Elijah Andrushenko

CPTS 317

04/30/2018

Data Mining

Project Report

1. Introduction

This project set out to find planets that were similar to Earth, for the purpose of seeing which planets that we know would be most habitable, by humans. Motivation for this project was my curiosity in astrology and to apply a computer science approach to a NASA engineering question. Data mining questions that were set out to be answered was how effect would the K-means algorithm work for finding planets similar to Earth. Also I observed how changing the K variable would affect the data and computational speed.

Problems encountered were finding data, many data points for the thousands of samples were missing so making accurate decisions would be a bit more difficult.

The results of this project showed me how the K variable in K-means affects clusters. As K increases the clusters contain smaller and smaller samples within each one. Thus similarity within clusters increases as K increases. I also learned that K-means is better at removing outliers rather than finding similarities. The algorithm left me with 100 planets that were similar to earth, which isn’t too useful, what was more useful is that 2700 planets were removed as candidates as potential planets.

2. Data Mining Task

Key challenges of this project were applying a Euclidean method to data that may not accept it. Descriptions such as mass, radius, distance from the sun, and temperature applied to a Euclidean method may not be the best when it comes to grouping planets since many of those fields differ.

Input was a CSV file with thousands of planets and several descriptors included in the README.txt file.

Output was a python lists of clusters, the code output was to check if Kepler-22b and Earth were in the same cluster, which they were. Thus the project in that area was a success.

3. Technical Approach

A way to approach the Euclidean problem would be use a similarity score instead of Euclid.

K-means algorithm was applied, first I opened the file and made a list of lists. Each list was a planets and all of its descriptions. Then a list for centers was made to keep track as to which planets went into which centers. Earth and many Milky Way planets were chosen as centers since they are different, but not too different when it comes to the entire universe of planets. Thus we would have respectable clusters and not just all the planets into one big cluster and a few in the rest. Another list was meant to hold the cumulative scores of each cluster. Global variables would hold the Euclidean scores of those clusters. Then K-means was implemented. Placing each planet into its most similar cluster, and then checking if planets needed to be moved to a different cluster each iteration.

4. Evaluation Methodology

A major challenge was several missing data points, thus I made all missing points Zeros which may have an impact on the data.

Another challenge is understanding the data, many terms I had to research since I had never heard of them. Examples would be Periastron, and semi-major axis.

5. Results

What worked was removing outliers and data point that did not fit.

What did not work was finding similar planets to Earth as there were too many in that cluster. Thus next time it would be best to increase K.

6. What I Learned

I learned how K-means as well as how to apply it to a data set. I also learned how to gather large data on the internet and how to create conclusions via data mining algorithms.

7. Sources

Data Source:

<http://www.openexoplanetcatalogue.com/>

K-Means help:

<http://mnemstudio.org/clustering-k-means-example-1.htm>

<https://piazza-resources.s3.amazonaws.com/jc15191lz542/jfv58b6ggrh6kr/Clustering.pdf?X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=ASIAIURRRFYUPZF5KRJA%2F20180501%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20180501T020810Z&X-Amz-Expires=10800&X-Amz-SignedHeaders=host&X-Amz-Security-Token=FQoDYXdzEPv%2F%2F%2F%2F%2F%2F%2F%2F%2F%2FwEaDIBDrhjwm7UPTV3bEyK3A3d5sqxyQ7AK4IMAE4mJRwTscdP%2FyxiD4wMZrtznea0tiil9BfPMu%2BxVzSldLlkmnQ9jiE50ad4Up25r6IFjPvfyI9woQaSjTiZOo8zmzBLlr7dXpUbx9Ty40nTluZgpl3uJWP7xAu4q2uQ9FkVqKrnrrds2YHsi52dZOxx13L5CCMfZbA3YaVBBNZBGOo0ac%2FacKMVDP6RPxTFxkA%2FCKQkcSFejPfzIj9bk2A6b%2BA6L1LYZlhA%2FwVEApTcuxsjSYckettAUVChubYTN8WVvVbJIBTh0fnz9cP51mT9Stu7Wj%2Fl1zmvi0Fjlh0qcRF7Mnubb956zccxTSxJwijL2hH3Lih2bpd0t3RtWX7xDwlXsC0sMRo22i0yajXtxo57qae9nomLjWvo8dOhxNyF8mAGyoPXJNCdzXW3hWmkGVqWYSgkrRWaJczzW2Q2Mz4hrOCmz53clZcHkKOwTshZQ5d8f0um5wgiZIDRQ5UQvylEeZ8XrIiMPMuA11herCjfaqpj2iShyIn2Mj3RSzZTppikh7fWV%2BvGHdz1lNuHM013p2gPWLXBfYw%2B0VWCUbHFs1lSn2mebM6Io15Cf1wU%3D&X-Amz-Signature=46ce714f96692bf00593aea19145d046ad6aa5f099c28eba8bc53e80696f54e2>