models (since they are all wrong) and charming for including a picture of the author's family as an illustration of face recognition algorithms—I cannot say I found the book more lacking in foundations or in the breadth of methods and concepts it covers than a "standard" statistics book. In short, this is a perfectly acceptable statistics book! Furthermore, it has a relevant and comprehensive selection of references (sometimes favoring "machine learning" references over "statistics" references). Even the vocabulary seems pretty standard to me.

All this makes me wonder why we distinguish between the two domains, following Larry Wasserman's view that the difference is mostly in the eye of the beholder (i.e., in which department one teaches). This was my perspective before I read the book, but it comforted me even further. And the author agrees as well ("The probabilistic approach to machine learning is closely related to the field of statistics, but differs slightly in terms of its emphasis and terminology," p.1). Let us unite!

Chapter 3 on Bayesian updating or learning (a most appropriate term) for discrete data is well done, if a bit stretched (which is easy to do with 1,000 pages left). I like the remark (Section 3.5.3) about the log-sum-exp trick. While lengthy, the chapter (Chap. 4) on Gaussian models has the appeal of introducing LDA. The true chapter about Bayesian statistics (Chap. 5) only comes later, which seems a wee bit late to me, but it mentions the (overlooked) paper by Druilhet and Marin (2007) about the dependence of the MAP estimator on the dominating measure. The Bayesian chapter covers the Bayesian interpretation of false discovery rates and decision theory (shared with the following chapter on frequentist statistics). This later chapter also covers the pathological features of p-values. The chapter on regression has a paragraph on the g-prior and its extensions (p.238). There are chapters on DAGs, mixture models, EM (which even mentions the early MCEM of Celeux and Diebolt, 1980!), factor and principal component analyses, Gaussian processes, CART models, HMMs and state-space models, MFRs, variational Bayes, belief and expectation propagations, and more. Most of the methods are implemented within a MATLAB package called PMTK (probabilistic modeling toolkit) that I did not check (just because it is MATLAB!).

There are two (late!) chapters dedicated to simulation methods, Monte Carlo Inference (Chap. 23), and MCMC Inference (Chap.24). I am somewhat unhappy with the label "inference" in those titles, as those are simulation methods. They cover the basics and more, including particle filters to some extent (but missing some of the most recent stuff such as Del Moral, Doucet, and Jasra, 2006, or Andrieu, Doucet, and Hollenstein, 2010). When introducing the Metropolis-Hastings algorithm, the author states the condition that the union of the supports of the proposal should include the support of the target, but this is a rather formal condition as the Markov chain may still fail to be irreducible in that case. My overall feeling is that too much is introduced in too little space, potentially confusing the student. [See the half-page Section 24.3.7 (p.855) on reversible jump MCMC. Or the other half-page on Hamiltonian MCMC (p.868).] However, an interesting entry is the study of the performances of the original Gibbs sampler of Geman and Geman (1984), which started the field (to some extent). It states that, unless the hyperparameters are extremely well calibrated, the Gibbs sampler suggested therein fails to produce a useful segmentation algorithm!

The section on convergence diagnoses is rather limited and refers to oldish methods, rather than suggesting a multiple-chain empirical exploratory approach. Similarly, there is only one page (p.872) of introduction to marginal likelihood approximation techniques, half of which is wasted on the harmonic mean "worst Monte Carlo method ever." And the other half is spent on criticizing Besag's candidate method exploited by Chib (1995).

Now, a wee bit more into detailed nitpicking… First, the mathematical rigor is not always "there" and the handling of Dirac masses and conditionals and big-Oh (Exercise 3.20) is too hand waving for my taste (see p.39 for an example). I also dislike the notion of the multinoulli distribution (p.35), first because it is a poor pun on Bernoulli's name and second because sufficiency makes this distribution somewhat irrelevant when compared to the multinomial distribution. Although the book covers the dangers and shortcomings of MAP estimators fairly well in Section 5.2.1.3 (p.150), this method remains the default solution. As a marginalia, Monte Carlo is not "a city in Europe known for its plush gambling casinos" but the district of the city-state Monaco where the casino stands. And it writes "Monte-Carlo" in the original. The approximation of π by Monte Carlo is the one suggested by Buffon, but it would have been nice to know the number of iterations (p.54).

The book unnecessarily and vaguely refers to Taleb about the black swan paradox (p.77). The first introduction of Bayesian tests is to use the HPD interval and check whether the null value is inside with a prosecutor's fallacy in conclusion (p.137). BIC, then AIC, are introduced (p.162), and the reader remains uncertain about which one to use, if any. The fact that the MLE and posterior mean differ (p.165) is not a sign of informativeness in the prior. The processing of the label-switching problem for mixtures (p.841) is confusing in that the inference problem (invariance by permutation that prohibits using posterior means) is compounded by the simulation problem (failing to observe this behavior in simulations).

The Rao-Blackwellization Theorem (p.841) does not readily apply to cases other than two-stage Gibbs sampling, but this is not clear from the text. The adaptive MCMC amcmc package of Jeff Rosenthal is not mentioned (because it is in R?). The proof of detailed balance (pp.854–855) should take a line. Having so many references (35 pages) is both a bonus and a nuisance in a textbook, in which students dislike the repeated occurrence of "see so-and-so." I also dislike references being given within parentheses at all times, as in "See (Doucet et al. 2001) for details." And definitely the least important remark given my own record in the matter, the quotes at the beginning are not particularly novel or relevant: The book could do without them. (Same thing for the "no free lunch theorem," which is not particularly helpful as presented.)

In conclusion, *Machine Learning, a Probabilistic Perspective* offers a fairly wide, unifying, and comprehensive perspective on the field of statistics—aka machine learning—that can be used as the textbook in a master's program where this is the only course on statistics (aka machine learning). Having not thoroughly read other machine learning books, I cannot judge how innovative it is. The beginning is trying to build the intuition of what the book is about before introducing the models. Just not my way of proceeding, but mostly a matter of taste and maybe audience. The computational aspects are not treated in enough depth for my taste and courses, but there are excellent books on those aspects. The Bayesian thread sometimes runs a bit thin, but it remains a thread nonetheless throughout the book, thus a nice textbook for the appropriate course and a reference for many.