Statistical Rethinking: Notes and Selected Exercises

Somnath Sikdar

June 23, 2021

Contents

Chapter 2

Small Worlds and Large Worlds

Exercise 2.1. Suppose there are two species of panda bear. Both are equally common in the wild and live in the same places. They look exactly alike and eat the same food, and there is yet no genetic assay capable of telling them apart. They differ however in their family sizes. Species A gives birth to twins 10% of the time, otherwise birthing a single infant. Species B births twins 20% of the time, otherwise birthing singleton infants. Assume these numbers are known with certainty, from many years of field research. Now suppose you are managing a captive panda breeding program. You have a new female panda of unknown species, and she has just given birth to twins. What is the probability that her next birth will also be twins?

Solution. We have to estimate $\Pr\{\text{twins again}|\text{twins before}\}$. We may write this conditional probability as follows:

$$\begin{aligned}\Pr\{\text{twins again}|\text{twins before}\} &= \Pr\{\text{twins again}|A,\text{twins before}\}\cdot\Pr\{A|\text{twins before}\}+\\ &\quad\Pr\{\text{twins again}|B,\text{twins before}\}\cdot\Pr\{B|\text{twins before}\}\\ &= \Pr\{\text{twins again}|A\}\cdot\Pr\{A|\text{twins before}\}+\\ &\quad\Pr\{\text{twins again}|B\}\cdot\Pr\{B|\text{twins before}\} \qquad (2.1)\end{aligned}$$

The last equality follows, since when we condition on any given species, the probability of having twins again is independent of whether twins were born before. Strictly speaking, this is also an assumption but a reasonable one. As such, we do not use the "before" and "again" qualifiers. To evaluate the last expression, we calculate $\Pr\{\cdot|\text{twins}\}$ for both species A and B. Using Bayes' Theorem, we obtain:

$$\begin{aligned}\Pr\{A|\text{twins}\} &= \frac{\Pr\{A|\text{twins}\}\cdot\Pr\{A\}}{\Pr\{A|\text{twins}\}\cdot\Pr\{A\}+\Pr\{B|\text{twins}\}\cdot\Pr\{B\}}\\ &= \frac{0.10\times 0.5}{0.5\times(0.1+0.20)}\\ &= \frac{1}{3}.\end{aligned}$$

A similar calculation yields $\Pr\{B|\text{twins}\}=\frac{2}{3}$. The final expression in (2.1) evaluates to

$$0.10\times\frac{1}{3}+0.20\times\frac{2}{3}=\frac{1}{6}=0.17.$$

This probability lies between 0.10 and 0.20 as expected and lies closer to 0.20 than 0.10. This is also clear since if the panda has had twins before, it is more likely to be of species B than species A. Hence it is more likely that she will birth twins again.

Chapter 3

Notes on WAIC and LOO

These notes are based on the paper *Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC* by Aki Vehtari, Andrew Gelman and Jonah Gabry.

Consider data $y = (y_1, \ldots, y_n)$ modeled as independent given parameters θ. We can then write: $p(y \mid \theta) = \prod_i p(y_i \mid \theta)$. Firstly, why is this assumption even made in Bayesian statistics? To understand this, assume that we have a prior distribution $p(\theta)$ and consider the case when we have used some data y to obtain the posterior $p(\theta \mid y)$. Let's suppose that we have some more data z. The new posterior is $p(\theta \mid y, z)$. Does this depend on the *order* in which we have seen the data? Intuitively, it shouldn't. Consider the expression for $p(\theta \mid y, z)$:

$$p(\theta \mid y, z) = \frac{p(y, z \mid \theta) p(\theta)}{p(y, z)}.$$

If we assume that the data are independent given the parameters, then the right-hand side simplifies to:

$$\frac{p(y \mid \theta) p(z \mid \theta) p(\theta)}{\int p(y \mid \theta') p(z \mid \theta') p(\theta') \mathrm{d}\theta'},$$

which does not depend on the order in which the data y and z arrive. Thus there are two primary reasons for making the data independence assumption given the parameters: first, it allows us to write the joint distribution as a product of the marginals and second, it guarantees that the posterior is the same irrespective of the order in which the data is seen.

The other distribution of interest here is the *posterior predictive distribution* $p(\tilde{y} \mid y)$. This is

the distribution of the unobserved values $\tilde{y}$ given the observed values y.

$$\begin{aligned} p(\tilde{y}\,|\,y) &= \int p(\tilde{y}, \theta\,|\,y)\mathrm{d}\theta \\ &= \int p(\tilde{y}\,|\,\theta, y)p(\theta\,|\,y)\mathrm{d}\theta \\ &= \int \frac{p(\tilde{y}, y\,|\,\theta)}{p(y\,|\,\theta)}p(\theta\,|\,y)\mathrm{d}\theta \\ &= \int \frac{p(\tilde{y}\,|\,\theta)p(y\,|\,\theta)}{p(y\,|\,\theta)}p(\theta\,|\,y)\mathrm{d}\theta \\ &= \int p(\tilde{y}\,|\,\theta)p(\theta\,|\,y)\mathrm{d}\theta. \end{aligned}$$

Note that we used the data independence property in the above derivation to write $p(\tilde{y}, y\,|\,\theta)$ as $p(\tilde{y}\,|\,\theta)p(y\,|\,\theta)$.

Chapter 4

Generalized Linear Models

This is a re-derivation of the result that the normal distribution with variance σ^2 has the largest entropy amongst all distributions defined on $[-\infty, +\infty]$ with variance σ^2. Let $p(x)$ be the pdf of the normal distribution with mean μ and variance σ^2. Let $q(x)$ be a pdf with the same variance. Since the entropy of a distribution does not depend on its mean, we may assume that the mean of $q(x)$ is μ.

The entropy $H(p)$ of the normal distribution is:

$$\begin{aligned}
H(p) &= -\int_{-\infty}^{+\infty} p(x)\log p(x)\mathrm{d}x \\
&= -\int_{-\infty}^{+\infty} p(x)\log((2\pi\sigma^2)^{-1/2})\mathrm{d}x + \int_{-\infty}^{+\infty} p(x)\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\mathrm{d}x \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\left(\frac{x-\mu}{\sigma}\right)^2\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)^2\right\}\mathrm{d}x \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma}\int_{-\infty}^{+\infty}\frac{\sigma}{\sqrt{2\pi}}z^2\exp\left\{-\frac{z^2}{2}\right\}\mathrm{d}z \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\mathrm{E}[Z^2] \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}(\mathrm{Var}(Z) + (\mathrm{E}[Z])^2) \\
&= \frac{1}{2}\log(2\pi e\sigma^2).
\end{aligned}$$

Now that we have the entropy of the normal, we use the KL-divergence metric to measure the distance of $q(x)$ from $p(x)$.

$$D_{\mathrm{KL}}(q,p) = \int_{-\infty}^{+\infty} q(x)\log\frac{q(x)}{p(x)}\mathrm{d}x = -H(q) + H(q,p).$$

At this point, we use the fact that $D_{\mathrm{KL}}(q,p) \geq 0$ for all distributions q and p. This gives us: $H(q,p) \geq H(q)$. We do not know what $H(q)$ is but the expression for $H(q,p)$ can be evaluated

quite easily.

$$
\begin{aligned}
H(q,p) &= -\int_{-\infty}^{+\infty} q(x)\log p(x)\mathrm{d}x \\
&= -\int_{-\infty}^{+\infty} q(x)\left[\log\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]\mathrm{d}x \\
&= -\log\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty} q(x)\mathrm{d}x + \frac{1}{2\sigma^2}\int_{-\infty}^{+\infty} q(x)(x-\mu)^2\mathrm{d}x \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\,\mathrm{E}_q[(x-\mu)^2] \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\sigma^2 \\
&= \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} \\
&= \frac{1}{2}\log(2\pi e\sigma^2) \\
&= H(p).
\end{aligned}
$$

So $H(q,p) = H(p)$ and combining this with $H(q,p) \geq H(q)$, we obtain $H(p) \geq H(q)$.

4.1 Binomial Distributions and Maximum Entropy

Just as the normal distribution is the one with maximum entropy amongst all distributions that have a constant variance, the binomial is a distribution with the maximum entropy amongst all distributions defined on an experiment with just two outcomes and with a constant mean.

To make this precise, consider binary sequences $S = \{s_i\}_{i=0}^{2^n-1}$ of length n sampled from a binomial distribution Binom(n,p). Let x_i denote the number of 1s in s_i; let $p_i = \Pr\{s_i\} = p^{x_i}(1-p)^{n-x_i}$ and define $\mathrm{p} = [p_0,\ldots,p_{2^n-1}]$. Let $\mathrm{q} = [q_0,\ldots,q_{2^n-1}]$ be some distribution over S. We wish to show that $H(p) \geq H(q)$. As in the case of the normal distribution, we start with the KL-divergence metric. We have

$$
D_{\mathrm{KL}}(\mathrm{q},\mathrm{p}) = \sum_{i=0}^{2^n-1} q_i \log\frac{q_i}{p_i} = -H(\mathrm{q}) + H(\mathrm{q},\mathrm{p}) \geq 0,
$$

which yields that $H(\mathrm{q},\mathrm{p}) \geq H(\mathrm{q})$.

Now consider the expression $H(\mathrm{q},\mathrm{p})$. We may write:

$$\begin{aligned}
H(\mathrm{q},\mathrm{p}) &= -\sum_{i=0}^{2^n-1} q_i \log p_i \\
&= -\sum_{i=0}^{2^n-1} q_i \log\left(p^{x_i}(1-p)^{n-x_i}\right) \\
&= -\sum_{i=0}^{2^n-1} q_i \left[x_i \log\frac{p}{1-p} + n\log(1-p)\right] \\
&= -\left(\log\frac{p}{1-p}\right)\sum_{i=0}^{2^n-1} q_i x_i - n\log(1-p)\sum_{i=0}^{2^n-1} q_i \\
&= -\left(\log\frac{p}{1-p}\right)\bar{q} - n\log(1-p).
\end{aligned}$$

Here $\bar{q}$ is the expected value of the distribution q. If we assume that $\bar{q} = \sum_{i=0}^{2^n-1} p_i x_i$, then we can retrace the steps in the last derivation and show that $H(\mathrm{q},\mathrm{p}) = H(\mathrm{p})$. Combining this with the inequality $H(\mathrm{q},\mathrm{p}) \geq H(\mathrm{q})$, we obtain that $H(\mathrm{p}) \geq H(\mathrm{q})$. That is, any distribution q on binary sequences of length n with the same expected value $\sum_{i=0}^{2^n-1} p_i x_i$ as p, has entropy at most that of p.

As a matter of fact,

$$\begin{aligned}
\sum_{i=0}^{2^n-1} p_i x_i &= \sum_{k=0}^{n} \binom{n}{k} k p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} k p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \\
&= np \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k} \\
&= np \sum_{j=0}^{n-1} \frac{(n-1)!}{j!(n-1-j)!} p^j (1-p)^{n-1-j} \\
&= np\,(p+(1-p))^n \\
&= np,
\end{aligned}$$

which matches with intuitive result that if the probability of a 1 is p, then the expected number of 1s in n trials is np.

4.2 The Poisson and Related Families of Distributions

The Poisson distribution is closely connected to distributions such as the binomial, exponential and the discrete Gamma (known as the Erlang distribution). To understand this connection, we begin with a description of Poisson process. A Poisson process is an arrival process that is described by a function $P(k,\tau)$ that gives the probability of k arrivals in a time interval of length τ and which satisfies the following conditions:

1. $P(k,\tau)$ is the same for all intervals of length τ;
2. the number of arrivals during a particular time interval is independent of the history of arrivals outside this time interval;
3. there exists a constant λ such that:
   - $P(0,\tau) = 1-\lambda\tau+o(\tau)$
   - $P(1,\tau) = \lambda\tau+o_1(\tau)$
   - $P(k,\tau) = o_k(\tau)$ for $k \in \{2,3,\ldots\}$

The last condition states that in the regime of small time intervals, the probability of a single arrival in a time interval of length τ is proportional to the length of the time interval. The functions $o(\tau), o_k(\tau)$ satisfy the conditions:

$$\lim_{\tau\to 0}\frac{o(\tau)}{\tau} = 0, \quad \lim_{\tau\to 0}\frac{o_k(\tau)}{\tau} \text{ for } k \in \{1,2,\ldots\}.$$

The functions $o(\tau)$ and $o_k(\tau)$ can be thought of as the higher-order terms in a Taylor series expansion of function of τ.

4.2.1 Number of arrivals

In order to compute a closed-form expression for $P(k,\tau)$, imagine dividing the time interval τ in n equal disjoint pieces intervals, each of length δ. If n is large enough, then the probability of a single arrival in a time interval of length δ is approximately $\lambda\delta$ and that of two or more arrivals is 0. Since the arrivals in each time window of length δ is independent of the arrivals in the other intervals, we may approximate this using a Bernoulli distribution with success probability $p = \lambda\delta$ and $n = \tau/\delta$ trials. The point to note is that $np = \lambda\tau$ is a constant. Thus the probability of k successes is:

$$\begin{aligned}
P(k,\tau) &= \binom{n}{k}p^k(1-p)^{n-k} \\
&= \frac{n!}{(n-k)!k!}\left(\frac{\lambda\tau}{n}\right)^k\left(1-\frac{\lambda\tau}{n}\right)^{n-k} \\
&= \frac{n(n-1)\cdots(n-k+1)}{n^k}\cdot\frac{1}{k!}\cdot(\lambda\tau)^k\cdot\left(1-\frac{\lambda\tau}{n}\right)^{-k}\cdot\left(1-\frac{\lambda\tau}{n}\right)^{n} \\
&= \frac{n}{n}\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)\cdot\frac{1}{k!}\cdot(\lambda\tau)^k\cdot\left(1-\frac{\lambda\tau}{n}\right)^{-k}\cdot\left(1-\frac{\lambda\tau}{n}\right)^{n}.
\end{aligned}$$

Fix k and let $n \to \infty$. In the limit we obtain:

$$P(k,\tau) = \frac{(\lambda\tau)^k}{k!} e^{-\lambda\tau}.$$

Note that $P(0,\tau) = e^{-\lambda\tau}$ and $P(1,\tau) = \lambda\tau e^{-\lambda\tau}$. Recall that the Taylor series for an infinitely differentiable function at $x = 0$ can be written as:

$$f(0+\delta) = f(0) + f^{(1)}(0)\delta + \frac{f^{(2)}(0)}{2!}\delta^2 + \cdots + \frac{f^{(n)}(0)}{n!}\delta^n + \cdots$$

The Taylor series expansion for $e^{-\lambda\tau}$ is:

$$e^{-(0+\lambda\tau)} = 1 - \lambda\tau + \frac{(\lambda\tau)^2}{2!} - \frac{(\lambda\tau)^3}{3!} + \cdots + (-1)^n \frac{(\lambda\tau)^n}{n!} + \cdots$$

Using this, we obtain that $P(0,\tau) = 1 - \lambda\tau + o(\tau)$ and $P(1,\tau) = \lambda\tau + o_1(\tau)$, consistent with the small-interval probability specifications.

4.2.2 Mean and Variance

The mean of a Poisson distribution can be easily calculated to be $\lambda\tau$. The variance of a Poisson distribution is also $\lambda\tau$. This can also be seen from the limiting process under which we derived the Poisson. The mean of a Binomial is np and the variance is $np - np^2$. If we consider a situation in which $n \to \infty$, $p \to 0$ and np is a constant, then the variance also tends to $np = \lambda\tau$.

4.2.3 Time of First Arrival and Interarrival Times

Let Y_1 denote the time of the first arrival. Then

$$\begin{aligned}\Pr\{Y_1 > t\} &= \Pr\{0 \text{ arrivals in } [0,t]\} \\ &= e^{-\lambda t}.\end{aligned}$$

Therefore $1 - F_{Y_1}(t) = e^{-\lambda t}$ and this yields that $f_{Y_1}(t) = \lambda e^{-\lambda t}$. Thus the time of the first arrival is exponentially distributed. Now after the first arrival, one can think of the Poisson process "restarting" so that the time till the next arrival is also exponentially distributed. This is so because of the memorylessness of the Poisson process. Let Y_k denote the time of the kth arrival and let T_k be the time interval between the $(k-1)$st and the kth arrival (also called the kth inter-arrival time). Then

$$\begin{aligned}&Y_k = T_1 + \cdots + T_k \text{ for } k \in \{1,2,\ldots\} \\ &T_k = Y_k - Y_{k-1} \text{ for } k \in \{2,3,\ldots\} \\ &T_k \overset{\text{iid}}{\sim} \text{Exp}(\lambda) \text{ for } k \in \{1,2,\ldots\}.\end{aligned}$$

What this means is that if the arrival process is assumed to be Poisson, then the inter-arrival times are exponentially distributed. Intrestingly, the converse is also true.

Consider an arrival process with independent inter-arrival times $T_1, \ldots, T_k$ that are all exponentially distributed with parameter λ. Fix a time interval $[0, \tau]$. Then the probability of 0 arrivals in $[0, \tau]$ is the same as $\Pr\{T_1 > \tau\}$. This is given by:

$$\begin{aligned}\Pr\{T_1 > \tau\} &= \int_{\tau}^{\infty} \lambda e^{-\lambda t} \mathrm{d}t \\ &= e^{-\lambda\tau}.\end{aligned}$$

But this exactly the expression for 0 arrivals in a time interval τ in a Poisson process with rate λ. Next, consider the case of exactly one arrival in the interval $[0, \tau]$. The event of exactly one arrival is the intersection of the events that $T_1 = t$, where $0 \leq t \leq \tau$, and $T_2 > \tau - t$.

$$\begin{aligned}\Pr\{\text{1 arrival in time } [0,\tau]\} &= \int_{t=0}^{\tau} \Pr\{T_1 = t \text{ and } T_2 > \tau - t\} \, \mathrm{d}t \\ &= \int_{t=0}^{\tau} \lambda e^{-\lambda t} \left(\int_{\tau - t}^{\infty} \lambda e^{-\lambda x} \mathrm{d}x \right) \mathrm{d}t \\ &= \int_{t=0}^{\tau} \lambda e^{-\lambda t} e^{-\lambda(\tau - t)} \mathrm{d}t \\ &= \int_{t=0}^{\tau} \lambda e^{-\lambda\tau} \\ &= \lambda\tau e^{-\lambda\tau}.\end{aligned}$$

Again, this matches the expression for 1 arrival in a Poisson process with rate λ in an interval of length τ. We will not show the full derivation here.

4.2.4 Time of kth Arrival

The time of the kth arrival Y_k follows a discrete Gamma or Erlang distribution. Let f_{Y_k} be the pdf of the distribution. Imagine that the kth arrival happens in a time interval $[y, y + \delta]$. Now the kth arrival takes place in $[y, y + \delta]$ iff

- A: $k - 1$ arrivals happen in the time interval $[0, y]$, and
- B: exactly one arrival happens in $[y, y + \delta]$.

$$\begin{aligned}\Pr\{y \leq Y_k \leq y + \delta\} &= \Pr\{A\} \cdot \Pr\{B\} \\ &= \frac{(y\lambda)^{k-1}}{(k-1)!} e^{-\lambda y} \cdot y\delta \\ &= \delta \cdot \frac{y^k \lambda^{k-1}}{(k-1)!} e^{-\lambda y}.\end{aligned}$$

Thus

$$\Pr\{Y_k = y\} = \frac{y^k \lambda^{k-1}}{(k-1)!} e^{-\lambda y}.$$

This is the discrete version of the Gamma distribution which is usually stated in terms of the parameters α and β, instead of k and λ.

$$\text{Gamma}(y \mid \alpha, \beta) = \frac{\beta^{\alpha-1}}{\Gamma(\alpha)} y^{\alpha} e^{-\beta y} \quad y \geq 0, \alpha, \beta > 0.$$

When stated in this continuous form, it is helpful to remember that α stands for the "number of events" and that β stands for the "rate" at which those events occur.