Statistical Data Analysis Discussion notes – week 5

- Problem sheet 2
- Some comments on Machine Learning

Problem sheet 2

oops, this was not on your PS1

,

Exercise 1 [10 marks]: Consider (as in Problem Sheet 1) the joint pdf for the continuous random variables x and y

$$f(x,y) = \begin{cases} \frac{1}{\pi R^2} & x^2 + y^2 \le R^2\\ 0 & \text{otherwise.} \end{cases}$$

Define the new variables

$$u = \sqrt{x^2 + y^2},$$

 $v = \tan^{-1}(y/x).$

That is, u corresponds to the radius and v to the azimuthal angle in plane polar coordinates, with $u \ge 0$ and $0 \le v < 2\pi$.

1(a)

(a) [5] Find the joint pdf of u and v. (Use the inverse of the transformation $x = u \cos v$, $y = u \sin v$.) Are u and v independent? Justify your answer.



1(b) (b) [5] Find the marginal pdfs for u and v.



Exercise 2 [5 marks]: Consider *n* random variables $\vec{x} = (x_1, \ldots, x_n)$ that follow a joint pdf $f(\vec{x})$ and constants c_0, c_1, \ldots, c_n .

(a) [1 mark] Starting from the definition of the expectation value for continuous random variables, show that



(b) [4 marks] Using the result from (a), show that the variance is

$$V\left[c_0 + \sum_{i=1}^n c_i x_i\right] = \sum_{i,j=1}^n c_i c_j \operatorname{cov}[x_i, x_j] .$$

For the variance above, find what this reduces to in the case where the variables x_1, \ldots, x_n are uncorrelated.



2(b) (cont.)



2 cont.) the x; are uncorrelated T $cov[x_i, x_j] = \delta_i \sigma_i^2$ $c_{o} + \hat{\Sigma}' c_{i} x_{i}$ $z \hat{\Sigma}' e_{i} c_{j} c_{o} v [x_{i}, x_{j}]$ E cici dijoi 1,5=1 = 1=1

Exercise 3 [5 marks]: Consider two random variables x and y and a constant α . From the previous exercise we have (no need to rederive)

$$V[\alpha x + y] = \alpha^2 V[x] + V[y] + 2\alpha \operatorname{cov}[x, y] = \alpha^2 \sigma_x^2 + \sigma_y^2 + 2\alpha \rho \sigma_x \sigma_y ,$$

where $\sigma_x^2 = V[x]$, $\sigma_y^2 = V[y]$, and the correlation coefficient is $\rho = \operatorname{cov}[x, y]/\sigma_x \sigma_y$. Using this result, show that the correlation coefficient always lies in the range $-1 \leq \rho \leq 1$. (Use the fact that the variance $V[\alpha x + y]$ is always greater than or equal to zero and consider the cases $\alpha = \pm \sigma_y/\sigma_x$.)



3 (cont.)



A simple example (2D)

Consider two variables, x_1 and x_2 , and suppose we have formulas for the joint pdfs for both signal (s) and background (b) events (in real problems the formulas are usually notavailable).

 $f(x_1|x_2) \sim$ Gaussian, different means for s/b, Gaussians have same σ , which depends on x_2 , $f(x_2) \sim$ exponential, same for both s and b, $f(x_1, x_2) = f(x_1|x_2) f(x_2)$:

$$f(x_1, x_2 | \mathbf{s}) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1 - \mu_{\mathbf{s}})^2 / 2\sigma^2(x_2)} \frac{1}{\lambda} e^{-x_2/\lambda}$$
$$f(x_1, x_2 | \mathbf{b}) = \frac{1}{\sqrt{2\pi}\sigma(x_2)} e^{-(x_1 - \mu_{\mathbf{b}})^2 / 2\sigma^2(x_2)} \frac{1}{\lambda} e^{-x_2/\lambda}$$
$$\sigma(x_2) = \sigma_0 e^{-x_2/\xi}$$

Joint and marginal distributions of x_1, x_2



G. Cowan / RHUL Physics

Likelihood ratio for 2D example

Neyman-Pearson lemma says best critical region is determined by the likelihood ratio:

$$t(x_1, x_2) = \frac{f(x_1, x_2|\mathbf{s})}{f(x_1, x_2|\mathbf{b})}$$

Equivalently we can use any monotonic function of this as a test statistic, e.g.,

$$\ln t = \frac{\frac{1}{2}(\mu_{\rm b}^2 - \mu_{\rm s}^2) + (\mu_{\rm s} - \mu_{\rm b})x_1}{\sigma_0^2 e^{-2x_2/\xi}}$$

Boundary of optimal critical region will be curve of constant $\ln t$, and this depends on x_2 !

G. Cowan / RHUL Physics

Contours of constant MVA output



G. Cowan / RHUL Physics

Contours of constant MVA output



Training samples: 10⁵ signal and 10⁵ background events

G. Cowan / RHUL Physics

ROC curve



ROC = "receiver operating characteristic" (term from signal processing).

Shows (usually) background rejection $(1-\varepsilon_b)$ versus signal efficiency ε_s .

Higher curve is better; usually analysis focused on a small part of the curve.

2D Example: discussion

Even though the distribution of x_2 is same for signal and background, x_1 and x_2 are not independent, so using x_2 as an input variable helps.

Here we can understand why: high values of x_2 correspond to a smaller σ for the Gaussian of x_1 . So high x_2 means that the value of x_1 was well measured.

If we don't consider x_2 , then all of the x_1 measurements are lumped together. Those with large σ (low x_2) "pollute" the well measured events with low σ (high x_2).

Often in HEP there may be variables that are characteristic of how well measured an event is (region of detector, number of pile-up vertices,...). Including these variables in a multivariate analysis preserves the information carried by the well-measured events, leading to improved performance.

G. Cowan / RHUL Physics

K. Cranmer, U. Seljak and K. Terao, *Machine Learning*, in R.L. Workman et al. (PDG), Prog. Theor. Exp. Phys. 2022, 083C01 (2022); https://pdg.lbl.gov/

Convolutional Neural Networks

Designed for image data (pixels) \rightarrow number of input variables $\gtrsim 10^{6}$.

Intermediate layers include "convolutions" of an area in previous layer, i.e., transformed pixel is a linear combination of pixels in local neighborhood in previous layer

 \rightarrow far fewer connections than a fully connected MLP.



CNNs widely used for image classification.

G. Cowan / RHUL Physics

K. Cranmer, U. Seljak and K. Terao, *Machine Learning*, in R.L. Workman et al. (PDG), Prog. Theor. Exp. Phys. 2022, 083C01 (2022); https://pdg.lbl.gov/

Recurrent Neural Networks

Designed for sequential data (time series).



Figure 41.6: Pictorial description of a RNN (on the left) which takes an input and produces an output at every step with a hidden-to-hidden connection. The right diagram is unrolled over discrete steps. The yellow box represents a cell: a set of operations unique to each architecture.

RNNs used, e.g., in natural language processing.

Jonathan Shlomi, Peter Battaglia, Jean-Roch Vlimant, *Graph Neural Networks in Particle Physics*, 2021 *Mach. Learn.: Sci. Technol.* **2** 021001, 2021; https://arxiv.org/abs/2007.13681.

Graph Neural Networks

GNNs work with graph-structured input data, e.g., signals from particles in tracking detector:

Graph = set of nodes plus set of edges:





Part of a larger field called "geometric deep learning":

CNN is a type of GNN, graph relates pixel to its neighbors.

Transformer is a GNN that uses a mechanism called "attention", used in natural language processing (T of ChatGPT).