MTH 7241 Fall 2022: Prof. C. King

Notes 7: Bayesian Inference

0.1 Bayesian vs frequentist

The distinction arises from the interpretation of the probability of an event. We write $\mathbb{P}(A)$ in both cases, but the meaning is a little different.

0.1.1 Frequentist 'classical' meaning

$\mathbb{P}(A)$ is the long-run fraction of occurrences of the event A under repeated independent sampling. This leads to, for example, the weak law of large numbers: make independent measurements of a random variable $X_1, \ldots, X_n, \ldots$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mathbb{E}[X]$$

$\mathbb{P}(A|B)$ is the long-run fraction of occurences of event A under repeated independent sampling when restricted to outcomes where B occurs.

0.1.2 Bayesian meaning

$\mathbb{P}(A)$ is your subjective belief of how likely it is that event A will occur.

$\mathbb{P}(A|B)$ is your updated belief of how likely it is that event A will occur given that event B has occurred.

Recall Bayes rule:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)} \tag{1}$$

0.1.3 Frequentist meaning

A and B are two random events, and it makes sense to consider both $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ (so long as both $\mathbb{P}(A)$ and $\mathbb{P}(B)$ are not zero). Bayes rule is the formula that gives the relation between these conditional probabilities. The order of events A, B in the conditional probabilities has no particular significance.

0.1.4 Bayesian meaning

$\mathbb{P}(A)$ is your belief in the likelihood of event A, and $\mathbb{P}(A|B)$ is your updated belief based on the fact that event B has occurred. Bayes rule tells you how to update your belief based on the measured data. The order of events A, B in the conditional probabilities is significant.

Bayesian application of Bayes rule:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\,\mathbb{P}(A)}{\mathbb{P}(B)} \tag{2}$$

In most applications, the event B is measurement of some data, for example some independent measurements of a random variable $X_1, \ldots, X_n$. So we use $\mathcal{D}$ to indicate the data measured. The event A is our guess about the value of some set of parameters θ which determine the distribution of X. So $\mathbb{P}(A) = \mathbb{P}_0(\theta)$ is called the *prior* distribution and represents our belief about θ before the measurements are taken. Then $\mathbb{P}(B|A) = \mathbb{P}(\mathcal{D}|\theta)$ is called the *likelihood* and $\mathbb{P}(A|B) = \mathbb{P}_1(\theta|\mathcal{D})$ is the *posterior*. The normalizing constant $\mathbb{P}(B) = \mathbb{P}(\mathcal{D})$ is called the *evidence*. The formula is used in cases where distributions are discrete, continuous, or a mixture of the two. So the formula is

$$\mathbb{P}_1(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\,\mathbb{P}_0(\theta)}{\mathbb{P}(\mathcal{D})}$$

The normalizing constant can be computed by summing over possible values for θ:

$$\mathbb{P}(\mathcal{D}) = \sum_j \mathbb{P}(\mathcal{D}|\theta_j)\,\mathbb{P}_0(\theta_j)$$

Example: Let's look at an example where we want to decide between two competing hypotheses. Suppose your friend claims that she can predict the outcome of a coin toss with 100% certainty.
H_0: your friend is lying, and she has 50% probability of guessing the outcome.
H_1: your friend is telling the truth.
We want to decide between H_0 and H_1. Our prior distribution is

$$\mathbb{P}(H_0) = 0.99 = 1 - \mathbb{P}(H_1)$$

We measure data: a coin is tossed 5 times, and your friend predicts the outcome every time. What is your updated belief in H_0?

Bayesian inference provides a systematic way to answer this question. Let $\mathcal{D}$ denote the outcomes of the 5 coin tosses, then

$$\mathbb{P}(\mathcal{D}|H_0) = \left(\frac{1}{2}\right)^5 = \frac{1}{32}$$

$$\mathbb{P}(\mathcal{D}|H_1) = 1$$

Using Bayes rule we get

$$\mathbb{P}(H_0|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|H_0)\,\mathbb{P}(H_0)}{\mathbb{P}(\mathcal{D})}$$

To compute $P(\mathcal{D}$ we use the total probability formula:

$$P(\mathcal{D}) = \mathbb{P}(\mathcal{D}|H_0)\,\mathbb{P}(H_0) + \mathbb{P}(\mathcal{D}|H_1)\,\mathbb{P}(H_1)$$

Putting it together we find

$$\mathbb{P}(H_0|\mathcal{D}) = \frac{0.99/32}{0.99/32 + 0.01} = 0.756$$

So our confidence that your friend is lying has been reduced to 75.6% based on her successes.

Bayesian inference works just as well when we investigate continuous parameters. Here is the typical setting: the prior is a pdf for the parameter θ, say $f_0(\theta)$. The posterior f_1 is computed using some data $\mathcal{D}$ whose distribution depends on θ:

$$f_1(\theta|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\theta)\, f_0(\theta)}{\mathbb{P}(\mathcal{D})}$$

Here the evidence is computed using an integral with the prior pdf:

$$P(\mathcal{D}) = \int P(\mathcal{D}|\theta)\, f_0(\theta)\, d\theta$$

Example: You are given a coin. It is biased with $\mathbb{P}(H) = q$ but you know nothing about the value of q. You toss the coin n times and it comes up Heads k times, and Tails $n - k$ times. Find the posterior distribution of q. What is the probability of getting Heads on the next toss?

We will assume that the 'prior' distribution for q is uniform on $[0,1]$, indicating our complete lack of knowledge. We will indicate this by $U(q)$. Let $\mathcal{D}$ denote the observed sequence of k Heads in n tosses, then the Bayesian update for q is

$$f_1(q|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|q)U(q)}{\int_0^1 dq\mathbb{P}(\mathcal{D}|q)U(q)} \tag{3}$$

Now

$$P(\mathcal{D}|q) = \binom{n}{k} q^k(1-q)^{n-k} \tag{4}$$

and $U(q) = 1$. So the normalization on the right side (aka the evidence) is

$$\int_0^1 dq\mathbb{P}(\mathcal{D}|q)U(q) = \binom{n}{k} \int_0^1 q^k(1-q)^{n-k}dq = \binom{n}{k} \frac{k!(n-k)!}{(n+1)!} \tag{5}$$

So the posterior for q is the Beta distribution

$$f_1(q|\mathcal{D}) = \frac{\binom{n}{k} q^k(1-q)^{n-k}}{\int_0^1 dq\mathbb{P}(\mathcal{D}|q)U(q)} = \frac{(n+1)!}{k!(n-k)!} q^k(1-q)^{n-k}$$

Let H be the event to get Heads on the next toss. Then

$$\mathbb{P}(H|\mathcal{D}) = \int_0^1 \mathbb{P}(H|\mathcal{D}, q)\, f_1(q|\mathcal{D})dq = \int_0^1 \mathbb{P}(H|q)\, f_1(q|\mathcal{D})dq \tag{6}$$

Since $\mathbb{P}(H|q) = q$ we get

$$\begin{aligned} \mathbb{P}(H|\mathcal{D}) &= \frac{(n+1)!}{k!(n-k)!} \int_0^1 q\, q^k (1-q)^{n-k} dq & (7) \\ &= \frac{(n+1)!}{k!(n-k)!} \frac{(k+1)!(n-k)!}{(n+2)!} & (8) \\ &= \frac{k+1}{n+2} & (9) \end{aligned}$$

This is known as Laplace's Rule.

Point estimate: given the posterior distribution $f_1(\theta|\mathcal{D})$, what is the best estimate for parameter θ? Of course there is no single answer to this, but a popular choice is the posterior mean, namely

$$\theta_{mean} = \int \theta \, f_1(\theta|\mathcal{D}) \, d\theta$$

Another popular choice is the maximum likelihood estimator θ_{MLE}: this is the value of θ which maximizes the posterior distribution $f_1(\theta|\mathcal{D})$.

Example: Return to the previous coin tossing example. The posterior is

$$f_1(q|\mathcal{D}) = \frac{(n+1)!}{k!(n-k)!}\, q^k (1-q)^{n-k}$$

We already computed the mean:

$$q_{mean} = \int_0^1 q\, f_1(q|\mathcal{D})\, dq = \frac{k+1}{n+2}$$

The MLE is found by maximizing $f_1(q|\mathcal{D})$ over q:

$$\frac{d}{dq} f_1(q|\mathcal{D}) = 0 \Leftrightarrow q_{MLE} = \frac{k}{n}$$

Note: when the prior distribution is uniform the estimator θ_{MLE} is the same as the usual MLE computed using the likelihood function; but if the prior is not uniform then it is different in general.

Credibility interval: going beyond the point estimate, we often want an interval estimate for θ. This is not the same as the confidence interval we encounter in usual hypothesis testing, but they are often quite close. Suppose that θ is a real parameter. We say that the interval (a, b) is a $100(1-\alpha)\%$ credibility interval for θ if

$$\mathbb{P}(a < \theta < b|\mathcal{D}) = 1 - \alpha$$

where the probability on the left side is computed using the posterior pdf $f_1(\theta|\mathcal{D})$. Note that here θ is a random variable, and we are finding numbers a, b so that the probability that θ will lie between a, b is 0.95.

Example: Contrast credibility interval with confidence interval for the previous coin tossing example. Suppose for concreteness that $n = 25$ and $k = 12$. The posterior is

$$f_1(q|\mathcal{D}) = \binom{25}{12} 26\, q^{12}\,(1-q)^{13} \quad 0 \le q \le 1$$

The point estimates are

$$q_{mean} = \frac{13}{27} = 0.481, \qquad q_{MLE} = \frac{12}{25} = 0.48$$

Can check that

$$\int_{0.297}^{0.663} f_1(q|\mathcal{D})\, dq = 0.95$$

So $(0.297, 0.663)$ is a (symmetric) 95% credibility interval for q given the data $\mathcal{D}$. [This was found by trial and error].

What would we say about a confidence interval for q given the same data? In this case we would be saying that q is some fixed number q_0 which is unknown to us, and we are trying to estimate q_0 by taking measurements. Our point estimate would be

$$\widehat{q} = \frac{k}{n} = \frac{12}{25} = 0.48$$

To get the 95% confidence interval for q, we would find numbers c, d so that

$$\mathbb{P}(c < \widehat{q} < d) = 0.95$$

The meaning is: if we repeat the measurement many times (ie tossing the coin 25 times and counting how many Heads each time), then in the long-run q_0 will lie in this interval 95% of the time. Now by the usual normal approximation we have

$$\widehat{q} \sim \mathcal{N}(q_0, \frac{\sigma^2}{n}), \quad \sigma^2 = \frac{q_0(1-q_0)}{n}$$

We approximate σ^2 by using $\widehat{q}$ in place of q_0. Then the 95% confidence interval is

$$\widehat{q} \pm 1.96 \left(\frac{\widehat{q}(1-\widehat{q})}{25} \right)^{1/2} = 0.48 \pm 0.196$$

So $(0.284, 0.676)$ is the 95% confidence interval for q_0. Notice that the intervals are different, but close. This is often the case, but there are instances where they can be very different.

There are some classes of models where analytical solutions are available. We will discuss a few of these.

The multinomial model generalizes coin tossing. There are K possible outcomes $\{1, 2, \ldots, K\}$ with probabilities $Q = (q_1, \ldots, q_K)$, satisfying $\sum_{i=1}^{K} q_i = 1$. The likelihood for a sequence of IID measurements $\mathcal{D} = (X_1, \ldots, X_n)$ is

$$\mathbb{P}(\mathcal{D}|Q) = \prod_{i=1}^{n} q_{X_i}$$

This can be put in a more useful form by describing the data with indicator random variables, as follows: define

$$Y_{ij} = \begin{cases} 1 & \text{if } X_i = j \\ 0 & \text{else} \end{cases}, \qquad i = 1, \ldots, n, \qquad j = 1, \ldots, K.$$

Then we can write

$$q_{X_i} = \prod_{j=1}^{K} q_j^{Y_{ij}}$$

So the likelihood function for all the data is

$$\begin{aligned} \mathbb{P}(\mathcal{D}|Q) &= \prod_{i=1}^{n} \prod_{j=1}^{K} q_j^{Y_{ij}} \\ &= \prod_{j=1}^{K} q_j^{\sum_{i=1}^{n} Y_{ij}} \end{aligned}$$

Dirichlet distribution
Define the probability simplex in $\mathbb{R}^K$:

$$\Delta_K = \{x = (x_1, \ldots, x_k) \in \mathbb{R}^K \,|\, 0 \leq x_i \leq 1,\ \sum_{j=1}^{K} x_j = 1\}$$

Every point in Δ_K is a probability distribution for a set of size K. We will describe a type of probability distribution on Δ_K called a Dirichlet distribution. Let $\alpha = (\alpha_1, \ldots, \alpha_K) \in \mathbb{R}^K$ where $\alpha_i \geq 0$ for all $i = 1, \ldots, K$. The Dirichlet distribution defined by α is

$$\pi_\alpha(x) = N(\alpha) \prod_{j=1}^{K} x_j^{\alpha_j - 1}, \quad x \in \Delta_K$$

The number $N(\alpha)$ is the normalizing constant which makes this a pdf. We won't really need it, but the value is

$$N(\alpha) = \frac{\Gamma(\sum_{j=1}^{K} \alpha_j)}{\prod_{j=1}^{K} \Gamma(\alpha_j)}$$

where Γ is the gamma-function. The particular case $K = 2$ corresponds to coin tossing.

The mean of the Dirichlet is

$$\mathbb{E}[x_i] = \int_{\Delta_K} x_i \, \pi_\alpha(x) \, dx = \frac{\alpha_i}{\sum_{i=1}^K \alpha_j}$$

The special case $\alpha = (1, 1, \ldots, 1)$ gives the uniform distribution on Δ_K. Also the product of two Dirichlet distributions is (up to normalization) again Dirichlet:

$$\begin{aligned} \pi_\alpha(x) \, \pi_\beta(x) &= N(\alpha) \, N(\beta) \prod_{j=1}^K x_j^{\alpha_j + \beta_j - 2} \\ &\propto \pi_{\alpha+\beta-1} \end{aligned}$$

Returning to the multinomial we see that the likelihood function is a Dirichlet distribution. It is common practice to also use a Dirichlet distribution for the prior, so that

$$f_0(Q) = \pi_\alpha(Q)$$

for some $\alpha = (\alpha_1, \ldots, \alpha_K)$. In this case the posterior is again a Dirichlet:

$$\begin{aligned} f_1(Q|\mathcal{D}) &\propto \prod_{i=1}^{n} \prod_{j=1}^{K} q_j^{Y_{ij}} \prod_{j=1}^{K} q_j^{\alpha_j - 1} \\ &\propto \prod_{j=1}^{K} q_j^{\sum_{i=1}^{n} Y_{ij} + \alpha_j - 1} \\ &\sim \text{Dirichlet}\left(\alpha_1 + \sum_{i=1}^{n} Y_{i1}, \ldots, \alpha_K + \sum_{i=1}^{n} Y_{iK}\right) \end{aligned}$$

So for example we immediately get the posterior means:

$$\mathbb{E}[q_j|\mathcal{D}] = \frac{\alpha_j + \sum_{i=1}^{n} Y_{ij}}{\sum_{l=1}^{K} \alpha_l + n}$$

Example: Suppose $K = 3$, so we have a multinomial distribution for three types, with probabilities $q = (q_1, q_2, q_3)$, where $q_1 + q_2 + q_3 = 1$. Assume that the prior distribution is uniform on the simplex, so this is Dirichlet with $\alpha = (1, 1, 1)$. The pdf is

$$f_0(q) = \begin{cases} 2 & \text{for } q_1 + q_2 + q_3 = 1 \\ 0 & \text{else} \end{cases}$$

Suppose we make 10 measurements and get the results

$$\mathcal{D} = (X_1, \ldots, X_{10}) = (1, 3, 3, 2, 3, 1, 1, 2, 1, 3).$$

So we have for example

$$Y_{11} = 1, Y_{12} = 0, Y_{13} = 0, Y_{21} = 0, Y_{22} = 0, Y_{23} = 1, \ldots$$

Then the posterior is again a Dirichlet distribution with parameters

$$\alpha' = \left(1 + \sum_{i=1}^{10} Y_{i1}, 1 + \sum_{i=1}^{10} Y_{i2}, 1 + \sum_{i=1}^{10} Y_{i2}\right) = (5, 3, 5)$$

So the posterior pdf is

$$f_1(q \,|\, \mathcal{D}) = N\, q_1^{\alpha_1' - 1}\, q_2^{\alpha_2' - 1}\, q_3^{\alpha_3' - 1} = 8316\, q_1^4\, q_2^2\, q_3^4$$

We can calculate the point estimates: the mean is

$$q_{1,mean} = \frac{5}{13}, \quad q_{2,mean} = \frac{3}{13}, \quad q_{3,mean} = \frac{5}{13}$$

and the MLE is

$$q_{1,MLE} = \frac{4}{10}, \quad q_{2,MLE} = \frac{2}{10}, \quad q_{3,MLE} = \frac{4}{10}$$

The normal model also allows some explicit computations. In this case we have a normal r.v. $X \sim \mathcal{N}(\theta, \sigma^2)$ where θ is the unknown mean, and we assume that the variance σ^2 is known. The data is IID measurements

$$\mathcal{D} = (X_1, \ldots, X_n)$$

For the prior we take a normal for θ;

$$\theta \sim \mathcal{N}(a, b^2)$$

Then can check that the posterior is also normal:

$$\theta \Big| \mathcal{D} \sim \mathcal{N}(\overline{\theta}, \tau^2)$$

where the parameters are

$$\begin{aligned} \overline{\theta} &= w\,\overline{X} + (1-w)\,a \\ w &= \frac{nb^2}{\sigma^2 + nb^2} \\ \tau^2 &= \frac{\sigma^2}{n}\,w \end{aligned}$$

The 95% credibility interval is

$$\overline{\theta} \pm 1.96\,\tau$$

The 95% confidence interval is

$$\overline{X} \pm 1.96\,\sigma$$

However for many (most) practical applications it is infeasible to derive exact results, and it is necessary to use some numerical computational methods to compute the posterior.

Example 1 *This problem appears in MacKay's book: Unstable particles are emitted from a source and decay at a distance x, a real number that has an exponential probability distribution with characteristic length λ. Decay events can be observed only if they occur in a window extending from $x = 1$ cm to $x = 20$ cm. N decays are observed at locations $x_1, \ldots, x_N$. It is known that λ lies somewhere in the interval $[5, 25]$. What is the estimate of λ based on the data?*

The problem here is to find a distribution for λ based on the evidence, which is the observed set of positions. Bayes rule is

$$P(\lambda \,|\, \{x_1, \ldots, x_N\}) = \frac{P(\{x_1, \ldots, x_N\}|\lambda) P(\lambda)}{P(\{x_1, \ldots, x_N\})} \tag{10}$$

The conditional probability on the right side is

$$P(\{x_1, \ldots, x_N\}|\lambda) = P(x_1|\lambda) \cdots P(x_N|\lambda) \tag{11}$$

where

$$P(x|\lambda) = \begin{cases} \frac{1}{\lambda} \frac{1}{Z(\lambda)} e^{-x/\lambda} & 1 \le x \le 20 \\ 0 & \textit{otherwise} \end{cases} \tag{12}$$

The factor $Z(\lambda)$ is chosen to normalize the probability:

$$Z(\lambda) = \lambda^{-1} \int_1^{20} e^{-x/\lambda} dx = e^{-1/\lambda} - e^{-20/\lambda} \tag{13}$$

The prior density for λ is uniform on $[5, 25]$, so

$$P(\lambda) = 1/20 \tag{14}$$

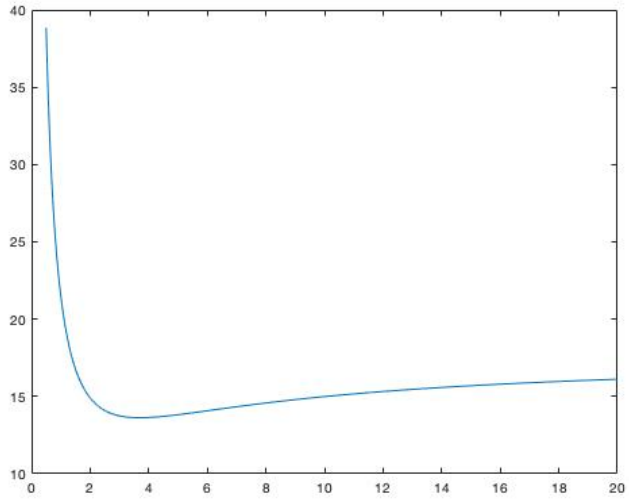
Thus we get

$$P(\lambda \,|\, \{x_1, \ldots, x_N\}) = \frac{1}{P(\{x_1, \ldots, x_N\})} \left(\frac{1}{\lambda Z(\lambda)} \right)^N e^{-\sum x_i/\lambda} \tag{15}$$

We get our estimate of λ by using the maximum likelihood method: we choose the value of λ which maximizes the right side of this expression. Taking the logarithm we see that we want to minimize the function

$$g(\lambda) = \frac{\sum x_i}{\lambda} + N \log(\lambda Z(\lambda)) \tag{16}$$

The figure below shows a plot of $g(\lambda)$ for the case $N = 6$, and measured values $\{1.5, 2, 3, 4, 5, 12\}$, so $\sum x_i = 27.5$.



The function has a unique minimum at $\lambda = 3.7$, so this is our maximum likelihood estimate.