S1: Principals of Data Science - Coursework

William Knottenbelt, wdk24

November 20, 2023

Part (a)

The total probability density function is given by:

$$p(M; f, \lambda, \mu, \sigma) = fs(M; \mu, \sigma) + (1 - f)b(M; \lambda),$$

where $s(M; \mu, \sigma)$ is the normal distribution and $b(M; \lambda)$ is the exponential decay distribution, which is only non-zero for $M \geq 0$:

$$b(M; \lambda) = \begin{cases} \lambda e^{-\lambda M} & \text{for } M \geq 0, \\ 0 & \text{for } M < 0. \end{cases}$$

The condition for $p(M; f, \lambda, \mu, \sigma)$ to be properly normalised over $M \in [-\infty, +\infty]$ is:

$$\int_{-\infty}^{+\infty} p \, dM = 1,$$

To show that p is properly normalised we will use the identity:

$$\int_{-\infty}^{+\infty} e^{-ax^2} = \sqrt{\frac{\pi}{a}}, \tag{1}$$

We have:

$$\int_{-\infty}^{+\infty} p \, dM = \int_{-\infty}^{+\infty} fs + (1 - f)b \, dM = f \int_{-\infty}^{+\infty} s \, dM + (1 - f) \int_{0}^{+\infty} b \, dM.$$

For the first term, we have:

$$\int_{-\infty}^{+\infty} s \, dM = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M - \mu)^2}{2\sigma^2}\right) dM.$$

We use the substitution $x = M - \mu$ such that $dx = dM$ and the limits are the same since $x(M \to \pm\infty) \to \pm\infty$ for any finite μ. Then we have:

$$\int_{-\infty}^{+\infty} s \, dM = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx.$$

Using identity (1) with $a = \frac{1}{2\sigma^2}$, we find that:

$$\int_{-\infty}^{+\infty} s \, dM = \frac{1}{\sigma\sqrt{2\pi}} \sqrt{2\sigma^2\pi} = 1.$$

To find the second term we have:

$$\int_0^c b\,dM = \int_0^c \lambda e^{-\lambda M}\,dM = (-e^{-\lambda M})|_0^c = 1 - e^{-\lambda c} \xrightarrow[c\to+\infty]{} 1.$$

Hence, we see that:

$$\int_{-\infty}^{+\infty} p\,dM = f\int_{-\infty}^{+\infty} s\,dM + (1-f)\int_0^{+\infty} b\,dM = f + (1-f) = 1.$$

Thus, the normalisation condition is satisfied for p over $M \in [-\infty, +\infty]$.

Part (b)

To ensure we have the right amount of the signal and background distributions, we must normalise them separately then sum them. By the 'right amount', I mean such that the signal contributes f to the total probability, while the background contributes (1-f).

Should I also shift the exponential to make it start from alpha???

Part (b)

When M is restricted to the range $M \in [\alpha, \beta]$, the probability density function is defined:

$$p(M;\boldsymbol{\theta}) = \begin{cases} A[fs(M) + (1-f)b(M)] & \text{for } M \in [\alpha,\beta], \\ 0 & \text{otherwise.} \end{cases}$$

where $\boldsymbol{\theta} \equiv (f, \lambda, \mu, \sigma)$ are the parameters and A is a normalisation factor. For p to be properly normalised, we must have:

$$\int_\alpha^\beta p(X)\,dX = \int_\alpha^\beta Afs(X) + A(1-f)b(X)\,dX = 1.$$

Then we have:

$$1 = \int_\alpha^\beta p(X)\,dX = \int_{-\infty}^\beta p(X)\,dX - \int_{-\infty}^\alpha p(X)\,dX$$

$$= Af\left(\int_{-\infty}^\beta s(X)\,dX - \int_{-\infty}^\alpha s(X)\,dX\right) + A(1-f)\left(\int_{-\infty}^\beta b(X)\,dX - \int_{-\infty}^\alpha b(X)\,dX\right),$$

thus:

$$1 = Af\left(F_s(\beta) - F_s(\alpha)\right) + A(1-f)\left(F_b(\beta) - F_b(\alpha)\right),$$

where F_s, F_b are the cumulative distribution functions of the (normal) signal distribution, s, and the (exponential decay) background distribution, b, respectively. These are given:

$$F_s(X) = \Phi(\frac{X-\mu}{\sigma})$$

$$F_b(X) = \begin{cases} 1 - e^{-\lambda X} & \text{for } X \geq 0, \\ 0 & \text{for } X < 0. \end{cases}$$

If we assume that α and β are positive, then we can solve for A to find:

$$A = \frac{1}{f\left(\Phi(\frac{\beta-\mu}{\sigma}) - \Phi(\frac{\alpha-\mu}{\sigma})\right) + (1-f)(e^{-\lambda\alpha} - e^{-\lambda\beta})}$$

Finally, the full expression for the total probability density function, assuming $\alpha, \beta > 0$, is:

$$p(M;\boldsymbol{\theta}) = \frac{\frac{f}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(M-\mu)^2}{2\sigma^2}\right) + (1-f)\lambda e^{-\lambda M}}{f\left(\Phi(\frac{\beta-\mu}{\sigma}) - \Phi(\frac{\alpha-\mu}{\sigma})\right) + (1-f)(e^{-\lambda\alpha} - e^{-\lambda\beta})}$$

Part (e)

Make sure to talk about the generation using the percentage point function works even though it is not the ppf of the normalised functions, since we find the upper and lower bound of the probabilities to feed into it, and only relative probabilities matter.