DS102 - Discussion 4

Wednesday, 23rd September, 2020

In this discussion we will investigate more examples of *conjugate priors*, that is, pairs of distributions (for the likelihood and the prior) such that the prior and posterior are from the same distribution, with possibly different parameters.

Recall that for observed data X, and prior distribution $p(\theta)$ on parameters θ, the *posterior probability* distribution on θ, after seeing the data, is given by[1]

$$\begin{aligned} p(\theta|x) &= \frac{p(x|\theta) \cdot p(\theta)}{p(x)} \\ &\propto p(x|\theta) \cdot p(\theta) \end{aligned}$$

where $\propto$ denotes "proportional to." We can always work this proportionality and solve for the proportionality constant at the end.

1. (Beta and Binomial) Say you've observed a sequence of coin flips, $X_1, ..., X_n$, all using the same coin, which has some probability of landing heads, p_h. Denote by H the total number of heads:

$$H = \sum_{i=1}^{n} \mathbb{I}\{X_i = \text{heads}\}$$

H follows a binomial distribution, with PDF

$$p(H = k) = \binom{n}{k} p_h^k (1 - p_h)^{n-k}$$

We didn't make this coin, it was given to us. We're willing to place a prior distribution on the probability of it landing heads, and we'll use the beta distribution to do so (for reasons we'll investigate). The beta distribution PDF is parameterized by shape parameters $\alpha > 0$ and $\beta > 0$, and is given by

$$f(z; \alpha, \beta) = \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} z^{\alpha-1}(1 - z)^{\beta-1}, \quad 0 < z < 1$$

(a) Show that the beta is a conjugate prior for the binomial distribution. What are the shape parameters for the posterior distribution?

Solution:

$$\begin{aligned} f(k|p_h) \cdot f(p_h; \alpha, \beta) &= \binom{n}{k} p_h^k (1 - p_h)^{n-k} \cdot \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} p_h^{\alpha-1}(1 - p_h)^{\beta-1} \\ &= \frac{n!}{(n-k)!k!} \cdot \frac{(\alpha + \beta - 1)!}{(\alpha - 1)!(\beta - 1)!} (p_h)^{k+\alpha-1}(1 - p_h)^{n-k+\beta-1} \end{aligned}$$

[1] The *prior* distribution on the parameters is given by $p(\theta)$ and the likelihood $p(x|\theta)$.

Where the second line follows from rearranging terms. In terms of p_h, the first two fractions are constant, so we can write

$$\begin{aligned} f(p_h|H=k) &\propto f(k|p_h)\cdot f(p_h;\alpha,\beta) \\ &\propto (p_h)^{k+\alpha-1}(1-p_h)^{n-k+\beta-1} \end{aligned}$$

From this we conclude that the posterior distribution has a beta distribution with shape parameters $(k+\alpha)$ and $(n-k+\beta)$.

(b) Now that we've gone through the mechanics, let's take a closer look at the beta distribution and its parameters. In particular, assume $\beta > 1$ and $\alpha > 1$.
(i) When $\alpha > \beta$, are small z (closer to zero) or large z (closer to 1) more likely under $f(z;\alpha,\beta)$? What about $\alpha = \beta$?

Solution: When $\alpha > \beta$, larger z are more likely. When $\beta > \alpha$, the reverse is true.

(ii) What is special about beta(1,1)?

Solution: beta(1,1) is the uniform distribution.

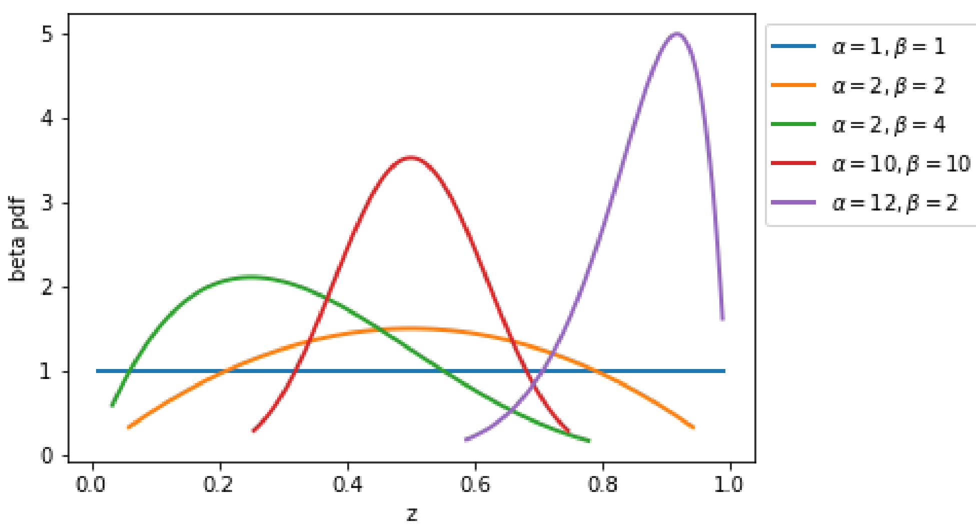

Figure 1: Beta distributions for various parameters

(c) Interpret the posterior distribution as an update to the prior distribution, after having seen the data. To get you started, suppose you started with a beta distribution prior on p_h, with $\alpha = 1, \beta = 1$, then after $n = 10$ flips, you observe $k = 5$ heads. What if there were $k = 8$ heads?

Solution: The posterior probability has parameters $\alpha_2 = k + \alpha$ and $\beta_2 = n - k + \beta$. If the prior has parameters $\alpha = 1, \beta = 1$, and half of the 10 flips are heads, then the resulting posterior distribution is beta(6,6), which is symmetric about 0.5, the observed mean of heads. If 8 of 10 flips are heads, the posterior distribution is beta(9,3), which puts more probability on z close to 1 (and is not symmetric).

What if the prior had set $\alpha = 2, \beta = 8$?

Solution: In this case, the posterior after seeing 5 heads is beta(7,13), and after seeing 8 heads would be beta(10,10).

Now comment more broadly on what you observe.

Solution: One way to think about this prior is as if you started the process with a phantom $\alpha - 1$ head flips and $\beta - 1$ tail flips, and proceeded from there. As we see more flips (bigger n), we are able to sway the distribution toward the opposite of the prior, if that is what the data suggests. Also as $n + \alpha + \beta$ gets big, the variance in the posterior distribution decreases, with a single mode in the distribution.

2. (Gaussian and Gaussian)

   Show that Gaussian is conjugate prior to itself with fixed variance, i.e. if $X \sim \mathcal{N}(\mu, \sigma_0^2)$, $\mu \sim \mathcal{N}(\mu_1, \sigma_1^2)$ follows two Gaussian distributions, where $\mu_1, \sigma_0, \sigma_1$ are constants, then $\mu|X$ follows a Gaussian distribution with new mean μ^* and σ^*.

   Solution: We have

$$
\begin{aligned}
p(\mu|X) &\propto p(X|\mu)p(\mu) \\
&\propto \exp(-\frac{(X-\mu)^2}{2\sigma_0^2} - \frac{(\mu-\mu_1)^2}{2\sigma_1^2}) \\
&\propto \exp(-\frac{(\mu-\mu^*)^2}{2\sigma^{*2}}),
\end{aligned}
$$

   where $\mu^* = \mu_1 + \frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2}(X - \mu_1)$, $\sigma^{*2} = \frac{\sigma_0^2\sigma_1^2}{\sigma_0^2+\sigma_1^2}$.

3. (Gamma and Exponential)

   A gamma distribution with parameters α, β has density function $p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ where $\Gamma(t)$ is the gamma function (see `https://en.wikipedia.org/wiki/Gamma_distribution`).

Show that gamma distribution is a conjugate prior for exponential distribution for multiple measurements, i.e. if we have samples $X_1, X_2, \cdots, X_n$ that are mutually independent given λ, and each $X_i \sim Exp(\lambda)$ and $\lambda \sim Gamma(\alpha, \beta)$, then $\lambda | X_1, X_2, \cdots, X_n \sim Gamma(\alpha^*, \beta^*)$ for some values α^*, β^*.

Solution:

$$\begin{aligned} p(\lambda | X_1, X_2, \cdots, X_n) &\propto p(X_1, X_2, \cdots, X_n | \lambda) p(\lambda) \\ &\propto \lambda^n e^{-\lambda \sum_i X_i} \lambda^{\alpha-1} e^{-\beta\lambda} \\ &\propto \lambda^{n+\alpha-1} e^{-(\sum_i X_i + \beta)\lambda}. \end{aligned}$$

So $\lambda | X_1, X_2, \cdots, X_n \sim Gamma(\alpha^*, \beta^*)$ with $\alpha^* = n + \alpha$, $\beta^* = \sum_i X_i + \beta$.