Bayesian Inference I

Matthew Edwards

Founder & Curriculum Developer, ml.edu
mr.edwards@utoronto.ca

August 19, 2021

Overview

Frequentist Approach

- Until now, much of what you have learned about classical statistics can be considered part of the **Frequentist** approach to probability
- In the **Frequentist** setting, we interpret probabilities over events as **long-run expected frequencies of occurrence**
- This often leads to intuitive estimation techniques. For instance, to determine $P(Heads)$ for a coin, we can simply flip the coin a large number of times, and count the number of heads.
- It is also why you must be very careful in interpretting a frequentist **confidence interval**
  - For a $(1-\alpha)\cdot 100\%$ CI, computed as $\hat{\theta} \pm Z_{\alpha/2} \cdot SE(\hat{\theta})$, we say that if we were to collect a large number of such intervals, then approximately $(1-\alpha)\%$ of them would contain the true θ
  - It is not correct to say that an individual interval contains the true value with $(1-\alpha)\%$ probability, because frequentist estimators only derive meaning through long-run repetition of experiments

Frequentist Approach (cont.)

  - Notice that this is a probability statement about the interval, not about θ or its estimator
- In the frequentist setting, individual realizations of random variables have no meaning (at least no useful one)
- Moreover, some probability statements don't easily fit into a long-run setting
  - E.g. What is the probability that it will rain tomorrow? What is the probability that Canada will win more than 10 gold medals at the Tokyo Olympics?
  - These events will likely never be repeated at all, let alone in large enough numbers to draw inference under the frequentist paradigm
  - Is there another way we can interpret probability that gives meaningful interpretation to all events without losing mathematical tractability and ease of estimation?

The Bayesian Approach

- The **Bayesian** approach treats probability as a **degree of belief** about whether an event will occur
- We maintain **prior** beliefs about the probabilities of events (called **a priori information**), and we use experimental evidence to update our beliefs according to **Bayes Rule**:

$$\pi(\theta|X) = \frac{\mathcal{L}(X|\theta)\pi(\theta)}{\pi(X)}$$

In the above equation:

- X is the **evidence**
- θ is event of interest
- $\pi(\theta)$ represents our **prior** degree of belief about θ
- $\mathcal{L}(X|\theta)$ is the **Likelihood** of observing the evidence given our event of interest has occurred

The Bayesian Approach (cont.)

- $\pi(\theta|X)$ represents our updated (a posteriori) beliefs about θ after observing evidence X
- $\pi(X)$ is a normalizing constant that ensures our posterior distribution integrates to 1 (ie ensures a proper posterior)
- In the frequentist setting, we treated X as the random variable, and θ as a set of fixed parameters
- In the Bayesian setting, we do the reverse - θ is a random variable, and our evidence is fixed
- The use of prior knowledge lets us treat Bayesian probabilities more subjectively - different people have different prior knowledge about a problem
- Let's see a simple example to illustrate these ideas...

Coin Flip Example

- Suppose we're flipping a two-sided coin (H, T), but we're not sure if the coin is fair
- There are two possibilities: $\theta \in \{fair, loaded\}$ corresponding to $P(H) = 0.5$, or $P(H) = 0.7$
- One way we can evaluate whether the coin is fair is by testing it. Suppose we flip it 5 times, and get 2H, 3T
- Let's assume the likelihood of receiving x heads in 5 flips follows a $Bin(5, ?)$ distribution

$$\mathcal{L}(x|\theta) = \binom{5}{x}(0.5)^5 I\{\theta = fair\} + \binom{5}{x}(0.7)^x(0.3)^{5-x} I\{\theta = loaded\}$$

$$\mathcal{L}(x = 2|\theta) = 0.3125 I\{\theta = fair\} + 0.1323 I\{\theta = loaded\}$$

- In the frequentist case, we choose the value of θ that is most likely to have generated our evidence (2H). Formally, $\hat{\theta}_{MLE} = argmin_\theta \mathcal{L}(x|\theta)$

Coin Flip Example

- This is called the **Maximum Likelihood Estimate** of θ, and for our likelihood, we can clearly see that $\hat{\theta}_{MLE} = \textit{fair}$, since

$$\mathcal{L}(2|\textit{fair}) = 0.3125 > 0.1323 = \mathcal{L}(2|\textit{loaded})$$

- But what if we knew in advance the coin was more likely to be loaded (ie suppose $P(\textit{loaded}) = 0.6$)? Under the Bayesian setting, we can use prior information to estimate θ:

$$\pi(\theta|X) = \frac{\mathcal{L}(X|\theta)\pi(\theta)}{\pi(X)} = \frac{\mathcal{L}(X|\theta)\pi(\theta)}{\sum_\theta \mathcal{L}(X|\theta)\pi(\theta)}$$

$$= \frac{\binom{5}{x}[(0.5)^5 I\{\theta = \textit{fair}\} \cdot (0.4) + (0.7)^x(0.3)^{5-x} I\{\theta = \textit{loaded}\} \cdot (0.6)]}{\binom{5}{x}[(0.5)^5(0.4) + (0.7)^x(0.3)^{5-x}(0.6)]}$$

Coin Flip Example

- Substituting our evidence:

$$\pi(\theta|X=2) = 0.612I\{\theta = \mathit{fair}\} + 0.388I\{\theta = \mathit{loaded}\}$$

- We can see that $P(\mathit{loaded}) = 0.388$. This answer is more intuitive than in the frequentist setting.
- What would happen if we were shown more evidence? **The posterior becomes the new prior**.
- Suppose in five more flips, we get 0H, 5T (recall our r.v. is the number of heads). We can update the old posterior with new evidence:

$$\pi(\tilde{\theta}|\tilde{x}) = \frac{\pi(\theta|x)\mathcal{L}(\tilde{x}|\tilde{\theta})}{\pi(\tilde{x})}$$

Coin Flip Example

$$= \frac{\binom{5}{0}[(0.5)^5 I\{\theta = \mathit{fair}\} \cdot (0.612) + (0.7)^0 (0.3)^{5-0} I\{\theta = \mathit{loaded}\} \cdot (0.388)]}{\binom{5}{0}[(0.5)^5 (0.612) + (0.7)^0 (0.3)^{5-0} (0.388)]}$$

$$\pi(\tilde{\theta}|\tilde{x}) = 0.953 I\{\theta = \mathit{fair}\} + 0.047 I\{\theta = \mathit{loaded}\}$$

- We can see that given the new evidence, our belief leans strongly towards the coin being fair
- This might seem confusing - Wouldn't a series of tails only be evidence of a loaded coin? But remember our definition of loaded was $P(H) = 0.7$
- Thus the absence of heads favors the lower of the two probabilities (fairness). For a more complete representation, we might give θ three possible values instead of two.
- Would the binomial likelihood still be appropriate in that case?

Conjugate Priors

- Sometimes, when we compute the posterior using the likelihood to account for new evidence, the resulting posterior distribution comes from the same family as the prior distribution
- When this happens, we say that the prior is **conjugate** for the likelihood $P(X|\theta)$
- This is very convenient mathematically, and happens both for discrete and continuous random variable distributions
- If the likelihood function is in the **exponential family**, then there always exists at least one conjugate prior, often also from the exponential family of distributions

Ex 1: Beta Prior, Bernoulli Likelihood

- Suppose our evidence is a set of N IID data points, $(x_1, \cdots, x_N)$, with each point sampled from a $Bern(\theta)$ distribution
- The joint likelihood of our evidence is thus:

$$\mathcal{L}(X|\theta) = \Pi_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum x_i}(1-\theta)^{N-\sum x_i}$$

- Suppose also that the prior distribution is $Beta(\alpha, \beta)$:

$$\pi(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}I(\theta \in [0,1])$$

  Note the presence of the indicator function that bounds θ in the prior.

- We can compute the posterior, using a special trick to first derive the normalizing constant:

Ex 1: Beta Prior, Bernoulli Likelihood

Note that since the beta distribution integrates to 1, we have that for the general beta r.v. t:

$$\int t^{a-1}(1-t)^{b-1} = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

- We can use this identity to isolate $\pi(X)$

$$\int \pi(\theta|X)d\theta = 1$$

$$\Rightarrow \frac{1}{\pi(X)} \int \mathcal{L}(X|\theta)\pi(\theta)d\theta = 1$$

$$\int \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} I(\theta \in [0,1])\theta^{\sum x_i}(1-\theta)^{N-\sum x_i} d\theta = \pi(X)$$

Ex 1: Beta Prior, Bernoulli Likelihood

$$\pi(X) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int \theta^{\alpha+\sum x_i - 1}(1-\theta)^{\beta+N-\sum x_i - 1} d\theta$$
$$= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot \frac{\Gamma(\alpha+\sum x_i)\Gamma(\beta+N-\sum x_i)}{\Gamma(\alpha+\beta+N)}$$

Back-substituting into Bayes Rule, we can derive the posterior:

$$\pi(\theta|X) = \frac{1}{\pi(X)} \cdot \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha+\sum x_i - 1}(1-\theta)^{\beta+N-\sum x_i - 1}$$
$$= \frac{\Gamma(\alpha+\beta+N)}{\Gamma(\alpha+\sum x_i)\Gamma(\beta+N-\sum x_i)} \theta^{\alpha+\sum x_i - 1}(1-\theta)^{\beta+N-\sum x_i - 1}$$
$$\sim Beta(\alpha + \sum x_i - 1, \beta + N - \sum x_i - 1)$$

Thus we see that the beta prior is conjugate for the Bernoulli likelihood.

Ex 2: Gamma Prior, Poisson Likelihood

- Suppose instead we had a *Gamma*(a, b) prior, meaning

$$\pi(\theta) = \frac{b^a}{\Gamma(a)} \theta^{a-1} e^{-b\theta} I(\theta > 0) \ ; \ a, b > 0$$

- Suppose too that our data are now IID according to a *Poisson*(θ):

$$\mathcal{L}(X|\theta) = \Pi_{i=1}^{N} \frac{\theta^{x_i} e^{-\theta}}{x_i!} = \frac{1}{\Pi_{i=1}^{N} x_i!} \theta^{\sum x_i} e^{-N\theta}$$

for $x_i \in \{0, 1, 2, \cdots\}$, meaning the Poisson distribution has a discrete support (non-negative whole numbers, or counts).

- As in the previous example, we can derive the posterior:

$$\pi(\theta|X) = \frac{1}{\pi(X)} \frac{1}{\Pi_{i=1}^{N} x_i!} \frac{b^a}{\Gamma(a)} \theta^{\sum x_i} e^{-N\theta} \theta^{a-1} e^{-b\theta} I(\theta > 0)$$

Ex 2: Gamma Prior, Poisson Likelihood

- We can use a similar trick to the one from the beta binomial example:

$$\int \pi(\theta|X) = 1$$

$$\Rightarrow \pi(X) = \frac{1}{\prod_{i=1}^N x_i!} \frac{b^a}{\Gamma(a)} \int \theta^{\sum x_i + a - 1} e^{-(N+b)\theta} I(\theta > 0) d\theta$$

$$\Rightarrow \pi(X) = \frac{1}{\prod_{i=1}^N x_i!} \frac{b^a}{\Gamma(a)} \frac{\Gamma(a + \sum x_i)}{(b + N)^{a + \sum x_i}}$$

Where we rely on the definition of the gamma distribution to convert the integral in the second line. Back-substituting,

$$\pi(\theta|X) = \frac{1}{\pi(X)} \frac{1}{\prod_{i=1}^N x_i!} \frac{b^a}{\Gamma(a)} \theta^{\sum x_i + a - 1} e^{-(N+b)\theta} I(\theta > 0)$$

Ex 3: Gamma Prior, Poisson Likelihood

$$
\begin{aligned}
&= \frac{(b+N)^{a+\sum x_i}}{\Gamma(a+\sum x_i)} \theta^{\sum x_i + a - 1} e^{-(N+b)\theta} I(\theta > 0) \\
&\sim Gamma(a + \sum x_i, b + N)
\end{aligned}
$$

- Thus the Gamma prior is conjugate for the Poisson likelihood.
- Note that the mean of a gamma distribution is simply the ratio of its parameters ($\frac{\alpha}{\beta}$)
- So our gamma posterior's mean (which is our updated belief about the average value of θ) will be high when:
  - prior beliefs indicate a high value ($a >> b$)
  - Evidence suggests θ has a high value ($\sum x_i / N >> 0$)

Ex 3: Normal Prior, Normal Likelihood

- Now suppose $x_1, \cdots, x_N \sim_{IID} N(\theta, 1)$ and $\theta \sim N(\mu_0, \sigma_0^2)$, where the prior parameters are unknown constants

$$\mathcal{L}(X|\theta) = \prod_{i=1}^{N} f_{x_i}(x_i|\theta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x_i - \theta)^2}{2} \right\}$$
$$= (2\pi)^{N/2} \exp\left\{ -\sum_{i=1}^{N} \frac{(x_i - \theta)^2}{2} \right\}$$
$$\pi(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left\{ -\frac{(\theta - \mu_0)^2}{2\sigma_0^2} \right\}$$

- To make our lives easier, we ignore the coefficients on the distributions (including the normalizing constant):

Ex 3: Normal Prior, Normal Likelihood

$$
\begin{aligned}
\pi(\theta|X) &\propto \exp\left\{-\sum_{i=1}^{N}\frac{(x_i-\theta)^2}{2}-\frac{(\theta-\mu_0)^2}{2\sigma_0^2}\right\} \\
&= \exp\left\{\frac{\sum x_i^2+2\theta\sum x_i+N\theta^2}{2}-\frac{\theta^2-2\mu_0\theta+\mu_0^2}{2\sigma_0^2}\right\} \\
&= \exp\left\{\frac{-\sigma_0^2(\sum x_i^2+2\theta\sum x_i+N\theta^2)+\theta^2+2\mu_0\theta+\mu_0^2}{2\sigma_0^2}\right\} \\
&= \exp\left\{\frac{\theta^2(1+N\sigma_0^2)-2\theta(\mu_0+\sum\sigma_0^2x_i)-(\mu_0^2+\sigma_0^2\sum x_i^2)}{2\sigma_0^2}\right\}
\end{aligned}
$$

Any term in our exponent that does not involve θ can be seen as part of the normalizing constant, and can be ignored

Ex 3: Normal Prior, Normal Likelihood

$$\pi(\theta|X) \propto \exp\left\{ \frac{\theta^2(1 + N\sigma_0^2) - 2\theta(\mu_0 + \sum \sigma_0^2 x_i)}{2\sigma_0^2} \right\}$$

$$\propto \exp\left\{ \frac{\theta^2 - 2\theta \frac{(\mu_0 + \sum \sigma_0^2 x_i)}{(1+N\sigma_0^2)}}{\frac{2\sigma_0^2}{(1+N\sigma_0^2)}} \right\}$$

We can complete the square $(ax^2 + bx + c)$, converting the interior from standard quadratic form to vertex form:

$$\propto \exp\left\{ -\frac{1}{2} \frac{(\theta - \frac{(\mu_0 + \sum \sigma_0^2 x_i)}{(1+N\sigma_0^2)})^2}{\frac{\sigma_0^2}{(1+N\sigma_0^2)}} \right\} \sim N(\mu = \frac{(\mu_0 + \sum \sigma_0^2 x_i)}{(1 + N\sigma_0^2)}, \tau^2 = \frac{\sigma_0^2}{(1 + N\sigma_0^2)})$$

Thus the normal prior is conjugate for the normal likelihood. Note that this result still holds for a non-unit likelihood variance (try and show this yourself!)

Non-informative Priors

- You may be wondering, how exactly do we choose a specification for our prior when we have no evidence?
- This usually depends on the parameter of interest θ
- When we don't have strong prior beliefs and want to let the evidence speak for itself, we often use a **non-informative** or **diffuse** prior
- Technically all priors carry at least some information about the parameter, so it is a misnomer, but the idea with this type of prior is to eliminate parameter values that will never occur, while maintaining as flat a distribution as possible
- For example, we may use a $Unif(0,1)$ prior to describe a percentage, or a $Gamma(a,b)$ distribution for the price of a hamburger (strictly non-negative)

Weakly Informative Priors

- Usually we do have some knowledge about the scale of our parameter of interest
- We want our prior to be **weakly informative**; it rules out unreasonable values, but still allows for all possible values that could occur (even if they rarely do)
- Roughly speaking (though not always) the information contained within a prior is inversely proportional to its variance (thus wider and flatter distributions are less informative)
- A common weakly informative prior is the $N(0, 1)$ distribution, where the units of the parameter of interest are appropriately scaled
- It is also common to use a $Cauchy(0, \gamma)$ prior, or the student-t (which interpolates between Normal and Cauchy), truncating the distribution when the parameter is strictly positive

Informative Priors

- Sometimes priors can be used to encorporate important information into the model
- Ex: Suppose we want to estimate tomorrow's trading volume of Apple stock on the NYSE
- We can examine historical trading data to determine the average trading volume m, and sample variance s^2
- A reasonable prior in this case would be $N(m, s^2)$; In general such numerical information may come from literature reviews, or previous analysis
- Note that the above computations ignore the fact that trading data tends to have high serial correlation (non-IID), so simply computing the sample variance may not be a good estimator (the example is purely illustrative)

Improper Priors

- Recall our Beta prior, Bernoulli likelihood example earlier
- Instead of using a mathematically convenient prior, we might ask: What is the least informative prior for the Bernoulli likelihood?
- Intuition might tell you to use a $U(0,1)$. This yields a posterior distribution of $\pi(\theta|y) = Beta(1 + \sum y_i, 1 + n - \sum y_i)$ (try to show this yourself)
- The Maximum likelihood estimate is simply $\frac{\sum y_i}{n}$, whereas the posterior mean is $\frac{\sum y_i + 1}{n+2}$, so clearly the standard uniform prior still carries information!
- The general form of the posterior mean for a $Beta(\alpha, \beta)$ prior is

$$E(\theta|x) = \frac{\alpha + \sum y_i}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} \cdot \frac{\alpha}{\alpha + \beta} + \frac{n}{\alpha + \beta + n} \cdot \frac{\sum y_i}{n}$$

Improper Priors

$$posterior\ mean = prior\ weight \cdot prior\ mean + data\ weight \cdot data\ mean$$

- Notice that in order for the prior to have no effect on the posterior (ie to carry no information), we require $\alpha = \beta = 0$
- This corresponds to a prior of $Beta(0, 0) = \frac{1}{\theta(1-\theta)}$, which is known as the **Haldane Prior**
- However the limiting Beta coefficient on the Haldane Prior is infinite, thus $\int \pi(\theta) d\theta > 1$
- Prior distributions that do not integrate to 1 are called **improper**, and can still be used successfully as long as the resulting posterior is proper (as was shown above)

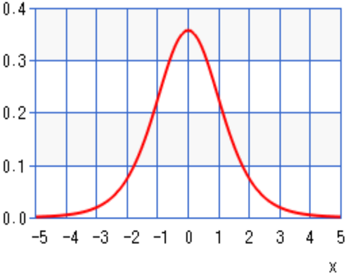
Jeffreys Priors

- Suppose we have a flat prior (ie $\theta \sim U(0,1)$)
- If we are ignorant about θ, then we should also be ignorant about $\phi = \log \frac{\theta}{1-\theta}$
- By method of CDF, if $F_\theta(t) = t$:

$$\begin{aligned}
F_\phi(t) &= Pr(\phi \le t) = Pr(\log\left(\frac{\theta}{1-\theta}\right) \le t) \\
&= Pr(\frac{\theta}{1-\theta} \le e^t) = Pr(\theta \le e^t - \theta e^t) \\
&= Pr(\theta \le \frac{e^t}{1+e^t}) = Pr(\theta \le \frac{1}{1+e^{-t}}) \\
&= F_\theta(\frac{1}{1+e^{-t}}) = \frac{1}{1+e^{-t}} \\
&\sim Logistic(0,1)
\end{aligned}$$

Jeffreys Priors

- The above distribution is not at all flat. It carries much more information (see below)



- This is because flat priors are not well defined. They are not **transformation invariant**
- **Jeffreys Prior**: Use $\pi(\theta) \propto I(\theta)^{1/2}$, where $I(\theta)$ is the **Fisher Information** of θ. This will be transformation invariant.

Jeffreys Prior Example: Exponential Distribution

- Suppose our likelihood follows an exponential distribution: $f(x|\theta) = \theta e^{-\theta x}$ (for non-negative x)
- Recall the **score function**:

$$s(\theta) = \frac{\partial}{\partial \theta} \log f(x|\theta) = \frac{1}{\theta} - x$$

- When the log-likelihood is twice differentiable, the Fisher information is the negative expectation of its second derivative:

$$I(\theta) = -E[\frac{\partial^2}{\partial \theta^2} \log f(x|\theta)] = -E[\frac{\partial}{\partial \theta} s(\theta)] = \frac{1}{\theta^2}$$

- **Jeffreys Rule**: Use $\pi(\theta) \propto \sqrt{\frac{1}{\theta^2}} = \frac{1}{\theta}$
- We will not prove transformational invariance here, but I encourage you to try and do so

Posterior Inference

- So now that we've identified the posterior distribution, what can we do with it?
- The first (and most obvious) calculations to find are point estimates, usually that summarize the center
  - Mean: $E(\theta|x)$
  - Median: $\hat{\theta} : \int_{-\infty}^{\hat{\theta}} P(\theta|x)d\theta = 0.5$
  - Mode: $argmax_{\theta} P(\theta|x)$
- We can also compute intervals with the posterior distribution
- These intervals are called **Bayesian Credible Intervals**
- A $(1-\alpha)\%$ interval is a credible interval if the probability that θ is contained in the interval is $1-\alpha$

Credible Intervals

- Such credible intervals are not unique on a posterior distribution. So how do we choose the end points?
- **Equi-tailed Interval**: Choose the interval such that the posterior probability of being below the interval is identical to the probability of being above ($\alpha/2$ in each tail)
- **Highest Posterior Density**: Choose the narrowest interval, which for a unimodal posterior means choosing the values with the highest posterior density (this includes the mode)
- We could also simply construct an interval centred around the posterior mean
- Regardless of the method, notice that unlike with confidence intervals, credible intervals are probability statements about θ

Posterior Sampling

- We've discussed a lot in these slides, mostly about deriving posterior distributions and conducting inference with them
- What happens when we cannot analytically derive a posterior? What do we do?
- Turns out we don't need to find the exact form of the posterior - We only need to be able to collect a sample from it!
- This will be the subject of the next set of slides: **Bayesian Inference II**