SI140 Discussion 12

Li Zeng, Tao Huang, Xinyi Liu

ShanghaiTech University, China
{zengli,huangtao1,liuxy10}@shanghaitech.edu.cn

1 Bayes Inference

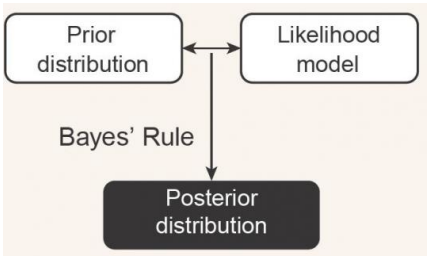

Fig. 1. The basic framework of bayes inference

Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more evidence or information becomes available. Bayesian inference derives the posterior probability as a consequence of two antecedents: a prior probability and a "likelihood function" derived from a statistical model for the observed data. Bayesian inference computes the posterior probability according to Bayes' theorem:

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)} \sim P(E \mid H) \cdot P(H)$$

With a conjugate prior the posterior is of the same type, e.g. for binomial likelihood the beta prior becomes a beta posterior. Conjugate priors are useful because they reduce Bayesian updating to modifying the parameters of the prior distribution (so-called hyperparameters) rather than computing integrals.

2 Beta-Binomial Conjugacy

2.1 Beta Distribution

From the last discussion, we summarise the basic of Beta distribution:

The Probability Density Function (PDF) for a Beta $X \sim \text{Beta}(a, b)$ is:

$$f(X = x) = \begin{cases} \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$$

$$E[X] = \frac{a}{a+b} \qquad \text{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

It has following properties:

- Beta Integral: $\beta(a, b) = \int_0^1 x^{a-1}(1-x)^{b-1} dx$
- Story. Bayes' billiards: For any integers k and n with $0 \le k \le n$: $\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}$

2.2 Beta-Binomial Conjugacy

Beta distribution is a conjugate prior for the binomial distribution. This means that if the likelihood function is binomial and the prior distribution is beta then the posterior is also beta.

Exercise 1 (Data Update!). Suppose that the likelihood follows a binomial (N, θ) distribution where N is known and θ is the (unknown) parameter of interest. We also have that the data x from one trial is an integer between 0 and N. Then fill in the blanks in the table below.

	hypothesis	data	prior	likelihood	posterior
distribution	θ	x	beta (a, b)	binomial (N, θ)	
formulation	θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$c_2\theta^x(1-\theta)^{N-x}$	

Solution 1.

	hypothesis	data	prior	likelihood	posterior
distribution	θ	x	beta (a, b)	binomial (N, θ)	beta $(a+x, b+N-x)$
formulation	θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$c_2\theta^x(1-\theta)^{N-x}$	$c_3\theta^{a+x-1}(1-\theta)^{b+N-x-1}$

2.3 *Beta-Geometric Conjugacy

Recall that the geometric(θ) distribution describes the probability of x successes before the first failure, where the probability of success on any single independent trial is θ. The corresponding pmf is given by $p(x) = \theta^x(1-\theta)$

Now suppose that we have a data point x, and our hypothesis θ is that x is drawn from a geometric (θ) distribution. From the table we see that the beta distribution is a conjugate prior for a geometric likelihood as well:

hypothesis	data	prior	likelihood	posterior
θ	x	beta (a, b)	geometric (θ)	beta $(a+x, b+1)$
θ	x	$c_1\theta^{a-1}(1-\theta)^{b-1}$	$\theta^x(1-\theta)$	$c_3\theta^{a+x-1}(1-\theta)^b$

Exercise 2. While traveling through the Mushroom Kingdom, Mario and Luigi find some rather unusual coins. They agree on a prior of $f(\theta) \sim$ beta (5,5) for the probability of heads, 18.05 class 15, Conjugate priors: Beta and normal, Spring 2014 3 though they disagree on what experiment to run to investigate θ further.

(a) Mario decides to flip a coin 5 times. He gets four heads in five flips.
(b) Luigi decides to flip a coin until the first tails. He gets four heads before the first tail.

Show that Mario and Luigi will arrive at the same posterior on θ, and calculate this posterior.

Solution 2. We will show that both Mario and Luigi find the posterior polf for θ is a beta (9,6) distribution.
Mario's table

Luigi's table

since both Mario and Luigi's posterior has the form of a beta (9,6) distribution that's what they both must be. The normalizing factor is the same in both cases because it's determined by requiring the total probability to be 1 .

hypothesis	data	prior	likelihood	posterior
θ	$x=4$	beta (5,5)	binomial $(5,\theta)$	???
θ	$x=4$	$c_1\theta^4(1-\theta)^4$	$\binom{5}{4}\theta^4(1-\theta)$	$c_3\theta^8(1-\theta)^5$

hypothesis	data	prior	likelihood	posterior
θ	$x=4$	beta (5,5)	geometric (θ)	???
θ	$x=4$	$c_1\theta^4(1-\theta)^4$	$\theta^4(1-\theta)$	$c_3\theta^8(1-\theta)^5$

3 Dirichlet-Multinomial Conjugacy

3.1 Multinomial/Dirichlet Distribution

Story: Each of n objects is independently placed into one of k categories. An object is placed into category j with probability p_j, where the p_j are non-negative and $\sum_{j=1}^k p_j = 1$. Let X_1 be the number of objects in category 1, X_2 the number of objects in category 2, *etc.*, so that $X_1 + ... + X_k = n$. Then $\mathbf{X} = (X_1, ..., X_k)$ is said to have the **Multinomial distribution** with parameters n and $\mathbf{p} = (p_1, ..., p_k)$. We write this as $\mathbf{X} \sim Mult_k(n, p)$.

Theorem 1 (Multinomial Joint PMF). *If $\boldsymbol{X} \sim Mult_k(n, p)$, then the joint PMF of $\boldsymbol{X}$ is*

$$P(X_1 = n_1, ..., X_k = n_k) = \frac{n!}{n_1!n_2!...n_k!} \cdot p_1^{n_1} p_2^{n_2} ... p_k^{n_k}$$

for $n_1, ..., n_k$ satisfying $n_1 + ... + n_k = n$.

Definition 1 (Dirichlet Distribution). *The Dirichlet distribution is parameterized by a vector α of positive real numbers. The PDF is:*

$$f(p_1, p_2, ..., p_k; \alpha_1, \alpha_2, ..., \alpha_k) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Pi_{i=1}^k \Gamma(\alpha_i)} \Pi_{i=1}^k p_i^{\alpha_i - 1}$$

where $p_1 + ... + p_k = 1$ and $0 < p_i < 1$.

3.2 Dirichlet-Multinomial Conjugacy

Story: Each of n objects is independently placed into one of k categories. An object is placed into category j with probability p_j, where the p_j are non-negative and $\sum_{j=1}^k p_j = 1$. Let X_1 be the number of objects in category 1, X_2 the number of objects in category 2, *etc.*, so that $X_1 + ... + X_k = n$. The prior distribution of $\mathbf{p} = (p_1, ..., p_k)$ is a Dirichlet distribution, *i.e.*, $\mathbf{p} \sim Dir(\alpha)$. Denote $\mathbf{X} = (X_1, ..., X_k)$, then

$$X|p \sim Mult_k(n, p).$$

Let $f(p)$ to be the prior distribution of p. The observations of the experiment is $N = (n_1, ..., n_k)$, then

$$\begin{aligned} f(p|X = N) &= \frac{P(X = N|p)f(p)}{P(X = N)} \\ &= \frac{\frac{n!}{n_1!...n_k!} p_1^{n_1} ... p_k^{n_k} \cdot \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Pi_{i=1}^k \Gamma(\alpha_i)} \Pi_{i=1}^k p_i^{\alpha_i - 1}}{P(X = N)} \\ &\propto p_1^{n_1 + \alpha_1 - 1} \cdots p_k (n_k + \alpha_k - 1) \sim Dir(\alpha + N) \end{aligned}$$

Thus we see that

$$\text{prior } Dir(\alpha) \rightarrow \text{ posterior } Dir(\alpha + N),$$

$$\alpha_i \rightarrow \alpha_i + n_i.$$

If we have a Dirichlet prior distribution on $\mathbf{p}$ and data that are conditionally Multinomial given $\mathbf{p}$, then when going from prior to posterior, we don't leave the family of Dirichlet distributions. We say that the Dirichlet is the conjugate prior of the Multinomial.

4 Gamma Distribution

4.1 Gamma Function

Definition 2 (Gamma Function). *The gamma function Γ is defined by*

$$\Gamma(\alpha) = \int_0^\infty x^\alpha e^{-x} \frac{dx}{x}$$

for real numbers $\alpha > 0$.

Property 1. For all $\alpha > 0$

$$\Gamma(\alpha + 1) = a\Gamma(\alpha)$$

Proof.

$$\begin{aligned}
\Gamma(\alpha) &= \int_0^\infty x^{\alpha-1} e^{-x} dx = \int_0^\infty x^{\alpha-1} d\left(-e^{-x}\right) \\
&= x^{\alpha-1}\left(-e^{-x}\right)\Big|_0^\infty - \int_0^\infty \left(-e^{-x}\right)(\alpha-1)x^{\alpha-2} dx \\
&= (\alpha-1)\int_0^\infty x^{(\alpha-1)-1} e^{-x} dx \\
&= (\alpha-1)\Gamma(\alpha-1)
\end{aligned}$$

Property 2. If n is a positive integer

$$\Gamma(n) = (n-1)!$$

4.2 Gamma Distribution

Consider the Gamma function, if we divide both sides by $\Gamma(\alpha)$ we get

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)} x^{\alpha-1} e^{-x} dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} dy$$

where we made a change of variables $x = \beta y$. Therefore, if we define

$$f(x \mid \alpha, \beta) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

then $f(x \mid \alpha, \beta)$ will be a probability density function since it is non-negative and it integrates to one.

Definition 3 (Gamma Distribution). *An r.v. Y is said to have the Gamma distribution with parameters α and $\lambda, \alpha > 0$ and $\lambda > 0$, if its* PDF *is*

$$f(y) = \frac{1}{\Gamma(\alpha)} (\lambda y)^\alpha e^{-\lambda y} \frac{1}{y}, y > 0$$

We write $Y \sim \text{Gamma}(\alpha, \lambda)$. Gamma distribution is a generalization of the exponential distribution.

4.3 Moments of Gamma Distribution

To compute the k-th moment of gamma distribution, Let us compute the k th moment of gamma distribution. We have,

$$\begin{aligned}
\mathbb{E}X^k &= \int_0^\infty x^k \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{(\alpha+k)-1} e^{-\beta x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} \int_0^\infty \frac{\beta^{\alpha+k}}{\Gamma(\alpha+k)} x^{\alpha+k-1} e^{-\beta x} dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\beta^{\alpha+k}} \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\beta^k} \\
&= \frac{(\alpha+k-1)\Gamma(\alpha+k-1)}{\Gamma(\alpha)\beta^k} \\
&= \frac{(\alpha+k-1)(\alpha+k-2)\dots\alpha\Gamma(\alpha)}{\Gamma(\alpha)\beta^k} \\
&= \frac{(\alpha+k-1)\cdots\alpha}{\beta^k}
\end{aligned}$$

Therefore, the mean is

$$\mathbb{E}X = \frac{\alpha}{\beta}$$

the second moment is

$$\mathbb{E}X^2 = \frac{(\alpha+1)\alpha}{\beta^2}$$

and the variance

$$\mathrm{Var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{(\alpha+1)\alpha}{\beta^2} - \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha}{\beta^2}$$

(Recall that the exponential distribution has a similar form, i.e. $E = \frac{1}{\lambda}, Var = \frac{1}{\lambda^2}$)

Example 1 (Inverse Gamma Distribution). Let $X \sim \mathrm{Gamma}(\alpha, \beta)$. What is the distribution of $Z = 1/X$?

Solution 3. The Jacobian $|dx/dz|$ of the transformation $z = 1/x$ is $1/z^2$. Multiplying the gamma density

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

by this Jacobian we obtain the density of Z :

$$p(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} e^{-\beta/z}$$

We refer to the distribution having this density as the inverse gamma distribution, denoted $\mathrm{IG}(\alpha, \beta)$.

Exercise 3. Calculate the moments, expectation and variance of $\mathrm{IG}(\alpha, \beta)$.

Solution 4. If $\alpha > n$,

$$\begin{aligned}
E\left(X^n\right) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^n x^{-\alpha-1} \exp(-\beta/x) dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{n-\alpha-1} \exp(-\beta/x) dx \\
&= \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha-n)}{\beta^{\alpha-n}} \\
&= \frac{\beta^n \Gamma(\alpha-n)}{(\alpha-1)\cdots(\alpha-n)\Gamma(\alpha-n)} \\
&= \frac{\beta^n}{(\alpha-1)\cdots(\alpha-n)}
\end{aligned}$$

In particular, for $\alpha > 1$

$$E(X) = \frac{\beta}{\alpha - 1}$$

and for $\alpha > 2$

$$E\left(X^2\right) = \frac{\beta^2}{(\alpha-1)(\alpha-2)}$$

and so for $\alpha > 2$

$$\mathrm{Var}(X) = E\left(X^2\right) - E(X)^2 = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$$

4.4 Gamma related to Exponential

Theorem 2. *Let $X_1, \ldots, X_n$ be i.i.d.* Expo(λ). *Then*

$$X_1 + \cdots + X_n \sim \mathrm{Gamma}(n, \lambda)$$

To move a step further (as exponential is a special case of gamma), we can also have

Theorem 3. *If we have a sequence of independent random variables*

$$X_1 \sim \Gamma\left(\alpha_1, \beta\right), \ldots, X_n \sim \Gamma\left(\alpha_n, \beta\right)$$

then $X_1 + \ldots + X_n$ has distribution $\Gamma\left(\alpha_1 + \ldots + \alpha_n, \beta\right)$

Proof. If $X \sim \Gamma(\alpha, \beta)$ then a moment generating function (m.g.f.) of X is

$$\begin{aligned}\mathbb{E}e^{tX} &= \int_0^\infty e^{tx}\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}dx = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-(\beta-t)x}dx \\ &= \frac{\beta^\alpha}{(\beta-t)^\alpha}\underbrace{\int_0^\infty \frac{(\beta-t)^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-(\beta-t)x}dx}_{1}\end{aligned}$$

The function in the last (underbraced) integral is a PDF of gamma distribution $\Gamma(\alpha, \beta - t)$ and, therefore, it integrates to 1. We get,

$$\mathbb{E}e^{tX} = \left(\frac{\beta}{\beta - t}\right)^\alpha$$

Moment generating function of the sum $\sum_{i=1}^n X_i$ is

$$\mathbb{E}e^{t\sum_{i=1}^n X_i} = \mathbb{E}\prod_{i=1}^n e^{tX_i} = \prod_{i=1}^n \mathbb{E}e^{tX_i} = \prod_{i=1}^n \left(\frac{\beta}{\beta-t}\right)^{\alpha_i} = \left(\frac{\beta}{\beta-t}\right)^{\sum \alpha_i}$$

and this is again a m.g.f. of Gamma distibution, which means that

$$\sum_{i=1}^n X_i \sim \Gamma\left(\sum_{i=1}^n \alpha_i, \beta\right)$$

4.5 Beta-Gamma Connection

When we add independent Gamma r.v.s X and Y with the same rate λ the total $X + Y$ has a Gamma distribution, the fraction $\frac{X}{X+Y}$ has a Beta distribution, and the total is independent of the fraction.

Remark 1. Relate to the famous bank-post office problem (in the lecture), it tells us *the total wait time is independent of the fraction of time that we wait at the bank.*

4.6 *Poisson Gamma Conjugacy

Suppose now that $X \mid \theta$ has a Poisson (θ) distribution, or more generally, that $X_1, \ldots, X_n \mid \theta$ is an iid sample from a Poisson (θ) distribution. Then X has conditional density

$$p(x \mid \theta) = \prod_{j=1}^{n} \frac{\theta^{x_j} e^{-\theta}}{x_j!} \propto \theta^{\sum_j x_j} e^{-n\theta}$$

A conjugate prior is the gamma (α_1, α_2) distribution, with density

$$p\left(\theta \mid \alpha_1, \alpha_2\right) \propto \theta^{\alpha_1 - 1} e^{-\alpha_2 \theta}$$

where the normalizing constant is $\alpha_2^{\alpha_1} / \Gamma\left(\alpha_1\right)$. The expectation of this distribution may be calculated using a method similar to that we used for the Dirichlet distribution, and we find that $E\left[\theta \mid \alpha_1, \alpha_2\right] = \frac{\alpha_1}{\alpha_2}$. The posterior distribution has density

$$P(\theta \mid x, \alpha) \propto \theta^{\sum_j x_j + \alpha_1 - 1} e^{-(\alpha_2 + n)\theta}$$

so that

$$E[\theta \mid x, \alpha] = \frac{\sum_j x_j + \alpha_1}{n + \alpha_2} = \kappa \frac{\alpha_1}{\alpha_2} + (1 - \kappa) \frac{\sum_j x_j}{n}$$

where $\kappa = \alpha_1 / \left(\alpha_2 + n\right)$. Again, we see that this is a convex combination of the prior mean and maximum likelihood estimate, and that it is asymptotically equivalent to the MLE.