Capstone #1 Proposal: Predicting Galaxy Handedness Based on Observed Physical Properties

For my first capstone project, I would like to create a program that determines a galaxy's spin orientation based on observables such as signal intensity, size, and distance.

What is the question and why is it important?

Spiral galaxies can either spin clockwise (CW) or counterclockwise (CCW), relative to us. A galaxy's orientation will indicate which part of the galaxy is moving towards the Earth, and which part is moving away. This is especially helpful when calculating a star/planet/cluster's Doppler shift[1], an important physical property for objects in space.

Who would use it and why?

The potential clients for this product would be astronomers. This would aid them in automatically determining the handedness of an observed galaxy. Their results could help them confirm one of the biggest theories in astronomy: the present-day universe is isotropic[2] (the same in all directions). If there are a disproportionate number of CW and CCW galaxies in the sky, this could directly challenge the prevailing theory.

About the data

My dataset will be taken from this archive. The data has already been collected and cleaned for the most part, so I expect my data wrangling to be minimal. The handedness of each entry was sourced from the Galaxy Zoo 2 project. This project allows anyone to view images of galaxies and classify their shape and structure. Because this classification is done by non-experts, only galaxies with a 90% or higher rate of agreement were included. The data was then "supercleaned" by a galaxy image analyzer and visually inspected, resulting in an error-free dataset. The photometric data (observables) of each galaxy was acquired from the findings of the Sloan Digital Sky Survey, a vast collection of data on millions of objects in the sky[3]. Again, all of this has already been accomplished and compiled into one CSV file.

Methods

First, I would conduct exploratory data analysis to summarize, visualize, and hypothesize my dataset. Then, I would determine which machine learning algorithms to use. This appears to be a classification problem, so I expect to use supervised machine learning. Because the number of observables is so high, I would single out the factors with the greatest impact and use those for analysis.

Deliverables

I will present my findings as a slideshow, write a report, and have my code available on GitHub.

Endnotes

1. For the curious: an object's Dopper shift determines the rate at which it is moving towards/away from the observer. If the object has a redshift, this means that it is moving away from the observer at such a speed that the light waves expand in length by the time it reaches the observer, causing it to appear "redder." Similarly, if the object has a blueshift, this means it is moving towards the observer, causing the light waves to contract and appear "bluer."
2. The current belief is that the pre-inflation universe was uniform. Then, when inflation occurred, everything must have been pushed out equally with no lumps or irregularities. If the current universe is not isotropic, this is indicative of irregularities in the early universe and should be further studied.
3. Technicality: the list of galaxies was further narrowed to only include objects where the right ascension was between 6h and 18h, and with a redshift of $z<0.25$; that is, all data points are located in one hemisphere of the sky, and they are all close to Earth.