Understanding the Bayesian Approach: A Nondogmatic Perspective

Kathryn Blackmond Laskey
Department of Systems Engineering
George Mason University
klaskey@gmu.edu

October 10, 1997
Questions

• What is probability?
• What is this Bayesian stuff anyway?
• What’s in it for me?
Views of Probability

• **Classical** - Probability is a ratio of favorable cases to total equipossible cases

• **Frequentist** - Probability is the limiting value as the number of trials becomes infinite of the frequency of occurrence of a random event

• **Logical** - Probability is a logical property of one’s state of knowledge about a phenomenon

• **Subjectivist** - Probability is an ideal rational agent’s degree of belief about an uncertain event

**Probability is none of these things!**
What is Probability?

• The “religious debate” is misdirected
• Probability is a body of mathematical theory
  – Elegant and well-understood branch of mathematics
  – Applied to problems of reasoning with uncertainty
• We can be more constructive if we focus on:
  – What problems can be modeled with probability
  – How to apply it sensibly to these problems
• Probability can be used as a model for:
  – Ratios of favorable to total outcomes
  – Frequencies
  – States of knowledge
History

• People have long noticed that some events are imperfectly predictable

• Mathematical probability first arose to describe regularities in problems with natural symmetries:
  – e.g., games of chance
  – equipossible outcomes assumption is justified

• People noticed that probability theory could be applied more broadly:
  – physical (thermodynamics, quantum mechanics)
  – social (actuarial tables, sample surveys)
  – industrial (equipment failures)
Hierarchy of Generality

- Classical theory is restricted to equipossible cases
- Frequency theory is restricted to repeatable, random phenomena
- Subjectivist theory applies to any event about which the agent is uncertain

**Thesis:**
Categorically ruling out third category is unsupported
The Frequentist

• Probability measures an objective property of real-world phenomena
• Probability can legitimately be applied only to repeatable, random processes
• Probabilities are associated with collectives not individual events
The Subjectivist

• Probability measures rational agent’s degrees of belief
  – No one “correct” probability
  – Viewpoints vary on whether “objective probabilities” exist
  – Use of probability is justified by axioms of rational belief

• Dawid’s theorem: Given feedback
  – rational agents will come to agree on probabilities for convergent sequences of trials
  – these probabilities will correspond to frequencies

• DeFinetti’s theorem: Formal equivalence between
  – subjective probabilities on exchangeable sequences
  – iid trials with prior on unknown “true” probability
deFinetti’s Theorem

- Establishes formal equivalence between exchangeable sequences and iid trials
  - A sequence $X_1, X_2, \ldots, X_n$ of Bernoulli trials is exchangeable if its probability distribution is independent of permutations of indices
  - A sequence is infinitely exchangeable if $X_1, X_2, \ldots, X_n$ is exchangeable for every $n$

- If $X_1, X_2, \ldots$ is infinitely exchangeable then:
  - $\frac{S_n}{n} \to p$ almost surely, where $S_n = \sum_{i=1}^{n} X_i$
  - $P(S_n = k) = \int_{0}^{1} \binom{n}{k} p^k (1-p)^{n-k} f(p) dp$

Infinitely exchangeable sequences behave like iid trials with common unknown distribution.
Views on Statistical Inference

- **Parametric statistics (of any persuasion)**
  - Assume data $X$ follow distribution $f(X|\theta)$
  - Goal: infer $\theta$ from $X$

- **Frequentist inference**
  - Parameter $\theta$ is unknown, data $X$ have distribution $f(X|\theta)$
  - Base inferences on distribution $f(X|\theta)$

- **Bayesian inference**
  - Parameter $\theta$ is uncertain, has distribution $g(\theta)$
  - Data $X$ are unknown before observation, predictive (marginal) distribution $f(X)$
  - Data $X$ are known after observation
  - Inference consists of conditioning on $X$ to find $g(\theta|X)$
  - **Bayesians** condition on knowns and put probabilities on unknowns
Decision Theory

- Inference cannot be separated from decision
- Elements of decision problem
  - Options
  - Consequences
  - Probability distribution expresses knowledge about consequences
  - Utility function expresses preferences for consequences
- Optimal choice is option with maximum expected utility
- Framework for:
  - Information gathering (experimental design, sequential decision)
  - Estimation and hypothesis testing
  - Model selection (Occam’s razor)
Why Be a Bayesian?

• Unified framework for rational inference and decision under uncertainty
  – Spectrum of problems from data-rich to data-poor
  – Spectrum from pure inference to pure decision

• Intuitive plausibility of models

• Understandability of results
  – “If an experiment like this were performed many times we would expect in 95% of the cases that an interval calculated by the procedure we applied would include the true value of $\theta$”
  – “Given the prior distribution for $\theta$ and the observed data, the probability that $\theta$ lies between 3.7 and 4.9 is 95%”

• Straightforward way to treat problems not easily handled in other approaches
Shrinkage toward the Prior

- Triplot: prior, posterior and normalized likelihood plotted on same axes
Subjectivity

• All models have subjective elements
  – Distributional assumptions
  – Independence assumptions
  – Factors included in model

• The prior distribution is just another element of a statistical model

• How to keep yourself honest:
  – Justify assumptions
  – Evaluate plausibility of assumptions in the light of data
  – Report sensitivity of analysis to assumptions
Where is the Payoff?

• Verities from STAT 101
  – Data mining is a bad word
  – Don’t grub through data without *a priori* hypotheses
  – Never estimate more than a few parameters at a time
  – Never use models with a “large” number of parameters relative to your data set

• The “dirty little secret”
  – *There is NEVER enough data!!!*
  – Everybody “peeks” at the data
  – Models always grow in complexity as we get more data

• Hierarchical Bayesian models
  – Formally sound and practical methodology for high-dimensional problems
  – Multiple levels of randomness allow adaptation of model to intrinsic dimensionality of the data set
Example

• Educational testing
  – Test scores for 15 classrooms
  – Between 12 and 28 students per class
  – Objective: estimate mean and error interval for each class

• Simple hierarchical model
  – Classrooms are exchangeable
  – Students within class are exchangeable
  – Scores follow normal distribution
Graphical Models

- Intuitively natural way to encode independence assumptions

- Directed and undirected graphs
  - Bayesian networks
  - Markov graphs
  - Hybrids

- Causal and correlational models

- Estimation and inference algorithms that make use of graph structure
  - e.g., Gibbs sampling and other Markov Chain Monte Carlo methods
Hierarchical Model

- Joint distribution $h(\alpha) \prod g(\theta_i \mid \alpha_i) \prod f(X_{ij} \mid \theta_i)$
- Prior on $\alpha$ can be vague
- Model adapts to dimensionality of data
- Empirical reports that hierarchical models improve out-of-sample performance on high-dimensional problems
Challenges

• Overfitting hasn’t gone away
  – Priors that adapt to effective dimensionality of data
  – Robust semi-parametric models

• Computational complexity
  – Monte Carlo
  – Extracting tractable submodels
  – Analytical approximations

• Prior specification
  – Semantics, elicitation
  – Exploring behavior of “typical” datasets/parameter manifolds generated by prior
  – Exploring behavior of posterior for “typical” and “nontypical” datasets
  – Visualization
Bayesian Model Choice

• Uncertainty about model structure

\[ P(X) = \sum_S P(S) \int f(X|S, \theta_S) d\theta_S \]

• Bayesian updating of structural uncertainty

\[ P(X_{new}|X) = \sum_S P(S|X)P(X_{new}|X, S) \]

\[ = \sum_S P(S|X) \int P(X_{new}|X, \theta_S S) f(\theta_S) d\theta_S \]

• This sum cannot be computed explicitly
  – Heuristic search
  – Markov Chain Monte Carlo Model Composition (MC\(^3\))
Occam’s Razor and Model Choice

• Occam’s razor says “prefer simplicity”
• As a heuristic it has stood the test of time
• It has been argued that Bayes justifies Occam’s razor. More precisely, if:
  – you put a positive prior probability on a sharp null hypothesis
  – the data are generated by a model “near” the null model
  – the sample size is not too large

Then (usually) the posterior probability of the null hypothesis is larger than its prior probability
Occam’s Razor (cont.)

• Of course we don’t really believe the null hypothesis!
• We don’t believe the alternative hypothesis either!
• When predictive consequences of $H_0$ and $H_A$ are similar:
  – $H_0$ is robust to plausible departures from $H_0$
  – When $H_A$ has many parameters in relation to the amount of data available we may do much worse by using $H_A$
  – $H_0$ is robust to (likely) misspecification of parameters $\theta_A$ of $H_A$
• But Occam’s razor only works if we’re willing to abandon simple hypotheses when they conflict with observations
Decision Theory and Occam’s Razor

• Occam’s razor is really about utility and not probability
  – Choose the simplest model that will give you good performance on problems you haven’t seen

• Decision theoretic justification
  – The simple model is not “correct”
  – Adding more parameters to fit the data is often not the way to make it correct
  – Too-complex models give false sense of precision and are difficult to apply
  – Occam’s razor is a heuristic for finding high-utility models
Another Level to the Hierarchy

• Statistics is about designing procedures that work well for large classes of problems
  – Problems to which it applies
  – Diagnosing when it doesn’t apply

• Decision theory can help us think about this problem
  – Inference procedures that usually work well
  – Inference procedures that are robust to plausible departures from model specification
  – Ways to diagnose situations in which procedures don’t work

• Is the best object-level procedure necessarily Bayesian?
Summary

• Bayesian decision theory is a unified framework for
  – Thinking about problems of inference and decision making under uncertainty
  – Designing statistical procedures that are expected to work well on large classes of problems
  – Analyzing behavior of statistical procedures on a class of problems

• Promising technologies:
  – Bayesian hierarchical models
    » Adaptive dimensionality
    » Few “truly free” parameters
  – Bayesian model selection

• Religious dogma is detrimental to good statistics