Other views and ideas

Dr. Wray Buntine
Heuristicrats Research, Inc.

wray@Heuristicrat.COM
http://www.Heuristicrat.COM/wray/

1678 Shattuck Avenue, Suite 310 • Berkeley, CA, 94709-1631
Tel: +1 (510) 845-5810 • Fax: +1 (510) 845-4405

Section in the tutorial at *Maximum Entropy and Bayesian Methods*, Sante Fe, New Mexico, June 31st, 1995.

Outline

- Justifications for probability
  - various axiomatic bases; are they believable?
- Quotable quotes
  - fallacies about probabilities routinely launch whole new fields of science
- Some paradoxes
  - Bayesian methods are consistent by construction, so most paradoxes are due to an error or slight of hand (e.g., see Jaynes, 80)
  - sometimes these errors can difficult to spot!
- Other communities dealing with uncertainty

  Claim: all "good" methods for dealing with uncertainty have some Bayesian counterpart
  - many listed interact with Bayesians, good WWW sites are listed
- Classical and applied statistics versus Bayesian
- Maximum entropy versus Bayesian

Justifications of probability

Axiom systems have been developed by many mapping coherency schemes to probability or utility:

Frequency axioms: Kolmogorov axiomatizes probability "as frequency" for arbitrary measure spaces.

Belief axioms: Cox axiomatizes probability "as belief" with axioms such as "events must be well-defined", "belief on events can be represented as a real number", etc. See Jaynes, 96.

Relative belief: qualitative, relative likelihood schemes from which quantitative probabilities are derived, see DeGroot, 70, BS94

Utility from probability and choice of action: Von Neumann and Morgenstern used bets together with coherent preference schemes, see DeGroot, 70.

Betting schemes; (Dutch books of de Finetti) where people without coherent probabilities can be bet against profitably, see BS94.

They all lead to the same conclusion: use probabilities and utilities!

Are these justifications believable?

- All these schemes implicitly assume
  - infinite introspection by the user (e.g., to compare endless series of bets or elicit prior on a complex space)
  - infinite computation (i.e., to compute the required expected values)
  - a single agent is performing the inference
- Gelman, Carlin, Stern and Rubin, '95 say:
  - "these considerations suggest that probabilities may be a reasonable approach"
  - "the ultimate proof is in the success of the applications"
- A notable MaxEnt Bayesian has said:
  - "the only valid reason for not using Bayesian methods is incompetence "
  - other reasons: resource/software/time/user-training constraints

Quotable quotes

(sources de-identified to protect the ignorant)

"Note that the general probability that a bird can fly may be irrelevant, because we are interested in the facts that influence our opinion about whether a particular bird can fly in a particular situation."

.....(thinks probabilities are frequencies and hence are "useless" in most realistic situations, this launched non-monotonic reasoning)

"Probabilities are virtually useless in medical applications, because the conclusions that one can draw from such probability values almost never justify the expense and inconvenience to the patient necessary to obtain them."

....(whereas, value of information calculations actually allow the most patient-friendly diagnosis, this launched certainty factors and other uncertainty-calculi for rule-based expert systems)

"In our view the raven [Hempel's] paradox and the grue [Goodman's] paradox are not mere problems to be solved by some refined syntactic account of induction [probabilistic methods] but rather are symptoms of the fundamental inadequacy of such accounts. ... we will attempt to resolve these paradoxes within our pragmatic framework."

....(these kinds of arguments launched genetic algorithms)

Quotable quotes (cont.)

"What is the probability that your average, 52 ohms, is in error by more than 1 ohm? ... The question as stated is ... what is

$$Pr\{ | \mu - 52 | < 1 \} ?$$

This is not an appropriate question! The population mean μ ... is just a constant ... the sample mean is just a constant, and probability statements about them are not appropriate."

.....(probability as frequency doesn't allow asking *interesting* questions!)

"analytic studies have lead to the somewhat surprising findings about consistency or the lack thereof."

.....(referring to Diaconis and Freedman, '86, they first dismiss nonparametric Bayesian methods, and subsequently they conclude with ...")

"the bias/variance dilemma [the overfitting problem] can be circumvented if one is willing to give up generality, that is, *purposely* introduce bias. ... Of course, one must insure that the bias is in fact *harmless for the problem at hand*."

.....(attempting to learn from smaller samples, staunch anti-Bayesians introduce a prior disguised as a "purposely introduced, harmless bias")

Paradoxes

"On the consistency of Bayes estimates", Diaconis and Freedman, *Annals of Statistics*, vol. 14, 1986.

- Authors show that for certain classes of priors, Bayes estimators in a non-parametric setting can be inconsistent, i.e., for large samples the estimator can converge to the wrong answer. They conclude:

  "... that is why we advise against the mechanical use of Bayesian nonparametric techniques."
- Several discussants (Barron, Berger, Lindley) point out that the prior used in the theorems and examples (the Dirichlet) places probability 1 on the class of discrete distributions, whereas the data is drawn from a continuous distribution. i.e., the "truth" is not in the hypothesis space
- Therefore, a better conclusion might be:

  Inconsistent use of Bayes theorem can lead to inconsistency. Inconsistent use can be difficult to detect in nonparametric settings.
- Some authors have since (falsely) referred to this paper as a justification that Bayesian methods are fundamentally flawed in nonparametric settings.

Other communities

Uncertainty in Artificial Intelligence (UAI): applies probabilistic reasoning to problems in intelligent systems, expert systems, planning, diagnosis, learning; the home of Bayesian networks, has a computational perspective (see Henrion, Breese and Horvitz, 91):

http://www.Heuristicrat.com/wray/uaiconnections.html – general community page

Pattern Recognition: vision, speech recognition, natural language, robotics, etc.; huge communities becoming increasingly probabilistic

Statistics: while the staunch anti-Bayesians are decreasing in numbers, the method of choice for the applied non-Bayesian statistician is resampling methods such as cross validation and bootstrap (see Efron and Tibshirani, 91)

http://www.isds.duke.edu/l – page for Inst. of Statistics and Decision Sciences, Duke Univ., contains pointers to lots of good Bayesian stuff

Other communities

Neural networks: borrows from the full range of uncertainty methods in a "connectionist" context; has become very sophisticated with probabilities including leading edge work on priors, computational methods, borrowing methods from statistical physics:

http://www.cs.cmu.edu:8001/afs/cs/project/cnbc/nips/NIPS.html – quality conference

ftp://131.111.48.8/pub/mackay/README.html – pointers to MacKay's favorites

Decision theory: what to do with probabilities once you've got them; community in OR, management science, largely Bayesian

http://www.rahul.net/lumina/DA.html – Decision/Risk Analysis page

Probabilistic networks: an emerging merger of UAI, neural networks, and graphical models from statistics

http://www.Heuristicrat.com/wray/graphbib.ps.Z – survey paper introducing the area

http://www.Heuristicrat.com/wray/lwgmJAIR.ps.Z – methodological outline

Other communities

Computational learning theory: theoretical computer science concerned with data analysis, includes many styles (Bayesian, MDL) and some fields listed below, see:

http://www.dsi.unimi.it/COLT – new home page, pointing to conferences, etc.

Inductive inference: asymptotic results for noise free learning from repeated trials, 1960-70's (see Angluin and Smith, 83)

Uniform Convergence, PAC, and PAB: sample and prior independent, worst-case, large-sample bounds, grew out of pattern recognition and computer science, late 1980's on (see Haussler, 92).

Statistical Physics: adapting mathematical techniques from statistical physics, late 1980's on; sometimes using the techniques (Silver, 93), and sometimes offering statistical physics as a new theory of learning from repeated trials (see Seung, Sompolinsky and Tishby, 93)

Other communities

Stochastic Complexity: also minimum description length (MDL), uses Kraft inequality etc., to replace probability with information theory; has large following of engineers and computer scientists because methods are claimed by some to be "objective" (which is of course absurd); roughly, are Bayesian MAP methods with robust priors: see (Li and Vitanyi, 92; Rissanen, 89; Wallace, 89)

http://www.cs.monash.edu.au/~jono/ – Oliver points to a sequence of tutorial papers

Knowledge discovery: a recent merge of machine learning, statistics and database search; seeing large profitable applications in industry, does "data mining" often scorned by classical statisticians

http://info.gte.com/~kdd/ – a community home page

Other communities

- other fields
  - fuzzy logic, rough sets, Dempster-Shafer methods, and other general "uncertainty calculi" posed as substitutes for probability
  - various applied statistical and decision theory sciences such as economic statistics, geostatistics, medical informatics, computational molecular biology, information theory, and so forth; many tend to be more conservative (read as "less Bayesian")
- multi-disciplinary studies
  - often done with the premise "let's prove our field is better than theirs"
  - empirical comparative studies are fraught with difficulty and best taken with a pound of salt
  - non-Bayesian methods are sometimes highly competitive or superior to existing Bayesian methods because of pragmatic effects, better "implicit" priors and modeling, or because they are Bayesian under a clever disguise
  - some interesting collections are (Wolpert, 93; Michie, Spiegelhalter and Taylor, 94; Ripley, 94)

Classical and Applied Statistics versus Bayesian

(see Appendix B of Bernardo and Smith, 94)

Hypothesis testing: one sided hypothesis testing often corresponds to its Bayesian counterpart but point hypothesis testing is known to have serious problems (e.g., is $\mu \neq 0$?)

Asymptotic tests: most asymptotic results converge for Bayesian, classical, and MDL

Significance testing: based on the sampling distribution have known absurd consequences, but in many typical cases they correspond to their Bayesian counterparts

Resampling methods: bootstrap and cross-validation work well for a range of applied problems, but also have known examples were they produce spurious results; Bayesian variations are discussed in Bernardo and Smith, 94; these remain popular because they are apparently "free of priors" and have been "proven" in applications

Summary: every method not based on probability theory alone has known problems

The Maximum Entropy method

The MaxEnt perspective:

B \ A	true	false
true	θ_1	θ_2
false	θ_3	θ_4

Given the constraint $p(A=true|B=false) = 0.2$, what are "good" values for θ_1, θ_2, θ_3 and θ_4 ?

Entropy is $I(\theta) = -\sum_i \theta_i \log \theta_i$

MaxEnt: choose θ maximizing entropy subject to the constraints

How does this fit in with probabilistic reasoning?

Maximum Entropy versus Bayesian methods

- Because MaxEnt settles on a single value for θ it suffers the same problems as all other statistical estimators:
  - we have no idea of our confidence in the single value
  - we cannot subsequently update our single value if additional evidence is obtained, i.e. is inapplicable in dynamic contexts
- MaxEnt = Maximum A Posterior reasoning using a prior in the form

  $p(\theta) \propto e^{\alpha I(\theta)}$ ← **NB.** this is the (approx.) prior developing using Jaynes' monkeys argument

- MaxEnt embodies one particular from of prior, which may or may not be appropriate — i.e. it is not a "universal" method for obtaining estimates