S1: Principles of Data Science

Problem Sheet 1

MPhil in Data Intensive Science

Matt Kenzie
mk652@cam.ac.uk

Michaelmas Term 2023

Problem Sheet 1

Lectures 1 – 6
Topics covered: datasets, visualisation, data structures, performance criteria, metrics, random variables, probability density, changing variables, marginal and conditional p.d.f.s, Bayesian inference, expectation values, common distributions

1. Get hold of the example dataset stored in the module `gitlab` page `https://gitlab.developers.cam.ac.uk/phy/data-intensive-science-mphil/s1_principles_of_data_science/-/blob/main/datasets/ps1.pkl`

   (a) Make some pair plots of the distributions in the dataset and try to become a bit familiar with what the data might be showing.

   (b) Make a publication quality plot which shows the two-dimensional distribution between the variables `vx` and `vz`.

   (c) Compute the correlation and covariance matrices between all variables in the dataset.

2. Show that the average deviation from the mean, rather than the squared deviation (*i.e.* the variance), is not a useful measure of the spread of a distribution. This may lead you to think a decent definition would be the absolute the deviation, can you think why we use the variance (squared deviation) rather than the absolute deviation?

3. The Monty-Hall problem with N doors. In the lectures we saw that in the Monty Hall game show problem with 3 doors it is always advantageous to switch. Sticking on the initial door has a win outcome with $p = 1/3$, whereas switching has $p = 2/3$.

   (a) Write a numerical simulation of the Monty Hall problem to show that the chance of winning if you switch is approximately 2/3 but only 1/3 if you stick

   (b) Now imagine there are 100 doors (still with only one car) and that after your initial choice the game show host opens 98 of them to reveal goats. Compute the probabilities of winning if you stick or switch. Perhaps this will help convince you (if Bayes' theorem didn't already) that switching is a good strategy.

(c) It should now be trivial for you to write down an expression (but it's also instructive to prove it) for the case where there are N doors and the host opens $N-2$.

(d) Show that if there are N doors and the host opens p of them that the probability of winning if you switch is

$$\frac{1}{N}\frac{N-1}{N-p-1}. \tag{1}$$

You can always go back to your simulation to check this.

(e) **An extension that you may decide to give up on.** Show that if there are N boxes, containing k prizes and the host opens p of them, then the probability of winning if you switch is

$$\frac{k}{N}\frac{N-1}{N-p-1}. \tag{2}$$

You can always go back to your simulation to check this.

(f) **An extension you may decide to give up on.** Show that if there are N doors, containing k prizes, where the contestant intially selects m doors and the host opens p boxes which reveal r prizes, then you should switch doors providing the proportion of prizes revealed by the host is smaller than the proportion in the intial configuration, *i.e.* if the following condition is met:

$$\frac{r}{p} < \frac{k}{N}. \tag{3}$$

It is easier to first consider the case where there is only a single initial door chosen, $m = 1$. You can then consider the case where $m > 1$ and a "switch" in this context means switching all of the initial choice for a new set of m doors (this does require that there are sufficient doors remaining so that this switch can be done, *i.e.* that $m \leq N - p - m$, but you don't need this constraint to solve the problem). You may finally consider the case where one is allowed to switch some subset of m, call this ℓ, doors. In all of these cases the same equality condition applies that you should switch if $r/p < k/N$.

4. Show that for a binomial distribution with n trials and a success rate of p that the sum of all possible outcomes is unity. In other words, prove that the binomial p.d.f. is properly normalised.

5. Imagine trying to "track" the path of some projectile using a set of cameras at fixed positions. You need at least three measurements of the projectile to accurately predict its motion (because it is accelerating). The "hit" efficiency for a single camera is 95% (in other words it will completely miss the projectile 5% of the time).

   (a) If we require a minimum of three hits, how efficient would the system be with three cameras?

   (b) Would having four or five give a significant improvement (once again requiring at least 3 hits)?

6. Show that the sum of two independent Poisson distributed variables is also Poisson distributed. In other words show that if X is distributed according to $\text{Poiss}(X;\lambda_X)$

and Y is distributed according to $\text{Poiss}(Y; \lambda_Y)$ then $Z = X+Y$ is distributed according to $\text{Poiss}(Z; \lambda_Z)$, and find the value of λ_Z.