

Week 11 Lecture: Concept Check Exercises

Bayesian Methods and Regression

1. Recall that the gamma function $\Gamma(x)$ is defined to be

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

We say that $X \sim \text{Gamma}(a, b)$ (for $a, b > 0$) if the PDF f_X is given by

$$f_X(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

for $x > 0$ and 0 otherwise. Suppose X has prior distribution $\text{Gamma}(a, b)$ and, conditional on X , the distribution of Y is Poisson with parameter X . What is the posterior distribution of X given Y ?

Solution. Note that

$$f_{X|Y}(x|y) \propto f_{Y|X}(y|x) f_X(x) \propto e^{-x} \frac{x^y}{y!} x^{a-1} e^{-bx} \propto x^{y+a-1} e^{-(b+1)x}$$

for $y \in \mathbb{Z}_{\geq 0}$, $x > 0$. Thus $X|Y = y \sim \text{Gamma}(y + a, b + 1)$ and has PDF

$$f_{X|Y}(x|y) = \frac{(b+1)^{y+a}}{\Gamma(y+a)} x^{y+a-1} e^{-(b+1)x}.$$

Note that the terms $(b+1)^{y+a}$ and $\Gamma(y+a)$ were determined without any calculations by comparing with the form of the Gamma PDF. This shows that the Gamma distribution is a conjugate prior for the Poisson likelihood.

2. (From DeGroot and Schervish) Let θ denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of θ is unknown, and two statisticians A and B assign to θ the following different prior PDFs $\xi_A(\theta)$ and $\xi_B(\theta)$, respectively:

$$\begin{aligned} \xi_A(\theta) &= 2\theta & \text{for } 0 < \theta < 1, \\ \xi_B(\theta) &= 4\theta^3 & \text{for } 0 < \theta < 1. \end{aligned}$$

In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

- (a) Find the posterior distribution that each statistician assigns to θ .

- (b) Find the Bayes estimate of θ (minimizer of posterior expected loss) for each statistician based on the squared error loss function.
- (c) Show that after the opinions of the 1000 registered voters in the random sample had been obtained, the Bayes estimates for the two statisticians could not possibly differ by more than 0.002, regardless of the number in the sample who were in favor of the proposition.

Solution. Note that both prior distributions are from the Beta family.

- (a) We have

$$\xi_A(\theta|x) \propto f(x|\theta)\xi_A(\theta) \propto \theta^{711}(1-\theta)^{290}$$

and

$$\xi_B(\theta|x) \propto f(x|\theta)\xi_B(\theta) \propto \theta^{713}(1-\theta)^{290}.$$

Thus the posteriors from A and B are both beta with parameters $(712, 291)$ and $(714, 291)$, respectively.

- (b) The respective means are $\frac{712}{1003}$ and $\frac{714}{1005}$.
- (c) In general the two means are given by

$$\frac{a+2}{1003} \quad \text{and} \quad \frac{a+4}{1005}.$$

The difference is less than $2/1000 = .002$.

3. Two statistics students decide to compute 95% confidence intervals for the distribution parameter θ using an i.i.d. sample X_1, \dots, X_n . Student B uses Bayesian methods to find a 95% credible set $[L_B, R_B]$ for θ . Student F uses frequentist methods to find a 95% confidence interval $[L_F, R_F]$ for θ . Both conclude that parameter θ is in their respective intervals with probability at least .95. Who is correct? Explain.

Solution. The frequentist student is totally incorrect, since they have misunderstood what a frequentist confidence interval is. Using frequentist methodology, θ is not a random variable, so it doesn't make sense to say it lies in some fixed interval $[L_F, R_F]$. The correct interpretation is that if independent experiments like this were repeated, then at least 95% of the time $[L_F, R_F]$ will contain θ . That is, the interval is random not θ .

We can say that the Bayesian student is consistent. Recall that to compute the credible set, the Bayesian student had to introduce some prior distribution π on θ . What we can say is if someone believes π is correct, then it is rational, given the data, to conclude that θ will lie in the posterior credible set with probability 95%.

4. Suppose θ has prior distribution $\text{Beta}(a, b)$ for some $a, b > 0$. Given θ , suppose we make independent coin flips with heads probability θ . Find values of a, b and the coin

flips so that the posterior variance is larger than the prior variance. [Hint: Recall that a Beta(a, b) random variable has variance given by

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

Try $b = 1$.]

Solution. As hinted, let's try $a = 10$, $b = 1$ and 9 coin flips all landing tails. The prior variance is given by

$$\frac{10 \cdot 1}{(10+1)^2(10+1+1)} = \frac{5}{726} \approx .0069$$

while the posterior variance is given by

$$\frac{10 \cdot 10}{(10+10)^2(10+10+1)} = \frac{1}{84} \approx .0119.$$

5. Fix $\sigma^2 > 0$. Let w , taking values in \mathbb{R}^d , have prior distribution $\mathcal{N}(\mu_0, \Sigma_0)$. Conditional on w and $x_1, \dots, x_n \in \mathbb{R}^2$ suppose that y_1, \dots, y_n are i.i.d. with $y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$. Let $\mathcal{N}(\mu_1, \Sigma_1)$ denote the posterior distribution of w given the data $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$.

- (a) Given a new x -value you want to forecast y to minimize the expected square loss. That is, we want to find

$$\hat{y} = \arg \min_y \mathbb{E}_{y'} (y - y')^2,$$

where y' has the predictive distribution given x and \mathcal{D} . What is \hat{y} , and what is the associated expected loss $\mathbb{E}_{y'} (\hat{y} - y')^2$?

- (b) What types of values for σ , Σ_0 , n will lead to the prior exerting a lot of influence on our prediction?
- (c) We saw that the Bayesian approach to Gaussian linear regression corresponds to ridge regression. What values in the Bayesian approach correspond to a large amount of regularization?

Solution.

- (a) We have $\hat{y} = \mu_1^T x$ with expected loss $x^T \Sigma_1 x + \sigma^2$, the mean and variance of the predictive distribution.
- (b) i. High σ meaning low certainty in data.
 ii. Small Σ_0 meaning high certainty in prior. A covariance matrix is small if its eigenvalues are small.
 iii. Small n meaning not a lot of data to learn from.
- (c) Small Σ_0 meaning high certainty in prior.

6. Suppose you are using Bayesian techniques to fit a Poisson regression model. Conditional on x, w , we have $y \sim \text{Pois}(e^{w^T x})$. A colleague, working with his own data set and prior, has given you a function f that returns i.i.d. samples from his posterior distribution on w . Give pseudocode that, given x , lets you sample from the predictive distribution of y given x .

Solution. Pseudocode follows:

- (a) Draw w from f .
- (b) Draw y from $\text{Pois}(e^{w^T x})$.
- (c) Return y .