Bayesian Inference Notes

Statistics learning notes.

---

1. Bayes Theorem

1.1. Bayes theorem

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta) \tag{1}$$

1.2. Prior predictive distribution

$$p(y) = \int_{\Theta} (y|\theta)p(\theta)\,\mathrm{d}\theta \tag{2}$$

1.3. Posterior predictive distribution

$$\begin{aligned} p(\tilde{y}|y) &= \int_{\Theta} p(\tilde{y}, \theta|y)\,\mathrm{d}\theta \\ &= \int_{\Theta} p(\tilde{y}|\theta, y)p(\theta|y)\,\mathrm{d}\theta \\ &= \int_{\Theta} p(\tilde{y}|\theta)p(\theta|y)\,\mathrm{d}\theta \end{aligned} \tag{3}$$

2. Fundamental Distributions

Name	PDF/PMF	Mean	Variance	Mode
Beta $(y\|\alpha,\beta)$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{\alpha-1}{\alpha+\beta-2}$
Binomial $(y\|n,p)$	$\binom{n}{y}p^y(1-p)^{n-y}$	np	$np(1-p)$	
Exponential $(y\|\lambda)$	$\lambda e^{-\lambda y}$	$\frac{1}{\lambda}$	$\frac{\ln 2}{\lambda}$	0
Erlang $(y\|\lambda,k)$	$\frac{\lambda^k y^{k-1}e^{-\lambda y}}{(k-1)!}$	$\frac{k}{\lambda}$	$\frac{k}{\lambda^2}$	$\frac{1}{\lambda}(k-1)$
ExGauss $(y\|\mu,\sigma,\lambda)$	$\frac{\lambda}{2}\exp\left(\frac{\lambda}{2}(2\mu+\lambda\sigma^2-2y)\right)\operatorname{erfc}(\frac{\mu+\lambda\sigma^2-y}{\sqrt{2}\sigma}$			
Gamma $(y\|\alpha,\beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y}$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\alpha-1}{\beta}$
InvGamma $(y\|\alpha,\beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)}y^{-\alpha-1}e^{-\beta/y}$	$\frac{\beta}{\alpha-1}$, $\alpha>1$	$\frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$	$\frac{\beta}{\alpha-1}$
LogNormal $(y\|\alpha,\beta)$	$\frac{1}{y\sigma\sqrt{2\pi}}e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}}$	$e^{\mu+\frac{\sigma^2}{2}}$	$\left(e^{\sigma^2}-1\right)e^{2\mu+\sigma^2}$	$e^{\mu-\sigma^2}$
Possion $(y\|\lambda)$	$\frac{\lambda^y e^{-\lambda}}{y!}$	λ	λ	
NegBinomial $(k\|r,p)$	$\binom{k+r-1}{k}(1-p)^k p^r$	$\frac{r(1-p)}{p}$	$\frac{r(1-p)}{p^2}$	
Normal $\left(y\|\mu,\sigma^2\right)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	μ	σ^2	μ
Student $(y\|\nu)$	$\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{y^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	0	$\frac{\nu}{\nu-2}$	0
Uniform $(y\|a,b)$	$\frac{1}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	

Table 1: Single Variate Distributions

3. Functions

3.1. Beta Function

$$B(z_1, z_2) = \int_0^1 t^{z_1-1}(1-t)^{z_2-1}\,\mathrm{d}t \tag{4}$$

Properties:

- $B(z_1, z_2) = B(z_2, z_1)$
- $B(z_1, z_2) = \frac{\Gamma(z_1)\Gamma(z_2)}{\Gamma(z_1+z_2)}$
- $B(m, n) = \frac{(m-1)!(n-1)!}{(m+n-1)!} = \frac{m+n}{mn} / \binom{m+n}{m}$

4. Conjugate Prior

The idea of **conjugate prior** is that for a give likelihood we choose a prior distribution such that, after observing data and applying Bayes' theorem, the posterior distribution belongs to the same family as the prior.

That is, if $p(\theta)$ and $p(\theta \mid y)$ have the same distributional form, then the prior is called a **conjugate prior** for the likelihood model.

This is useful because it makes Bayesian updating analytically tractable. Instead of performing difficult integration or numerical approximation, we can often derive the posterior parameters in closed form.

5. Conjugate Prior for Exponential Families

Note general exponential family:

$$p(y_i|\theta) = \underbrace{f(y_i)}_{\text{base measure}} \exp\big(\underbrace{\phi(\theta)^\mathsf{T}}_{\text{natural parameter}} \underbrace{u(y_i)}_{\text{sufficient statistic}} - \underbrace{g(\theta)}_{\text{log normalizer}}\big) \tag{5}$$

Likelihood of a sequence of i.i.d.samples:

$$p(y|\theta) = \left(\prod_{i=1}^{n} f(y_i)\right) \exp\left(\phi(\theta)^\mathsf{T} \overbrace{\sum_{i=1}^{n} u(y_i)}^{t(y)=\sum_{i=1}^{n} u(y_i)} - n g(\theta)\right) \tag{6}$$

So conjugate prior for that likelihood is

$$p(\theta) \propto \exp\big(\phi(\theta)^\mathsf{T}\nu - n_0 g(\theta)\big) \tag{7}$$

Posterior is

$$p(\theta|y) \propto \exp\big(\phi(\theta)^\mathsf{T}(\nu + t(y)) - (n_0 + n)g(\theta)\big) \tag{8}$$

6. Proper and Improper Prior Distributions

A prior is called **proper** if it is a valid probability distribution:

$$p(\theta) \geq 0, \forall \theta \in \Theta, \int_{\theta \in \Theta} p(\theta)\, \mathrm{d}\theta = 1 \tag{9}$$

And **improper** if

$$p(\theta) \geq 0, \forall \theta \in \Theta, \int_{\theta \in \Theta} p(\theta)\, \mathrm{d}\theta = \infty \tag{10}$$

- If a prior is proper, so must the posterior.
- If a prior is improper, the posterior could be proper or improper.

In theory, all priors are acceptable, as long as the posterior is proper.

7. Fisher Information Matrix

$$\vec{I}\left(\vec{\theta}\right) = E\Big(\left(\nabla \log p\left(y|\vec{\theta}\right)\right)\left(\nabla \log p\left(y|\vec{\theta}\right)\right)^{\mathsf{T}}\Big) = -E\left(\nabla^2 \log p\left(y|\vec{\theta}\right)\right) \tag{11}$$

8. Jeffreys' Prior

$$\pi_J\left(\vec{\theta}\right) \sim \sqrt{\left|\vec{I}\left(\vec{\theta}\right)\right|^2} \tag{12}$$

9. Pivotal Quantities

For the binomial and other single-parameter models, different principles give (slightly) different noninformative prior distributions. But for two cases—location parameters and scale parameters—all principles seem to agree[1].

9.1. Location Parameter

$$p(\theta) \sim 1 \tag{13}$$

9.2. Scale Parameter

$$p(\theta) \sim \frac{1}{\theta} \tag{14}$$

10. Predictive Accuracy

People care about the accuracy in two different ways. First to assume that the model is all we known and check posterior predictions. The second is to compare several candidate models. Even if all of the models being considered have mismatches with the data, it can be informative to evaluate their predictive accuracy, compare them, and consider where to go next[2].

11. KL Divergence

12. Linear Algebra

12.1. Convex Combination

A subset A of a vector space V is said to be convex if $\lambda\vec{x} + (1-\lambda)\vec{y}$ for all vectors $\vec{x}, \vec{y} \in A$, and all scalars λ in $[0, 1]$.

Via induction, this can be seen to be equivalent to the requirement that $\lambda\vec{x}_1, \lambda\vec{x}_2, ..., \lambda\vec{x}_n \in A$ for all vectors $\vec{x}_1, \vec{x}_2, ..., \vec{x}_n \in A$, and for all scalars $\lambda_1, \lambda_2, ..., \lambda_n \geq 0$ such that $\sum k_i = 1$.

Bibliography

[1] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and others, *Bayesian Data Analysis*, Third. Boca Raton, Florida: Crc, 2013. [Online]. Available: https://stat.columbia.edu/~gelman/book/

[2] A. Gelman, J. Hwang, and A. Vehtari, "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, vol. 24, no. 6, pp. 997–1016, Nov. 2014, doi: 10.1007/s11222-013-9416-2.