Homework 8: Bayesian Statistics

Due Thursday, November 7, 2019

This homework will cement your understanding of the material in Rice Chapter 8, Section 8.6, as well as some of the concepts from the first few sections of Chapter 9. The assignment is due before the end of class on the date listed above. You may discuss problems with your classmates, but you must turn in your own writeup, and list all students with whom you discussed the assignment. You must show **all your work** in order to receive full credit. Stop by office hours if you need additional help or hints.

Problem 1 (Warmup: checking an integral)**.** Recall the Gamma function,

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt. \tag{1}$$

In lecture, we stated without proof that for α, ν both greater than 0,

$$f(\lambda \mid \alpha, \nu) = \begin{cases} \frac{\nu^\alpha \lambda^{\alpha-1} e^{-\nu\lambda}}{\Gamma(\alpha)} & \text{if } \lambda \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

was a probability density. Show that this is the case by verifying that

$$\int_0^\infty \lambda^{\alpha-1} e^{-\nu\lambda} d\lambda = \frac{\Gamma(\alpha)}{\nu^\alpha}.$$

Hint: try a change of variables and then use the definition of the Gamma function.

Problem 2 (Free Money: the Rao-Blackwell Theorem)**.** Suppose that $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$ is an estimator of θ, satisfying $\mathbb{E}_\theta \hat{\theta}^2 < \infty$ for all θ, where $\mathbb{E}_\theta$ denotes expectation with respect to the joint distribution of $X_1, X_2, \ldots, X_n \sim F_\theta$ (note that the observations need not be independent!). The Rao-Blackwell Theorem (Theorem A in Section 8.8.2 of Rice) states that if $T = T(X_1, X_2, \ldots, X_n)$ is a sufficient statistic for θ, then we can improve our estimator $\hat{\theta}$ by replacing it with $\tilde{\theta} = \mathbb{E}(\hat{\theta} \mid T)$.

(a) Read Theorem A in Section 8.8.2 of Rice. The main result of the theorem is that for all θ,

$$\mathbb{E}(\tilde{\theta} - \theta)^2 \leq \mathbb{E}(\hat{\theta} - \theta)^2,$$

and that this inequality is strict unless $\hat{\theta} = \tilde{\theta}$, which is to say,

$$\Pr\left[\hat{\theta}(X_1, X_2, \ldots, X_n) \neq \tilde{\theta}(X_1, X_2, \ldots, X_n)\right] = 0.$$

Rice goes on to describe the theorem as saying that "If an estimator is not a function of a sufficient statistic, then it can be improved." Explain in words (a sentence or two is fine) why Rice's summary is correct.

(b) Read the proof of Theorem A in Section 8.8.2 of Rice. Rice writes that

$$\mathbb{E}\tilde{\theta} = \mathbb{E}\left[\mathbb{E}(\hat{\theta} \mid T)\right] = \mathbb{E}\hat{\theta}.$$

(i) Explain why each of these two equalities is true (a sentence or two is fine).

(ii) In the term $\mathbb{E}\left[\mathbb{E}(\hat{\theta} \mid T)\right]$, there are two different expectations being taken. Explain what variable or variables each of these expectation is with respect to.

(c) Rice writes that

$$\operatorname{Var}\hat{\theta} = \operatorname{Var}\tilde{\theta} + \mathbb{E}[\operatorname{Var}(\hat{\theta} \mid T)],$$

and concludes from this that $\operatorname{Var}\hat{\theta} \geq \operatorname{Var}\tilde{\theta}$. Why is this true?

(d) The statement of the theorem requires that T be sufficient for θ. Look over the proof again. Where did we use this sufficiency? On first glance, it isn't entirely clear (at least not to me!). If you go and read about this theorem elsewhere (e.g., in David Blackwell's original paper[1] or on Wikipedia[2]), you will find that the important property is that our improved estimator $\tilde{\theta}$ be *observable*, by which we mean that we do not need to know the true value of θ to compute $\tilde{\theta} = \mathbb{E}[\hat{\theta} \mid T]$. Explain why sufficiency of T for θ implies that $\tilde{\theta}$ is observable.

(e) When I was taught this theorem, it was described to me as "free money", because it gives us a straightforward way to improve *any* estimator. Unfortunately, most of the time, either

(i) $\hat{\theta}$ is sufficient for θ, or

(ii) $\mathbb{E}[\hat{\theta} \mid T]$ is *really* hard to work out.

Explain (a sentence or two is fine) why either of these two possibilities renders the Rao-Blackwell Theorem less useful.

Problem 3 (Coinflips and the Beta Distribution)**.** Let $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli random variables with success parameter $\theta_0 \in [0, 1]$. Suppose that we have a prior distribution for the true value of the parameter given by a Beta distribution,

$$\pi(\theta; \alpha_{\text{Prior}}, \beta_{\text{Prior}}) = \frac{\theta^{\alpha_{\text{Prior}}-1}(1-\theta)^{\beta_{\text{Prior}}-1}}{B(\alpha_{\text{Prior}}, \beta_{\text{Prior}})},$$

where, $B(\alpha, \beta)$ denotes the Beta function[3] (note the distinction between the Beta distribution and the Beta function)

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

with $\Gamma(z)$ denoting the Gamma function as in Equation (1) above.

[1] Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*. 18(1): 105–110.

[2] `https://en.wikipedia.org/wiki/Rao-Blackwell_theorem`

[3] `https://en.wikipedia.org/wiki/Beta_function`

(a) Let $f_X(X_1, X_2, \ldots, X_n)$ denote the marginal distribution of the data. Note that since we are being Bayesian in this problem, θ is itself a random variable, and thus the marginal distribution for the data should *not* depend on θ or θ_0, though it *will* depend on our prior. Explain why

$$f_X(x_1, x_2, \ldots, x_n) = \int_0^1 \left(\prod_{i=1}^n t^{x_i}(1-t)^{1-x_i} \right) \pi(t; \alpha_{\text{Prior}}, \beta_{\text{Prior}}) dt,$$

and then use this to show that

$$f_X(x_1, x_2, \ldots, x_n) = \frac{B(\alpha_{\text{Prior}} + \sum_i x_i, \beta_{\text{Prior}} + n - \sum_i x_i)}{B(\alpha_{\text{Prior}}, \beta_{\text{Prior}})}. \tag{2}$$

(b) For simplicity, let $k = \sum_i x_i$. Use Bayes' theorem and Equation (2) to show that

$$f_{\theta|X_1,\ldots,X_n}(\theta \mid x_1, x_2, \ldots, x_n) = \frac{\theta^{\alpha_{\text{Prior}}+k-1}(1-\theta)^{\beta_{\text{Prior}}+n-k-1}}{B(\alpha_{\text{Prior}} + k, \beta_{\text{Prior}} + n - k)}.$$

That is, the posterior distribution of θ given the data is itself a Beta random variable, with parameters $\alpha_{\text{Post}} = \alpha_{\text{Prior}} + \sum_i x_i$ and $\beta_{\text{Post}} = \beta_{\text{Prior}} + n - \sum_i x_i$. You should recognize this property from lecture—the Beta distribution is the conjugate prior of the Bernoulli distribution.

(c) Suppose that $n \geq 2$. Show that, provided $k \geq 1$, the maximum a posteriori estimate for θ is

$$\hat{\theta}_{\text{MAP}} = \frac{\alpha_{\text{Post}} - 1}{\alpha_{\text{Post}} + \beta_{\text{Post}} - 2}.$$

Hint: taking the first derivative and setting equal to zero is straightforward. To verify that this is a maximum, use the fact that since α_{Prior} and β_{Prior} are both greater than 0 (by definition of the Beta distribution), α_{Post} and β_{Post} are both greater than 1.

(d) Pierre Simon Laplace[4] was a French mathematician and philosopher (back then, the distinction between math, science and philosophy was a bit hazier), credited as being among the inventors of probability theory [5]. Laplace is famous for his *rule of succession*[6]. Laplace was interested in the probability that the sun will rise tomorrow, given that it has been observed to rise for the past few thousand years. Abstracting this idea, Laplace argued that if we have $n+1$ independent Bernoulli trials with success probability θ that are independent conditional on θ, then conditioned on the fact that the first n trials are all successes, the probability that the $(n+1)$-th trial is a success is

$$\Pr[X_{n+1} = 1 \mid X_1 = X_2 = \cdots = X_n = 1] = \frac{n+1}{n+2}.$$

Show that Laplace's rule of succession corresponds to a Beta prior with parameters $\alpha_{\text{Prior}} = \beta_{\text{Prior}} = 1$. Pretty good for a philosopher from 200 years ago!

[4] `https://en.wikipedia.org/wiki/Pierre-Simon_Laplace`

[5] I highly recommend Keith Devlin's (no relation) book, *The Unfinished Game: Pascal, Fermat, and the Seventeenth-Century Letter that Made the World Modern*, if you want to learn a bit about the origins of probability theory and how it connects to modern mathematics, insurance, gambling, etc.

[6] `https://en.wikipedia.org/wiki/Rule_of_succession`

(e) If Laplace were alive today, he would have lived for about 71600 days, and thus seen about 71600 successful sunrises. Use Laplace's rule of succession to compute what would be Laplace's subjective probability that the sun will rise tomorrow given that he has seen the sun rise every morning since his birth (you are of course free to use a calculator).

Problem 4 (Bayesian estimation of the normal)**.** Consider the normal distribution, with mean μ and variance σ^2. In what follows, it will be easier to work with the *precision*, $\xi = 1/\sigma^2$ (you should be able to convince yourself that this is an appropriate name for the inverse of the variance—when σ^2 is small, our observations will be close to the mean). Then the normal distribution has a likelihood given by

$$f_{x|\mu,\xi}(x \mid \mu, \xi) = \sqrt{\frac{\xi}{2\pi}} \exp\left\{ \frac{-\xi}{2}(x-\mu)^2 \right\}.$$

For this problem, let us assume that ξ is known, say $\xi = \xi_0$, (See Example B in Section 8.6 of Rice for the case where ξ is also unknown). Suppose that we wish to estimate $\mathbb{E}X = \mu_0$ and our prior f_μ takes the form of a normal $\mathcal{N}(\mu_{\text{Prior}}, \xi_{\text{Prior}}^{-1})$, where μ_{Prior} is our a priori guess as to the mean μ_0. Note that ξ_{Prior} is the inverse of a variance. Thus, if ξ_{Prior} is large, our prior is tightly centered around μ_{Prior}, meaning we are more certain that the true value of the mean is near μ_{Prior}. On the other hand, if ξ_{Prior} is small, our prior distribution will be comparatively "flat", encoding our prior uncertainty about the true mean. In what follows, for ease of notation, we will use X to denote the data, $X_1, X_2, \ldots, X_n$.

1. Use the fact that

$$f_{\mu|X}(\mu \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \propto f_{X|\mu}(x_1, x_2, \ldots, x_n \mid \mu) f_\mu(\mu)$$

to show why

$$f_{\mu|X}(\mu \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \propto \sqrt{\frac{\xi_0}{2\pi}} \exp\left\{ \frac{-\xi_0}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \sqrt{\frac{\xi_{\text{Prior}}}{2\pi}} \exp\left\{ \frac{-\xi_{\text{Prior}}}{2}(\mu - \mu_{\text{Prior}})^2 \right\}.$$

Hint: recall that $\propto$ denotes proportionality, where we ignore scalar multiplicative terms that do not depend on the parameter of interest (in our case, μ).

2. Continuing along the same lines, show that

$$f_{\mu|X}(\mu \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \propto \exp\left\{ \frac{-1}{2} \left[\xi_0 \sum_{i=1}^n (x_i - \bar{x})^2 + \xi_0 n(\bar{x} - \mu)^2 + \xi_{\text{Prior}}(\mu - \mu_{\text{Prior}})^2 \right] \right\},$$

where $\bar{x} = n^{-1} \sum_i x_i$.

3. Since the $\xi_0 \sum_i (x_i - \bar{x})^2$ term does not depend on μ, we have shown that

$$f_{\mu|X}(\mu \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \propto \exp\left\{ \frac{-1}{2} Q(\mu) \right\},$$

where

$$Q(\mu) = \xi_0 n(\bar{x} - \mu)^2 + \xi_{\text{Prior}}(\mu - \mu_{\text{Prior}})^2.$$

Define $\xi_{\text{Post}} = n\xi_0 + \xi_{\text{Prior}}$ and

$$\mu_{\text{Post}} = \frac{n\bar{x}\xi_0 + \xi_{\text{Prior}}\mu_{\text{Prior}}}{n\xi_0 + \xi_{\text{Prior}}}.$$

Show, by rearranging terms and completing the square, that $Q(\mu)$ can be written as

$$Q(\mu) = \xi_{\text{Post}}(\mu - \mu_{\text{Post}})^2 + Z,$$

where Z does not depend on μ.

4. Use the above parts, along with the fact that $f_{\mu|X}$ is a probability, to conclude that

$$f_{\mu|X}(\mu \mid X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = C \exp\left\{\frac{-\xi_{\text{Post}}}{2}(\mu - \mu_{\text{Post}})^2\right\},$$

where $C > 0$ is some constant, not depending on μ (but possibly depending on $\xi_0, \xi_{\text{Prior}}, \mu_{\text{Prior}}$ and the data). That is, our posterior distribution for μ_0 is a normal. What are the mean and precision of this normal?

Problem 5 (Adapted from Rice, Problem 9.1)**.** A coin with Pr[Heads] $= p_0 \in [0, 1]$ is tossed n times to test the hypothesis $H_0 : p_0 = 1/2$ against the alternative $H_A : p_0 \neq 1/2$. Suppose that our test rejects H_0 if the number of heads is at most k or at least $n - k$, where $k \leq n/2$ is an integer.

1. Is H_0 a simple hypothesis, or composite? What about H_A?
2. Write, as a function of n and k, the significance level of the test.
3. Write, as a function of n, k and p, the power of the test when $p_0 = p$.
4. When $n = 10$ and $k = 0$, what is the significance level of the test?
5. If, under the same setting of $n = 10, k = 0$, the true probability of heads is $p_0 = 1/10$, what is the power of the test?