SURP 2021 Computing Project - Juan Alfonzo

SURP Supervisor: Kartheik Iyer

May 5, 2021

In this assignment, we will explore the basic techniques needed to work with our dataset (galaxies from SDSS-IV MaNGA) and analyse them to look for the link between their spatial and temporal star formation activity. This is mostly a launching pad to grow familiar with the techniques we will be using over the summer, so please feel free to contact me on slack in case anything isn't clear or you run into roadblocks along the way!

1 The SDSS-IV MaNGA galaxy sample

- Install Marvin in a python environment.
- Use marvin.cube to download the datacube for a couple of galaxies (MaNGA ids: '12-84731', '1-547210') and plot (i) the V-band images and (ii) the spectrum of the central spaxel of these galaxies. You can use either matplotlib or marvin's mapplot function for this.
- Use the MaNGA DR15 catalog to create a sample of galaxies with (i) stellar mass $M_* > 10^9 M_\odot$, (ii) good quality flags, and (iii) redshift $0 < z < 0.1$. How many galaxies fall in your sample?
- (stretch goal) If you use the DR15 Galaxy Zoo classifications, how many of these galaxies are face-on disk galaxies?

2 Summary statistics for Morphology and Star-Formation activity

- What are commonly used summary statistics for whether a galaxy is actively forming stars or not? Can you come up with a discrete/categorical and a continuous example? (*Contact me if you'd like further elaboration on this*) Bluck et al. 2020 is a good reference to get started.
- Can you compute these statistics for the sample from part (1).(iii)?
- What are commonly used summary statistics to talk about the morphology of a galaxy?
- Can you compute these statistics for the sample from part (1).(iii)?
- Whitaker et al. (2015) found a correlation between the star forming state of a galaxy and it's morphology. If you repeat this with the statistics you computed in the previous parts of this section, do you see similar behaviour? Generate a plot that summarises your results.

3 A brief stab at a Machine Learning Approach

- Set up a simple linear regression network using PyTorch. (*Google is your friend, and also contact me if you're running into any issues*). We want to use this to try to try and predict a galaxy's star formation activity from tracers of it's morphology (that we visualized in the previous section).
- Set up a dataloader in pytorch to cycle through your sample and train the network. Try splitting your sample into 60-20-20 for the train-test-validate part of this procedure.
- Choose the Adam optimizer and an appropriate loss function (this will depend on whether you're training it to predict a categorical or continuous variable; *contact me if you have doubts!*).
- Train the network, and quantify it's performance using the loss function chosen in the previous step.
- Plot the loss function as a function of epoch (as the network trains), for the train and test sets. Does the network do as well on the validation set as it did on the test set?
- Now repeat this procedure with a network that has a single hidden layer. How does the performance change?
- (Stretch Goal) Instead of summary statistics for a morphology, what happens if we directly use the images of the galaxy (+ a few convolutional layers) to predict summary statistics of its star formation activity?