CDS: Machine Learning 2023 // Tutorial Week 1

Jochem Pannekoek, Daria Mihaila & Jasper Pieterse

September 13, 2023

MacKay 2.10

We want to find the conditional probability $P(u = A|N_B = 1)$, which is the probability that urn A is selected given that one black ball is drawn. We use Bayes' theorem to find this posterior:

$$P(u = A|N_B = 1) = \frac{P(N_B = 1|u = A) \cdot P(u = A)}{P(N_B = 1)}$$

The likelihood is the probability of drawing one black ball from urn A, which is $\frac{1}{3}$. The prior is $P(u = A) = \frac{1}{2}$ since one of the two urns is selected at random. The evidence is calculated as follows:

$$P(N_B = 1) = P(u = A) \cdot P(N_B = 1|u = A) + P(u = B) \cdot P(N_B = 1|u = B) = \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{2}$$

Using these expressions we can calculate:

$$P(u = A|N_B = 1) = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

1 Extra Exercise 2.1

We want to find the probability that urn u was selected $P(U = u|N_B \text{ is even})$ using Bayes' theorem:

$$P(U = u|N_B \text{ is even}) = \frac{P(N_B \text{ is even}|U = u) \cdot P(U = u)}{\sum_{u=0}^{10} P(N_B \text{ is even}|U = u) \cdot P(U = u)} \tag{1}$$

The prior $P(U = u) = \frac{1}{11}$ since there are 11 urns that are all equally likely to be chosen. To compute the likelihood, we first compute the probability of drawing n_b black balls from urn u:

$$P(N_B = n_b|U = u) = \binom{N}{n_b} f_u^{n_b} (1 - f_u)^{N - n_b}$$

where $f_u = \frac{u}{N}$ is the fraction of black balls in urn u. We can now compute the probability of drawing an even number of black balls from urn u:

$$P(N_B \text{ is even}|U = u) = \sum_{n_b = \text{even}} \binom{N}{n_b} f_u^{n_b} (1 - f_u)^{N - n_b}$$

We can also compute the evidence $P(N_B \text{ is even})$:

$$P(N_B \text{ is even}) = \sum_{u=0}^{10} P(U = u) \cdot P(N_B \text{ is even}|U = u) = \frac{1}{11} \sum_{u=0}^{10} \binom{N}{n_b} f_u^{n_b} (1 - f_u)^{N - n_b}$$

Using equation (1) we can now compute the posterior:

$$P(U = u|N_B \text{ is even}) = \sum_{n_b=\text{even}} \frac{\binom{N}{n_b} f_u^{n_b}(1-f_u)^{N-n_b}}{\sum_{u=0}^{10}\binom{N}{n_b} f_u^{n_b}(1-f_u)^{N-n_b}}$$

2 Extra Exercise 2.2

a.)

We again compute the posterior using Bayes' theorem:

$$p(\mu|x) = \frac{p(x|\mu)p(\mu)}{p(x)} \tag{2}$$

We assume the prior $p(\mu)$ to be a constant, since any μ is equally probable. The likelihood is given by the Gaussian distribution:

$$p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

The evidence is computed using a Gaussian integral $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$:

$$p(x) = \int_{-\infty}^{\infty} p(x|\mu)p(\mu)d\mu = \frac{p(\mu)}{\sqrt{2\pi\sigma}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} d\mu = p(\mu)$$

We can now compute the posterior by substituting the expressions in (2):

$$p(\mu|x) = \frac{p(\mu)}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \frac{1}{p(\mu)} = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

We can see that the posterior is again a Gaussian distribution with $\sigma = 1$ and $\mu = x$.

b.)

We calculate the posterior of N data points $x_1, x_2, ..., x_N$, given by Bayes' theorem:

$$p(\mu|x_1, x_2, ..., x_N) = \frac{p(x_1, x_2, ..., x_N|\mu)p(\mu)}{p(x_1, x_2, ..., x_N)} \tag{3}$$

We again assume that the prior $p(\mu)$ is a constant. The likelihood is given by the product of the likelihoods of the individual data points:

$$p(x_1, x_2, ..., x_N|\mu, N) = \prod_{i=1}^{N} p(x_i|\mu)p(\mu)d\mu = \frac{p(\mu)}{(\sqrt{2\pi\sigma})^N} \prod_{i=1}^{N} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} d\mu$$

We can absorb the product into the exponential as a summation:

$$p(x_1, x_2, ..., x_N|\mu, N) = \frac{p(\mu)}{(\sqrt{2\pi\sigma})^N} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2} d\mu$$

We can then rewrite the exponent using the mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$:

$$p(x_1, x_2, ..., x_N|\mu, N) = \frac{p(\mu)}{(\sqrt{2\pi\sigma})^N} e^{-\frac{N}{2\sigma^2}(\bar{x}-\mu)^2} d\mu$$

We can compute the evidence as follows:

$$\begin{aligned} p(x_1, x_2, ..., x_N) &= \int_{-\infty}^{\infty} p(x_1, x_2, ..., x_N|\mu, N)d\mu \\ &= \frac{p(\mu)}{(\sqrt{2\pi}\sigma)^N} \int_{-\infty}^{\infty} e^{-\frac{N}{2\sigma^2}(\bar{x}-\mu)^2} d\mu \\ &= \frac{p(\mu)}{(\sqrt{2\pi}\sigma)^N} \sqrt{\frac{2\pi\sigma^2}{N}} \\ &= \frac{(\sqrt{2\pi}\sigma)^{1-N}}{\sqrt{N}} p(\mu) \end{aligned}$$

Combining our results in (3) we can now compute the posterior:

$$p(\mu|x_1, x_2, ..., x_N) = \frac{\sqrt{N}}{(\sqrt{2\pi}\sigma)} e^{-\frac{N}{2\sigma^2}(\bar{x}-\mu)^2} d\mu$$

If we then fill in $\sigma = 1$ we get that the posterior is a Gaussian distribution with $\sigma = \frac{1}{\sqrt{N}}$ and $\mu = \bar{x}$.

3 Extra Exercise 2.3

a.)

We aim to write $p(x|\mu, \Sigma)$ in the canonical form of the exponential family distributions following

$$p_\theta = \exp(\langle \theta, \phi(x) \rangle - A(\theta)),$$

where $A(\theta)$ is the log-partition function. We can rewrite $p(x|\mu, \Sigma)$ as

$$\begin{aligned} \log p(x|\mu, \Sigma) &= -\frac{1}{2} \log(2\pi^n \det \Sigma) - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \\ &= [x^T \Sigma \mu - \frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2}(n \log 2\pi + \log \det \Sigma - \mu \Sigma^{-1} \mu)] \\ &= [\langle \Sigma^{-1}\mu, x \rangle - \frac{1}{2} x^T \Sigma^{-1} x - \frac{1}{2}(n \log 2\pi + \log \det \Sigma - \mu \Sigma^{-1} \mu)]. \end{aligned}$$

$x^T \Sigma^{-1} x$ can be written as

$$\begin{aligned} x^T \Sigma^{-1} x &= \begin{bmatrix} x_0 & x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \sum_{i=0}^n \Sigma_{0,i}^{-1} x_i \\ \sum_{i=0}^n \Sigma_{1,i}^{-1} x_i \\ \vdots \\ \sum_{i=0}^n \Sigma_{n,i}^{-1} x_i \end{bmatrix} = \sum_{j=0}^n \sum_{i=0}^n x_j \sum_{ji}^{-1} x_i \\ &= \begin{bmatrix} \Sigma_{00}^{-1} & \cdots & \sum_{0i}^{-1} & \cdots & \sum_{ji}^{-1} \end{bmatrix} \begin{bmatrix} x_0 x_0 \\ \vdots \\ x_0 x_i \\ \vdots \\ x_j x_i \end{bmatrix} = \langle \text{vec}(\Sigma^{-1}), \text{vec}(xx^T) \rangle. \end{aligned}$$

Thus,

$$p(x|\mu, \Sigma) = \exp(\langle \Sigma^{-1}\mu, x \rangle + \langle \text{vec}(-\frac{1}{2}\Sigma^{-1}), \text{vec}(xx^T) \rangle - \frac{1}{2}(n \log 2\pi + \log \det \Sigma - \mu \Sigma^{-1} \mu)).$$

From this we can see that the log partition sum, parameter vector, and the sufficient statistics are given by

$$A(\theta) = \frac{1}{2}(n \log 2\pi + \log \det \Sigma - \mu \Sigma^{-1} \mu)$$
$$\theta = \begin{bmatrix} \Sigma^{-1}\mu \\ \text{vec}(-\frac{1}{2}\Sigma^{-1}) \end{bmatrix}$$
$$\phi(x) = \begin{bmatrix} x \\ \text{vec}(xx^T) \end{bmatrix}$$

b.)

Since the Gaussian distribution is a maximum entropy solution and the distribution is an exponential family distribution, we can use theorem 6.7 and its corresponding constraints:

For $\theta \in \mathbb{R}^d$, let P_θ have density

$$p_\theta(x) = \exp\left(\langle h_\theta, \phi(x)\rangle - A(\theta)\right), \quad A(\theta) = \log Z \exp\left(\langle h_\theta, \phi(x)\rangle\right) d\mu(x),$$

with respect to the measure μ. If $\mathbb{E}_{P_\theta}[\phi(X)] = \alpha$, then P_θ maximizes $H(P)$ over P^α_{lin}; moreover, the distribution P_θ is unique.

Therefore, the constraints are: $\lambda \geq 0$ for $p(x) \geq 0$,$\theta_0 \in \mathbb{R}$ for the normalization constraint that $P(X) = 1$, and θ_i for the constraint $\mathbb{E}_P[\phi_i(X)] = \alpha_i$.

4 Extra Exercise 2.4

a.)

We can compute the integral as follows:

$$\begin{aligned}
\int dz_1 q_1(z_1) \log q_1(z_1) &= \int dz_1 \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2} z_1^2\right) \log\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2} z_1^2\right)\right) \\
&= \frac{1}{\sqrt{2\pi\sigma_1^2}}\Big[- \int dz_1 \exp(-\frac{1}{2\sigma_1^2} z_1^2) \log \sqrt{2\pi\sigma_1^2} - \int dz_1 \exp(-\frac{1}{2\sigma_1^2} z_1^2) \frac{1}{2\sigma_1^2} z_1^2 \Big] \\
&= -\log \sqrt{2\pi}\sigma_1 - \frac{2\sigma_1^2}{\sqrt{2\pi\sigma_1^2}}\Big(z \left[\frac{-\exp(-az_1^2)}{2a}\right]_{-\infty}^{\infty} - \int dz_1 - \frac{\exp(-az_1^2)}{2a}\Big)
\end{aligned}$$

where we denote:

$$a = \frac{1}{2\sigma_1^2}$$

The evaluation from $[\infty, -\infty]$ will equal zero and using Gauss' formula on the last integral, we get to the desired result:

$$\int dz_1 q_1(z_1) \log q_1(z_1) = -\log \sqrt{2\pi\sigma_1^2} - \frac{1}{2}.$$

And analogously this can be showed for $q_2(z_2)$

b.)

We can compute the KL divergence as follows:

$$\begin{aligned}
KL(q \mid p) &= \int dz_1 dz_2 q\left(z_1, z_2\right) \log \frac{q\left(z_1, z_2\right)}{p\left(z_1, z_2\right)} \\
&= \int dz_1 dz_2 q(z_1) q(z_2) \big[\log q_1 \log q_2 - \log p(z_1, z_2) \big] \\
&= \int dz_1 q_1(z_1) \log q_1(z_1) \int dz_2 q_2(z_2) - \int dz_2 q_2(z_2) \log q_2(z_2) \int dz_1 q_1(z_1) \\
&+ \frac{1}{2} \Big[\int dz_1 q_1(z_1) a z_1^2 \int dz_2 q_2(z_2) + \int dz_2 q_2(z_2) a z_2^2 \int dz_1 q_1(z_1) + 2b \int dz_1 q_1(z_1) z_1 \int dz_2 q_2(z_2) z_2 \Big]
\end{aligned}$$

We can ignore the last two integrals as they are uneven. We can compute the first integral using Gauss' formula:

$$I = \int dz_1 q_1(z_1) a z_1^2 = \frac{1}{\sqrt{2\pi\sigma_1^2}} \int dz_1 \exp \frac{-z_1^2}{2\sigma_1^2} z_1^2 = \sigma_1^2$$

Using the relation from part a.) and the fact that:

$$\int dz_i q_2(z_i) = 1$$

We can rewrite the tedious expression as follows:

$$\begin{aligned}
KL(q \mid p) &= -\frac{1}{2} - \log \sqrt{2\pi\sigma_1^2} - \frac{1}{2} - \log \sqrt{2\pi\sigma_2^2} + \frac{1}{2}\big(a\sigma_1^2 + a\sigma_2^2\big) \\
&= -1 - \log \sqrt{2\pi\sigma_1^2} - \log \sqrt{2\pi\sigma_2^2} + \frac{1}{2} a(\sigma_1^2 + \sigma_2^2)
\end{aligned}$$

c.)

The variational solution is given by taking the partial derivatives of the KL divergence with respect to σ_1 and σ_2:

$$\frac{\partial KL}{\partial \sigma_1} = \frac{\partial(-\log \sqrt{2\pi\sigma_1^2} + \frac{1}{2} a\sigma_1^2)}{\partial \sigma_1} = 0$$

So, taking derivatives and then solving for a:

$$a\sigma_1 - \frac{1}{\sigma_1} = 0$$

$$a\sigma_1 = \frac{1}{\sigma_1}$$

The expression derived in b.) is symmetric in σ_1 and σ_2, so we can conclude that the solution for σ_2 is the same. Therefore, the overall solution is given by $\sigma_i^2 = \frac{1}{a}$.

d.)

We want to minimize the reverse KL divergence:

$$KL(p \mid q) = \int dz_1 dz_2 p(z_1, z_2) \log \frac{p(z_1, z_2)}{q(z_1, z_2)} \tag{4}$$

$$= \int dz_1 dz_2 p(z_1, z_2) \log p(z_1, z_2) - \int dz_1 dz_2 p(z_1, z_2) \log q(z_1) q(z_2) \tag{5}$$

The first integral [Entropy term] cannot be simplified further to our knowledge. The same goes for the second integral [Cross entropy term]. Therefore, we cannot minimize the reverse KL divergence. We would love to know if there is a way to do this.

e.)

Since $E_p z_i z_j = (\Lambda^{-1})_{ij}$ and

$$\Lambda^{-1} = \frac{1}{a^2 - b^2} \begin{bmatrix} a & -b \\ -b & a \end{bmatrix}, \tag{6}$$

we can easily see that

$$\sigma_i^2 = E_p z_i^2 = \frac{a}{a^2 - b^2}.$$

Thinking about the use of forward and reversed probabilities, we expect that minimising the forward $KL(p|q)$ would stretch our variational distribution and the reversed $KL(q|p)$ would squeeze it. Thus, we can expect the variance of the forward distribution to be smaller than the one for the reversed as illustrated by the equation:

$$\frac{1}{a} < \frac{a}{a^2 - b^2}$$