Principle of data science coursework

Alberto Plebani
ap2387

1 Exercise 1

The statistical model is presented in Equation (1), where the background follows an exponentially decaying distribution $b(M;\lambda) = \Theta(M)^1 \lambda e^{-\lambda M}$ and the signal is a Gaussian distribution with mean μ and variance σ^2.

$$p(M; f, \lambda, \mu, \sigma) = f \cdot s(M; \mu, \sigma) + (1 - f) \cdot b(M; \lambda) \tag{1}$$

a) Prove that the probability distribution is normalised in the range $M \in [-\infty, \infty]$.

Proof. $\int_{\mathbb{R}} p(M; f, \lambda, \mu, \sigma) dM = (1-f)\lambda \int_{\mathbb{R}^+} e^{-\lambda M} dM + \frac{f}{\sqrt{2\pi}\sigma} \int_{\mathbb{R}} \exp\left(\frac{-(M-\lambda)^2}{2\sigma^2}\right)$. The first integral is $1 - f$, because the integral of the exponential from 0 to ∞ is $-1/\lambda$, whereas the integral of the second function is f, because the Gaussian distribution is normalised in $\mathbb{R}$. Therefore, the sum of the two integrals is $f + 1 - f = 1$. □

b) Because $M \in [5, 5.6]$, we need to change the normalisation factor so that the total probability equals 1. In order to do so, recall the cumulative density function (cdf) $F(X) = \int_{\infty}^{X} f(X')dX'$. Given that for the exponential $F(X) = 1 - e^{-\lambda X}$ and for the Gaussian $F(X) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{X-\mu}{\sqrt{2}\sigma}\right)\right)$, we can normalise the two different distributions separately, as in Equations (2) and (3), respectively for signal and background.

$$N_S^{-1}(\lambda, \mu, \sigma; \alpha, \beta) = \frac{1}{2}\left(\text{erf}\left(\frac{\beta - \mu}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{\alpha - \mu}{\sqrt{2}\sigma}\right)\right) \tag{2}$$

$$N_B^{-1}(\lambda, \mu, \sigma; \alpha, \beta) = e^{-\lambda\alpha} - e^{-\lambda\beta} \tag{3}$$

Therefore the resulting pdf, displayed in Equation (4), is normalised in the range $M \in [\alpha, \beta]$

$$\text{pdf}(M; \boldsymbol{\theta}, \alpha, \beta) = N_S(\lambda, \mu, \sigma; \alpha, \beta) \cdot f \cdot s(M; \mu, \sigma) + N_B(\lambda, \mu, \sigma; \alpha, \beta) \cdot (1 - f) \cdot b(M; \lambda) \tag{4}$$

c) In order to prove that the pdf is correctly normalised, I used the code `src/solve_part_c.py`. The details to run it are in the README file. I generated 10000 random values of

[1] $\Theta(x)$ is the Heaviside step function, which returns 0 for values smaller than 0, and 1 for values greater

Number of models	α	β	Mean	Variance
10000	5	5.6	1.0	6.7×10^{-33}
10000	0	10	1.0	9.5×10^{-33}
10000	-5	5	1.0	7.9×10^{-33}
49	5	5.6	1.0	5.0×10^{-33}
49	0	10	1.0	6.8×10^{-33}

Table 1: Some example results obtained when running the code, which display that the pdf defined in Equation (4) is normalised.

$\boldsymbol{\theta} = (f, \lambda, \mu, \sigma)$ uniformly distributed in a specified range of values, and then I numerically evaluated the integralof the pdf in $[\alpha, \beta]$ with the `np.trapz` function. I tested different values of α and β, and for all values the average of the 10000 integrals was 1, with a very low variance, as displayed in Table 1.