STA 602 - Intro to Bayesian Statistics

Lecture 8

Li Ma

Duke University

Gaussian sampling model

- Sampling model for n readings given the mean θ and variance σ^2 is
$$X_1, X_2, \ldots, X_n \,|\, \theta, \sigma^2 \overset{\text{iid}}{\sim} \mathrm{N}(\theta, \sigma^2)$$
- A Gaussian prior for the mean μ
$$\theta \,|\, \sigma^2 \sim \mathrm{N}(\mu_0, \tau_0^2)$$
- Note that here I emphasize the conditioning on σ^2, which we have been implicitly doing all along by assuming σ^2 is known.

The conditional posterior of θ

By Bayes' theorem.

$$
\begin{aligned}
p(\theta|\mathbf{x},\sigma^2) &\propto p(\theta|\sigma^2)p(\mathbf{x}|\theta,\sigma^2) \\
&= \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{(\theta-\mu_0)^2}{2\tau_0^2}} \times \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}} \\
&\propto e^{-\frac{1}{2}\left[\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{\sum_{i=1}^n (x_i-\theta)^2}{\sigma^2}\right]}
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\frac{(\theta-\mu_0)^2}{\tau_0^2}+\frac{\sum_i (x_i-\theta)^2}{\sigma^2} \\
&=\theta^2\underbrace{(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2})}_{A}-2\theta\underbrace{(\frac{\mu_0}{\tau_0^2}+\frac{n\bar{x}}{\sigma^2})}_{B}+C \\
&=A\theta^2-2B\theta+C \\
&=A\left(\theta^2-2B/A\cdot\theta\right)+C \\
&=A(\theta-B/A)^2+C' \\
&=\frac{(\theta-\mu_n)^2}{\tau_n^2}+C'
\end{aligned}
$$

where

$$
\mu_n=\frac{B}{A}=\frac{\mu_0/\tau_0^2+n\bar{x}/\sigma^2}{1/\tau_0^2+n/\sigma^2} \quad \text{and} \quad \tau_n^2=\frac{1}{A}=\frac{1}{1/\tau_0^2+n/\sigma^2}.
$$

- Thus
$$p(\theta|\mathbf{x},\sigma^2) \propto e^{-\frac{(\theta-\mu_n)^2}{2\tau_n^2}} \quad \text{for } -\infty < \theta < \infty.$$
- This is the same as the p.d.f of a Normal(μ_n, τ_n^2) distribution up to a normalizing constant. Therefore we must have
$$p(\theta|\mathbf{x},\sigma^2) = \frac{1}{\sqrt{2\pi}\tau_n} e^{-\frac{(\theta-\mu_n)^2}{2\tau_n^2}} \quad \text{for } -\infty < \theta < \infty.$$
- In particular, if we specify our conditional prior for θ as equivalent to have κ_0 observations, that is, $\tau_0^2 = \frac{\sigma^2}{\kappa_0}$.
- Then
$$\mu_n = \frac{\kappa_0}{\kappa_n}\mu_0 + \frac{n}{\kappa_n}\bar{x}$$
where $\kappa_n = \kappa_0 + n$ and
$$\tau_n^2 = \sigma^2/\kappa_n$$
or equivalently
$$\frac{1}{\tau_n^2} = \frac{\kappa_n}{\sigma^2} = \frac{\kappa_0}{\sigma^2} + \frac{n}{\sigma^2}.$$

Non-informative priors

- In Bayesian inference, one expresses one's prior belief about the underlying distribution using a prior distribution.
- What if one wants to express a "lack of prior knowledge" yet remain in the Bayesian paradigm.
- One possibility is to choose a "vague" prior that spread probability over a while range of values.
- With large prior uncertainty, let the empirical evidence from the data dominate inference.

A non-informative, improper, prior on θ

- Now suppose I have "no idea" what the value of θ might be *a priori*.
- What would be a natural choice of prior for θ?
- Idea: let $\tau_0^2 \to \infty$, or equivalently $\kappa_0 \downarrow 0$.
- In the limit the prior becomes a constant

$$p(\theta) \propto 1.$$

  Note that this is not a probability density any more, as it doesn't integrate to 1 over $\mathbb{R}$.

The corresponding posterior

- Nevertheless, if we still carry out the computation under Bayes theorem

$$
\begin{aligned}
p(\theta \,|\mathbf{x}) &\propto p(\mathbf{x} \,|\, \theta) p(\theta) \\
&\propto p(\mathbf{x} \,|\, \theta) \\
&\propto e^{-\frac{\sum (x_i - \theta)^2}{2\sigma^2}} \\
&\propto e^{-\frac{(\theta - \bar{x})^2}{2(\sigma^2/n)}}.
\end{aligned}
$$

  That is, we still have

$$
\theta \,|\, \mathbf{x}, \sigma^2 \sim \mathrm{N}(\mu_n, \tau_n^2)
$$

  where $\mu_n = \bar{x}$ and $\tau_n^2 = \sigma^2/n$.
- These are exactly the limits of μ_n and τ_n^2 as $\tau_0^2 \uparrow \infty$ or $\kappa_0 \downarrow 0$.

A more general view of Bayes theorem

- One can view Bayes theorem as applying a weight function (the prior) to the likelihood which summerizes the empirical evidence from the data for different values of θ.
- When the weight function is a probability density (after normalization), it provides a natural probabilistic interpretation as the marginal distribution of θ.
- When the weight function cannot be normalized to a density (i.e., integrates to ∞), the reweighting could still work if the product $p(\theta)p(\mathbf{x}|\theta)$ can be normalized to a probability distribution (the posterior).

A caveat

- One must be careful in using a *improper* prior, as it may not lead to a posterior distribution!
  - Example: Suppose $p(\theta) \propto e^{\theta^2}$ in the Gaussian example.
  - Example, for the political poll example, suppose we observed $x = 40$ out of $n = 100$, then if we use a prior

$$p(\theta) \propto \theta^{-70}(1-\theta)^{-70}.$$

Jeffrey's prior

- In choosing a *non-informative* prior, one may want to enforce some basic properties.
- In particular, the prior probability should be *invariant with respect to change-of-variable*.
  - Measurements in particles per square inch versus particles per square cm, inches versus feet?

- That is, if we put a prior on θ, then if we reparametrize the model into ξ such that $\theta = h(\xi)$ for some bijection $h(\cdot)$ (without loss of generality, assume that h is differentiable and increasing), we apply the same principle in assigning a prior on ξ, the resulting prior probability

$$\mathrm{P}_\theta(\theta_0 < \theta < \theta_0 + d\theta) = \mathrm{P}_\xi(\xi_0 < \xi < \xi + d\xi))$$

where $\theta_0 = h(\xi_0)$ and $\theta_0 + d\theta = h(\xi + d\xi)$. Thus,

$$p_\theta(\theta)d\theta = p_\xi(\xi)d\xi.$$

Thus

$$p_\theta(\theta) = p_\xi(\xi)|d\xi/d\theta|.$$

- This needs to hold for any transform $h(\cdot)$ between parameters.
- In other words, our strategy for specifying our prior in terms of θ and in terms of ξ should not matter—specify one and apply a change-of-variable should result in the other.

Fisher's information

- The *Fisher information* the data contains about θ is defined to be

$$I(\theta) = \mathrm{E}\left\{ \left[\frac{d}{d\theta} \log p(\mathbf{X}|\theta) \right]^2 |\theta \right\} = -\mathrm{E}\left[\frac{d^2}{d\theta^2} \log p(\mathbf{X}|\theta)|\theta \right],$$

  where

$$\frac{d}{d\theta} \log p(\mathbf{X}|\theta) = \frac{\frac{d}{d\theta} p(\mathbf{X}|\theta)}{p(\mathbf{X}|\theta)}.$$

- The expectation is over $\mathbf{X}$ under *repeated sampling*. So it is a frequentist property of the sampling model.
- This is a frequetist property of the sampling model $p(x|\theta)$.
- The meaning of Fisher's information.

Example

- $X \sim \mathrm{N}(\theta, \sigma^2)$ where σ is known.
- Then
$$p(x|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(x-\theta)^2 \right\}$$
and so
$$\begin{aligned} \log p(x|\theta) &= -\log(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(x-\theta)^2 \\ \frac{d}{d\theta}\log p(x|\theta) &= \frac{1}{\sigma^2}(x-\theta) \\ \left[\frac{d}{d\theta}\log p(x|\theta)\right]^2 &= \frac{(x-\theta)^2}{\sigma^4}. \end{aligned}$$

- Therefore

$$I(\theta) = \mathrm{E}\left\{ \left[\frac{d}{d\theta} \log p(x|\theta) \right]^2 \middle| \theta \right\} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

- In this case, $I(\theta)$ doesn't depend on θ. So the amount of information an observation X from $p(x|\theta)$ has about its mean doesn't depend on θ.
- Exercise 1: Use the alternative way to compute Fisher's information.
- Exercise 2: Find the Fisher's information $I(\sigma)$ for σ if instead θ is known and σ is the unknown parameter.

Some facts related to Fisher's information

- Additivity property:
  - If $X \sim p_\theta$ with Fisher's information $I_X(\theta)$, and $Y \sim q_\theta$ with Fisher's information $I_Y(\theta)$.
  - Assume X and Y are independent. Then

$$I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta).$$

- If $X_1, X_2, \ldots, X_n$ are i.i.d.$\sim p(x|\theta)$ and $I(\theta)$ is the information any X_i contains about θ. Then the information $(X_1, X_2, \ldots, X_n)$ contains about θ is

$$I_n(\theta) = nI(\theta).$$

Dependence on parametrization

- The Fisher's information depends on the parameterization of the model.
- If $\theta = h(\xi)$, where h is "nice"—one-to-one and differentiable. Then the information X contains about ξ is

$$I_\xi(\xi) = I_\theta(\theta)|d\theta/d\xi|^2 = I_\theta(h(\xi)) \cdot [h'(\xi)]^2.$$

  This follows immediately from applying the chain rule to the definition of Fisher's information.

Relationship to Jeffrey's prior:

- Under Jeffrey's reasoning, we want a prior invariant to change-of-parametrization:

$$p_\theta(\theta) \cdot \left| \frac{d\theta}{d\xi} \right| = p_\xi(\xi).$$

- Because we have $I(\xi) = I(\theta) \cdot (d\theta/d\xi)^2$, we can let

$$\pi_\theta(\theta) \propto I_\theta(\theta)^{1/2}.$$

Examples

- For the Gaussian sampling model with unknown mean θ and known variance σ^2, the Fisher's information is
$$I(\theta) = \frac{1}{\sigma^2} \propto 1.$$
Hence the corresponding Jeffrey's prior is exactly the *improper* flat prior we used
$$p(\theta) \propto I(\theta)^{1/2} =\propto 1.$$
- For the pollitical poll example, the model is Binomial(n, θ). The Fisher information from is
$$I_n(\theta) = \frac{n}{\theta(1-\theta)}$$
So the Jeffrey's prior on θ is
$$p(\theta) \propto I(\theta)^{1/2} \propto \theta^{-1/2}(1-\theta)^{-1/2}.$$
This is a *proper* prior—the Beta$(1/2, 1/2)$ distribution.
- The uniform prior on $(0,1)$ is actually not invariant to different parametrizations!

Limitations of Jeffrey's prior

- It works only for univariate models in the sense that it will easily be dominated by the data (i.e., being non-informative).
- For multi-parameter models, the amount of prior information imposed by Jeffrey's prior (based on the multivariate Fisher's information) is actually very strong.
- So we need to generalize Jeffrey's prior without using Fisher's information.

Reference priors (optional materials)

- A theoretical formulation of "uninformative prior" by maximizing the *"information distance"* (aka Kullback-Leibler divergence) between the prior $p(\theta)$ and the posterior $p(\theta|\mathbf{x})$, average under the marginal distribution of $\mathbf{x}$. That is,

$$
\begin{aligned}
I(p(\theta), p(\theta|\mathbf{X})) &= \mathrm{E}\, D_{\mathrm{KL}}(p(\theta), p(\theta|\mathbf{X})) \\
&= \mathrm{E} \int p(\theta \,|\, \mathbf{X}) \log \frac{p(\theta|\mathbf{X})}{p(\theta)} d\theta \\
&= \int p(\mathbf{x}) \int p(\theta \,|\, \mathbf{x}) \log \frac{p(\theta|\mathbf{x})}{p(\theta)} d\theta d\mathbf{x} \\
&= \int \int p(\theta, \mathbf{x}) \log \frac{p(\theta, \mathbf{x})}{p(\theta) p(\mathbf{x})} d\theta d\mathbf{x}
\end{aligned}
$$

  where the expectation is taking over the marginal distribution (i.e., prior predictive distribution) of $\mathbf{x}$.
- The above quantity is called *intrinsic discrepancy*.

A formal definition of being “non-informative”

- The intrinsic descripancy quantifies the information gap between prior and posterior, i.e., the amount of information that one can gain from observing the data relative to the prior knowledge.
- Generally speaking, the stronger the prior information, the smaller this gap.
- A non-informative (or “reference”) prior is defined to be the prior that maximizes this gap:

$$p_r(\boldsymbol{\theta}) = \text{argmax}_{p(\boldsymbol{\theta})} \int \int p(\boldsymbol{\theta}, \mathbf{x}) \log \frac{p(\boldsymbol{\theta}, \mathbf{x})}{p(\boldsymbol{\theta}) p(\mathbf{x})} d\boldsymbol{\theta} d\mathbf{x}.$$

- One can show that it is equivalent to Jeffrey’s prior in univariate models, but differs in multivariate models.