STATISTICS 225 – Fall 2018

Homework 1 - solutions

HW is graded at 5 points per problem. I will try to provide detailed solutions. I will provide my R code for solving the computational problems – you should not assume that my code is the most efficient way to get the answer.

1. **Loss functions** – This was reasonably straightforward. One minor point is that you should check that your solution does, in fact, **minimize** the expected loss.

   (a) $L(\theta, a) = (\theta - a)^2$:

   $$\begin{aligned} E(L|y) &= \int (\theta - a)^2 p(\theta|y) d\theta \\ \frac{dE(L|y)}{da} &= -2\int (\theta - a) p(\theta|y) d\theta = 2a - 2E(\theta|y) \end{aligned}$$

   Setting the first derivative to zero yields $a = E(\theta|y)$. Note that the 2nd derivative is positive so this is a minimizer.

   (b) $L(\theta, a) = |\theta - a| = (\theta - a)I(\theta \geq a) + (a - \theta)I(\theta < a)$:

   $$\begin{aligned} E(L|y) &= \int_{-\infty}^{a} (a - \theta) p(\theta|y) d\theta + \int_{a}^{\infty} (\theta - a) p(\theta|y) d\theta \\ \frac{dE(L|y)}{da} &= \int_{-\infty}^{a} p(\theta|y) d\theta - \int_{a}^{\infty} p(\theta|y) d\theta \\ &= 2\int_{-\infty}^{a} p(\theta|y) d\theta - 1 \quad (\text{ add and subtract } \int_{-\infty}^{a} p(\theta|y) d\theta) \end{aligned}$$

   Note that the derivatives here are complicated because a appears in the limits of integration and in the integrand. Fortunately all the messy stuff drops out. Setting the derivative to zero indicates that we should find a such that posterior cdf is 1/2, i.e., a posterior median. The second derivative is not informative here but we can see that the first derivative is negative for smaller values of a and positive for bigger values of a which makes this a minimum.

2. **Prior distributions for the Poisson**

   (a) The calculations below demonstrate that the kernel of the posterior distribution is that of a Gamma($\sum_i y_i + \alpha, n + \beta$) distribution so the gamma is a conjugate prior distribution.

   $$p(\lambda|y) \propto p(y|\lambda)p(\lambda) \propto \lambda^{\sum_i y_i} e^{-n\lambda} \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda^{\sum_i y_i + \alpha - 1} e^{-(n+\beta)\lambda}$$

   (b) There is not a unique right answer. In the conjugate (Gamma) family, the mean of the distribution is α/β. The expert tells us that they would guess the rate is around 1 which suggests choosing $\alpha = \beta$. The expert also believes that it would not be surprising to find the rate is between (.1 and 10). Thus we want most of our prior probability in this range (how much? Perhaps .80 or .90 or .95.). You can play around with the gamma distribution in R. **pgamma** is the cdf function. It takes a shape parameter and a scale or rate parameter. The form of the density that I've given above has β as a rate parameter (and thus $1/\beta$ as a scale parameter). The Gamma(1,1) distribution has probability .9 between .1 and 10 so may be appropriate. Many of you noticed though that just about all of the probability here is well less than 10 so this prior essentially rules out values bigger than 10. Choosing a smaller shape parameter (say .1) provides for a longer tail. Unfortunately it also puts a lot of probability near zero.

   I played around a bit (more than I would expect you to) and discovered that $\alpha = \beta$ may not be ideal after all. A gamma distribution with shape $\alpha = .7$ and inverse scale (rate) $\beta = .3$ has .025 percentile equal to .014, median equal to 1.27 and .975 percentile equal to 8.92. That comes somewhat close to the expert's views.

   (c) Noninformative prior distributions

      i. The flat prior distribution is not a proper distribution since it does not have a finite integral over the positive half of the real line. If $p(\lambda) \propto 1$, then $p(\lambda|y) \propto \lambda^{\sum_i y_i} e^{-n\lambda}$ which is the kernel of a Gamma($\sum_i y_i + 1, n$) distribution. Thus the flat prior leads to a proper posterior distribution.

ii. To find Jeffrey's prior we need the Fisher information. We need to compute $E(-\frac{d^2 logp(y|\lambda)}{d\lambda^2}|\lambda)$. The log of the Poisson distribution is $\sum_i y_i \log\lambda - n\lambda$. The negative 2nd derivative is $\sum_i y_i/\lambda^2$. Finally $J(\lambda) = E(\sum_i y_i/\lambda^2|\lambda) = n/\lambda$. Jeffrey's prior is proportional to the square root of the Fisher information which is $\lambda^{-1/2}$.

iii. If $\phi = \sqrt{\lambda}$, then the inverse transformation is $\lambda = \phi^2$. The density of ϕ is derived as

$$p_\phi(\phi) = p_\lambda(\phi^2)|d\lambda/d\phi| \propto (\phi^2)^{-1/2}2\phi \propto \text{constant}$$

iv. The Jeffreys' prior, $p(\lambda) \propto 1/\sqrt{\lambda}$, is not a proper prior distribution. Once again it does not integrate to a finite number on the positive half of the real line.

v. If we use $p(\lambda) \propto 1/\sqrt{\lambda}$, then the posterior distribution is Gamma$(\sum_i y_i + 1/2, n)$ which is a proper distribution. (My hint was incorrect. Sorry about that.)

3. **Weibull (conjugate case)**:

(a) The calculations below demonstrate that the kernel of the posterior distribution is that of a Gamma$(n+\alpha, \sum_i y_i^2 + \beta)$ distribution so the gamma is a conjugate prior distribution.

$$p(\theta|y) \propto p(y|\theta)p(\theta) \propto \theta^n e^{-\sum_i y_i^2\theta}\theta^{\alpha-1}e^{-\beta\theta} = \theta^{n+\alpha-1}e^{-(\sum_i y_i^2+\beta)\theta}$$

(b) The problem statement tells you that the y_i's are conditionally independent given θ. Now consider the marginal distribution (where I'm no longer ignoring constants!):

$$\begin{aligned} p(y) &= \int p(y|\theta)p(\theta)d\theta = \int 2^n\theta^n(\prod_i y_i)e^{-\sum_i y_i^2\theta}\frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}d\theta \\ &= 2^n\prod_i y_i\frac{\beta^\alpha}{\Gamma(\alpha)}\int \theta^{\alpha+n-1}e^{-\theta(\beta+\sum_i y_i^2)}d\theta = \frac{2^n(\prod_i y_i)\beta^\alpha\Gamma(\alpha+n)}{(\beta+\sum_i y_i^2)^{n+\alpha}\Gamma(\alpha)} \end{aligned}$$

where the final equality comes from completing the gamma density. The marginal can not be factored as a product of the marginal distribution of each y_i hence the y_i's are not independent in the marginal distribution.

(c) R commands/programs and output are provided on a separate sheet. A key point is that the Gamma(1.4,2) prior distribution appears to concentrate on θ from 0 to 3 (95% central prior interval runs from .04 to 2.24; 99% central prior interval runs from .01 to 3.10). Note that $\theta = 3$ corresponds to a Weibull distribution for which the mean of y is equal to 0.5 (the lower limit suggested by our prior information) and that $\theta = .03$ corresponds to a Weibull distribution with mean of y equal to 5 (the upper limit suggested by our prior information). Thus the chosen prior makes sense.

(d) Here we can use theory rather than simulation. The posterior distribution of θ is Gamma$(11.4, 11.377)$ which has mean 1.002 and variance 0.088. The 95% posterior interval for θ (see output) is $(.507, 1.662)$.

(e) The best prediction of the lifetime of a new part depends on the loss function (as we showed in the first problem). I didn't specify a loss function here. Most people usually think of the posterior mean as the natural estimate. Here we actually want the posterior mean of the expected lifetime $E(Y)$ which is a function of θ (recall that $E(Y) = .886\theta^{-0.5}$). This is easiest to obtain by simulation. I simulated 1000 draws from the posterior distribution of θ, computed $.886\theta^{-0.5}$ for each one, and then took the mean of these values to come up with an estimate of .9155. To get a predictive interval, simulate 1000 draws from the posterior distribution of θ and then for each draw simulate a random variable y from the Weibull distribution with that θ (see code). The predictive interval from my simultations is $(.139, 2.086)$.

4. **Mixtures of conjugate priors**

(a) From the definition:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int p(y|\theta)p(\theta)} = \frac{\sum_{m=1}^k \lambda_m p(y|\theta)p_m(\theta)}{\sum_{m=1}^k \int \lambda_m p(y|\theta)p_m(\theta)d\theta} = \frac{\sum_{m=1}^k \lambda_m p_m(y)p_m(\theta|y)}{\sum_{m=1}^k \lambda_m p_m(y)}$$

where the last equality comes from recognizing that $p(y|\theta)p_m(\theta) = p_m(\theta|y)p_m(y)$ and that $p_m(y) = \int p(y|\theta)p_m(\theta)d\theta$. Notice that the final expression is a mixture of $p_m(\theta|y), m = 1, \ldots, k$ which is of the same form as the prior (since each $p_m(\theta|y)$ is of the same form as $p_m(\theta)$).

(b) From the last expression it is clear that the mixture proportion for component m is $\lambda_m p_m(y)/(\sum_k \lambda_k p_k(y))$.

(c) i. This prior distribution reflects the facts provided in the introductory paragraph. Specifically, normal coins (approximately 80% of the total) land heads more often than tails (75/25) which is conveyed in the Beta(6,2) mixture component. The other 20% are counterfeit and land tails more than heads (conveyed in the Beta(2,6) mixture component). The prior distribution is displayed in the accompanying R output.

(c) ii. The sampling distribution is binomial ... $p(y|\theta) \propto \theta^y(1-\theta)^{n-y}$. If $p_m(\theta) = \text{Beta}(a,b)$, then $p_m(\theta|y) = \text{Beta}(a+4, b+6)$ and $p_m(y) = \binom{10}{4}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+4)\Gamma(b+6)}{\Gamma(a+b+10)}$. From parts (a) and (b) the posterior is a mixture of the Beta posteriors that correspond to these data with mixing proportions propotional to $\lambda_m p_m(y)$. Let $c_1 = \binom{10}{4}\frac{\Gamma(8)}{\Gamma(6)\Gamma(2)}\frac{\Gamma(10)\Gamma(8)}{\Gamma(18)} = 0.04535171$ and $c_2 = \binom{10}{4}\frac{\Gamma(8)}{\Gamma(2)\Gamma(6)}\frac{\Gamma(6)\Gamma(12)}{\Gamma(18)} = 0.1187783$. Then

$$p(\theta|y) = \left(0.8c_1\text{Beta}(10,8) + 0.2c_2\text{Beta}(6,12)\right)/(0.8c_1 + 0.2c_2) = .604\ \text{Beta}(10,8) + .396\ \text{Beta}(6,12)$$

The posterior distribution is now peaked around 0.5 representing the compromise between our prior opinion ($\theta \approx 0.75$) and the data (which suggests $\theta \approx 0.4$).

(c) iii. Let $\tilde{y} = 1$ if the next toss is a head and zero otherwise. We want $\Pr(\tilde{y} = 1|y) = \int p(\tilde{y} = 1|\theta, y)p(\theta|y)d\theta = \int \theta p(\theta|y)d\theta$ where the last equality indicates that given θ the probability of heads on the next trial is θ regardless of what data y are observed. The last calculation is $\Pr(\tilde{y} = 1|y) = .604 * (10/18) + .396 * (6/18) = 0.47$.