Statistics and Data Analysis

Evaluation Exercise for Hypothesis testing part

We have a detector that measures hit positions in an (x, y) plane in the range ([-1,1],[-1,1]) and the energy (E) in the range [0,10], in arbitrary units.

Using such a detector, we perform an experiment with the aim of statistically establishing the detection of events of type *signal*, in the presence of events of type *background*. Signal and background events have different p.d.f.'s for x, y and E. The result of our experiment is provided in the input file `data_On.txt`.

In an independent experiment, using the same detector, we measure in conditions such that only background events are collected. We measure for 3 times longer and therefore we expect 3 times more background events in the Off data sample than in the On data sample. The results are provided in the input file `data_Off.txt`

In order to study the best way to separate signal from background using the spatial information we use two independent simulated samples (train and test) for signal and background events, respectively. Those are `train_signal.txt`, `train_bkg.txt`, `test_signal.txt` and `test_bkg.txt`.

The joint p.d.f.'s for the spatial coordinates for signal and background events are, respectively:

$$f_s(x, y) \propto e^{-(ax^2+by^2+2c\,xy)}$$

with $a = b = 6$ and $c = -5$, for signal, and

$$f_b(x, y) \propto e^{-\frac{1}{2}\left(\frac{r-r_0}{\sigma_r}\right)^2}$$

with $r = \sqrt{x^2 + y^2}$, $r_0 = 0.6$ and $\sigma_r = 0.4$, for background.

In addition, we know that the spectrum of signal and background events have, respectively, the following shapes:

$$\frac{dN_s}{dE} \propto e^{-\frac{1}{2}\left(\frac{E-E_0}{\sigma_E}\right)^2}$$

with E_0 known to be in the range [0,10] and $\sigma_E = 1$, for signal, and

$$\frac{dN_b}{dE} \propto 2 + \gamma E$$

with $\gamma > -1/5$, for background.

1. Select, out of the data and background-control samples, signal-enriched subsamples following these steps:

   (a) Using the train sample and/or the known spatial p.d.f.'s for signal and background events, consider/construct the following test statistics:

   - The radial distance $r = \sqrt{x^2 + y^2}$.
   - A Fisher discriminant using as input the polar coordinates (r,θ), with $\theta = \tan^{-1}(y/x)$. Comment why using r and θ should be better than using x and y.

- The exact likelihood ratio $\lambda(x, y) = \frac{f_b(x,y)}{f_s(x,y)}$.
- The likelihood ratio estimated from the train sample.
- A neural network (use, e.g. scikit-learn MLPRegressor class, more information at `scikit-lern.org`)

(b) For each considered test, you can compute its value T for all train and test events. Compare the distributions of T for the train and test samples by plotting them together (consider the signal and background cases separately). In addition, using a Kolmogorov or least-square test, compute the p-values for the compatibility of train and test T-distributions. Comment the cases for which the p-value is very low (i.e. $p < 0.01$).

(c) For each test, use the test sample to estimate and plot the values of $(1 - \alpha)$, β and $(1 - \alpha)/\sqrt{\beta}$ all vs. T_{cut}, where:

$$\alpha = \int_{T_{\text{cut}}}^{\infty} f_s(T)\, dT$$

and

$$\beta = \int_{0}^{T_{\text{cut}}} f_b(T)\, dT$$

(d) If we select events fulfilling $T < T_{\text{cut}}$, then $(1 - \alpha)/\sqrt{\beta}$ is the *signal-to-noise ratio*, which is approximately proportional to the statistical significance of the signal in the selected subsample. For each test statistics, compute T_{cut} as the value of T maximizing the signal-to-noise ratio, and its corresponding $(1 - \alpha)/\sqrt{\beta}$ value.

(e) Compare the performance of the different test statistics by plotting together their signal-to-noise ratio vs. $(1 - \alpha)$ curves. Comment on the result: which test is better and why do we get the order we get.

(f) For each test statistic, draw the boundaries of the critical region defined by T_{cut}, in the (x, y) plane. Comment about the reason for the shape of the critical region in each case.

(g) Using the optimal test statistic, select signal-enriched subsamples out of the data (`data_On.txt`) and background-control (`data_Off.txt`) samples. What are the estimated number of signal and background events in the data subsample? What is the statistical significance for the presence of the signal?

2. In a second step, we use the spectral information to try to increase the significance of signal detection.

(a) Write the expression for the extended/full joint likelihood function for data and background-control samples, which takes into account the spectral information as well as the number of observed events in each of the samples.

(b) Consider s (the number of signal events in the data sample) as a free parameter, and E_0 and γ as nuisance parameters. Compute the significance of the signal using the likelihood profile ratio test evaluated for events fulfilling $T < T_{\text{cut}}$.

(c) What are the best fit values for s, E_0 and γ?