CDS: Machine Learning 2023 // Tutorial Week 2

Jochem Pannekoek, Daria Mihaila & Jasper Pieterse

September 20, 2023

MacKay 3.12

We denote W as the hypothesis that the original counter in the bag was white and B to denote the original counter was black. After observing data D (drawing a white counter), we want to know the posterior probability $P(W \mid D)$. We can use Bayes' theorem to find this probability. Before putting the counter in the bag, we have a prior $P(W) = P(B) = 1/2$. The likelihood of the data for each hypothesis is given by:

$$P(D \mid W) = 1 \text{ and } P(D \mid B) = \frac{1}{2}$$

The evidence is given by:

$$P(D) = P(D \mid W)P(W) + P(D \mid B)P(B) = 1 \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$$

The posteriors are then given by Bayes' theorem:

$$P(W \mid D) = \frac{1 \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{2}{3} \text{ and } P(B \mid D) = \frac{\frac{1}{2} \cdot \frac{1}{2}}{\frac{3}{4}} = \frac{1}{3}$$

So the probability of drawing a white counter is $2/3$. Even though the bag is in the exactly identical state as before, the probability of drawing a white counter is higher than before. This is because we have gained information about the counter in the bag by drawing a white counter.

MacKay 28.1

To compute the evidence $P(\mathcal{H}_i \mid x)$ we use the evidence framework:

$$P(\mathcal{H}_i \mid x) = \frac{P(x \mid \mathcal{H}_i)P(\mathcal{H}_i)}{P(x)}$$

The priors for each hypothesis are equal $P(\mathcal{H}_i) = 1/2$. The likelihoods are given by:

$$P(x \mid \mathcal{H}_0) = \prod_{i=1}^{5} \frac{1}{2} = \frac{1}{32}$$

$$P(x \mid \mathcal{H}_1) = \prod_{i=1}^{5} \frac{1}{2}(1 + mx_i) = \frac{1}{32}(1 + 0.3m)(1 + 0.5m)(1 + 0.7m)(1 + 0.8m)(1 + 0.9m)$$

The evidence is given by:

$$P(x) = P(x \mid \mathcal{H}_0)P(\mathcal{H}_0) + P(x \mid \mathcal{H}_1)P(\mathcal{H}_1) = \frac{1}{64}\left[1 + \prod_{i=1}^{5}(1 + mx_i)\right]$$

Where we used shorthand notation for sake of clarity. The posteriors are then given by Bayes' theorem:

$$P(\mathcal{H}_0 \mid x) = \frac{\frac{1}{32} \cdot \frac{1}{2}}{\frac{1}{64}\left[1 + \prod_{i=1}^{5}(1 + mx_i)\right]} = \frac{1}{1 + \prod_{i=1}^{5}(1 + mx_i)}$$

and

$$P(\mathcal{H}_1 \mid x) = \frac{\frac{1}{32}\prod_{i=1}^{5}(1 + mx_i) \cdot \frac{1}{2}}{\frac{1}{64}\left[1 + \prod_{i=1}^{5}(1 + mx_i)\right]} = \frac{\prod_{i=1}^{5}(1 + mx_i)}{1 + \prod_{i=1}^{5}(1 + mx_i)}$$

Extra Exercise 3.1

For hypothesis $\mathcal{H}_0$ the probability of heads is assumed to be $\frac{1}{2}$ and for hypothesis $\mathcal{H}_1$ the probability is assumed to be λ. Using this we can establish the following likelihoods for N_W balls:

$$P(N_W \mid N, \mathcal{H}_0) = f^{N_W}(1-f)^{N-N_W} = \binom{N}{N_W}\left(\frac{1}{2}\right)^N$$

and

$$P(N_W \mid \lambda, N, \mathcal{H}_1) = \lambda^{N_W}(1-\lambda)^{N-N_W}$$

We assume equal prior probabilities on the hypotheses $P(\mathcal{H}_0) = P(\mathcal{H}_1) = \frac{1}{2}$. The prior $P(\lambda \mid \mathcal{H}_1) = 1$ because we assume a uniform distribution for λ. The probabilities of the data given the hypothesis are given by:

$$P(N_W \mid N, \mathcal{H}_0) = \frac{1}{2}^{N_H}\frac{1}{2}^{N-N_H} = \frac{1}{2}^{N}$$

and

$$P(N_W \mid \lambda, N, \mathcal{H}_1) = \frac{N_H!(N-N_H)!}{(N+1)!}$$

As described by MacKay (3.12). The evidence is then given by:

$$P(N_W \mid N) = P(N_W \mid N, \mathcal{H}_0)P(\mathcal{H}_0) + P(N_W \mid N, \mathcal{H}_1)P(\mathcal{H}_1) = \frac{1}{2}^{N+1} + \frac{N_H!(N-N_H)!}{2(N+1)!}$$

The posterior probabilities are then given by Bayes' theorem:

$$P(\mathcal{H}_0 \mid N_W, N) = \frac{P(N_W \mid N, \mathcal{H}_0)P(\mathcal{H}_0)}{P(N_W \mid N)} = \frac{\frac{1}{2}^{N+1}}{\frac{1}{2}^{N+1} + \frac{N_H!(N-N_H)!}{2(N+1)!}} = \frac{1}{1 + \frac{2^N N_H!(N-N_H)!}{(N+1)!}}$$

and

$$P(\mathcal{H}_1 \mid N_W, N) = \frac{P(N_W \mid N, \mathcal{H}_1)P(\mathcal{H}_0)}{P(N_W \mid N)} = \frac{\frac{1}{2}\frac{N_H!(N-N_H)!}{(N+1)!}}{\frac{1}{2}^{N+1} + \frac{N_H!(N-N_H)!}{2(N+1)!}} = \frac{1}{1 + \frac{(N+1)!}{2^N N_H!(N-N_H)!}}$$

For $N_W = 0$ and $N = 2$ the posteriors are given by:

$$P(\mathcal{H}_0 \mid N_W = 0, N = 2) = \frac{1}{1 + \frac{4 \cdot 0!(2-0)!}{3!}} = \frac{3}{7}$$

and

$$P(\mathcal{H}_1 \mid N_W = 0, N = 2) = \frac{1}{1 + \frac{3!}{4 \cdot 0!(2-0)!}} = \frac{4}{7}$$

For $N_W = 1$ and $N = 2$ the posteriors are given by:

$$P(\mathcal{H}_0 \mid N_W = 1, N = 2) = \frac{1}{1 + \frac{4 \cdot 1!(2-1)!}{3!}} = \frac{3}{5}$$

and

$$P(\mathcal{H}_1 \mid N_W = 1, N = 2) = \frac{1}{1 + \frac{3!}{4 \cdot 1!(2-1)!}} = \frac{2}{5}$$

For $N_W = 2$ and $N = 2$ the posteriors are given by:

$$P(\mathcal{H}_0 \mid N_W = 2, N = 2) = \frac{1}{1 + \frac{4 \cdot 2!(2-2)!}{3!}} = \frac{3}{7}$$

and

$$P(\mathcal{H}_1 \mid N_W = 2, N = 2) = \frac{1}{1 + \frac{3!}{4 \cdot 2!(2-2)!}} = \frac{4}{7}$$

b.)

You will find that for $N_H = 0, 2$ model H_1 is more likely and for $N_H = 1$ model H_0 is more likely. Explain these results.

$\mathcal{H}_0$ assumes that there is an exactly half of the observations are head, since $f = 0.5$. Therefore, when $N_W = 1, N = 2$ this assumption is met and the $\mathcal{H}_0$ is more likely. When $N_H = 0, 2$, $\mathcal{H}_1$ is more likely because it takes a uniform distribution for the probability of heads: λ. For example, there could be a chance that there is a 100% or 0% probability for heads. This could be explained with $\mathcal{H}_1$.

Extra Exercise 27.1

a.)

The Laplace approximation is given by:

$$p_1(\lambda \mid r) = \frac{p(r \mid \lambda)p(\lambda)}{p(r)} \approx \frac{p(r \mid \lambda)p(\lambda)}{p(r \mid \lambda_0)p(\lambda_0)}$$

Where λ_0 is the maximum of the posterior distribution. The posterior is proportional to:

$$p(\lambda \mid r) \propto p(r \mid \lambda)p(\lambda) = \frac{e^{-\lambda}\lambda^{r-1}}{r!}$$

The negative log of the posterior then becomes:

$$\begin{aligned} Q &= -\ln p(\lambda \mid r) \\ &= -\left[\ln\left(e^{-\lambda}\right) + \ln\left(\lambda^{r-1}\right) - \ln(r!)\right] + \ln p(r) \\ &\propto \lambda + (1-r)\log(\lambda) \end{aligned}$$

Where we dropped the constant term $\ln(r!)$ and the term $-\ln p(r)$ because they are constant with respect to λ. We compute the maximum of Q as follows:

$$\begin{aligned} \frac{\partial Q}{\partial \lambda} &= 1 + \frac{1-r}{\lambda} \\ 0 &= 1 + \frac{1-r}{\lambda} \\ \lambda_0 &= -(1-r) = r - 1 \end{aligned}$$

We can now compute A as the second derivative of Q:

$$A = \left.\frac{\partial^2 Q}{\partial \lambda^2}\right|_{\lambda=\lambda_0} = \left.\frac{r-1}{\lambda^2}\right|_{\lambda=\lambda_0}$$

Substituting λ_0 into A gives:

$$A = \frac{r-1}{(r-1)^2} = \frac{1}{r-1}$$

We then can compute $p_1(\lambda \mid r)$ as follows:

$$\begin{aligned} p_1(\lambda \mid r) &= \sqrt{\frac{A}{2\pi}} \exp\left\{-\frac{A}{2}(\lambda - \lambda_0)^2\right\} \\ &= \frac{1}{\sqrt{2\pi(r-1)}} \exp\left\{-\frac{1}{2(r-1)}(\lambda - (r-1))^2\right\} \end{aligned}$$

We see that this is a Gaussian distribution with $\mu_\lambda = r - 1$ and $\sigma^2 = r - 1$.

b.)

If we apply the coordinate transformation $y = \log \lambda$, we obtain the following expressions:

$$dy = \frac{\delta y}{\delta \lambda} d\lambda = \frac{d\lambda}{\lambda} \tag{1}$$

$$p(y) = \frac{p(\lambda) d\lambda}{dy} = \frac{\frac{1}{\lambda} d\lambda}{\frac{d\lambda}{\lambda}} = 1 \tag{2}$$

c.)

The posterior distribution can be transformed as follows:

$$\begin{aligned} p(y \mid r) &= p(\lambda \mid r) |\frac{d\lambda}{dy}| \\ &= p(\lambda|r)\lambda \end{aligned}$$

We can write the negative log-posterior as:

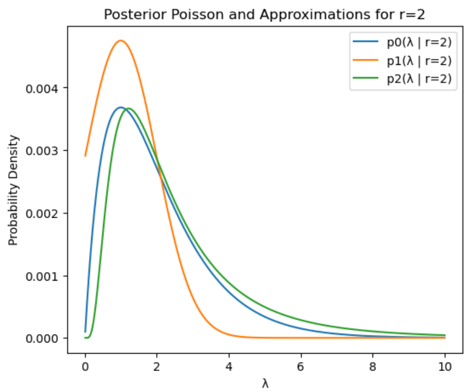
$$\begin{aligned} Q &\propto -\ln p(\lambda|r)\lambda \\ &\propto -\ln\left[\exp(-\lambda)\lambda^r\right] \\ &\propto -\ln\left[\exp(-e^y) + e^{yr}\right] \\ &\propto e^y - yr \end{aligned}$$

Where we used the transformation $\lambda = e^y$ when going from 2nd to 3rd line and we dropped constant terms. We compute the maximum of Q as follows:

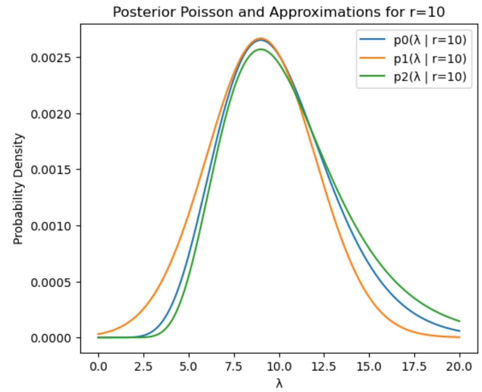
$$\begin{aligned} \frac{\partial Q}{\partial y} &= e^y - r \\ 0 &= e^y - r) \\ y_0 &= \ln(r) \end{aligned}$$

We can now compute A as the second derivative of Q:

$$A = \left.\frac{\partial^2 Q}{\partial y^2}\right|_{y=y_0} = e^y|_{y=y_0} = r$$



(a) $r = 2$



(b) $r = 10$

Figure 1: The posterior distributions for λ for $r = 2$ and $r = 10$. The blue line is the exact posterior, the orange line is the Laplace approximation and the green line is the approximation using the coordinate transformation.

We then can compute $p_2(y \mid r)$ as follows:

$$\begin{aligned} p_2(y \mid r) &= \sqrt{\frac{A}{2\pi}} \exp\left\{-\frac{A}{2}\left(y - y_0\right)^2\right\} \\ &= \sqrt{\frac{r}{2\pi}} \exp\left\{-\frac{r}{2}(y - \ln(r))^2\right\} \end{aligned}$$

We see that this is a Gaussian distribution with $\mu_y = \ln(r)$ and $\sigma^2 = \frac{1}{r}$.

d.)

Transforming $p_2(y \mid r)$ back to $p_2(\lambda \mid r)$ is done as follows:

$$\begin{aligned} p_2(\lambda \mid r) &= p_2(y \mid r)|\frac{dy}{d\lambda}| \\ &= \sqrt{\frac{r}{2\pi}} \exp\left\{-\frac{r}{2}(y - \ln(r))^2\right\} \frac{1}{\lambda} \\ &= \sqrt{\frac{r}{2\pi}} \exp\left\{-\frac{(r)}{2}(\ln\lambda - \ln(r))^2\right\} \frac{1}{\lambda} \end{aligned}$$

We will plot the distributions:

$$\begin{aligned} p(\lambda \mid r) &= \frac{e^{-\lambda}\frac{\lambda^{r-1}}{r!}}{p(r)} \\ p_1(\lambda \mid r) &= \frac{1}{\sqrt{2\pi(r-1)}} \exp\left\{-\frac{1}{2(r-1)}(\lambda - (r-1))^2\right\} \\ p_2(\lambda \mid r) &= \sqrt{\frac{r}{2\pi}} \exp\left\{-\frac{(r)}{2}(\ln\lambda - \ln(r))^2\right\} \frac{1}{\lambda} \end{aligned}$$

The results are shown in Figure 1b. We see that the for smaller values like $r = 2$, $p_2(\lambda \mid r)$ is the better approximation. This is because the logarithmic transformation could better capture the exponential nature of the Poisson distribution. For larger r values, as $r = 10$, $p_1(\lambda \mid r)$ is the better approximation because the Poisson distribution becomes more Gaussian-like.

Extra Exercise 28.1

a.)

The likelihood of the data given the parameters w_0 and w_1 can be written as:

$$p(t_i \mid x_i, w_0, w_1) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(t_i - w_0 - w_1 x_i)^2\right)$$

The likelihood of the entire data set D is given by the product of the likelihoods of each data point:

$$\begin{aligned} p(D \mid w_0, w_1) &= \prod_{i=1}^{N} p(t_i \mid x_i, w_0, w_1) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(t_i - w_0 - w_1 x_i)^2\right) \end{aligned}$$

We can absorb the product in the exponent as a summation and split the term into a function of w_0 and a function of w_1:

$$\begin{aligned} \sum_{i=1}^{N}(t_i - w_0 - w_1 x_i)^2 &= \sum_{i=1}^{N}(t_i^2 + w_0^2 + w_1^2 x_i^2 - 2t_i w_0 - 2w_1 t_i x_i + 2w_0 w_1 x_i) \\ &= \sum_{i=1}^{N}(t_i^2 - 2t_i w_0 + w_0^2 + w_1^2 x_i^2 - 2w_1 t_i x_i) \\ &= \sum_{i=1}^{N}(t_i - w_0)^2 + w_1 \sum_{i=1}^{N}(w_1 x_i^2 - 2t_i x_i) \end{aligned}$$

Where in going from the 1st to 2nd line we used that $\sum_{i=1}^{N} x_i = 0$. Finally, we can write the likelihood as:

$$p(D \mid w_0, w_1) = \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(t_i - w_0)^2\right)\right) \times \left(\exp\left(-\frac{w_1}{2\sigma^2}\sum_{i=1}^{N}(w_1 x_i^2 - 2t_i x_i)\right)\right)$$

b.)

Given that $\mathcal{H}_1$ states that $w_1 = 0$, its evidence is given by:

$$p(D \mid \mathcal{H}_1) = \int p(D \mid w_0) p(w_0) dw_0$$

Where $p(w_0)$ is the prior on w_0. For $\mathcal{H}_2$, the evidence is:

$$p(D \mid \mathcal{H}_2) = \int \int p(D \mid w_0, w_1) p(w_0) p(w_1) dw_0 dw_1$$

With $p(w_0)$ and $p(w_1)$ the priors. We can express $p(D \mid w_0, w_1)$ using the previously derived result, which was a product of a function of w_0 and a function of w_1.

$$p(D \mid w_0, w_1) = f(w_0) g(w_1)$$

Using this, the double integral $p(D \mid \mathcal{H}_2)$ factorizes and our expression simplifies to:

$$p(D \mid \mathcal{H}_2) = \int f(w_0)p(w_0)dw_0 \int g(w_1)p(w_1)dw_1$$

When computing the odds ratio, the terms with w_0 cancel out because both models use the same prior for w_0. This leaves us with:

$$\frac{p(D \mid \mathcal{H}_2)}{p(D \mid \mathcal{H}_1)} = \frac{\int g(w_1)p(w_1)dw_1}{1}$$

Now, using the result from the previous part:

$$g(w_1) = \exp\left(-\frac{w_1}{2\sigma^2}\sum_{i=1}^{N}(w_1 x_i^2 - 2t_i x_i)\right) = \exp\left(-\frac{w_1}{2\sigma^2}(w_1 N\langle x^2\rangle - 2N\langle xt\rangle)\right)$$

Where we substituted $\frac{1}{N}\sum_i = 1^N x_i^2 = \langle x^2\rangle$ and $\frac{1}{N}\sum_i = 1^N t_i x_i = \langle xt\rangle$. The prior is given by:

$$p(w_1) = \mathcal{N}(0,1) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{w_1^2}{2}\right)$$

We can now plug these into the integral for $g(w_1)p(w_1)$:

$$\frac{p(D \mid \mathcal{H}_2)}{p(D \mid \mathcal{H}_1)} = \frac{1}{\sqrt{2\pi}}\int \exp\left(-\frac{w_1^2}{2}(1 - \frac{N\langle x^2\rangle}{\sigma^2}) - \frac{Nw_1\langle xt\rangle}{\sigma^2}\right) dw_1$$

We see that the odds ratio is an integral that depends on the input variance $\langle x^2\rangle$, the input-output correlation $\langle xt\rangle$ and N.

c.)

From the equation for the odds ratio we derived:

$$\frac{p(D \mid \mathcal{H}_2)}{p(D \mid \mathcal{H}_1)} = \frac{1}{\sqrt{2\pi}}\int \exp\left(-\frac{w_1^2}{2}(1 - \frac{N\langle x^2\rangle}{\sigma^2}) - \frac{Nw_1\langle xt\rangle}{\sigma^2}\right) dw_1$$

We are given that in the limit of large N, $\sigma^2 = \langle x^2\rangle = 1$. Plugging these values in, we get:

$$\frac{p(D \mid \mathcal{H}_2)}{p(D \mid \mathcal{H}_1)} = \frac{1}{\sqrt{2\pi}}\int \exp\left(-\frac{w_1^2}{2}(1 - N) - Nw_1\langle xt\rangle\right) dw_1$$

In the large N limit, the term $Nw_1^2/2$ in the exponent will dominate, and the integrand will be peaked sharply around $w_1 = 0$. The peak will be high (and hence the evidence for $\mathcal{H}_2$ will be high) if the term linear in w_1 is significant, which means if the correlation $\langle xt\rangle$ is significant. For model $\mathcal{H}_1$ to be preferred, the evidence for $\mathcal{H}_2$ (which includes the w_1 term) should be smaller. This means that the w_1 term in the exponent should not make a significant contribution:

$$\langle xt\rangle^2 \lesssim \frac{\log N}{N}$$

On the other hand, if $\langle xt\rangle^2 > \frac{\log N}{N}$, then the complex model $\mathcal{H}_2$ would be preferred.

Extra Exercise 28.2

(a)

Knowing that the die is fair for H_0 gives us the $p(x_i|H_0) = \frac{1}{k}$ with k number of sides of the die.
Therefore, also taking into account the combinatorial factor as the outcomes are independent,the probability of the data is given by:

$$
\begin{aligned}
p(D|H_0) &= \frac{n!}{n_1!n_2!...n_k!}\prod_{i=1}^{k}(\frac{1}{k})^{n_i} \\
&= \frac{n!}{n_1!n_2!...n_k!}\left(\frac{1}{k}\right)^{n_1+n_2+...+n_k} \\
&= \frac{n!}{n_1!n_2!...n_k!}\left(\frac{1}{k}\right)^{n}
\end{aligned}
$$

(b)

For H_1, the probability of the data given the probabilities $\vec{p}$ and the hypothesis is similar to $p(D|H_0)$

$$
p(\vec{n}|\vec{p}, H_1) = \frac{n!}{n_1!n_2!...n_k!}\prod_{i=1}^{k}(p_i)^{n_i}.
$$

We also assume that the priors for these probabilities are from a uniform distribution $p(\vec{p}|H_1) = 1$. We can combine these to obtain $p(\vec{n}|H_1)$ by using the normalization function $B(\alpha)$.

$$
\begin{aligned}
p(\vec{n}|H_1) &= \int_0^1 d\vec{p}p(\vec{p}|H_1)p(\vec{n}|\vec{p}, H_1) \\
&= \int_0^1 dp_1...dp_k\frac{n!}{n_1!n_2!...n_k!}\prod_{i=1}^{k}(p_i)^{n_i} \\
&= \frac{n!}{n_1!n_2!...n_k!}\frac{\prod_{i=1}^{k}\Gamma(n_i+1)}{\Gamma(\sum_{i=1}^{k}n_i+1)} \\
&= \frac{n!}{n_1!n_2!...n_k!}B(\vec{n}+1)
\end{aligned}
$$

For H_1 we use the Dirichlet distribution and also the combinatorial factor. First we realize that the likelihood function $p(\mathbf{n}|\mathbf{p}, H_1) = \frac{n!}{n_1!n_2!...n_k!}\prod_{i=1}^{k}(\frac{1}{k})^{n_i}$

$$
p(D|H_1) = \frac{n!}{n_1!n_2!...n_k!}\frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma\left(\alpha_i\right)}\prod_{i=1}^{n}p_i^{\alpha_i-1}
$$

(c)

For the posterior probability of the models $\mathcal{H}_{0,1}$ assuming equal priors, we calculate

$$
\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(D|H_0)}{p(D|H_1)}.
$$

For the first and second dataset this yields respectively

$$\frac{p(H_0|D)}{p(H_1|D)} = 22.4086$$

$$\frac{p(H_0|D)}{p(H_1|D)} = 15653.4.$$