Statistics
MDI220
3. Bayesian Statistics

Thomas Bonald
Institut Polytechnique de Paris

2023 – 2024

We consider the case where some *prior* information is known about the parameter θ of the statistical model. The parameter θ is now itself random.

1 Bayes estimator

Let π be the probability density function of θ. It is called the *prior*.

> The prior π corresponds to the distribution of θ before any observation.

Example. *For the Bernoulli model with a uniform prior, we get:*

$$\pi(\theta) = 1_{[0,1]}(\theta) \quad X|\theta \sim \mathcal{B}(\theta).$$

The *posterior* distribution of θ (after observation) follows from Bayes' theorem:

$$\pi(\theta|x) = \frac{\pi(\theta)p(x|\theta)}{p(x)}.$$

Observe that we use the notation $p(x|\theta)$ instead of $p_\theta(x)$ for the probability density function of X, since θ is now a random variable.

Example. *For the Bernoulli model with a uniform prior, we have $p(x|\theta) = \theta^x(1-\theta)^{1-x}$ so that:*

$$\pi(\theta|x=1) \propto 1_{[0,1]}(\theta)\theta$$
$$\pi(\theta|x=0) \propto 1_{[0,1]}(\theta)(1-\theta).$$

An estimation of the parameter θ follows from the expectation of the posterior distribution.

> The Bayes estimator of θ is given by $\hat{\theta}(x) = \mathrm{E}(\theta|x)$.

Example. *For the Bernoulli model with a uniform prior, we get:*

$$\hat{\theta}(1) = \frac{2}{3}$$
$$\hat{\theta}(0) = \frac{1}{3}.$$

2 Bayes risk

Consider the quadratic risk $R(\theta, \hat{\theta})$ for an estimator $\hat{\theta}$. Since θ is random, this is a random variable. By taking the expectation, we obtain the Bayes risk.

The Bayes risk of the estimator $\hat{\theta}$ is $r(\hat{\theta}) = \mathrm{E}(R(\theta, \hat{\theta}))$.

The optimality of Bayes estimator follows from the equality:

$$r(\hat{\theta}) = \mathrm{E}((\hat{\theta}(X) - \theta)^2) = \mathrm{E}(\underbrace{\mathrm{E}((\hat{\theta}(X) - \theta)^2 | X)}_{\text{minimized for } \hat{\theta}(x) = \mathrm{E}(\theta|x)})$$

Example. *For the Bernoulli model with a uniform prior, the bias and variance of Bayes estimator are respectively given by:*

$$\begin{aligned} b(\theta, \hat{\theta}) &= \mathrm{E}(\hat{\theta}(X)|\theta) - \theta, \\ &= \theta\frac{2}{3} + (1-\theta)\frac{1}{3} - \theta, \\ &= \frac{1-2\theta}{3} \end{aligned}$$

and

$$\mathrm{var}(\hat{\theta}(X)|\theta) = \mathrm{var}\left(\hat{\theta}(X) - \frac{1}{3}|\theta\right) = \frac{1}{9}\theta(1-\theta).$$

We deduce the quadratic risk:

$$R(\hat{\theta}, \theta) = b(\theta, \hat{\theta})^2 + \mathrm{var}(\hat{\theta}(X)|\theta) = \frac{3\theta^2 - 3\theta + 1}{9}$$

and the Bayes risk:

$$r(\hat{\theta}) = \mathrm{E}(R(\hat{\theta}, \theta)) = \frac{1}{18}.$$

3 Conjugate prior

The posterior distribution is not always easy to compute. This is the case when the prior and posterior distributions are *conjugate*, in the sense that they belong to the same family.

The distribution π is a conjugate prior if the posterior distribution belongs to the same family as π.

Example. *Consider the Bernoulli model with n observations. The Beta distribution is a conjugate prior. Assume the prior is a Beta distribution with parameters (a, b):*

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1} \quad \theta \in [0, 1].$$

Then,

$$\pi(\theta|x) \propto \pi(\theta)\theta^s(1-\theta)^{n-s} \propto \theta^{s+a-1}(1-\theta)^{b-1+n-s}$$

where $x = (x_1, \ldots, x_n)$ and $s = \sum_{i=1}^n x_i$. The posterior distribution also has a Beta distribution.

4 Exponential family

The exponential family is a set of distributions with explicit conjugate prior distributions.

A probability density function p_θ is in the exponential family if $p_\theta(x) = h(x)e^{\eta(\theta)^T T(x) - A(\theta)}$ for some functions h, η, T and A.

Example. *The Bernoulli distribution belongs to the exponential family, with:*

$$h(x) = 1 \quad \eta(\theta) = \begin{pmatrix} \log\theta \\ \log(1-\theta) \end{pmatrix} \quad T(x) = \begin{pmatrix} x \\ 1-x \end{pmatrix} \quad A(\theta) = 0.$$

The conjugate prior of a statistic model of the exponential family is $\pi(\theta) \propto e^{\eta(\theta)^T \alpha + \beta A(\theta)}$ for some parameters α, β.

The proof follows on observing that:

$$\pi(\theta|x) \propto \pi(\theta)p(x|\theta) \propto e^{\eta(\theta)^T(\alpha + T(x)) - (\beta+1)A(\theta)}.$$

Example. *The conjugate prior of the Bernoulli model is the Beta distribution:*

$$\pi(\theta) \propto e^{\eta(\theta)^T \alpha} \propto \theta^{\alpha_1}(1-\theta)^{\alpha_2} \quad \theta \in [0,1].$$

Statistical model	Conjugate prior
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Multinomial	Dirichlet
Gaussian (mean unknown)	Gaussian
Gaussian (variance unknown)	Inverse Gamma
Exponential	Gamma

Table 1: Some models of the exponential family and their conjugate priors

5 Jeffreys prior

When no prior information is known about the parameter θ, it is still possible to do Bayesian statistics by choosing a *non-informative* prior π. The idea is that the prior distribution remains the same for any reparametrization of θ. We first assume that $\theta \in \mathbb{R}$.

The Jeffreys prior, defined by $\pi(\theta) \propto \sqrt{I(\theta)}$, is not informative.

Example. *The Jeffreys prior of the Bernoulli model is the Beta distribution with parameters* $(\frac{1}{2}, \frac{1}{2})$*:*

$$\pi(\theta) \propto \sqrt{I(\theta)} = \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \quad \theta \in [0,1].$$

To prove that the Jeffreys prior is not informative, consider some new parameter ϕ, viewed as a function of θ, and p_ϕ the corresponding probability density function of the observation x. Since $p_\theta = p_\phi \circ \theta$, the score is given under proper regularity conditions by

$$\frac{\partial \log p_\theta}{\partial \theta} = \phi'(\theta) \frac{\partial \log p_\phi}{\partial \phi}$$

and the Fisher information by:

$$I(\theta) = \phi'(\theta)^2 I(\phi).$$

By a change of variable, the prior π on ϕ satisfies:

$$\pi(\phi) \mathrm{d}\phi = \underbrace{\pi(\phi(\theta))|\phi'(\theta)|}_{\propto \sqrt{I(\theta)}} \mathrm{d}\theta,$$

so that

$$\pi(\phi) \propto \frac{\sqrt{I(\theta)}}{|\phi'(\theta)|} = \sqrt{I(\phi)}.$$

Example. *Consider the Bernoulli model* $\mathcal{P} = \{P_\phi \sim \mathcal{B}(\frac{\phi}{\phi+1}),\ \phi \geq 0\}$. *Then*

$$p_\phi(x) = \frac{\phi^x}{\phi + 1}.$$

We deduce the score:

$$\frac{\partial \log p_\phi}{\partial \phi}(X) = \frac{X}{\phi} - \frac{1}{\phi + 1}$$

and the Fisher information:

$$I(\phi) = \frac{1}{\phi(\phi+1)^2}.$$

The Jeffreys prior on ϕ *is:*

$$\pi(\phi) \propto \sqrt{I(\phi)} = \frac{1}{\sqrt{\phi}(\phi+1)}.$$

By the change of variable $\theta = \frac{\phi}{\phi+1}$, *we check that the prior is the same as before:*

$$\pi(\theta) = \frac{\pi(\phi)}{|\theta'(\phi)|} \propto \frac{\phi+1}{\sqrt{\phi}} \propto \frac{1}{\sqrt{\theta(1-\theta)}}.$$

In the vectorial case $\theta \in \mathbb{R}^k$, the Jeffreys prior becomes:

$$\pi(\theta) \propto \sqrt{\det(I(\theta))}.$$

By a change of variable, we can check similarly that this prior is not informative.