Model Comparison

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

02/22/2021

Multiple hypotheses/models

- Bayesian framework does not treat H_0 and H_a differently
- Methodology may be extended to more than two conclusions
- Instead of "hypotheses", compare evidence for "models"
- For data $\mathbf{y}$, models $M_1, \ldots, M_m$:
  - M_i: $\mathbf{y} \sim p(\mathbf{y} \mid \theta_i, M_i)$, with prior $\theta_i \sim p(\theta_i \mid M_i)$
  - With prior probabilities $P(M_i)$:

$$P(M_1) + \ldots + P(M_m) = 1.$$

Multiple hypotheses/models

- The posterior probability of model i is

$$p(M_i \mid \mathbf{y}) = \frac{P(M_i)p(\mathbf{y} \mid M_i)}{\sum_{j=1}^{m} P(M_j)p(\mathbf{y} \mid M_j)}$$

where

$$p(\mathbf{y} \mid M_i) = \int p(\mathbf{y} \mid \theta_i, M_i)p(\theta_i \mid M_i)\, d\theta_i .$$

Model choice

- Actions $\mathcal{A} = \{M_1, \ldots, M_m\}$
- Under "$0 - 1$" loss,

$$l(M_i, d(\mathbf{y})) = \mathbb{1}_{\{d(\mathbf{y}) \neq M_i\}}$$

  - Choose M_i with highest posterior probability $P(M_i \mid \mathbf{y})$
- Under "$0 - c_i$" loss,

$$l(M_i, d(\mathbf{y})) = c_i \mathbb{1}_{\{d(\mathbf{y}) \neq M_i\}}$$

  - Posterior risk for choosing M_i is

$$= \sum_{j=1}^{m} c_j \mathbb{1}_{\{M_i \neq M_j\}} \cdot P(M_j | y)$$

$$\rho(p_\theta, a = M_i) = \sum_{j \neq i} c_j P(M_j \mid \mathbf{y})$$

  - Choose M_i with highest weighted posterior $c_i P(M_i \mid \mathbf{y})$

Note $\rho(p_\theta, m_a) < \rho(p_\theta, m_b)$

$$\Longleftrightarrow \sum_{j \neq a} c_j P(m_j | y) < \sum_{j \neq b} c_j P(m_j | y)$$

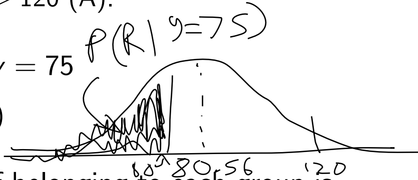
$$\Longleftrightarrow c_b P(m_b | y) < c_a P(m_a | y)$$

Example: IQ

- Human IQs have a Normal$(100, 225)$ distribution
- A given IQ test has normal error with variance 64.
- Observe the test score y for a student
  - $p(y \mid \mu) = \text{Normal}(\mu, 64)$
  - $p(\mu) = \text{Normal}(100, 225)$
- The posterior distribution for their true IQ is
  - $p(\mu \mid \mathbf{y}) = \text{Normal}(22.15 + 0.779\, y\,,\ 49.83)$

Example: IQ

- A given student belongs to the
  - remedial learning group if IQ$<$ 80 (R)
  - standard learning group if $80 <$ IQ < 120 (S)
  - accelerated learning group if IQ$>$ 120 (A).
- Assume that a student has score $y = 75$
  - $p(\mu \mid \mathbf{y}) = \text{Normal}(80.56\,,\, 49.83)$



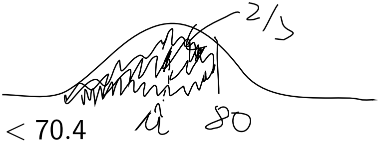
- Then, their posterior probability of belonging to each group is
  - $P(R \mid y = 75) = 0.468$
  - $P(S \mid y = 75) = 0.532$
  - $P(A \mid y = 75) \approx 0$

  http://www.ericfrazerlock.com/Model_Comparison_Rcode1.r

Example: IQ

- Assign loss functions
  - $l(R, d(\mathbf{y})) = \mathbb{1}_{\{d(\mathbf{y}) \neq R\}}$
  - $l(S, d(\mathbf{y})) = 2 \cdot \mathbb{1}_{\{d(\mathbf{y}) \neq S\}}$
  - $l(A, d(\mathbf{y})) = \mathbb{1}_{\{d(\mathbf{y}) \neq A\}}$
- For $y = 75$:
  - $2P(S \mid y = 75) = 1.064 > P(R \mid y = 75) = 0.468$, and
  - $2P(S \mid y = 75) = 1.064 > P(A \mid y = 75) \approx 0$, so
  - So choose the standard group (S).

- Decision rule for arbitrary y:



$$d(y) = \begin{cases} R & \text{if } y < 70.4 \\ S & \text{if } 70.4 \le y \le 129.6 \\ A & \text{if } y > 129.6 \end{cases}$$

Choose R if $P(R|y) > 2 \cdot \underbrace{P(S|y)}_{\approx 1 - P(R|y) \text{ for } y \le 75}$

$\rightarrow P(R|y) > \frac{2}{3}$

$$\frac{\overbrace{\mu - 22.15 - 0.779y}^{\hat{\mu}}}{\sqrt{49.83}} \sim N(0,1)$$

$$\frac{80 - 22.15 - 0.779y}{\sqrt{49.83}} > \underbrace{z_{2/3}}_{0.431}$$

$\rightarrow y < 70.4$

Bayes factors for model comparison

- Recall the Bayes factor for model M_1 over model M_2 is

$$BF = \frac{p(\mathbf{y} \mid M_1)}{p(\mathbf{y} \mid M_2)}$$

- A likelihood ratio test is based on maximum for each model:

$$\Lambda = \frac{\max_{\theta_1} p(\mathbf{y} \mid \theta_1, M_1)}{\max_{\theta_2} p(\mathbf{y} \mid \theta_2, M_2)}$$

- Under point models $M_1 : \theta = \theta^{(1)}$ and $M_2 : \theta = \theta^{(2)}$:

$$BF = \Lambda = \frac{p(\mathbf{y} \mid \theta^{(1)})}{p(\mathbf{y} \mid \theta^{(2)})}$$

Bayesian Information Criterion

- Let p_i be number of parameters in model M_i
- Let n be the data sample size
- A heuristic for assessing the fit of a model is the *Bayesian Information Criterion* (BIC):

$$BIC(M_i) = -2\log(\max_{\theta_i} p(\mathbf{y} \mid \theta_i, M_i)) + p_i \log n,$$

  - Smaller values are preferred
  - log likelihood, with penalty for the dimension of the model

Bayesian Information Criterion

- Likelihood ratio test usually based on transformed ratio

$$W = -2\log\left[\frac{\max_{\theta_1} p(\mathbf{y} \mid \theta_1, M_1)}{\max_{\theta_2} p(\mathbf{y} \mid \theta_2, M_2)}\right]$$

- The difference in BIC can be expressed in terms of W:

$$\Delta BIC = W - (p_2 - p_1)\log n,$$

  - Δ denotes change (from M_1 to M_2)
  - The likelihood ratio statistic corrected for dimension of each model

Bayesian Information Criterion

- For $\mathbf{y} = y_1, y_2, \ldots y_n$ iid, as $n \to \infty$,

$$-2\log(BF) \approx \Delta BIC$$

  under mild assumptions.

- Derivation: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.723.8015&rep=rep1&type=pdf
- ΔBIC may be easier to compute than the BF
- ΔBIC does not depend on prior distributions
- *BIC* also called *Schwarz information criterion* for G. Schwarz
  - Original article: http://projecteuclid.org/euclid.aos/1176344136

Partial Bayes factors

- If $p(\theta_i \mid M_i)$ is improper, then so is

$$p(\mathbf{y} \mid M_i) = \int p(\mathbf{y} \mid \theta_i, M_i) p(\theta_i | \mathrm{M_i}) d\theta_i$$

  so Bayes factors involving M_i not well defined.

- Possible solution:
  - Assume $p(\theta_1 \mid \mathbf{y}_1)$ is proper for $\mathbf{y}_1 = (y_1, \ldots, y_i)$
  - Find conditional Bayes factor for $\mathbf{y}_2 = (y_{i+1}, \ldots, y_n)$

$$BF(\mathbf{y}_2 \mid \mathbf{y}_1) = \frac{p(\mathbf{y}_2 \mid \mathbf{y}_1, M_1)}{p(\mathbf{y}_2 \mid \mathbf{y}_1, M_2)}$$

  - This is a *Partial Bayes factor*

Example: traffic accidents

- Would like to estimate weekly accident rate at new traffic intersection.
- Each week observe $y \sim \text{Poisson}(\lambda)$ accidents
- M_1:Elicited prior from city planner: $p_1(\lambda) = \text{Gamma}(3, 2)$.
- M_2: Compare with (improper) uniform prior $p_2(\lambda) = 1$.
- Observe data for 5 weeks:
  - $y_1 = 3,\ y_2 = 6,\ y_3 = 2,\ y_4 = 4,\ y_5 = 2$

Poisson-Gamma marginal

- If $y_1, \ldots, y_n \overset{iid}{\sim}$ Poisson(λ) and $p(\lambda) =$ Gamma(α, β),

$$p(\mathbf{y}) = \frac{\beta^\alpha \Gamma(\sum y_i + \alpha)}{\Gamma(\alpha) \prod y_i! (\beta + n)^{\sum y_i + \alpha}}$$

$$P(\vec{y}|\lambda) = \frac{\lambda^{\Sigma y_i} e^{-n\lambda}}{\prod_{i=1}^{n} y_i!} \qquad P(\lambda) = \lambda^{\alpha-1} e^{-\beta\lambda} \frac{\beta^\alpha}{\Gamma(\alpha)}$$

$$P(\vec{y}) = \int_0^\infty P(\vec{y}|\lambda) \cdot P(\lambda) d\lambda = \frac{1}{\prod y_i!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \underbrace{\int \lambda^{\Sigma y_i + \alpha - 1} e^{(-n-\beta)\lambda} d\lambda}_{\text{Gamma}(\Sigma y_i + \alpha,\ \beta+n)}$$

$$= \frac{1}{\prod y_i!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\Sigma y_i + \alpha)}{(\beta+n)^{\Sigma y_i + \alpha}}$$

Example: traffic accidents

- $p(\mathbf{y} \mid M_2)$ is improper

$$\int p(y|M_2)\,dy = \iint p(y|\theta, M_2)\,p(\theta|M_2)\,d\theta\,dy$$

"Fubini" $$= \iint (\cdot)\,dy\,d\theta$$

$$= \int p(\theta|M_2) \underbrace{\int p(y|\theta, M_2)\,dy}_{1}\,d\theta$$

$$= \int p(\theta|M_2)\,d\theta = \infty$$

- Condition on y_1:
  - $p(\cancel{\theta}\ \lambda \mid M_1, y_1) = \text{Gamma}(y_1 + 3, 3)$
  - $p(\cancel{\theta}\ \lambda \mid M_2, y_1) = \text{Gamma}(y_1 + 1, 1)$

$$\hookrightarrow p(\lambda|M_2, y_1) \propto \frac{\lambda^{y_1} e^{-\lambda}}{y_1!} \cdot 1$$

$$\propto \text{Gamma}(y_1+1, 1)$$

Example: traffic accidents

- Compute partial Bayes factor, conditioned on y_1:
  - $p(y_2 = 6, y_3 = 2, y_4 = 4, y_5 = 2|M_1, y_1 = 3) = 0.000133$
  - $p(y_2 = 6, y_3 = 2, y_4 = 4, y_5 = 2|M_2, y_1 = 3) = 0.000224$
  - The partial BF for M1 over M2 is

$$BF(y_2, y_3, y_4, y_5 \mid y_1) = 0.596$$

http://www.ericfrazerlock.com/Model_Comparison_Rcode2.r

- Modest evidence that the elicited prior is not better than flat prior

Intrinsic Bayes factors

- Compute n partial Bayes factors:

$$BF(\{y_j\}_{j \neq i} \mid y_i)$$

  for $i = 1, \ldots, n$

- The average of these partial BFs is the *intrinsic Bayes factor*
  - Could take arithmetic or geometric average
  - If $BF(\{y_j\}_{j \neq i} \mid y_i)$ does not exist, condition on larger subsets instead

- The traffic accident example has arithmetic intrinsic Bayes factor 1.64.
  http://www.ericfrazerlock.com/Model_Comparison_Rcode2.r