
Model Confidence

**Kathryn Blackmond Laskey
Department of Systems Engineering
George Mason University**

**Summer Institute on Probability in AI
July, 1994**

Models

A model is a representation of a system which can be used to answer questions about the system

Prototypical question: What is $P(y|x)$?

- for a decision problem y is utility and x is action
- for a learning problem x is a previous sample of similar cases and y is set of future case(s)
- for diagnosis x is a set of symptoms and y represents possible causes of the symptoms

Typical response:

- Construct model M relating y to x
- Compute or approximate $P(y|x,M)$
- Hope that $P(y|x,M)$ is a good approximation to $P(y|x)$

Examples of Models

- **Bayes network**
 - Given values of some variables, find probability distributions for other variables
- **Bayes network with uncertain conditional probability matrices**
 - Learn conditional probability matrices from observations
- **Bayes network with unknown structure and conditional probability matrices**
 - Learn structure and conditional probability matrices from observations
 - Predict future observations
- **Other**
 - Neural network
 - Hidden Markov model
 - ...

The Art of Model Building

- **Models are constructed from:**
 - Past data on system or related systems
 - Judgment of subject matter experts
 - Judgment of experienced model builders
- **A trained and experienced analyst can:**
 - select model that is likely to be a good approximation for the purpose at hand
 - recover if it turns out not to be

**“All models are wrong
but some are useful”**

Exploratory and Consolidative Modeling

(Bankes, 1993)

- **Consolidative modeling uses the model as a surrogate for the system**
 - consolidates known facts about the system
 - for purposes of analysis the model adequately represents the system
- **Exploratory modeling explores how the world would behave if various hypotheses were correct**
 - many details and mechanisms of system are uncertain
 - model is not a reliable image of the world
 - goal is to explore ensemble of plausible models

Goals of Exploratory Modeling

- **Identify an ensemble of plausible models and modeling assumptions**
- **Identify the range of outputs predicted by plausible models under plausible assumptions**
- **Identify the relationship between modeling assumptions and model outputs**
- **Find assumptions that have a large impact on model outputs; and**
- **Identify predictions that are robust across different modeling assumptions.**

Measuring Quality of Model

- **Some measures:**

- **Calibration:** $p\%$ of events for which $P(x|M) = p$ happen
- **Refinement:** $P(x|M)$ makes extreme predictions
- **Coherence:** $P(x|M)$ satisfies rules of probability calculus
- **Robustness:** $P(x|M)$ gives good results under M' near M

- **Comments:**

- **A well calibrated model need not be very useful**
 - » you will be perfectly calibrated if every day you forecast climatological probability for rain
- **A poorly calibrated model can be very useful**
 - » a forecaster who is always dead wrong is quite useful!
- **Coherence alone is a weak criterion**
 - » “Beware of consistent, all-embracing systems of prejudice”

- **Model quality depends on purpose of model**

What is Model Confidence?

- You are confident in a model if the model would not change dramatically given a reasonable amount of data not strongly inconsistent with model
- You assess $P(\text{Heads}) = .5$. Someone tosses the coin and it comes up heads. What is $P(\text{Heads}_2 | \text{Heads}_1)$?
- Cases:
 - Coin is an ordinary coin you received as change in a drugstore yesterday
 - You bought coin in a magic shop and don't know how it's weighted
 - You bought coin in a magic shop. It has either 2 heads or 2 tails and you don't know which

Types of Meta-Reasoning for Model Confidence

- **Construct model given current knowledge**
- **Detect problem with current model**
- **Select immediate response to identified problem**
- **Learn a better model**
- **Allocate resources among**
 - Problem solving
 - Diagnosing problems with current model
 - Learning a better model

Robust Models

- **Models are often chosen**
 - because everyone uses them
 - because the computer package supports them
 - because they are mathematically tractable
- **Criteria for robust model:**
 - Under most plausible departures M' from M , $P(y|x,M)$ is a good enough approximation for the purpose at hand
 - Most M' for which this is not the case can be detected before it's too late to respond
- **Issues:**
 - Meaning of “good enough” (utility function) may differ for different applications of model
 - What are plausible departures from M ?

Some Examples

**of heuristics that appear to give robust models
in certain commonly occurring situations**

- **Low dimensional models**
- **Uniform priors**
- **Maximum entropy**
- **Hierarchical models**

**Beware: Robustness is
context dependent!**

How to Achieve Robustness

- **Invest effort in analyzing the behavior of model model under plausible alternate assumptions**
- **Examine what your model says about features you didn't design it to fit**
- **Examine your model's behavior on improbable cases**

Examining operating Characteristics of Models (Rubin, 1984)

- **Helps to decide whether a model is good enough to use (even if not correct)**
- **Procedure:**
 - **Simulate data randomly from several plausible alternative models**
 - **Analyze data using the candidate model**
 - **Evaluate quality of results**
- **Robust models perform well when data are generated by plausible alternative models**

Formalizing Model Uncertainty

- Many models can be decomposed $M=(S, \theta)$
 - S = structural assumptions (conditional independence, normality, etc.)
 - θ is a structure-specific parameter (link probability matrix, mean and covariance of normal distribution, etc.)
- Statisticians often use the data to pick the “best” S^* and estimate parameters assuming S^* is correct
 - data-mining
 - overfitting
- Expert systems developers often elicit “best” S^* and θ^* from expert and use it
- Result can be serious miscalibration

Advantages of Paying Attention to Model Uncertainty

- **Error bounds on predictions incorporate both within-model and between-model variability**
- **Avoid overfitting and data mining**
- **Outputs are more robust if a range of plausible alternatives has been considered**

Parameter Uncertainty: Probabilistic Treatment

- **Higher order probability**
 - Probability distribution on unknown
 - $P(y|x)$ assumed equal to $E[P(y|x, \theta)]$
 - **Expected utility of including uncertainty**
 - » Zero for single case
 - » May be large for multiple cases
- **Expert assessment as “noisy measure” of**
 - Assess prior and likelihood functions for θ ; update after obtaining assessment
- **Interval probabilities**

Parameter Uncertainty: Other Approaches

- **Possibility theory**
 - Deals with imprecision: “John is tall”
- **Belief functions**
 - Deals with incompleteness of evidence: “Sensor may be malfunctioning”
- **Symbolic approaches**
 - Propagate symbolic descriptors of uncertainty

Structural Uncertainty

- **Most research has focused on parameter uncertainty**
- **Approaches to handling structural uncertainty:**
 - Qualitative treatment of multiple structures
 - Put probabilities on structures and weight outputs
 - “Noisy approximation”

Qualitative consideration of multiple models

- **Predictions (including uncertainty bounds) under several plausible models are computed**
- **No integration into a single model**
- **Measure of support of data for model (Bayes factor) may be computed**
- **Look for action that works reasonably well for range of plausible models**

Higher Order Uncertainty for Structures

- **Assume:**

$$\begin{aligned} P(y|x) &= \prod_{i=1} P(S_i|x) P(y|x, S_i) \\ &= \prod_{i=1} P(S_i|x) \prod_{S_i} P(y|x, S_i) f(S_i|S) \end{aligned}$$

- **The space of models is much too big!**
- **Approximate $P(y|x)$ by a (small) finite number of terms:**

$$P(y|x) \approx \sum_{i=1}^k \bar{P}(S_i|x) P(y|x, S_i)$$

Issues

- **Which models to include?**
 - high $P(x,y|S_i)$ for probable (x,y)
 - $P(y|x,S_i)$ is very different from the other $P(y|x,S_j)$ for probable (x,y)
- **Probabilities assigned to the models need not be the same in the finite and infinite sums**
- **Data can be used to update**
 - Distribution for S
 - Distribution for s

Using Data: The Bayes Factor

$$\frac{P(S_i | \mathbf{x})}{P(S_j | \mathbf{x})} = \frac{P(\mathbf{x} | S_i) P(S_i)}{P(\mathbf{x} | S_j) P(S_j)}$$

Prior odds ratio

Likelihood ratio

- Bayes factor measures relative support data give to models
- Computation of Bayes factor often involves difficult integrals

$$P(\mathbf{x} | S_i) = \int P(\mathbf{x} | \theta, S_i) g(\theta | S_i) d\theta$$

- Bayes factor is sensitive to choice of prior distribution for θ

Interpretation of $P(S)$

- **Straightforward interpretation: $P(S_i)$ is the probability that model S_i is correct**
- **For most problems the probability that model S_i is correct is zero!**
- **Alternative interpretations:**
 - **$P(S_i)$ is probability that model S_i is adequate for the objective**
 - » not mutually exclusive and exhaustive!
 - **$P(S_i)$ is the probability that model S_i is the best model**
 - » what justifies Bayesian updating and averaging?
 - » why should “most probably best” be weighted highest?

Another Interpretation

- Structures represent regions in model space
- My “real” model (if I knew it) would look like:

$$P(y|x) = \sum_i P(R_i) P(y|x, R_i)$$

- $P(S_i)$ is the probability that the region in model space represented by S_i is the region within which the correct model lies
- $P(y|x, S_i)$ is an estimate of the correct model for the region represented by the structure S_i

Advantages of the “Regions” Interpretation

- **It is a better formalization of our real goal**
 - Sample regions in such a way that we achieve a good enough estimate
 - NOT enumerate all possible models and compute the correct probability
- **It gives a framework for meta-level reasoning**
 - When I get data that are very unlikely under all the models, this indicates my sampling of regions is inadequate given my current information
 - We can think about using models we have constructed to estimate what a model we haven't constructed would say
 - We can do value-of-information calculations to allocate model-building and data-collection resources
 - It gives a theoretical justification for the intuition that Bayes rule may not be the best way to update $P(S_i)$

Models as “Noisy Measures”

- Define prior distribution for y given x
- Define likelihood function for $P(y|x,M)$ given x
- Update distribution for y using Bayes rule

Occam's Razor

- **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:
 - $P(y|x, H_0)$ may be close to $P(y|x, H_0) + (1 - \alpha)P(y|x, H_A)$ for any
 - So $P(y|x, H_0)$ 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 θ_A of H_A
- But Occam's razor only works if we're willing to abandon simple hypotheses when they conflict with observations

In Practice

- **Mixture distributions with a small number of terms often perform better than one “big” model with a diffuse prior over the unknown parameters**
- **Posterior probabilities on the models reflect relative support of the data for the models**
- **Estimates of variability can be decomposed**
 - uncertainty in data
 - uncertainty in parameters given structure
 - uncertainty in structure

Model Diagnosis

- **Posterior probabilities on models reflect support of observations for models we have explicitly articulated**
- **Model diagnosis attempts to warn us when our set of models may be too small**
- **Model diagnosis attempts to increase the *a posteriori* probability that $P(y|x, M)$ is near $P(y|x)$**

Basic Approach

- **Select a model monitoring statistic $T(x)$**
- **Compute distribution of $T(x|M)$ for hypothetical replications of current study**
- **Is observed value of $T(x|M)$ typical of this distribution? If not, model may be wrong**
- **Goal of model & test statistic selection:**
 - $P(x,y|M)$ is usually near $P(x,y)$
 - $T(x)$ detects the most likely situations in which $P(x,y|M)$ is far from $P(x,y)$
 - $T(x)$ gives clues about deficiencies in M
- **This approach may be much more efficient than mixing over more models initially**

Examples

- **Outlier detection in statistical model fitting**
- **Residual analysis in statistical model fitting**
- **Conflict detection for Bayes networks**
 - Jensen, et al., 1990 (UAI '90)
 - Laskey, 1991 (UAI '91)
 - Habbema, 1989 (cited in Jensen et al. and Laskey)
- **General points:**
 - Most published methods rely on “tail area” statistics
 - Conflict detection will be more efficient if you can anticipate direction of departure from current model and construct $T(x)$ to exploit

Conclusion

- **Advantages of explicitly treating model uncertainty**
 - counteracts overconfidence in estimating prediction intervals
 - helps select robust models and reject inadequate models
- **Bayesian theory provides formal framework for exploratory modeling**
 - scientifically justified, non ad-hoc methods
 - protects against dangers of data-mining
 - theory can be used to suggest and justify qualitative methods

References

Bankes, S. (1993) Exploratory Modeling for Policy Analysis *Operation* 41(3), p. 435-449.

Draper, David (1993) *Assessment and Propagation of Model Uncertainty* paper, Department of Mathematics, University of California at Los Angeles

Kass R.E. and Raftery, A.E. (1993) *Bayes Factors and Model Uncertainty* Pittsburgh, PA: Carnegie Mellon University Department of Statistics Technical Report # 571.

Laskey, K. and Lehner, P. Metareasoning and the Problem of Small V *IEEE Transactions on Systems, Man and Cybernetics*, to appear.

Rubin, D.B. (1984) Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician. *The Annals of Statistics*, 12(4) 1151-1172.