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 $b(M;\lambda) = \lambda e^{-\lambda M}$ 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 does not converge, whereas the integral of the second function is simply f, and therefore the sum of the two integrals is $f + 1 - f = 1$. $\square$ Check this

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) $F(X) = \int\limits_{\infty}^{X} f(X')dX'$. Because for the exponential $F(X) = 1 - e^{-\lambda X}$ and for the Gaussian $F(X) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{X-\mu}{\sqrt{2}\sigma}\right)\right)$, by forcing $\int\limits_{\alpha}^{\beta} N \cdot p(M; f, \lambda, \mu, \sigma)dM$ to be 1, meaning the pdf is normalised in the range $M \in [\alpha, \beta]$, we get Equation (2)

$$N(f, \lambda, \mu, \sigma; \alpha, \beta) = N(\boldsymbol{\theta}; \alpha, \beta) = \left[(1-f)(e^{-\lambda\alpha} - e^{-\lambda\beta}) + \frac{f}{2}\left(\text{erf}\left(\frac{\beta-\mu}{\sqrt{2}\sigma}\right) - \text{erf}\left(\frac{\alpha-\mu}{\sqrt{2}\sigma}\right)\right)\right]^{-1} \tag{2}$$