Exoplanet Analysis: The Search for Habitability

Benjamin Irving, Thomas Wilson

December 18, 2021

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. Using a support vector classifier, we were able to *liberally* classify 'habitable' and 'not habitable' exoplanets, though achieving a more discerning and sensitive model proved quite a challenge, and our neural network was not able to achieve the success we were hoping. However, a lot can be learned from the clustering and imputation done without labels that can be built upon in the next iterations of our research.

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 [3].
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 [3]. 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 [3]. 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. [5] Many relationships remain less clear.

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. 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 to throwing away rows (the majority, in this case) and to simply inserting 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 Classifier

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 discriminates heavily against the negative points. 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, optimized using stochastic gradient descent. 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.
Our original intention was to employ the use of the cross entropy loss function, in order to optimize the weights in the network over our iterations. The Pytorch API provides built in functionality to execute this Cross Entropy Loss, which takes the form of

$$\ell(x, y) = L = \{l_1, \dots, l_N\}^\top, \quad l_n = -\sum_{c=1}^{C} w_c \log \frac{\exp(x_{n,c})}{\exp(\sum_{i=1}^{C} x_{n,i})} y_{n,c}$$

Here, x can be described as the input variable, y the target, w is the weight, and c is the number of classes. N spans the mini-batch dimension, which we did not utilize for the training of our network.
The project is a variation of the common "imbalanced classification" problem. We first chose to try class-weighting, using a basic rule where the weight of the largest class served as the numerator for our weights, and the weight of class C served as the denominator for class C.

Due to the irregularity in our data, we also chose to employ the ADADELTA optimizer, with the aim of dynamically altering our learning rate according to first order information. The hyper parameters, in the scope of this problem, were hard to nail down, due to a lack of positive samples. As such, we did not want our learning rate dependent on these values.

input : $\gamma(lr)$, $\theta_0(params)$, $f(\theta)(objective)$, $\rho(decay)$, $\lambda(weightdecay)$
initialize : $v_0 \leftarrow 0$ $(squareavg)$, $u_0 \leftarrow 0$ $(accumulatevariables)$

for $t = 1$ **to** ... **do**
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$
$if\ \lambda \neq 0$
$g_t \leftarrow g_t + \lambda\theta_{t-1}$
$v_t \leftarrow v_{t-1}\rho + g_t^2(1-\rho)$
$\Delta x_t \leftarrow \frac{\sqrt{u_{t-1}+\epsilon}}{\sqrt{v_t+\epsilon}} g_t$
$u_t \leftarrow u_{t-1}\rho + \Delta x_t^2(1-\rho)$
$\theta_t \leftarrow \theta_{t-1} - \gamma\Delta x_t$

return θ_t

Essentially, this algorithm restricts the accumulation of the sum of the past gradients to a fixed size of w, storing the gradients as an exponentially decaying average [3]. It also makes use of the Hessian matrix, or an approximation of it, which is formed of the second partial derivatives of a scalar valued function, to make updates to the parameters.

2.5 Oversampling

Our Neural Network struggled severely with performance, due to the imbalance of size in the respective classes (habitable, uninhabitable). Thus, we chose to employ the technique of oversampling, which entails adding values from the minority dataset with replacement into the training data to ensure that the network meets with sufficient examples to make accurate classifications. [6]

3 Experimental Results

3.1 Kernel PCA

PCA.png 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

[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.

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

We were able to achieve an accuracy of 20% success among the positive class, for the Transit Strata. On the other hand, we were able to acheive 100% success for the positive class on the Radial Velocity Strata. However, the overall success rate was rather low (90% for Radial Velocity Strata), which means that the classifier is forced to overcompensate for the bounds of the positive class. Still, this is significant results, and we may be able to acheive better results with a more sophistocated approach, such as a Neural Network.

3.4 Neural Network

Unfortunately, our Neural Network was met with extremely limited success. The primary issue lay with our inability deal with the imbalanced classification. When using the cross-entropy loss optimizer, with our weights, the classification of our test sample remained stagnant. We achieved an accuracy of 63 percent for both layer formations. Our training data achieved an accuracy of 72 and 76 percent respectively. However, we did see marginal improvement upon changing the optimizer from Adam to the ADADELTA. Our training data achieved an accuracy of 83 percent, while our test sample accuracy remained stagnant. The primary the network appeared to be defaulting in its classifications, casting the value of 0 across all samples in the test data. The obvious cause was the sheer imbalance in the data, which was the result of the lack of classified "habitable" planets. As such, we attempted to amend the problem by using oversampling, adding planets with replacement from the habitable class at random into our data sets. We did this in batches of 1000, 2000, and even 5000. However, this failed to affect the performance of the network in a statistically significant manner, resulting in percentage changes within 0.1 percent of one another. There were obvious limitations to the structure of our model, particularly in the manner in which the Cross-entropy loss processed its correct, positive classifications.

4 Participants Contribution

Thomas Wilson - Preliminary Analysis, Kernel Principal Component Analysis, Iterative Imputation, Support Vector Classifier
Benjamin Irving - Concept, Scientific Research, Feed-Forward Neural Network

References

[1] NASA Exoplanet Archive, 2021. *https://exoplanetarchive.ipac.caltech.edu/cgi-bin/TblView/nph-tblView?app=ExoTblsconfig=PS*

[2] Planetary Habitability Laboratory at University of Puerto Rico Arecibo, 2021 *h*ttps://phl.upr.edu/projects/habitable-exoplanets-catalog

[3] Matthew D. Zeiler, *ADADELTA: AN ADAPTIVE LEARNING RATE METHOD*, (Google Inc, New York University, 2012). https://arxiv.org/pdf/1212.5701.pdf

[4] Dirk Schulze-Makuch, René Heller, and Edward Guinan *In Search for a Planet Better than Earth: Top Contenders for a Superhabitable World*, (German AeroSpace Agency, 2019).

[5] William C. Danchi1 and Bruno Lopez *EFFECT OF METALLICITY ON THE EVOLUTION OF THE HABITABLE ZONE FROM THE PRE-MAIN SEQUENCE TO THE ASYMPTOTIC GIANT BRANCH AND THE SEARCH FOR LIFE*, (Exoplanets and Stellar Astrophysics Laboratory, NASA Goddard Space Flight Center, 2012)

[6] Roweida Mohammed, Jumanah Rawashdeh and Malak Abdullah *Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results*, (Exoplanets and Stellar Astrophysics Laboratory, NASA Goddard Space Flight Center, 2012)