Statistics, Data Analysis, and Machine Learning for Physicists
Tejaswi Nerella
Winter 2023

Homework 1

This homework will introduce you to a few statistical concepts and analysis techniques that we have talked about in class, hopefully at a challenging level. Upload your solutions on Gauchospace by EOD Sunday 01-29-2023.

Problem 1: Supernova Neutrinos

Most of the energy in a supernova explosion is released in the form of neutrinos. Supernova 1987A, which exploded in the Large Magellanic Cloud, was the nearest supernova to the Earth in modern times, and the Kamiokande II detector in Japan observed 12 neutrinos from this explosion. We will model the neutrino event rate, $R(t)$, at Kamiokande as a function of three terms: $t_{\rm SN}$, the time when the supernova went off, τ, the exponential decay time of the neutrino signal and F_0, the product of the fluence and the "effective" cross-section of the detector:

$$R(t) = \frac{F_0}{\tau} \exp\left[-\frac{t - t_{\rm SN}}{\tau}\right] \Theta(t - t_{\rm SN}) \tag{1}$$

where $\Theta(x)$ is the Heaviside function: $\Theta(x) = 0$ for $x < 0$ and $\Theta(x) = 1$ for $x \geq 0$. The Heaviside function encodes the information that the event rate is 0 before the explosion.

We detect N events from the supernova that arrive at time t_1, t_2, ..., t_N, where $t_1 < t_2 < ... < t_N$.

a. Write down the log of the likelihood function, the probability of the data given the model (parametrized by F_0, τ, and $t_{\rm SN}$), by binning the data in bins of width Δt, chosen to be small enough such that within each bin, the event rate is constant. [*Note that you could solve the rest of the problem using binned data, but binning is never optimal. The useful method here, letting the interval size go to zero and using Poisson statistics, ends up simplifying a lot of problems;*]

b. Now take the limit as $\Delta t \to 0$ to get an un-binned log likelihood.

c. Determine analytic expressions for the maximum likelihood values of F_0, τ, and $t_{\rm SN}$ by differentiating.

d. This is a case where the likelihood function is very asymmetric. We can see this by computing the likelihood function as a function of $t_{\rm SN}$ and τ for realistic data.

   For this part of the problem, assume that the detector saw 12 events that arrived at the following times: 0, 0.1, 0.15, 0.3, 0.5, 0.9, 1.55, 1.7, 3, 5, 7 and 9.15 seconds, where time is measured relative to the first neutrino.

   Plot a contour plot for the likelihood function as a function of $t_{\rm SN}$ and τ; show the 1, 2, and 3σ contours (when stating the σ number for a contour, people count the probability enclosed within the contour, and ask at how many σs away from a mean, both ways, does a one-dimensional Gaussian enclose the same probability). In this case, you need to determine the contours of constant likelihood enclosing 68%, 95%, and 99.7% of the total, integrated likelihood.

e. Finally, marginalize (integrate) over F_0 and τ assuming uniform priors to obtain a one-dimensional probability distribution for $t_{\rm SN}$. Plot this distribution. Calculate its expectation value.

Problem 2: Detection in Gaussian Noise

Suppose you have data with x_i $(i = 1, \ldots, N)$ with N real-numbers. Your task is to determine whether this data is pure noise, or contains a known signal μ_i (with additive noise still present). In other words, if you think of the data as

$$x_i = \mu_i S + n_i, \tag{2}$$

where the noise terms n_i are all independent Gaussian random variables, each with zero mean, and known variances σ_i^2, you have to choose whether $S = 0$ or $S = 1$.

a. Derive the Neyman-Pearson detection statistic for making the choice between $S = 0$ (the null hypothesis) and $S = 1$ (the signal hypothesis). You can (and should) liberally use the property that monotonic functions of the test statistic lead to classifiers with the same characteristics.

b. Having found the detection statistic in part a, (say $\mathcal{T}$), derive its distribution under both the null hypothesis, and the signal hypothesis.

Hint: A linear superposition of a number of Gaussian random variables is itself distributed as a Gaussian random variable.

c. If you choose a threshold η for the test statistic value, derive expressions for the false alarm probability α and the false negative probability β of your classifier in terms of some functions that you can evaluate on a computer if you knew the values of all parameters.

d. Suppose $N = 512$, $\sigma_i = 1$ for all i, and

$$\mu_i = \frac{3}{(2\pi \times (20^2))^{1/4}} e^{-\frac{(i-256)^2}{4\times(20^2)}} \tag{3}$$

Explicitly simulate a large number of realizations of the null and signal hypotheses, and show that the distributions you derived in part b. agree with the empirical distributions of the test statistic.

Given a set of samples, you can plot their empirical distribution in Python using the `matplotlib` "hist" function with a suitable choice of number of bins, with the argument `density = True`.

e. Plot the ROC curve of your classifier for the parameters given in part d.

f. Your friend who hasn't taken Phys 240 tells you "Let's not do all this fancy math and let's just try to cook up a test that feels right. I've plotted the data for several realizations, and I sort of see that the values tend to be larger near the middle when $S = 1$, where μ_i peaks. Let's just use the 'power' of the inner part of the data: $\sum_{i=256-50}^{256+50} x_i^2$ as our test statistic. When x_i is large this should favor the signal hypothesis".

Empirically simulate the distribution of your friend's test statistic in the signal and null hypotheses, and over-plot the ROC curve for their classifier on top of your curve from part e. At a fixed false-alarm of 1%, how does their true positive rate compare to yours?

Note: It turns out you can also analytically compute the distributions of your friend's test statistic. We will see that later in the course.