Exoplanet Analysis: The Search for Habitability

Benjamin Irving, Thomas Wilson

Abstract

In recent years, there have been a variety of strategies employed to find potentially habitable environments elsewhere in the cosmos. The primary problem lies with the lack of viable samples, environments where life is known to exist. In fact, there is exactly one: Earth. Thus, all of our assumptions about extraterrestrial habitability come from conjecture based on Earth-bound life. However, researchers have been able to determine some features which are more important than others, and have made cautious estimations about the viability of certain planets. Causal patterns between these features, in relation to the existence of habitable environments, still lie in the realms of speculation, due to the vast gaps in humanities extraterrestrial knowledge. This project sought to analyze these features through a feed-forward Neural Network to find functional relationships in the data that is available in order to classify the sample of planets as either habitable or not. (complete the experiment before continuing)

1 Introduction

At the dawn of the Space Age, humanity looks to the skies more than ever. What lies above now seems closer some how, particularly with the deployment of the Tess and Kepler space telescopes, among others, by NASA. These satellites were launched with the directive of finding Earth-like planets amongst the stars. To no surprise, this search has proved difficult. Obviously, there is no data pertaining to extraterrestrial life. Thus, when thinking about the viability of extraterrestrial habitats, we instead must use conjecture based on Earth-based species, in analyzing their structures, the history of their emergence, and, most importantly, the planetary conditions which allowed life to blossom in the first place.

Through the intense examination of terrestrial life, researchers have been able to make preliminary analyses of how certain observable planetary features lend themselves to the emergence of life. The primary factor seems to be the existence of liquid water. In terms of observable characteristics that lend themselves to this phenomena, researchers have determined an assortment of observable features that are vital. Interestingly, most of the characteristics are dependent on one another, as they all represent portions of the complex, dynamic systems that make up these worlds. For instance, the stellar mass is directly correlated to the stellar luminosity, for the magnitude of the energy output from a star can be inferred by the amount of "fuel" (in the form of hydrogen and helium) the star has left to carry out fusion. Many relationships remain less clear. (include a less clear relationship here)

The purpose of this project is to find statistical relationships between disparate planetary features. Our hope is for our model to extend to future planetary candidates.

2 Technical Approach

When beginning our approach to this problem, we made sure to take our time in building a strong statistical understanding of the patterns captured in the database provided by the NASA Exoplanets Archive, as this is the raw information we wish to understand, and any habitability classification that we may use to train our final model would be a somewhat subjective labeling applied by humans, rather than an indisputable fact of nature.

2.1 Principal Component Analysis

The first preliminary step that we took was to analyze the Gaussian mean and co-variance of the data. This is a reasonable way to look at the distribution , simply because we are looking at naturally occurring phenomena, and the cause-and-effect nature of the universe. In actuality, we learn that the nature of the

data is not modeled well by a Gaussian, though the majority of points exist near to the mean. Following this, the next preliminary method employed was some rounds of Principal Component Analysis, using various different kernels. From this we can gain a reductionist idea of the general shape of the distribution of characteristics of extra-solar planets.

2.2 Preprocessing and Training Approach

Now that we have a stronger mental model of the distribution, we prepare our data for the process of optimizing a model for the classification of 'habitable exoplanets'. There are two problems we must addess, initially: (1) The data is filled with holes – missing values in otherwise good points, and (2) we still lack a labeling on this data that we can use to train a model to learn the conditions of habitability. To address the first of these issues, we have decided to use a process of Iterative Imputation to fill in the missing values in the data table. This is preferable, in our opinions, than to throw away rows (the majority, in this case) or to simply insert a constant value (mean of that feature, etc.) and it helps us continue to build our understanding of the features that we are working with. To address the second problem, we first attempted to form our own metric of habitability, using the data available to us. However, for the purposes of training our model with respect to the most reliable scientific definition, we pull from the Planetary Habitability Lab at the University of Puerto Rico Arecibo, which provided information about what planets are currently considered by the scientific community to be possibly habitable for life. From the tables of potentially habitable planets provided by the PHL, we were able to form a binary class labelling for each point in the inputs.

2.3 Support Vector Machine

Because we want to discern the best hyper-plane to separate out this very small group of outliers in a large table of mostly one class, we must choose a classification model and hyper-parameters which is discerning enough to pick out the positive points, yet descriminates heavily against the negative points. For this we have tuned a Support Vector Machine, using hinge loss, that weights the 'not habitable' planets heavily – about 70:30, and which has an adaptive learning rate as well as a very strict convergence condition. It was very important that we use 'adaptive' learning rate, which self-adjusts closer the GD is to convergence.

2.4 Feed Forward Classifier Network (Multi Layer Perceptron)

We employ a Feed-Forward Neural Network, performing optimization with both a model with a single hidden layer and a model with three hidden layers. We use a ReLU activation for the hidden layers, and a sigmoid function for the final classification step. We optimize the cross-entropy loss over the network to acheive a

3 Experimental Results

Describe the datasets used for your experiments. Be precise in describing all information about the datasets, including, classes, number of samples per class, features used to represent data, and all pre/post processing of the datasets.
Describe the details about the implementation of each algorithm, e.g., how you perform training, validation, testing, values of the hyperparameters and your methods for hyperparameter tuning, training/validation/testing error on the dataset, and all useful plots/tables that help to better interpret your results and your work.

3.1 PCA

If we reduce the dimensionality of the data to 2D and plot it, generally what we can see is that the data mainly looks like a skewed Gamma distribution along one component, with a smaller trends off of this main cluster along the other component. In relation to the features expressed by the rest of the distribution, the data for Earth appears to be roughly near to the mean, though this may be partly affected by the relative measures that are included as some of the features in the data. Whether we choose to interpret this as meaning that the class of 'habitable' planets is located outside of the main distribution, acting as outliers, or as meaning that the class is a small group inside the "sweet spot" of the main cluster affects the decisions we must make in tuning the hyper parameters of our learning

model. There is truth to both of these lines of thinking, though our general approach to the following analysis was to look at the positive class as lying at the edge of the overall spread of data.

3.2 Imputation

The approach we took to dealing with missing values uses many iterations of regressions to 'guess' the missing features, as a function of similar points, using a Decision Tree estimator. We chose this estimator because it gave results that most accurately[1] modeled key features. We can learn a bit about the frequency of certain features from this, as well as how similar points express correlation among certain features. In our trials to find the best estimator for this, it could be observed that the less effective estimators tended to push estimated values closer to the normal distribution and mean. The planetary mass had some of the strongest effects on the distribution of the other features, and is a key pillar in most definitions of habitability.

3.3 SVC

The best accuracy that we were able to acheive with the SVM was 20%, specifically for the set of planets that we obtained from the PHL at the University of Puerto Rico, which make up our positive class. Not the best, however, the range of planets that it is labelling as habitable is getting closer to the actual thing. and we are of course still labelling the overall distribution with about 98% accuracy.

4 Participants Contribution

Please list the name of the participants. For each participant explain in details the role he/she played in the project: explain which methods was implemented by which member, which dataset was processed by which member, which experimental results were generated by which members, etc.

** Please do not change the size of the fonts.
** Please note that your submission must be at most 7 pages long, excluding references.

[1] Do note that the overall distribution *is* affected by this process, and we do create the potential for some inconceivable data being included in our table.