Predicting habitable exoplanets from NASA's Kepler mission data using Machine Learning

Rajeev Misra (ramisra@stanford.edu)

Overview

Purpose of this project was to process data produced by NASA's Kepler mission about exoplanets, build machine learning model using the planetary and stellar features of both potentially habitable and non habitable exoplanets. This trained model could then be used to predict habitability of exoplanets on new data.

Data

Data was collected from Kepler's public data repository [1]. It contained both confirmed and candidate exoplanet's list. Data about habitable planets was collected from "Planetary Habitability Laboratory" [2]. A recently published paper [3] identified several new habitable planets from Kepler's data.

126 habitable exoplanets and 2247 potentially non-habitable exoplanets were used in our training. This data contained planetary features such as "Planetary radius, equilibrium temperature, isolation flux" and Stellar features such as "Stellar radius, Stellar temperature".

Model

This was a binary classification problem. Support Vector Machines (SVM) was used to implement this model. SVM solves following primal problem [4]

$$\min_{w,b,\xi}\left(\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i\right)$$
$$s.t.\ \ y_i\left(w^T\phi\left(x^{(i)}\right) + b\right) \geq 1-\xi_i$$
$$\xi_i \geq 0,\quad i=1,..,m$$

and its dual form is

$$\min_{\alpha}\left(\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i y_i y_j K(x_i,x_j)\alpha_j - \sum_{i=1}^{m}\alpha_i\right)$$
$$s.t.\ \ \sum_{i=1}^{m}\alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq 1\ \ i=1,..,m$$
$$where\ \ K(x_i,x_j)=\phi(x_i)^T\phi(x_j)$$

Here K is kernel with training data mapped to higher dimension using function ϕ

Decision function used was

$$sgn\left(\sum_{i=1}^{m} y_i\alpha_i K(x_i,x)+\rho\right)$$

"Radial Basis Function" kernel was used which is of the form

$$K(x,x') = \exp\left(-\Upsilon\|x-x'\|^2\right)$$

Value of $\Upsilon=10$ was used in our SVM model.

A linear kernel with "Recursive elimination of features with cross validation" was also experimented.

scikit-learn [4] SVM API was used to implement this project.

Features

First feature selection process was manually identifying columns in Kepler data which were related with planetary or stellar features. Further feature selection was done through forward search algorithm with 50/20/30 % ratio of Training, Dev and Test set data.

To avoid one feature set over fitting just one set of dev data, multiple iterations of forward feature search was run after randomly shuffling mixture of training and dev set data. Feature set with lowest dev set error was finally selected.

Above algorithm selected "Planetary Radius" and "Isolation Flux" from data.

Error rates in percentage with above feature set for few runs were

Train Error	Dev Error	Test Error
0.35 ± 0.10	1.0 ± 0.50	1.0 ± 0.25

References

[1] Keplet Exoplanet Archive

[2] Planetary Habitability Laboratory

[3] "PLANETARY CANDIDATES OBSERVED BY Kepler.VIII" Susan E. Thompson et. al. 17 Oct 2017.

[4] scikit-learn

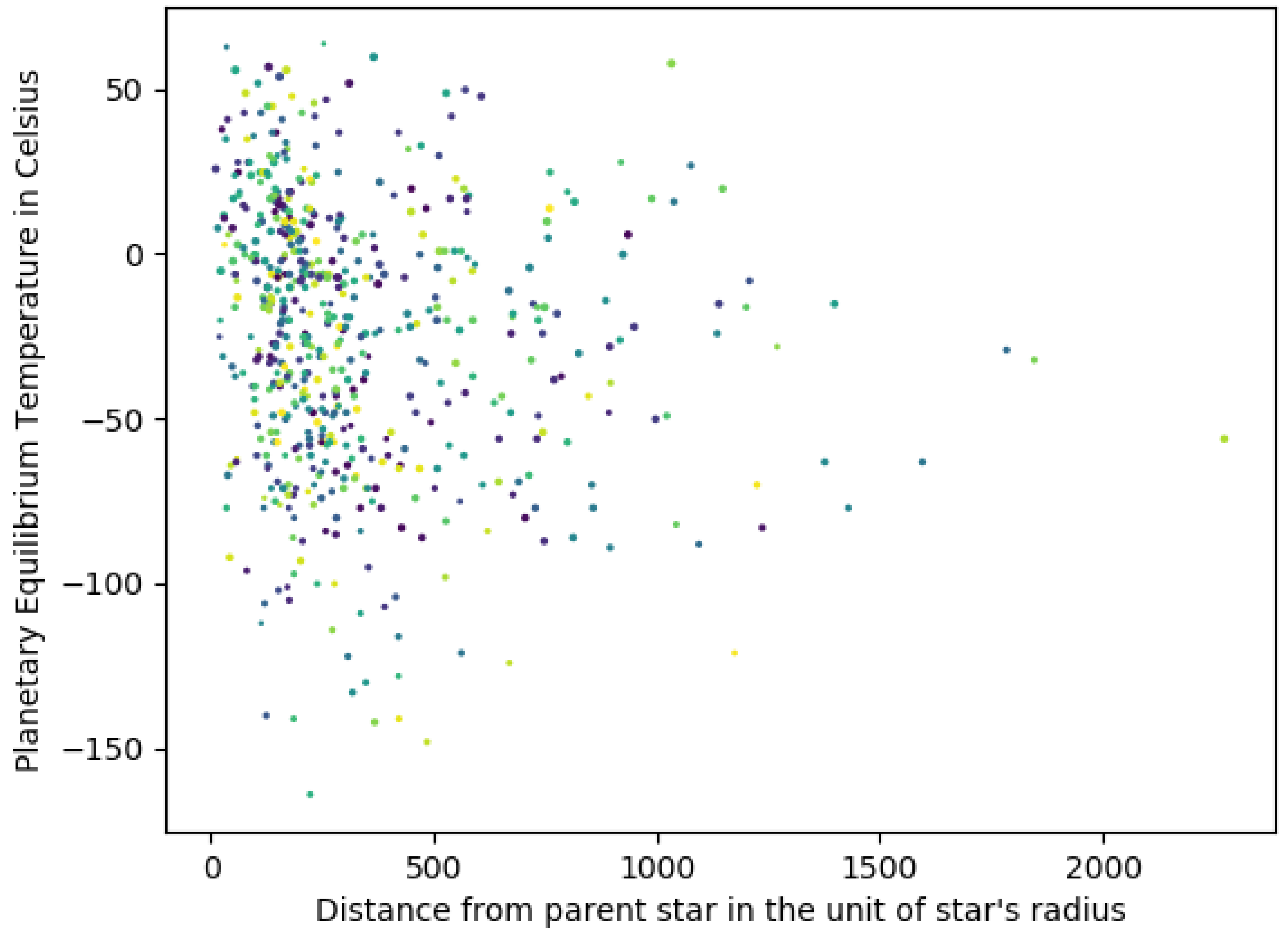
[5] API design for machine learning software: experiences from the scikit-learn project.

[6] predict_habitability.py

[7] Planetary Equilibrium Temperature

Results & Discussion

Once we developed our model [6] and trained it, we ran this model on a cumulative data set from Kepler [1] which contained all possible disposition of planets (confirmed, candidate, false positive etc) with total 9565 planets. This also included our training data. A scatter plot was created for habitable planets predicted with our model. X axis is planet's distance from parent star in the unit of parent star radius. Y axis is planetary equilibrium temperature [7] in degree Celsius. Size of scattered points are planet's radius compared to earth radius.



In this plot, every little small speck of dot is one almost earth size potentially habitable planet. We can see there is a high concentration of planets with equilibrium temperature of around 0 ± 50^oC and distance from parent star ranging from approximately 100 unit to 400 unit distance.

For comparison, Earth is 215 unit distance from Sun and equilibrium temperature of Earth is around -13^oC. The actual planetary surface temperature could be higher depending on amount of greenhouse gas effect similar to Earth where mean surface temperature is 27^oC due to greenhouse [7]. This means even equilibrium temperature below 0^oC may have higher surface planetary temperature and may support liquid water which is necessary for life.

Being relatively closer to parent star means, there is high probability that these are rocky planets as rocky planets generally form closer to their parent star. Size of these planets is similar to earth (Approx. 0.5 to 2 times Earth radius).

This means our model predicted exoplanets with features in close vicinity of Earth's features. This gives us confidence that conditions on these planets are probably similar to Earth which make them suitable for life.

Model's prediction matching real world observation gives us confidence that model developed was on right track.