Meta-Bayesian Analysis

A Bayesian decision-theoretic analysis of Bayesian inference under model misspecification

Jun Yang
joint work with Daniel Roy

Department of Statistical Sciences
University of Toronto

World Congress in Probability and Statistics
July 11, 2016

Motivation

"*All* models are wrong,
some are useful."
— George Box

"truth [...] is much too complicated to
allow anything but approximations."
– John von Neumann

- Subjectivism Bayesian:
  alluring but impossible to practice when model is wrong
- Prior probability = degree of Belief... in what?
  What is a prior?
- Is there any role for (subjective) Bayesianism?

Our proposal: More inclusive and pragmatic definition for "prior".
Our approach: Bayesian decision theory

Example: Grossly Misspecified Model

Setting: Machine learning

data are collection of documents: 

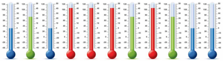
- Model: Latent Dirichlet Allocation (LDA)
  aka "topic modeling"
- Prior belief: $\tilde{\pi} \equiv 0$,
  i.e., no setting of LDA is faithful to
  our true beliefs about data.
- Conjugate priors $\pi(d\theta) \sim \mathrm{Dirichlet}(\alpha)$

What is the meaning of a prior on LDA parameters?

Pragmatic question: If we use an LDA model (for whatever reason), how should we choose our "prior"?

Example: Accurate but still Misspecified Model

Setting: Careful Science

data are experimental measurements: 

- Model: $(Q_\theta)_{\theta\in\Theta}$, painstakingly produced after years of effort
- Prior belief: $\tilde{\pi} \equiv 0$,
  i.e., no Q_θ is 100% faithful to
  our true beliefs about data.

What is the meaning of a prior ~~in a misspecified model~~?

(All models are misspecified.)

Pragmatic question: How should we choose a "prior"?

Standard Bayesian Analysis for Prediction

$Q_\theta(\cdot)$	Model on $\mathcal{X} \times \mathcal{Y}$ given parameter θ $\mathcal{X}$: what you will observe $\mathcal{Y}$: what you will then predict
$\pi(\cdot)$	prior on θ
$(\pi Q)(\cdot) = \int Q_\theta(\cdot)\pi(d\theta)$	Marginal distribution on $\mathcal{X} \times \mathcal{Y}$

Believe $(X, Y) \sim \pi Q$

The Task

1. Observe X.
2. *Choose* action $\hat{Y}$.
3. Suffer loss $L(\hat{Y}, Y)$

The Goal
Minimize expected loss

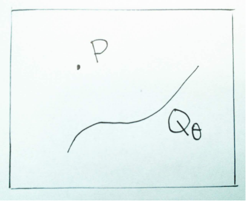
Bayes optimal action minimizes expected loss under the conditional distribution of Y given $X = x$, written $\pi Q(dy|x)$:

$$\mathrm{BayesOptAction}(\pi Q, x) = \arg\min_a \int L(a, y)\, \pi Q(dy|x).$$

- Quadratic loss $\longrightarrow$ posterior mean.
- Self-information loss (log loss) $\longrightarrow$ posterior $\pi Q(\cdot|x)$.

Meta-Bayesian Analysis

- $(Q_\theta)_{\theta \in \Theta}$: the model, i.e., a family of distributions on $\mathcal{X} \times \mathcal{Y}$.
- Don't believe Q_θ, i.e., model is misspecified
- P: represents our true belief on $\mathcal{X} \times \mathcal{Y}$.



Believe $(X, Y) \sim P$
But We Will Use Q_θ **to predict**

The Task

1. *Choose* (surrogate) prior π
2. Observe X.
3. Take action $\hat{Y} = \mathrm{BayesOptAction}(\pi Q, x)$
4. Suffer loss $L(\hat{Y}, Y)$

The Goal

Minimize expected loss **with respect to** P **<u>not</u>** πQ.

Meta-Bayesian Analysis

Key ideas:

- Believe $(X, Y) \sim P$
- But predict using $\pi Q(\cdot|X = x)$ for *some* prior π
- Prior π is an choice/decision/action.
- Loss associated with π and (x, y) is

$$L^*(\pi, (x, y)) = L(\mathrm{BayesOptAction}(\pi Q, x),\ y)$$

Meta-Bayesian risk

- Bayes risk under P of doing Bayesian analysis under πQ

$$R(P, \pi) = \int L^*(\pi, (x, y)) P(dx \times dy).$$

- Meta-Bayesian optimal prior minimizes meta-Bayesian risk:

$$\inf_{\pi \in \mathcal{F}} R(P, \pi),$$

where $\mathcal{F}$ is some set of priors under consideration.

Meta-Bayesian Analysis

Recipe

- Step 1: State P, Q_θ, and select a loss function L;
- Step 2: Choose prior π that minimizes meta-Bayesian risk.

Examples

- Log loss: minimizing the conditional relative entropy

$$\inf_{\pi} \int \mathrm{KL}\left(P^2(x,\cdot)||\pi Q(\cdot|x)\right) P^1(dx)$$

  where $P(dx, dy) = P^1(dx)P^2(x, dy)$.

- Quadratic loss: minimizing the expected quadratic distance between two posterior means $\pi Q(\cdot|x)$ and $P^2(x,\cdot)$:

$$\inf_{\pi} \int \|m_{\pi Q}(x) - m_{P^2}(x)\|_2^2 P^1(dx)$$

Meta-Bayesian Analysis

High-level Goals

- Meta-Bayesian analysis for Q_θ under P is generally no easier than doing Bayesian analysis under P directly.
- But P serves only as a placeholder for an impossible-to-express true belief.
- Our theoretical approach is to attempt to prove general theorems true of broad classes of "true beliefs" P.
- The hope is that this will tell us something deep about subjective Bayesianism.

Remaining results are some key findings.
Meta-Bayesianism sometimes violates traditional Bayesian tenets.

Meta-Bayesian 101: if true belief is realizable

When model is well-specified

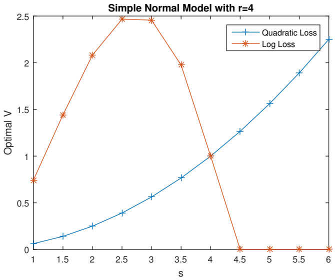
- There exists π such that $P = \int Q_\theta \pi(\mathrm{d}\theta)$ (i.e. $P = \pi Q$)
- Meta-Bayesian loss reduces to expected loss in traditional Bayesian
- Self-consistency: π is the meta-Bayesian optimal prior.

Meta-Bayesian Analysis reduces to traditional Bayesian Analysis when model is well-specified.

Meta-Bayesian Analysis for i.i.d. Normal Model

Example: i.i.d. Normal

- true belief P: $\mathcal{N}(\theta, r^2)$, with $\tilde{\pi}(\mathrm{d}\theta) \sim \mathcal{N}(0, 1)$.
- model $Q_\theta = \mathcal{N}(\theta, s^2)$ where $s^2 \neq r^2$.
- prior π: $\mathcal{N}(0, V)$ with one parameter V.
- $X \in \mathcal{R}^n$, $Y \in \mathcal{R}^k$.



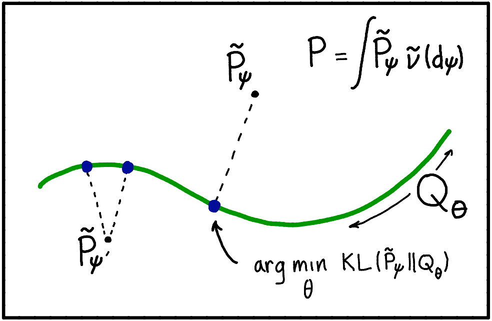
Results for $n = 1$ and $k = 1$

- Predictive of Y given $X = x$:
  P: $\mathcal{N}(\frac{x}{1+r^2}, r^2 + \frac{r^2}{1+r^2})$
  πQ: $\mathcal{N}(\frac{x}{1+s^2/V}, s^2 + \frac{s^2}{1+s^2/V})$
- Quadratic Loss: $V_{\mathrm{opt}} = \frac{s^2}{r^2}$
- Log Loss: V_{opt} balances predictive mean and variance.
- If well-specified ($s^2 = r^2$), $V_{\mathrm{opt}} = 1$ for both losses.

In general, the optimal prior depends on n, k and the *loss*!

General Results when P is a mixture of i.i.d.

Theorem (Berk 1966). Posterior distribution of θ concentrates asymptotically on point minimizing the KL divergence.



Conjecture

- For each $\psi \in \Psi$, assume there is a unique parameter $\phi(\psi) \in \Theta$ such that $Q_{\phi(\psi)}$ minimizes the KL divergence with $\tilde{P}_\psi$.
- Maybe "KL-projection" of prior, i.e., $\tilde{\pi} = \tilde{\nu} \circ \phi^{-1}$, is optimal.

General Results when P is a mixture of i.i.d.

- Let $\tilde{\pi} = \tilde{\nu} \circ \phi^{-1}$ and $\tilde{\nu}(\mathrm{d}\psi|\theta)$ be disintegration of $\tilde{\nu}$ along ϕ.
- We can transform true model over Ψ to one over Θ:

$$P_\theta = \int \tilde{P}_\psi \, \tilde{\nu}(\mathrm{d}\psi|\theta).$$

- Belief about first k observations: $P^{(k)} = \int_\Theta P_\theta^k \tilde{\pi}(\mathrm{d}\theta)$.

Theorem (Y.–Roy)

For every $\theta \in \Theta$, assume θ is the unique point in Θ achieving the infimum $\inf_{\theta' \in \Theta} \mathrm{KL}(Q_{\theta'}||P_\theta)$. Then

$$\Big| \underbrace{\mathrm{KL}(P^{(k)}||\pi_k^* Q^k)}_{R(P,\pi_k^*)} - \underbrace{\mathrm{KL}(P^{(k)}||\tilde{\pi} Q^k)}_{R(P,\tilde{\pi})} \Big| \to 0 \text{ as } k \to \infty.$$

True belief about asymptotic "location" of posterior distribution is an asymptotically optimal (surrogate) prior.

Meta-Bayesian Analysis for i.i.d. Bernoulli Model

Example

data are coin tosses: 10001001100001000100100

- true belief P: two state $\{0,1\}$ discrete Markov chain with transition matrix $\begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$.
- model $Q_\theta^k = \mathrm{Bernoulli}(\theta)^k$.
- true *prior belief*

$$\tilde{\nu}(\mathrm{d}p, \mathrm{d}q) = \tilde{\pi}(\mathrm{d}\theta)\, \tilde{\kappa}(\mathrm{d}\psi|\theta),$$

where

$$\theta = \frac{p}{p+q}$$

is the limiting relative frequency of 1's (LRF).

What does a prior on an i.i.d. Bernoulli model mean?

Conjecture

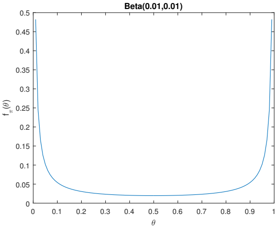
Optimal prior for the model Q_θ^k is our true belief $\tilde{\pi}(d\theta)$ on the LRF.

In general, false!

Counterexample

Assume we know $\theta = \frac{1}{2}$.

- Truth: Sticky Markov Chain:
  000000111111110000000111111111
- Model: i.i.d. sequence
  0010011101001011001001001001



If we make one observation ($n = 1$) and then make one prediction ($k = 1$) better off with $Beta(0.01, 0.01)$ prior than *true belief* $\delta_{\frac{1}{2}}$ on LRF.

What does a prior on an i.i.d. Bernoulli model mean?

Theorem (Y.–Roy)

1. *Let Q_θ^k be the i.i.d. Bernoulli model.*
2. *Let P be true belief and assume P believes in LRF.*
3. *Let $\tilde{\pi}(d\theta)$ be the true belief about the LRF and assume $\tilde{\pi}$ is absolutely continuous.*
4. *Let $\pi_k^* = \arg\min_\pi R(P, \pi)$ be an optimal surrogate prior.*

Then

$$\Big|\underbrace{\mathrm{KL}(P^{(k)}||\pi_k^* Q^k)}_{R(P,\pi_k^*)} - \underbrace{\mathrm{KL}(P^{(k)}||\tilde{\pi} Q^k)}_{R(P,\tilde{\pi})}\Big| \to 0 \text{ as } k \to \infty.$$

True belief about limiting relative frequency is an asymptotically optimal (surrogate) prior.

Conclusion and Future work

Conclusion

- Standard definition of a (subjective) prior too restrictive
- More useful definition using Bayesian decision theory.
- Meta-Bayesian prior is one you believe will lead to best results.

Future Work

- Beyond choosing priors: General Meta-Bayesian analysis (optimal prediction algorithms)
- Analysis of the rationality of non-subjective procedures (e.g, switching, empirical Bayes)