HOMEWORK 1

Due date: 10/04/2022 23:59 hrs

> Homeworks in this class are turned in as Jupyter Notebooks.

The data in the attached file, `sn-data.csv`, contains a mixture of real and fake data. The real data was obtained from the Super Nova Cosmology Project, which in turn provides a compilation from several sources (for details, check their website). The columns in the file, correspond to `redshift`, `Distance Modulus`, and `Distance Modulus Uncertainty`. Each row corresponds to the measurement of these quantities for an individual Super Nova (SN). Roughly speaking, redshift (usually symbolized as z) is a measure of the velocity with which the SN (and by extension the galaxy in which the SN occurred), is moving away from us: the larger the redshift, the faster it is moving away. Distance Modulus (symbolized as μ), is a quantity that is closely related to the distance and is based on how much fainter a source appears when it is farther away: the higher distance modulus, the farther away the SN and its galaxy were. The relation between distance and velocity is what allowed us to discover the expansion (in fact the accelerated expansion) of the Universe. We will explore this data set in the following 2 problems.

Problem 1

We will start by trying to model the relation between redshift and distance modulus. You must attempt a polynomial regression.

1. Use an MLE to fit a polynomial between the distance modulus and the redshift. Use what you leaned about cross-validation to decide what order polynomial to use. In particular, split the sample into a training and a testing sample (take a look at `sklearn.model_selection.train_test_split`).
2. For the order that was chosen in the previous part, perform a Ridge and a Lasso regression and compare the results.
3. The data set contains a few outliers. Perform a Hubber loss regression. How did you choose the hyper-parameter? How does it compare to the Ridge and Lasso regressions done before?

Problem 2

The model that you found in the previous problem does not provide much information about the physical relationship between z and μqp while. Our current standard Cosmological model provides a physically motivated way to model this relationship (James Peebles was awarded the Novel Prize in Physics in 2019 for his theoretical work that established the foundations of this model). The relevant equation is the following:

$$\mu(z) = -5\log_{10}\left((1+z)\frac{c}{H_0}\int_0^z \frac{dz}{(\Omega_m(1+z)^3+\Omega_\Lambda)^{1/2}}\right)$$

where c is the speed of light, and H_0, Ω_m, and Ω_Λ are free parameters of the model (they are the Hubble constant, the matter density, and the Cosmologigal Constant, respectively). today we know that these parameters are approximately: $H_0 = 71\text{km s}^{-1}\text{ Mpc}^{-1}$, $\Omega_m = 0.27$ and $\Omega_\Lambda = 0.73$.

Notice that, although it looks complicated, the above equation can be used as a generative model, that is, given a set of parameters $\{H_0, \Omega_m, \Omega_\Lambda\}$, you can generate values of μ for a given z.

Perform a full Bayesian regression to determine the joint probability distribution for the three free parameters given the observed data. You can assume that the column `Distance Modulus Uncertainty` represents Gaussian noise and you must take into account that there are outliers in the data.

> The discussion in Chapter 8 of "Statistics, Data Mining and Machine learning in Astronomy", Ivecic et al., can be useful for Problems 1 and 2.

Problem 3

It is important to develop an idea of the intricacies associated with the information provided by large surveys in their databases. Here we will estimate the depth of an SDSS image to get a sense for the uncertainties associated with the photometry that the SDSS database provides. To that end:

1. Download the r-band SDSS dr7 (data release 7) image centered at coordinates: `RA 13:39:55.92, DEC +00:50:10.02`. Download a region of 30×30 arcmin with a resolution of 0.3 arcseconds per pixel. The **python** package `astroquery.skyview` is recommended for this. Beware that the file size will be close to 137 Mb, it will take a while to download. If your computer has trouble handling it, download a smaller image and specify in your solution.

2. We will estimate the depth of the image in 3" diameter apertures. For that, you will have to place a large number of apertures in random positions of the sky and study the distribution of the fluxes within those apertures. Avoid the beautiful pair of galaxies in the center (this system is known as ARP 240). The **python** package **photutils** can be useful. Note that the image is not background-subtracted, so, for every aperture you must estimate the local background using an annulus (see here).

   Note: the transformation from the image units into magnitudes is not straightforward for these images. The zeropoint should be 28.2576 but there are several other complications that need to be taken into account to obtain accurate photometry (airmass at which the image was taken, for example). If you are interested, you can find all the information necessary here but to keep things simple, we will only care about the S/N, which should be independent of the units of the image.

3. Report the 5σ limiting *flux* (in whatever units the image is in) for 3 arcsec-diameter apertures.

4. Now pick a few *compact* sources (our circular apertures work best with stars) in the field and compare your estimates of their S/N with those reported by SDSS. To do this, you have to identify the position of the sources you are interested in and query the SDSS tables for their magnitudes and uncertainties. You can do this manually in the SkyServer webpage but you should try to do it programmatically.

5. Do you see differences? What explains the differences?