The Counter-intuitive Non-informative Prior for the Bernoulli Family

Skorik Sergey

MIPT, 2022

September 26, 2022

Introduction

Definition

Let $\mathbf{X}$ be a set of samples from $f(\mathbf{x}|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^k$. Then

$$\mathcal{I}(\boldsymbol{\theta}) \equiv -\mathbb{E}\left[\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\mathbf{X}, \boldsymbol{\theta})\right)^2\right], \quad L(\mathbf{X}, \boldsymbol{\theta}) = \prod_{x \in \mathbf{X}} f(x|\boldsymbol{\theta})$$

is the Fisher information (here L is the likelihood function)

Definition

In Bayesian probability, the Jeffreys prior is a non-informative (objective) prior distribution for a parameter space whose probability density is proportional to

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

Introduction

Property

Suppose $\pi_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ is Jeffreys prior and $\boldsymbol{\phi} = h(\boldsymbol{\theta})$ is re-parameterization of the problem, then

$$\pi_{\boldsymbol{\phi}}(\boldsymbol{\phi}) = \pi_{\boldsymbol{\theta}}(h^{-1}(\boldsymbol{\phi})) \left| \frac{d\boldsymbol{\theta}}{d\boldsymbol{\phi}} \right|$$

Proof.

(One-parameter case)

$$\mathcal{I}(\phi) = -\mathbb{E}\left[\frac{d^2}{d\phi^2} \log L(\mathbf{X}, \phi)\right] = \mathbb{E}\left[\frac{d^2}{d\theta^2} \log L(\mathbf{X}, \theta) \left(\frac{d\theta}{d\phi}\right)^2\right]$$

Therefore

$$\pi_{\phi}(\phi) \propto \mathcal{I}(\phi)^{\frac{1}{2}} = \mathcal{I}(\theta)^{\frac{1}{2}} \left| \frac{d\theta}{d\phi} \right| = \pi_{\theta}(\theta) \left| \frac{d\theta}{d\phi} \right|$$

Introduction

Problem statement

Let $\{X_i\}_{i=1}^n$ be i.i.d observations and $X_i \sim Bernoulli(p)$, $\sum_{i=1}^n X_i = x$. Then

$$f(X|p) = \binom{n}{x} p^x (1-p)^{n-x} \Rightarrow \frac{d^2}{dp^2} \log f(X|p) \propto -\frac{x}{p^2} - \frac{n-x}{(1-p)^2} \quad (1)$$

Thus

$$\pi(p) \propto \sqrt{\mathbb{E}\left[-\frac{x}{p^2} - \frac{n-x}{(1-p)^2}\right]} = \sqrt{\frac{np}{p^2} + \frac{n-np}{(1-p)^2}} \propto \frac{1}{\sqrt{p(1-p)}}$$

Problem statement

In this simple case, it is most intuitive to use the uniform distribution on $[0, 1]$ as a non-informative prior; it is non-informative because it says that all possible values of p are equally likely a *priori*.

Introduction

Problem statement

$$\pi(p) \propto \frac{1}{\sqrt{p(1-p)}} \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right) \tag{2}$$

Jeffrey's prior for this simple problem can be quite counter-intuitive. Under the prior (2), it appears that some values of p are more likely than others. Therefore intuitively, it appears that this prior is actually quite informative.

Heuristic

Since the maximum likelihood estimator (MLE) is not affected by any prior opinion, we simply ask: is there a prior which would produce a Bayesian estimate (e.g., posterior mean) that coincides with the MLE? If so, that prior could be regarded as non-informative since the prior opinion exerts no influence on the final estimate whatsoever.

The Bayesian Estimator

Maximum Likelihood Estimator

from (1) we can get

$$\frac{x}{p} - \frac{n-x}{1-p} = 0 \Rightarrow \hat{p}_{\text{MLE}} = \frac{x}{n} \tag{3}$$

Conjugate Prior

The posterior distribution of p is given by (Bayes' Theorem):

$$\pi_1(p|x_1, \ldots x_n) = \frac{f(x_1, \ldots x_n|p)\pi_0(p)}{\int_0^1 f(x_1, \ldots x_n|p)\pi_0(p)}$$

Definition: if the posterior distribution $p(\theta|x)$ is in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(x|\theta)$.

The Bayesian Estimator

Statement

the Beta distribution and the Bernoulli distribution form a conjugate pair

Proof.

Consider the $Beta(\alpha, \beta)$ distribution as the prior for p, i.e.

$$\pi_0(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

So,

$$\pi_1(p) \propto f(x_1, \dots x_n|p)\pi_0(p) = p^x(1-p)^{n-x}p^{\alpha-1}(1-p)^{\beta-1} =$$

$$= p^{\alpha+x-1}(1-p)^{\beta+n-x-1} \sim Beta(\alpha+x, \beta+n-x)$$

The Bayesian Estimator

Beta moments

If $p \sim Beta(\alpha, \beta)$, then

$$\mathbb{E}(p) = \frac{\alpha}{\alpha + \beta}, \quad \mathbb{D}(p) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

So, we can now write down a general formula for obtaining the Bayesian posterior mean in our case

$$\hat{p}_{\text{bayes}} = \frac{\alpha + x}{\alpha + \beta + n} \tag{4}$$

Note, that Jeffreys and Uniformal prior are a members of Beta distributions family

$$J(p) \sim Beta\left(\frac{1}{2}, \frac{1}{2}\right), \quad U(p) \sim Beta(1, 1) \Rightarrow \hat{p}_{\text{uni-bayes}} = \frac{1 + x}{2 + n}$$

The Bayesian Estimator

The Effects of Different Priors

Focus on a particular sub-family of Beta distributions with $\alpha = \beta = c$, i.e. $\pi_0(p) \sim Beta(c, c)$. Note, that $J(p)$ and $U(p)$ are also a members of this subfamily. The central moments of this dstributions

$$\mathbb{E}(p) = \frac{c}{c+c} = \frac{1}{2}, \quad \mathbb{D}(p) = \frac{c^2}{4c^2(2c+1)} = \frac{1}{4(2c+1)} \tag{5}$$

It is clear from (4) that the prior parameter c influences the posterior mean as if an extra $2c$ observations, equally split between zeros and ones, were added to the sample.

The Bayesian Estimator

Case 1 $c \to \infty$

It is easy to see from (4) that as $c \to \infty$, we have $\hat{p}_{\text{bayes}} = \frac{1}{2}$. In other words, our prior opinion of p is so strong that it can not be changed by the observed outcomes.

Case 2 $c \to 0$

Following the same logic, it is clear from (4) that using such a prior, the posterior mean would have been the same as the MLE.
But the $Beta(0, 0)$ distribution is not defined. Consider the distribution $Beta(\varepsilon, \varepsilon)$ for arbitrarily small $\varepsilon > 0$. To understand the behavior of this distribution, we can examine the limiting distribution as $c \to 0$

$$B_{0,0} = \lim_{c \to 0} Beta(c, c)$$

The Bayesian Estimator

Theorem 1

The limiting distribution $B_{0,0}$ consists of two equal point masses at 0 and 1

Proof.

Let $X \sim B_{0,0}$; let $f(x)$ be its probability function. Clearly $f(x)$ is symmetric about $\frac{1}{2}$. Now suppose $f(x)$ is not just two point masses at 0 and 1. Then there exist $0 < \varepsilon < \frac{1}{2}$ and $\delta > 0$ such that

$$g(\varepsilon) \equiv \int_{\varepsilon}^{1-\varepsilon} f(x)dx > \delta$$

Because $f(x)$ is symmetric about $\frac{1}{2}$ it follows that

$$\mathbb{D}(p) = \int_0^1 \left(x - \frac{1}{2}\right)^2 f(x)dx$$

The Bayesian Estimator

Proof.

$$\mathbb{D}(p) = 2\int_0^{\varepsilon}\left(x-\frac{1}{2}\right)^2 f(x)dx + \int_{\varepsilon}^{1-\varepsilon}\left(x-\frac{1}{2}\right)^2 f(x)dx \leqslant$$

$$\leqslant \frac{1}{2}\left(\frac{1-g(\varepsilon)}{2}\right) + \left(\varepsilon^2+\varepsilon+\frac{1}{4}\right)g(\varepsilon) = \frac{1}{4} - \varepsilon(1-\varepsilon)g(\varepsilon) <$$

$$< \frac{1}{4} - \varepsilon(1-\varepsilon)\delta$$

From (5) for $c \to 0$ is clear that $\mathbb{D}(p) = \frac{1}{4}$. So, we can write

$$\frac{1}{4} - \varepsilon(1-\varepsilon)\delta > \frac{1}{4} \Rightarrow \varepsilon(1-\varepsilon)\delta < 0$$

Since $0 < \varepsilon < \frac{1}{2}$ and $\delta > 0 \Rightarrow \varepsilon(1-\varepsilon)\delta > 0$, this is a contradiction. □

Remarks

Remark 1

Theorem 1 states that the limiting distribution $B_{0,0}$ is a $Bernoulli(\frac{1}{2})$ distribution. Moreover, note that if $B_{0,0}$ is actually used as a prior, then the posterior distribution is not defined unless all the observations $X1, X2, ..., Xn$ are identical.

Remark 2

Another common prior that is used in the literature for this problem is an improper prior of the form

$$\pi_0(p) = \propto \frac{1}{p(1-p)}$$

also called the Haldane prior. It is improper because $\int_0^1 \pi_0(p)dp = \infty$

Remarks

Remark 3

This heuristic in other situations as well to evaluate the non-informativeness of different priors. F.e. consider $X_1, \dots X_n \sim \mathcal{N}(\mu, \sigma^2)$, where σ^2 is known. Let $\pi(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$ be a prior on μ. Then it can be shown that posterior mean is

$$\theta_0 \left(\frac{1/\sigma_0^2}{1/\sigma_0^2 + n/\sigma^2} \right) + \overline{x} \left(\frac{n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2} \right)$$

It is then clear that the posterior mean agrees with the MLE $\overline{x}$ if and only if $\sigma_0 \to \infty$, i.e. if we put a prior on μ that is essentially flat.

Conclusion

Uninformative priors

That the posterior mean coincides with the MLE is not necessarily the right criterion to judge whether a prior is non-informative, but it provides an extremely simple and effective demonstration of why sometimes flat priors are not necessarily non-informative, while non-informative priors are not always flat. So, the term "uninformative prior" is somewhat of a misnomer.

Some attempts have been made at finding a priori probabilities, i.e. probability distributions in some sense logically required by the nature of one's state of uncertainty; these are a subject of philosophical controversy, with Bayesians being roughly divided into two schools: "objective Bayesians", who believe such priors exist in many useful situations, and "subjective Bayesians" who believe that in practice priors usually represent subjective judgements of opinion that cannot be rigorously justified.

Conclusion

Objective Bayesianism

Perhaps the strongest arguments for objective Bayesianism were given by Edwin T. Jaynes, based mainly on the consequences of symmetries and on the principle of maximum entropy.

Jaynes Sample

Let ball has been hidden under one of three cups, A, B, or C, but no other information is available about its location. In this case a prior $p \sim U[0, 1]$ seems intuitively like the only reasonable choice.
But more formally, we can see that the problem remains the same if we swap around the labels ("A", "B" and "C") of the cups. So, we have to choose a prior for which a permutation of the labels do not cause a change in our predictions. The uniform prior is the only one which preserves this invariance.