AStA Wirtsch Sozialstat Arch (2019) 13:269–278
https://doi.org/10.1007/s11943-019-00259-3

ORIGINALVERÖFFENTLICHUNG

Statistical literacy for classification under risk: an educational perspective

Laura Martignon · Kathryn Laskey

Received: 20 October 2019 / Accepted: 25 October 2019 / Published online: 21 November 2019

Abstract After a brief description of the four components of risk literacy and the tools for analyzing risky situations, decision strategies are introduced, These rules, which satisfy tenets of Bounded Rationality, are called fast and frugal trees. Fast and frugal trees serve as efficient heuristics for decision under risk. We describe the construction of fast and frugal trees and compare their robustness for prediction under risk with that of Bayesian networks. In particular, we analyze situations of risky decisions in the medical domain. We show that the performance of fast and frugal trees does not fall too far behind that of the more complex Bayesian networks.

Keywords Statistical literacy · Risk literacy · Components · Uncertainty · Decision trees

JEL-Classification A21 · A22 · C00 · I21

1 Introduction

Our quality of life as individuals and the functioning of our society depend on our ability to understand risks and act appropriately in response. Recognizing this need, the education community has placed increasing emphasis on education for risk literacy. Effective education for risk literacy draws on the theory of rational choice under uncertainty, behavioral research on how people perceive and respond

L. Martignon
Ludwigsburg University of Education, Ludwigsburg, Germany
E-Mail: martignon@ph-ludwigsburg.de

K. Laskey (✉)
George Mason University, Washington, USA
E-Mail: klaskey@gmu.edu

to risks, and educational research on how youngsters learn about risk. The aim is to develop educational approaches to guide the development of risk literacy and inculcate effective strategies for coping with risk.

Probability and decision theory emerged during the Enlightenment as a model of rational belief and decision-making in the presence of risk (Daston 1995). According to Laplace, "the theory of probabilities is nothing but common sense reduced to calculus; it enables us to appreciate with exactness that which accurate minds feel with a sort of instinct for which often they are unable to account," (Laplace 1812). That is, probability formalizes the intuitions of the enlightened man. In the nineteenth century, with the rise of positivism and the quest for objectivity in science, probability fell out of favor as a model of rational thought. The mid-twentieth century brought a resurgence, with the introduction of personal probability (Savage 1954; de Finetti 1934). The personalist theory explicitly allows rational individuals to disagree on the probability of an event. Its inherent subjectivity brought skepticism and even hostility from adherents of an objective view of science. Another line of attack against probability came from the cognitive and behavioral sciences, with a flurry of research demonstrating systematic ways in which human reasoning fails to conform to the probability calculus (Kahneman et al. 1982).

Some researchers (Gigerenzer et al. 1999) argued that systematic deviations of human reasoning from probability are rational in an ecological sense. That is, humans have evolved a toolbox of "fast and frugal" strategies to draw inferences and make decisions in an environment of bounded cognitive resources and limited time. These ecologically rational strategies give results that are nearly as good as optimal but infeasible probabilistic methods. Arguments for ecological rationality are supported by comparing the output of computer models inspired by human reasoning with that of explicitly probabilistic computer models. For many problems we face in everyday life, cognitively inspired "fast and frugal" strategies perform nearly as well and, often better than, much more computationally intensive probabilistic approaches.

Taking a somewhat different approach, the field of decision analysis has focused on developing "cognitive tools" (von Winterfeldt and Edwards 1986) to help people come closer to the decision theoretic norm. Decision support tools informed by cognitive research help people to construct and reason with models that explore the logical consequences of their intuitive judgments, using the computer to perform probability calculations that exceed their intuitive capacity.

In our view, these streams of research are complementary, and together suggest promising directions for education in risk literacy. Acknowledging the systematic deviations of human thinking from the probability calculus, we seek to exploit naturalistic human reasoning and cognitive development to develop pedagogical strategies that capitalize on human strengths while overcoming the weaknesses of unaided reasoning. In this paper, we focus on a simple but commonly occurring class of problems—using evidential cues to classify a situation into one of two categories. We examine "fast and frugal" heuristics proposed in the literature, and show that their performance compares favorably to more computationally intensive methods from statistics and machine learning. We discuss the role of normative theory in justifying the performance of the simpler models, and the educational value of

instilling a deep understanding of both the normative approaches and the simpler heuristics. We conclude with remarks on implications for mathematics curriculum.

2 Risk literacy

Risk literacy refers to *a person's ability to evaluate and understand risk,* for the purpose of informed decision making. In some sense, our ability to estimate risk depends on external factors like the design of risk communications (e.g., simple visual aids can promote or bias risk comprehension). As has been shown, one's practical understanding of mathematics tends to be the strongest single predictor of risk literacy and general decision-making skill.

Risk Literacy consists of four basic components (see Fig. 1, below):

a) Identifying risk and uncertainty
b) Analyzing and modeling the uncertain or risky situation
c) Pondering and comparing alternatives
d) Making decisions and acting.

The first component refers to the ability to identify risks and uncertainties properly. It is well known that humans tend to overestimate certain dangers and underestimate others, due to evolutionary biases and the effect of vivid news reports. Developing sound methods for estimating the impact of phenomena on our life requires plain statistical literacy (Martignon and Hoffrage 2019). The second component is one that requires elementary probabilistic training in classifying situations as dangerous or innocuous. Some tools for acquiring this component will be discussed in the next session. The third component also relies on mathematical skills for dealing with trade-off situations. Most human decisions are between alternatives, each one with certain advantages and disadvantages. Risk literacy requires an ability to evaluate the relative strength of advantages versus disadvantages. The fourth component has as a precondition being acquainted with the irreversibility of decisions: one has to cross the Rubicon and act. Here psychological conditions play important roles. Some people tend to make sudden, intuitive decisions, without even pondering about consequences, i.e., without passing through the third component.

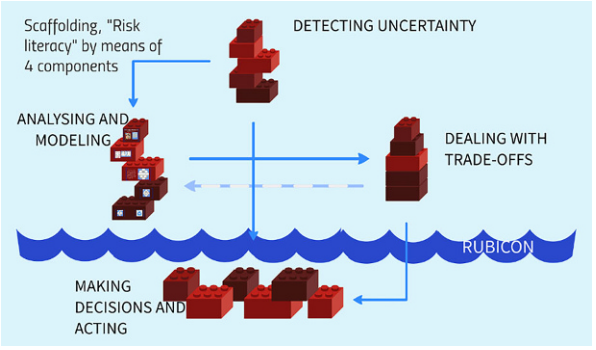

Fig. 1 Scaffolding Risk Literacy by means of four components. (Adapted from Martignon and Hoffrage 2019)

In this paper, we emphasize the second component and present mathematical tools, which may be acquired even in school, to support the ability to analyze and model risks.

Although the time allotted to probabilistic training in school varies across countries, there is consensus that the basic elements of probabilities, including Bayesian reasoning and expected values should be taught in secondary school. In some countries pre-service teachers for elementary school are now being trained in rudimentary probability based on icon arrays and simple proportions.

3 Analyzing and modelling: classification methods

Classification problems pervade everyday life. As an exemplar, we consider a doctor examining a patient. The doctor typically is presented with a set of cues: symptoms reported by the patient; background information such as age, weight and gender; test results; physician observations. The task faced by the physician is to arrive at a diagnosis.

For illustrative purposes, we simplify the problem as follows. First, we suppose the doctor is answering a simple yes/no question: does the patient have a specific condition under consideration? Second, we assume that the input is a set of binary cues, e.g., a symptom is or is not present; a test reading is high or low. Finally, we assume that the physician reports a single answer: yes (condition present) or no (condition absent), with no opportunity to hedge the result. We are interested in whether the answer given by the physician is correct or incorrect. For this simplified set of problems, we examine the performance of two different "fast and frugal" classification tree methods and compare them with several methods drawn from the statistics and machine learning literature.

3.1 The normative approach

The normative approach to this problem is known as Bayesian inference, also called inverse probability. The physician starts with a *prior* probability $P(D)$ that a disease D is present. The physician observes evidence E in the form of a set of symptoms, background information, tests, and other observations. The physician assesses the probability $P(E|D)$ that the evidence would occur if the disease is present and the probability $P(E|\bar{D})$ that the evidence would occur if the disease is not present. The physician then uses *Bayes theorem* to find the probability $P(D|E)$ that the disease is present given the observed evidence, also called the *posterior* probability of the disease given the evidence:

$$P(D|E) = \frac{P(E|D)P(D)}{P(E|D)P(D) + P(E|\bar{D})P(\bar{D})} \tag{1}$$

In the medical domain the evidence for a disease is usually provided by a symptom or the result of a test. If a test T, like, say a mammogram shows a positive result T_+ the doctor tends to believe that the patient has the disease. Bayes' Theorem helps

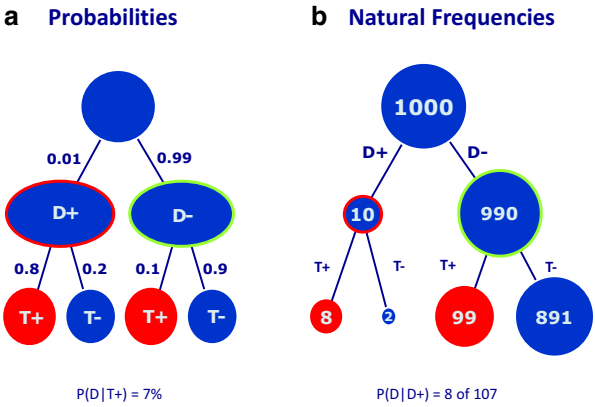

Fig. 2 Two trees for representing data in a Bayesian task: **a** data are probabilistic. **b** data are presented in natural frequencies

estimating the probability that the patient has the disease given that the test is positive. An important finding of cognitive psychology is that, whereas people have trouble estimating this probability using Eq. 1, they are at ease when provided with so called "natural frequencies" (e.g., Gigerenzer and Hoffrage 1995).

Fig. 2 contrasts the natural frequency representation (right-hand side) with the less intuitive probability representation (left-hand side). To reason with natural frequencies, the doctor imagines a population of fictitious people, say 1000 of them. She divides them into those who do and do not have the disease. For example, if the disease is present in only 1% of the population, she would partition her imaginary 1000 patients into 10 who have the disease and 990 who do not. Of those who have the disease, suppose 80% will test positive. In this case, the doctor imagines that 8 of the 100 ill patients will test positive and 2 will test negative. Now, suppose that 90% of those who do not have the disease will have a negative test result. In our doctor's imaginary population, this works out to 99 well patients who test positive and 891 who test negative. It has been demonstrated that people, especially those with little formal training in probability, find the natural frequency representation more intuitive and are able to reason more accurately with natural frequencies than with probabilities.

For a single symptom, using natural frequencies to calculate the probability that the patient has the disease given that the test is positive is a straightforward computation that can easily be performed with pencil and paper or a calculator. For a large number of symptoms, the general case is quite challenging. If there are n evidence items, one must consider the probability of all 2^n possible combinations given both presence and absence of the disease, for a total of 2^{n+1} probabilities. For 10 symptoms, one must consider $2^{11} = 2048$ probabilities, a daunting challenge.

3.2 Naïve Bayes

A commonly applied simplification to the fully general problem is to assume that the evidence items are independent conditional on presence or absence of the disease. In this case, for symptom E_k, we need to assess only two probabilities: and $P(E_k^H \mid D)$

and $P(E_k^H|\bar{D})$, the probability that the evidence is in its "high" state given that the disease is present or absent. By the laws of probability, the evidence is in its "low" state with probability $P(E_k^L|D) = 1 - P(E_k^H|D)$ if the disease is present and $P(E_k^L|\bar{D}) = 1 - P(E_k^H|\bar{D})$ if the disease is absent. Substituting into (1), we obtain the equation:

$$\left(D\,E_1^{j_1}, ..., E_n^{j_n}\right) = \frac{P\,(D)\prod_k P\left(E_k^{j_k} D\right)}{P\,(D)\prod_k P\left(E_k^{j_k} D\right) + P\left(\bar{D}\right)\prod_k P\left(E_k^{\bar{j}_k} \bar{D}\right)} \tag{2}$$

where j_k denotes either the "high" or "low" state of E_k, and $\bar{j}_k$ denotes the opposite state. The required probabilities may be assessed intuitively by an expert, or estimated from data.

The naïve Bayes model has drastically reduced the number of probability assessments from 2^{n+1} to $2n+1$, a reduction from 2048 to 11 in the case of ten symptoms. Still, the calculation (2) is beyond the reach of intuitive judgment and is very cumbersome with pencil and paper; a computer is all but required. This simplification is valid under the assumption that evidence items are conditionally independent given presence or absence of the disease. When this assumption is not met, naïve Bayes can give misleading results. Experience has shown that as long as dependencies among evidence items are not too great, naïve Bayes tends to perform very well. Although the model is beyond the reach of unaided human judgment, it is among the simplest of Bayesian models, can be easily applied if a computer is available, and is relatively robust to minor violations of the independence assumptions.

3.3 Logistic regression

Logistic regression dispenses with the attempt to perform Bayesian reasoning from prior to posterior probability in favor of modeling the posterior probability $P(D|E)$ directly. Again, a fully general model of $P(D|E)$ is infeasible for all but the smallest problems—one would need a probability for each of the 2^n possible evidence combinations. Instead, a simplified model is assumed. We assign a value of 0 to the "low" state of E_k and a value of 1 to the "high" state of E_k. The logistic regression equation is:

$$\frac{P\,(D|E_1,...,E_n)}{1-P\,(D|E_1,...,E_n)} = e^{\beta_0 + \sum_k \beta_k E_k} \tag{3}$$

where the parameters are typically estimated from data.

3.4 Fast and frugal classification trees

Trees are one of the earliest approaches to classification. At the root of the tree, one asks a question, e.g., "Is evidence E_k in its high or its low state?" One proceeds along one branch if the answer is "high" and the other if the answer is "low." One

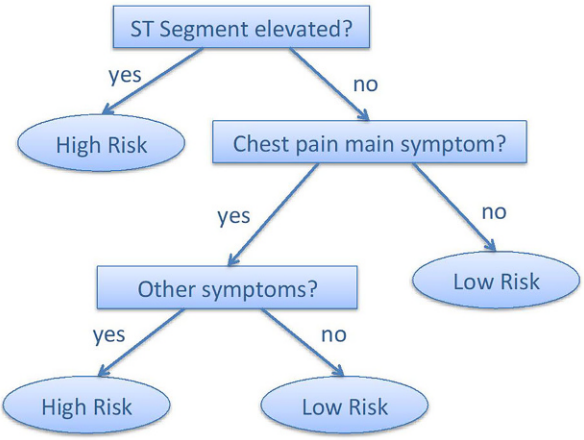

Fig. 3 Green and Mehr's fast and frugal tree for heart disease risk assessment

continues in this way, branching at each node of the tree, until arriving at a leaf node. Each leaf of the tree gives a diagnosis.

Fig. 3 shows an example of a classification tree, taken from (Green and Mehr 1997), for classifying patients as high or low risk for heart disease. This tree is "fast and frugal" by the definition given in (Martignon et al. 2012)—at each node of the tree, the choice is either to stop with a diagnosis or to continue to the next level. A fast and frugal classification tree provides a very simple procedure for performing the classification task. Green and Mehr found that diagnosis according to this fast and frugal tree was more accurate than both the physicians' clinical judgment and logistic regression.

4 Comparison of methods

We compared the performance of five classification methods on eleven data sets taken from the medical and veterinary domains (Laskey and Martignon 2014). Most of the data sets were taken from the UC Irvine Machine Learning Repository (Bache and Lichman 2013). Each data set consisted of a class variable and between five and 22 evidence variables. The number of observations ranged from a minimum of 62 to a maximum of 768. Many of the evidence variables were two-valued; others were numerical. The numerical variables were dichotomized by assigning values larger than the median to the "high" category and values less than or equal to the median to the "low" category. A classifier was estimated for each method on each data set by selecting a random subset of the data as a training sample, dichotomizing continuous variables (if any), applying the classifier fitting method to the training sample, and then classifying each element of the remaining test sample. This process was repeated 1000 times for each classifier. This whole process was carried out for training samples of 15%, 50% and 90% of the data set. Thus, training samples ranged from a minimum of 9 observations to a maximum of 691 observations. Classifying

a different set of observations from the ones used to train the classifiers is standard practice in classification research to test the ability of the method to generalize to new data. The five classification methods are:

- *Naïve Bayes*: The prior probability $P(D)$ and the evidence distributions $P(E_k^H|D)$ and $P(E_k^H|\bar{D})$ were estimated using the Beta-Binomial conjugate prior method. This method estimates the probability of an event as $(r+1)/(m+2)$, where r is the number of previous trials on which the event occurred out of m total trials. This is a simple Bayesian estimation method. It has the advantage that it avoids estimating a probability as zero when there the event does not occur in the sample or 100% when the event occurs for every case in the sample. Each case was classified as "yes" if the posterior probability $P(D|E)$ was greater than 0.5 and "no" otherwise.
- *Logistic regression*: We used a standard logistic regression method to estimate the regression coefficients from (2). Each case was classified as "yes" if the posterior probability $P(D|E)$ was greater than 0.5 and "no" otherwise.
- *CART*: CART (Breiman 1984) is a method for building trees for classifying categorical variables or predicting numerical variables. It uses a collection of rules designed to maximize information gain from each split of the tree, with splits terminating at a leaf node when an additional split would yield no further information gain. CART trees are not necessarily fast and frugal.
- *Fast and frugal trees with Zig-Zag rule*: This method constructs the tree by using positive and negative *cue validities*. Positive validity is the proportion of cases with a positive outcome among all cases with a positive cue value. Negative validity is the proportion of cases with a negative outcome among all cases with a negative cue value. The Zig-Zag method alternates between "yes" and "no" exits at each

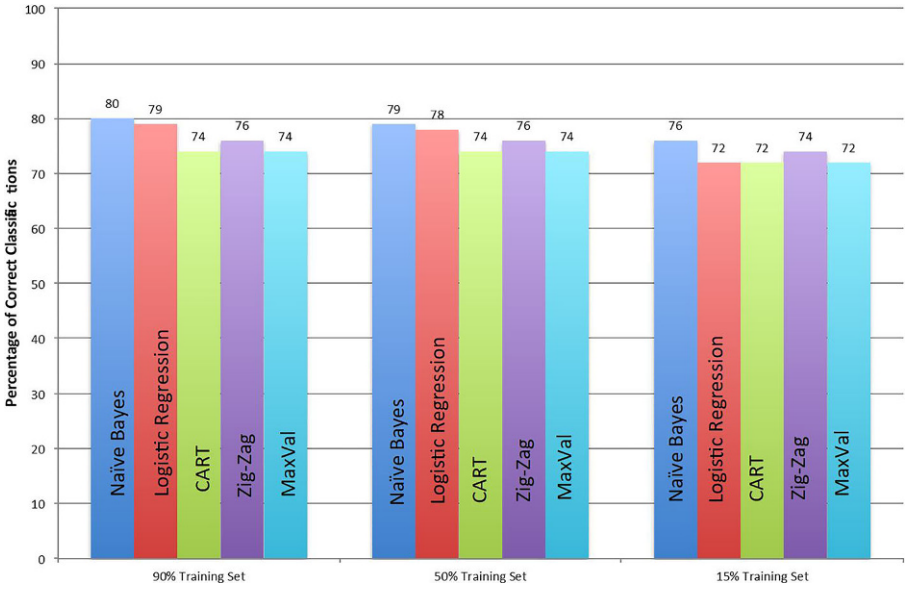

Fig. 4 Average Performance of Different Classifiers on Eleven Medical Data Sets

level, choosing according to the cue with the greatest positive (for "yes") or negative (for "no") validity among the cues not already chosen.

- *Fast and frugal trees with MaxVal rule*: This method also uses positive and negative cue validities. It begins by ranking the cues according to the higher of each cue's positive or negative validity. It then proceeds according to this ranking, applying the cues in order and exiting in the positive (negative) direction if the positive (negative) validity of the cue is higher. Ties in this process are broken randomly.

Our results for logistic regression and the three classification tree methods are taken from (Martignon et al. 2012); the naïve Bayes results were first presented by the authors at ICOTs 9. Unsurprisingly, accuracy of all methods increases as the training sample becomes larger. Performance of the five methods is remarkably similar. Naïve Bayes has the best performance across the board, living up to its reputation as a simple and robust benchmark. In fact, we tried other more complex Bayesian methods, but none performed better than naïve Bayes, most likely due to the relatively small (by machine learning standards) data sets. However, naïve Bayes performed only slightly better than the other methods. The fast and frugal trees are only slightly less accurate than the computationally more expensive naïve Bayes (Fig. 4).

5 Conclusion

Our results have important implications for risk education, especially for the debate over how best to respond to the well-established literature demonstrating that people's intuitive judgments do not live up to the Bayesian ideal. Fast and frugal trees are extremely simple computationally. Cue validities can be estimated using only counting and ratios. Once cue validities have been estimated, tree construction involves only a few simple rules. After the tree has been constructed, its use for classification involves only traversing the tree and answering one simple question at each node. This simple process yields a classifier almost as accurate as the Bayesian benchmark.

The most advanced operation required for constructing fast and frugal trees is estimating cue validities. It has been demonstrated that children as young as fourth grade can understand the concept of cue validity through enactive education approaches, manipulating towers of colored tinker cubes to represent the relationship between cues and outcomes (Martignon and Monti 2010). Children can apply their understanding to can answer questions on the validity of cues. Thus, even at a young age, children can acquire basic reasoning strategies for coping with risks, strategies that will serve them well as they reach adulthood.

Building on this foundation, students in secondary school can comprehend the concept of conditional independence, and are prepared to understand the naïve Bayes model as well as more complex Bayesian network models (Krauss et al. 2010). At this stage, students can evaluate trade-offs between computational cost and accuracy, and to choose an approach that balances these objectives appropriately for the sit-

uation. Students at the secondary level could perform studies of the kind described in this paper, comparing computation and accuracy of naïve Bayes with fast and frugal strategies. Thus, they would be able to conclude for themselves that fast and frugal approaches yield nearly the same accuracy as the Bayesian benchmark, while requiring far less computation.

References

Bache K, Lichman M (2013) UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml. Accessed 4 Apr 2013

Breiman L (1984) Classification and regression trees. Chapman & Hall, New York

Daston L (1995) Classical probability in the enlightenment. Princeton University Press, Princeton (Reprint edition)

De Finetti B (1934) Theory of probability: a critical introductory treatment, 2nd edn. Wiley, New York

Gigerenzer G, Hoffrage U (1995) How to improve Bayesian reasoning without instruction: frequency formats. Psychol Rev 102:684–704

Gigerenzer G, Todd P, The ABC Group (eds) (1999) Simple heuristics that make us smart. Oxford University Press, Oxford

Green L, Mehr DR (1997) What alters physicians' decisions to admit to the coronary care unit? J Fam Pract 45(3):219–226

Kahneman D, Slovic P, Tversky A (1982) Judgment under uncertainty: heuristics and biases. Cambridge University Press, Cambridge

Krauss S, Bruckmaier G, Martignon L (2010) Teaching young grownups how to use Bayesian networks (Presented at the ICOTS 8, Ljubljana, Slovenia)

Laplace PS (1812) Théorie analytique des probabilités. Ve. Courcier, Paris (http://archive.org/details/thorieanalytiqu01laplgoog)

Laskey K, Martignon L (2014) Comparing fast and frugal trees and Bayesian networks for risk assessment. In: Makar K (ed) Proceedings of the Ninth International Conference on Teaching Statistics. International Statistical Institute and International Association for Statistical Education, Flagstaff (http://iase-web.org/icots/9/proceedings/pdfs/ICOTS9_8I4_LASKEY.pdf)

Martignon L, Hoffrage U (2019) Wer wagt, gewinnt? Hogrefe, Göttingen

Martignon L, Monti M (2010) Conditions for risk assessment as a topic for probabilistic education (Presented at the ICOTS 8, Ljubljana, Slovenia)

Martignon LF, Katsikopoulos KV, Woike JK (2012) Naïve, fast, and frugal trees for classification. In: Todd PM, Gigerenzer G (eds) Ecological rationality: intelligence in the world. Oxford University Press, USA

Savage LJ (1954) The foundations of statistics. Wiley, New York

Von Winterfeldt D, Edwards W (1986) Decision analysis and behavioral research. Cambridge University Press, Melbourne